Bayesian Reduced-Rank Regression

Maria Fernanda Pintado Serrano

Supervised by
Alexander Y. Shestopaloff, Luca Rossini and Matteo Iacopini

School of Mathematical Sciences
Queen Mary University of London

Submitted in partial fulfillment of the requirements of the degree of
Doctor of Philosophy

May 2025

Statement of Originality

I, Maria Fernanda Pintado Serrano, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by other, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that Queen Mary University of London has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature: Maria Fernanda Pintado Serrano

Date: 24th May 2025

Details of collaborations and publications:

- Pintado, M.F., Iacopini, M., Rossini, L. and Shestopaloff, A. (2025) – Uncertainty Quantification in Bayesian Reduced-Rank Sparse Regressions. *Statistics and Computing*, 35(4):110.
- Pintado, M.F., Iacopini, M., Rossini, L. and Shestopaloff, A.Y. (2025) – Bayesian Partial Reduced-Rank Regression. *Journal of Computational and Graphical Statistics*, 1–12.
- Pintado, M.F., Iacopini, M., Rossini, L. and Shestopaloff, A. (202X) – Bayesian Markov-Switching Partial Reduced-Rank Regression. *(In progress).*

Abstract

Regression methods are widely used across various fields for data analysis. The multivariate linear regression model captures relationships between multiple response variables and predictors through a coefficient matrix. Reduced-rank regression is an effective technique that considers response interrelations by imposing a low-rank constraint on the coefficient matrix. Consequently, these linear restrictions result in a reduced number of parameters, thus improving parsimony and interpretability of the model. This thesis contributes to the existing body of literature on reduced-rank regression by introducing novel statistical methods within the framework of Bayesian analysis. First, a mixture prior on the regression coefficient matrix is proposed for rank estimation, along with a global-local shrinkage prior on its low-rank decomposition for variable selection, providing full uncertainty quantification. Then, the first Bayesian approach to partial reduced-rank regression is developed, where the response vector and the coefficient matrix are partitioned into low-rank and full-rank sub-groups, increasing flexibility. The latter methodology is extended to incorporate time-varying parameters, allowing the allocation of the groups to switch over time between different states driven by a hidden Markov chain, and a flexible nonparametric model is used for the remaining component.

Acknowledgements

The work presented in this thesis would not have been possible without the invaluable support I was fortunate to find around me, both in moments of celebration and times of challenge.

First and foremost, I wish to express my deepest gratitude to my supervisors: Alex Shestopaloff, Luca Rossini, and Matteo Iacopini. From the very beginning, their guidance, expertise, and encouragement have been essential these past years. It is often said to choose your supervisor wisely, since they shape not only your PhD journey but potentially your future. I sincerely believe I could not have made a better choice, and was fortunate enough to have not one, but three inspiring mentors. I am especially thankful for their patience and thoughtful advice, which have taught me valuable lessons about academia and have prepared me for the next step in my career.

I gratefully acknowledge Javier Rubio and Fabrizio Leisen, who served as the examiners of my thesis, for their constructive comments, which helped improve the quality and clarity of this work.

I have been incredibly fortunate to be part of the stimulating environment in the School of Mathematical Sciences at Queen Mary University of London, particularly within the Statistics and Data Science group, which provided a remarkable place to learn and grow. I am especially grateful to Hugo Maruri and Effie Solea for serving on my committee and for their insightful feedback. I would also like to thank Claudia Garetto, for her dedication to fostering a more inclusive academic environment, and Reem Yassawi for supporting the graduate students in the School.

My gratitude extends to all my colleagues and friends who shared this experience with me and turned these years into a truly joyful chapter of my life. I am thankful to the Maths PhD community at QMUL for making the office an exciting space to work and learn together. In particular, I thank Eve Mladenova, Norberto Lucero and Danilo Riccio for their friendship; Zain Kapadia for being a dedicated fellow Postgraduate Representative; and Kabiru Abubakari, Sam Brevitt, Adi Onus, Christo Morison, James Kiln, Luka Ilic and Marco Gramatica for their good company. Beyond QMUL, I had the pleasure of meeting other talented PhD researchers. I am especially grateful to Nicolas Bianco, who became a friend when we met at my first academic conference, and to Max Hird, whose friendship and insightful conversations brought joy and inspiration during my final year. I also thank my non-academic friends: Jesús González, Giuliano Luchetta, Nathalia Céspedes, Andrea Torres and Jesús Vázquez for helping me stay grounded and sane since the day we started our lives in a new city, and for making home feel a little closer. I extend my heartfelt gratitude to Davide Carta for being an exceptional companion throughout this period, for his encouragement, kindness, and unwavering support in every step.

Finally, I am profoundly grateful to my family: my parents, Kena and Carlos, my sister Kena, my aunt Leticia, and my uncles Sergio and Jorge. Their love sustained me throughout this journey and transcends any distance. I dedicate this thesis to my parents, who gave me the gift of education and the courage to pursue my dreams.

The work presented in this dissertation was funded by CONAHCyT (Mexico). This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT, and the Computational resources provided by the Core Facility INDACO, a project of High-Performance Computing at the University of Milan.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The multivariate linear regression (MLR) model estimates the relationship between a group of response variables and a common set of predictor variables, as expressed by a matrix of regression coefficients. For each observational unit $i = 1, \dots, n$ from a sample of size n, let $\mathbf{y}_i \in \mathbb{R}^q$ be a vector of response variables explained by p possible predictors $\mathbf{x}_i \in \mathbb{R}^p$. The multivariate linear regression model is defined as

$$\mathbf{y}_i = \mathbf{C}'\mathbf{x}_i + \mathbf{e}_i,$$

where the error terms $\mathbf{e}_i$ are independent and normally distributed with mean zero and $q \times q$ covariance matrix $\mathbf{\Sigma}$, while $\mathbf{C}$ is the $p \times q$ matrix of regression coefficients. The MLR model can be expressed in matrix form by stacking all observations into an $n \times q$ matrix of response variables $\mathbf{Y}$ with the ith row as $\mathbf{y}_i'$, and an $n \times p$ matrix of covariates $\mathbf{X}$ with the ith row as $\mathbf{x}_i'$. The resulting model is

$$\mathbf{Y} = \mathbf{XC} + \mathbf{E}, \qquad \mathbf{E} = (\mathbf{e}_1, \dots, \mathbf{e}_n)', \qquad \mathbf{e}_i \overset{iid}{\sim} \mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma}). \tag{1.1}$$

The interpretability of the MLR model makes it a powerful tool widely used in various fields (for example, economics, genetics, medicine and environmental science) and deeply studied, from which variants have arisen. In certain situations, we may encounter correlations among the response variables, which the usual MLR model fails to account for, since estimating the elements of the coefficient matrix $\mathbf{C}$ simultaneously is equivalent to fitting a separate regression for each response, regressed on all predictors. A method that leverages these interrelationships is the reduced-rank regression approach (e.g, see Anderson, 1951; Izenman, 1975; Geweke, 1996; Reinsel et al., 2022).

Reduced-rank regression (RR) postulates that the dependence structure between the responses and the covariates can be represented through a small number of linear combinations of the latter. For instance, Reinsel et al. (2022) used multivariate linear regression to study the influence of certain covariates on the chemical properties of a urine specimen in the context of reduced-rank regression. Jöreskog and Goldberger (1975) applied a similar model in sociology considering unobservable latent variables. Gudmundsson (1977) constructed linear combinations of variables in an econometric model of the United Kingdom to capture aspects of the economic situation. Specifically, a RR model assumes the coefficient matrix $\mathbf{C}$ to be rank deficient, implying a reduction in the number of parameters through linear restrictions on the entries of $\mathbf{C}$, where a smaller number of relevant linear combinations of the predictor variables explain the variation in all the responses. Consequently, a reduced-rank decomposition of the coefficient matrix allows to obtain $\mathbf{C} = \mathbf{BA}'$, where $\mathbf{A}$ and $\mathbf{B}$ are two full-rank matrices whose dimensions depend on the rank r of $\mathbf{C}$.

The first approach to this method was made by Anderson (1951), who proposed a class of regression models that restrict $\mathbf{C}$ to be rank deficient. Thereafter, Izenman (1975) introduced the term reduced rank regression for these models and provided a further study of the estimators. From a Bayesian perspective, Geweke (1996) pioneered the early work on reduced-rank regression by assigning independent Gaussian priors on the elements of the coefficient matrix conditioning on the rank, assumed to be known.

Reduced-rank regression has since become a widely used technique and remains an active area of research across different fields. This thesis contributes to the Bayesian literature on RR in three principal directions. First, we introduce a novel mixture prior for rank selection, alongside two indices for covariate selection in Chapter 2, where full uncertainty quantification about these estimations is guaranteed. Second, Chapter 3 focuses on a variant of RR known as partial reduced-rank regression, which partitions the response variables into low-rank and full-rank groups, and we infer these allocations from the data. Finally, based on the findings in the previous chapter, we extend the partial reduced-rank model to a dynamic setting in Chapter 4, allowing the response grouping to evolve over time. In addition, the relationship between covariates and the full-rank component is modelled nonparametrically, offering greater flexibility in the regression structure.

A closely related topic of research to reduced-rank regression is low-rank matrix completion, where the aim is to complete the missing entries of a matrix from a partial observation (see Alquier, 2013, for a revision of Bayesian methods). The similarity with reduced-rank regression lies in the estimation of a low-rank matrix, where the rank is user-defined (Lim and Teh, 2007; Salakhutdinov and Mnih, 2008) or implicitly estimated (Babacan et al., 2011). The reduced rank regression model is also connected to the econometrics literature on cointegration, where the rank r of a coefficient matrix identifies the number of long-run common trends among the endogenous response variables. In this framework, several attempts have been made to infer r from the data (e.g., see Kleibergen and Paap, 2002; Chua and Tsiaplias, 2018; Strachan, 2003) using techniques such as Bayes' factors to compare nested models with varying ranks, or estimating posterior probabilities of different ranks with approaches such as Markov Chain Monte Carlo (MCMC).

Similarly, the low-rank decomposition $\mathbf{C} = \mathbf{B}\mathbf{A}'$ is closely related to the principal components analysis (PCA) and sparse factor analysis (FA). Standard PCA can be obtained by setting $\mathbf{Y} = \mathbf{X}$ and introducing an intercept term (Izenman, 2008, ch.7), where r represents the number of principal components used to capture most of the variability of $\mathbf{X}$. Instead, in sparse FA, $\mathbf{A}$ is a vector of latent factors, and $\mathbf{B}$ corresponds to a sparse matrix of factor loadings, with r corresponding to the number of latent factors. Within these frameworks, a recent stream of literature considers the problem of rank estimation (i.e., determining the number of components or factors) in a Bayesian perspective, such as Šmídl and Quinn (2007); Sobczyk et al. (2017); Ning and Ning (2024) for PCA and Lopes and West (2004); Viroli (2009); Ghosh and Dunson (2009); Frühwirth-Schnatter et al. (2025) in FA, among others.

Despite different specifications linked to low-rank decomposition, the estimation of the rank parameter remains a central task. The value of the rank of the coefficient matrix has direct implications on the estimation of the model. A 0-rank matrix implies a lack of meaningful relationships between the predictor variables and the responses. Meanwhile, a full-rank matrix yields independent regressions for each response variable, where no dimensionality reduction is achieved. Hence, it becomes apparent the need to estimate an optimal value of the rank and quantify the uncertainty around this estimation. Further gains towards a more parsimonious model can be garnered by incorporating sparsity to reduce the number of coefficients. If we assume a relatively small share of nonzero entries of the coefficient matrix $\mathbf{C}$, then a sparse multivariate linear regression (sparse MLR) is obtained, (e.g., see Goh et al., 2017), and from a Bayesian perspective as Zhu et al. (2014); Chakraborty et al. (2019); Yang et al. (2022). To simultaneously address these issues, we propose the Bayesian Rank Estimation and Covariate Selection (BRECS) method, where the rank is estimated with a fully Bayesian approach, providing uncertainty quantification about its estimation. Additionally, two measures of uncertainty for covariate selection are developed under a sparse estimation setting based on the Signal Variance Adaptive Variable Selector (Ray and Bhattacharya, 2018, SAVS), which employs a sparsification step on a point estimate of the parameter of interest into exact zeros. BRECS successfully integrates both sparsity and rank reduction, while additionally offering uncertainty quantification. The description of this method is elaborated in Chapter 2.

Another important aspect of modelling is the prevalence of intrinsic group structures in certain data types, particularly in those involving complex information. These structures signify the presence of correlations among variables within these groups, and ignoring them can lead to an inefficient use of the available data. In the context of multivariate response-predictor analysis, a commonly adopted strategy is to perform covariate selection for each response variable, as is the case of proposed approach in Chapter 2 and in other methods addressing the predictor group structure (see Buch et al., 2023, for a review of such methods). However, the group structure among response variables is typically overlooked. Considering group structures within the response variables enhances our comprehension of the data in diverse fields such as macroeconomics (Reinsel and Velu, 2006) and genetics (Li et al., 2015; Luo and Chen, 2020).

We explore a generalisation of the standard RR regression model, where the reduced-rank coefficient structure applies to only an unknown subset of the response variables. In contrast, the remaining subset maintains a full-rank coefficient submatrix. This more flexible approach has been called partially reduced-rank (PRR) regression by Reinsel and Velu (2006). The PRR model adds flexibility to the regressions in accommodating complex relationships observed in real-world datasets, such as those found in economic contexts. The first Bayesian approach to PRR regression is developed in Chapter 3, where the low-rank and full-rank response groupings are unknown and directly inferred from the data under our proposed Bayesian Partial Reduced-Rank (BPRR) regression.

When the observations in the regression model are indexed by time, the relationship between covariates and responses could change over periods. A time-varying grouping structure in the response variables in RR regression is currently understudied. We address this limitation in Chapter 4, where the PRRR framework is generalised in two directions. First, we replace the full-rank linear part of the model with a more flexible nonparametric term using a Gaussian process. Second, we introduce a Markov-switching mechanism to allow for a time-varying and persistent clustering structure in a Hidden Markov regression model framework (Hamilton, 1990; Frühwirth-Schnatter, 2006). This choice allows for the clustering of the response variables based on the degree of complexity in their relationship with the covariates: one group retains a simple linear, low-rank specification, whereas the other considers a complex nonparametric regression. We refer to the proposed method as Markov-switching Bayesian Partial Reduced-Rank (MSPRR) regression.

Table 1.1 summarises the key features of various regression methodologies considered throughout this thesis. Traditional multivariate linear regression (MLR) lacks all of the listed characteristics, but its sparse counterpart (sparse MLR) focuses on variable selection, although is restricted in terms of dimension reduction or uncertainty quantification. Reduced-rank regression (RR) addresses dimension reduction and rank estimation but does not incorporate sparsity or provide measures of uncertainty. BRECS extends RR by integrating sparse estimation and offering full Bayesian uncertainty quantification on both rank and sparsity. Meanwhile, our proposed BPRR introduces the novel ability to cluster responses based on the complexity of their relationship to covariates, while retaining rank estimation and uncertainty quantification on that aspect. Lastly, MSPRR further generalises BPRR by incorporating time-varying clusters over the response variables. The three proposed methods in this thesis (BRECS, BPRR and MSPRR) offer additional features compared to traditional regression approaches such as MLR, sparse MLR, and RR, by addressing model flexibility, interpretability, and uncertainty quantification.

Feature	MLR	Sparse MLR	RR	BRECS	BPRR	MSPRR
Dimension reduction	✗	✗	✓	✓	✓	✓
Rank estimation	✗	✗	✓	✓	✓	✓
Sparse estimation	✗	✓	✗	✓	✗	✗
Unc. Quant. on rank	✗	✗	✗	✓	✓	✓
Unc. Quant. on sparsity	✗	✗	✗	✓	✗	✗
Response clustering	✗	✗	✗	✗	✓	✓
Time variation	✗	✗	✗	✗	✗	✓

Table 1.1: Comparison of desired features (rows) across different regression models (columns).

1.1 Notation

For ease of reference, Table 1.2 summarizes the key symbols and notation used throughout the thesis.

Symbol	Description
$\mathbf{M}$	bold capital letters denote matrices
$\mathbf{x}$	bold lower case letters denote vectors
$\mathbb{R}$	set of real numbers
$\otimes$	Kronecker product
$\propto$	proportional to
$\mathbb{I}(\cdot)$	indicator function
$\mathbf{M}'$	transpose of $\mathbf{M}$
$\lvert\mathbf{M}\rvert$	determinant of $\mathbf{M}$
$\mathbf{M}^+$	Moore-Penrose pseudoinverse of $\mathbf{M}$
$\mathbf{K}_{m,n}$	$mn \times mn$ commutation matrix
$\lVert\mathbf{x}\rVert_2$	Euclidean norm of $\mathbf{x}$
$\lVert\mathbf{M}\rVert_F$	Frobenius norm of $\mathbf{M}$
$\text{rank}(\mathbf{M})$	rank of $\mathbf{M}$
$\text{tr}(\mathbf{M})$	trace of $\mathbf{M}$
$\text{vec}(\mathbf{M})$	vectorisation of matrix $\mathbf{M}$
$\text{nbd}(\mathbf{x})$	neighbourhood of $\mathbf{x}$

Table 1.2: Common symbols used throughout the thesis, and their corresponding descriptions.

Chapter 2

Uncertainty Quantification in Bayesian Reduced-Rank Sparse Regressions

2.1 Introduction

In reduced-rank regression, as stated in Chapter 1, the decomposition of the coefficient matrix into the product of two full-rank matrices depends on the value of the rank r. The first Bayesian approach developed by Geweke (1996) assumes the rank to be known, and also proposes to use predictive odds ratios for regression models with different ranks when r is unknown to perform model selection. Typically, there is no guidance in fixing a specific rank for $\mathbf{C}$ in real data applications such as economics and finance, where the importance of RR regression becomes apparent (Cubadda and Hecq, 2021). Although further methods considering the rank, r, as unknown have been developed, they typically treat it as a parameter to be fixed before performing the inference (Chen and Huang, 2012; Goh et al., 2017).

The literature that assumes the rank as an unknown quantity relies on post-processing steps to estimate it, for example, by thresholding the singular values (Bunea et al., 2011; Chen et al., 2013; Chakraborty et al., 2019). Moreover, the performance of post-processing methods typically depends on some user-specified tuning parameters, whose choice is hardly justifiable, and it does not allow uncertainty quantification. Differently, the literature on nonparametric reduced-rank regression treats the rank as a tuning parameter chosen by cross-validation (Lian and Ma, 2013), estimates it by thresholding singular values along with parameters to be selected by the user (Mukherjee, 2013) or by imposing regularization penalties (Foygel et al., 2012). To overcome these limitations, we propose the Bayesian Rank Estimation and Covariate Selection (BRECS) method, with the crucial difference that the rank is estimated jointly with the other parameters in a single-step fully Bayesian approach. As such, BRECS allows for uncertainty quantification and removes the need for a post-processing scheme (i.e., two-step approach). The key step in our method is to assume a finite mixture prior on the coefficient matrix, with mixture components corresponding to fixed possible values of the rank variable.

Sparsity-inducing estimators of the coefficient matrix have been used to overcome the over-parametrization and potential over-fitting that typically characterise high-dimensional models. As stated in Chapter 1, the matrix of coefficients, **C**, is defined as the product of two full-rank matrices, **A** and **B**, which needs to be inferred by defining prior distributions. We impose a global-local shrinkage prior on the columns of **B**, which encourages sparsity in the matrix of coefficients (Bhattacharya et al., 2015), while on matrix **A** we rely on a standard Gaussian prior.

Global-local shrinkage priors ensure that the coefficients are pulled towards zero, but exact sparsification (i.e., estimated coefficients exactly equal to zero) is prevented by the continuity of the prior. This requires an additional step for coefficient selection, thus, the uncertainty about this mechanism becomes relevant. Ray and Bhattacharya (2018) proposed the Signal Adaptive Variable Selector (SAVS) to post-process a point estimate, such as the posterior mean, and group coefficients into exact zeros and non-zeros. Consequently, Huber et al. (2021) applied the SAVS to each MCMC draw of the parameter of interest, thus obtaining a posterior inclusion probability (PIP) for each coefficient. We adopt a similar strategy and apply SAVS to each element of the coefficient matrix **C**; then, we derive a new PIP uncertainty index to quantify uncertainty in coefficient selection. A related work by Yuchi et al. (2023) provides another way of defining a prior (conditional on the rank) and different uncertainty quantification measures for a low-rank matrix in the context of matrix completion.

In addition to the PIP uncertainty index, we also introduce the Relevance Index (RI). Differently from the PIP uncertainty index, that quantifies uncertainty about variable inclusion, the RI, which is the most important one, is a distribution representing the relevance of a covariate in terms of the share of response variables on which it has a significant impact. This index provides full uncertainty quantification. Moreover, we propose a rule of thumb for variable selection that summarises and complements the information embedded in the RI by means of its survival function.

Uncertainty quantification in both rank estimation and variable selection is currently underdeveloped in the literature. Yang et al. (2022) address this issue by using the Laplace approximation of the posterior distributions within a collapsed Gibbs sampler and obtaining complete sparse rows of **C** for covariate selection. By coupling our mixture prior on **C** with the shrinkage prior on **B**, our proposed method enables sparse and low-rank estimation of the coefficient matrix by sampling from the exact full-conditional posteriors, obviating the demand for an approximation. Also, our approach removes the need for post-processing while jointly incorporating the quantification of uncertainty. So, unlike current techniques, we do not rely on visual inspection of the plots and on user-specified thresholds to choose the rank. Instead, we allow for a simple and transparent interpretation based on the entire posterior distribution of the rank. Practically being able to quantify the uncertainty in the rank allows a user to assess how well the data guides in choosing a specific rank value in a concrete setting. Besides, differently from Yang et al. (2022), our approach is able to obtain a sparse estimate of **C** where either entire rows or only single entries are null. The PIP reports uncertainty quantification on the estimates of the rank and the coefficients, and then the RI is used to measure the uncertainty about variable selection when working with a multivariate response.

This chapter is organised as follows. Section 2.2 introduces the model, and presents the proposed priors for rank selection. Then, Section 2.3 demonstrates our sampling algorithm and provides different definitions of uncertainty quantification. Section 2.4 illustrates the performance of the proposed methods in simulated experiments, while Section 2.5 applies them to datasets on the chemical composition of tobacco and on the photometry of galaxies. Finally, Section 2.6 provides a discussion.

2.2 Reduced-rank regression model

The reduced-rank regression model identifies a smaller set of linear combinations of the variables that can explain a large proportion of the variation in the data under the MLR framework in Eq.(1.1). To consider this model, an assumption on the rank of the matrix of coefficients $\mathbf{C}$ is defined as $\text{rank}(\mathbf{C}) = r \leq \min(p, q)$. Such an assumption translates into fewer parameters, leading to a more parsimonious model. We consider the low-rank decomposition $\mathbf{C} = \mathbf{B}\mathbf{A}'$ with $\mathbf{B} \in \mathbb{R}^{p\times r}$ and $\mathbf{A} \in \mathbb{R}^{q\times r}$, where the rank r needs to be estimated.

2.2.1 Mixture prior for rank selection

In this chapter, we contribute to the literature on reduced rank regression by proposing a new Bayesian approach for rank estimation and related uncertainty quantification. Our proposal relies on finite mixture priors, which combine two or more prior probability distributions, namely the mixture components, each with its own set of parameters. The main aspect that favours a mixture prior is that rank selection, which corresponds to an automatic model choice, is made along with parameter estimation. In the proposed mixture prior, each component corresponds to a different rank. The Bayesian approach to inference coupled with data augmentation allows for obtaining a posterior distribution for the rank. Besides deriving a point estimate as the maximum a posteriori (MAP), our approach permits uncertainty quantification, a novel feature of this method in contrast to existing literature.

In detail, we define a finite mixture prior on $\mathbf{C}$ made by a number of components equal to $R = \min(p, q)$, assumed to be q. Employing the notation $\mathbf{C}_s$ for the matrix $\mathbf{C}$ under the restriction $\text{rank}(\mathbf{C}) = s$, the prior is expressed as

$$p(\mathbf{C}) = \sum_{s=1}^{q} w_s p(\mathbf{C}_s),$$

where w_s is the prior probability of $\text{rank}(\mathbf{C}) = s$. Denoting $\mathbf{w} = (w_1, \ldots, w_q)$, we assume a Dirichlet prior distribution with parameter $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_q)$, that is $\mathbf{w} \sim \mathcal{D}ir(\boldsymbol{\omega})$. We introduce a latent allocation variable u which assumes values in the set $\{1, \ldots, q\}$, and follows a prior Multinomial distribution with parameter $\mathbf{w}$, represented as

$$p(u \mid \mathbf{w}) = \prod_{s=1}^{q} w_s^{\mathbb{I}(u=s)},$$

where $\mathbb{I}(u = s)$ is the indicator function, taking value 1 if $u = s$ and 0 otherwise.

To define a prior on $\mathbf{C}$ conditional on its rank u, we rely on the low-rank representation $\mathbf{C} = \mathbf{B}\mathbf{A}'$, and assume prior independence between $\mathbf{A}$ and $\mathbf{B}$, which results in

$$p(\mathbf{C}_u) = p(\mathbf{A}_u)p(\mathbf{B}_u),$$

where $\mathbf{A}_u$ and $\mathbf{B}_u$ are the factors of the rank-u decomposition of $\mathbf{C}$, reminding that their dimensions are rank dependent, being $q \times u$ and $p \times u$, respectively.[1] For the entries of the matrix $\mathbf{A}_u$, we consider a standard normal prior $a_{jh} \sim \mathcal{N}(0, 1)$, with $j = 1, \ldots, q$ and $h = 1, \ldots, u$.

[1] As the factorisation $\mathbf{C} = \mathbf{B}\mathbf{A}'$ is not unique, the factor matrices are not separately identifiable. One possible strategy for identification consists in imposing that the first u rows of $\mathbf{A}$ are the identity matrix $\mathbf{I}_u$ (Geweke, 1996). However, since our interest lies in the matrix $\mathbf{C}$ that is always identified, we do not impose any such restriction.

Regarding the prior specification for $\mathbf{B}_u$, we use a global-local shrinkage prior on each column $\mathbf{b}_h = (b_{1h}, b_{2h}, \ldots, b_{ph})' \in \mathbb{R}^p$ of $\mathbf{B}_u$, for $h = 1, \ldots, u$. This family of distributions consists of a hierarchical scale mixture of (multivariate) Gaussian distributions of the type

$$b_{lh} \mid \tau_h, \phi_{lh} \sim \mathcal{N}(0, \tau_h \phi_{lh}), \qquad \tau_h \sim \pi_{\tau_h}(\tau_h), \qquad \phi_{lh} \sim \pi_{\phi_{lh}}(\phi_{lh}),$$

where τ_h and ϕ_{lh} are the global and local components of the variance, respectively, with distributions $\pi_{\tau_h}(\cdot)$ and $\pi_{\phi_{lh}}(\cdot)$.[2] We consider a Dirichlet-Laplace prior (Bhattacharya et al., 2015; Cross et al., 2020), which can be represented as

$$\begin{aligned} b_{lh} \mid \psi_{lh}, \tau_h, \phi_{lh} &\sim \mathcal{N}(0, \psi_{lh} \tau_h^2 \phi_{lh}^2), \quad l = 1, \ldots, p \\ \tau_h \mid \alpha_h &\sim \mathcal{G}a(\alpha_h p, 1/2), \\ \boldsymbol{\phi}_h \mid \alpha_h &\sim \mathcal{D}ir(\alpha_h, \ldots, \alpha_h), \\ \psi_{lh} &\sim \mathcal{E}xp(1/2), \\ \alpha_h &\sim \mathcal{U}(L_\alpha, U_\alpha), \end{aligned} \tag{2.1}$$

where $\mathcal{G}a(\cdot)$ and $\mathcal{D}ir(\cdot)$ denote the Gamma (with the shape-rate parametrization) and Dirichlet distributions, respectively. As usual with hierarchical priors, the performance of the DL prior depends on the hyperparameter values, particularly on α_h. Small values of α_h lead to an increase in the amount of global shrinkage while maintaining thicker tails on the local marginal distributions, thereby avoiding the incorrect elimination of variables that carry significant predictive signal. To address this issue, similarly to Cross et al. (2020), we assume a continuous uniform prior for α_h, with $L_\alpha = 0.12$ and $U_\alpha = 0.45$. Although Cross et al. (2020) consider the interval $[p^{-1}, 0.5]$, this choice leads to unstable results in our framework, while the fixed bounds generate a reasonable amount of shrinkage and bypass user-specified parameters. For each entry l in column h, the local shrinkage parameter is ψ_{lh}, and the vector of global shrinkage is $(\tau_h \phi_{1h}, \ldots, \tau_h \phi_{ph})$, where $\boldsymbol{\phi}_h = (\phi_{1h}, \ldots, \phi_{ph})'$ is constrained to lie in the $(p-1)$ simplex Δ^{p-1}. Finally, we impose no restrictions on the covariance matrix $\mathbf{\Sigma}$, and place an inverse Wishart prior, $\mathbf{\Sigma} \sim \mathcal{IW}(\nu, \Upsilon)$.

As a consequence of the mixture prior on the matrix of coefficients $\mathbf{C}$, the observed likelihood function is a mixture distribution with the same weights. Denoting $\mathbf{A} = \{\mathbf{A}_1, \ldots, \mathbf{A}_q\}$ and $\mathbf{B} = \{\mathbf{B}_1, \ldots, \mathbf{B}_q\}$, the likelihood is represented as

$$p(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \mathbf{\Sigma}, \mathbf{w}) = \sum_{s=1}^{q} w_s (2\pi)^{-\frac{n}{2}} |\mathbf{\Sigma}|^{-\frac{n}{2}} \exp\Big\{ -\frac{1}{2} \operatorname{tr}[\mathbf{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\mathbf{B}_s \mathbf{A}_s')' (\mathbf{Y} - \mathbf{X}\mathbf{B}_s \mathbf{A}_s')] \Big\}$$

and the complete data likelihood is

$$\begin{aligned} p(\mathbf{Y}, u \mid \mathbf{A}, \mathbf{B}, \mathbf{\Sigma}, \mathbf{w}) &= p(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \mathbf{\Sigma}, u, \mathbf{w}) p(u \mid \mathbf{w}) \\ &= \prod_{s=1}^{q} \bigg(\frac{1}{(2\pi)^{nq/2} |\mathbf{\Sigma}|^{n/2}} \exp\bigg\{ -\frac{1}{2} \operatorname{tr}[\mathbf{\Sigma}^{-1} (\mathbf{Y} - \mathbf{X}\mathbf{B}_s \mathbf{A}_s')' (\mathbf{Y} - \mathbf{X}\mathbf{B}_s \mathbf{A}_s')] \bigg\} w_s \bigg)^{\mathbb{I}(u=s)}. \end{aligned}$$

[2] By specifying the distributions of the two variance components and eventually introducing additional layers in the hierarchy, it is possible to generate a wide range of prior distributions that shrink the coefficients toward zero while allowing the data to inform about large deviations from the origin thanks to the heavy tails of the marginal prior (obtained integrating out the global component τ).

2.2.2 Alternative prior parametrizations

The previous section has introduced a generic prior structure for the matrices $\mathbf{A}_u$ and $\mathbf{B}_u$, conditional on the rank of $\mathbf{C}$ being u. However, no direct connection was assumed between $\mathbf{A}_u$ and $\mathbf{A}_v$, for $u \neq v$ (similarly for $\mathbf{B}_u$). In this section, we leverage on the particular structure imposed by the reduced-rank assumption for $\mathbf{C}$ to define two alternative prior parametrizations for the matrices $\mathbf{A}_u$ and $\mathbf{B}_u$. In particular, we propose two main parametrizations for $\mathbf{A}_u$ and $\mathbf{B}_u$: the naïve (RRn) and the column-sharing (RRcs). The first case provides the best unconditional approximation of the matrix of coefficients with the estimated rank, a parametrization that results in a computationally intensive algorithm with $O(q^3 + q^2 p)$ parameters. In contrast, RRcs is based on sharing information across columns, leading to the best conditional approximation, a reduced number of parameters of $O(q^2 + pq)$, and a computationally faster MCMC for posterior inference compared to the former approach.

The auxiliary variable u represents the (unknown) rank of $\mathbf{C}$, thus implicitly determining the number of columns of $\mathbf{A}$ and $\mathbf{B}$. Within the RRn parametrization, each value of u is associated with a specific collection of matrices $\mathbf{A}_u \in \mathbb{R}^{q\times u}$, $\mathbf{B}_u \in \mathbb{R}^{p\times u}$ and the corresponding parameters of the hierarchical prior for each column of $\mathbf{B}_u$, that is $(\phi, \tau, \psi, \alpha)$. As u ranges from 1 to q, there are in total q collections of parameters, also differing in the number of elements included in each collection (Figure 2.1a). Instead, in the RRcs parametrization, moving from u to $u + 1$ implies sharing the same parameters as in u, plus an additional column of the matrices $\mathbf{A}_{u+1}$, $\mathbf{B}_{u+1}$ (and the corresponding parameters ϕ, τ, ψ, α). Therefore, the sharing mechanism conditions the reduced-rank approximation of the matrix $\mathbf{C}_{u+1}$ to the *even lower* rank approximation $\mathbf{C}_u$ (Figure 2.1b).

To summarise, the crucial difference is the parametrization of the collection of matrices $\{\mathbf{A}_u, \mathbf{B}_u\}_{u=1}^{q}$. Denoting with $\mathbf{a}_h^{(u)}$ and $\mathbf{b}_h^{(u)}$ the hth column of the matrices $\mathbf{A}_u$ and $\mathbf{B}_u$, the RRn parametrization assumes:

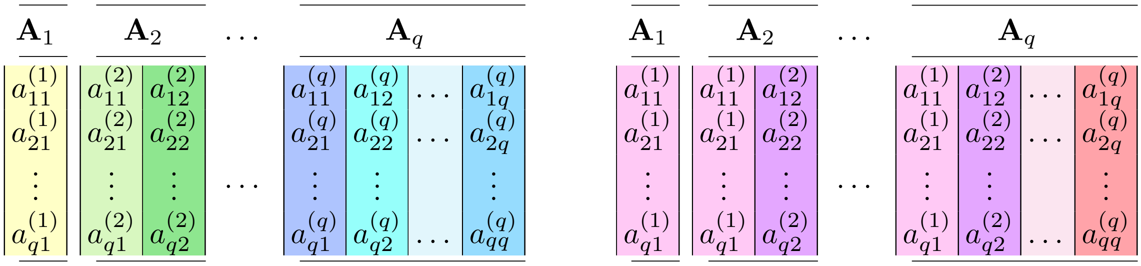
$$\begin{aligned}
&\mathbf{A}_1 = [\mathbf{a}_1^{(1)}] && \mathbf{B}_1 = [\mathbf{b}_1^{(1)}] \\
&\mathbf{A}_2 = [\mathbf{a}_1^{(2)}, \mathbf{a}_2^{(2)}] && \mathbf{B}_2 = [\mathbf{b}_1^{(2)}, \mathbf{b}_2^{(2)}] \\
&\quad\vdots && \\
&\mathbf{A}_q = [\mathbf{a}_1^{(q)}, \mathbf{a}_2^{(q)}, \dots, \mathbf{a}_q^{(q)}] && \mathbf{B}_q = [\mathbf{b}_1^{(q)}, \mathbf{b}_2^{(q)}, \dots, \mathbf{b}_q^{(q)}].
\end{aligned}$$

Conversely, the RRcs parametrization assumes:

$$\begin{aligned}
&\mathbf{A}_1 = [\mathbf{a}_1^{(1)}] && \mathbf{B}_1 = [\mathbf{b}_1^{(1)}] \\
&\mathbf{A}_2 = [\mathbf{a}_1^{(1)}, \mathbf{a}_2^{(2)}] && \mathbf{B}_2 = [\mathbf{b}_1^{(1)}, \mathbf{b}_2^{(2)}] \\
&\quad\vdots && \\
&\mathbf{A}_q = [\mathbf{a}_1^{(1)}, \mathbf{a}_2^{(2)}, \dots, \mathbf{a}_q^{(q)}] && \mathbf{B}_q = [\mathbf{b}_1^{(1)}, \mathbf{b}_2^{(2)}, \dots, \mathbf{b}_q^{(q)}].
\end{aligned}$$

The first parametrization produces a different estimate of the hth column of each low-rank matrix: $\mathbf{a}_h^{(u)} \neq \mathbf{a}_h^{(v)}$, for $u \neq v$, and $h \leq \min(u, v)$. In contrast, the second parametrization assumes that $\mathbf{a}_h^{(u)} = \mathbf{a}_h^{(v)}$. The same rationale applies to matrix $\mathbf{B}$.

Concerning the prior construction for the columns of either matrix, we assume the same distributions for the RRn and RRcs, that is $\mathbf{a}_h^{(u)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and Eq. (2.1) for each $\mathbf{b}_h^{(u)}$.


$\mathbf{A}_1$
$\mathbf{A}_2$
$\mathbf{A}_q$
$a_{11}^{(1)}$, $a_{21}^{(1)}$, $a_{q1}^{(1)}$
$a_{11}^{(2)}$, $a_{21}^{(2)}$, $a_{q1}^{(2)}$, $a_{12}^{(2)}$, $a_{22}^{(2)}$, $a_{q2}^{(2)}$
$a_{11}^{(q)}$, $a_{21}^{(q)}$, $a_{q1}^{(q)}$, $a_{12}^{(q)}$, $a_{22}^{(q)}$, $a_{q2}^{(q)}$, $a_{1q}^{(q)}$, $a_{2q}^{(q)}$, $a_{qq}^{(q)}$
$a_{11}^{(1)}$, $a_{21}^{(1)}$, $a_{q1}^{(1)}$, $a_{12}^{(2)}$, $a_{22}^{(2)}$, $a_{q2}^{(2)}$, $a_{1q}^{(q)}$, $a_{2q}^{(q)}$, $a_{qq}^{(q)}$

(a) RRn (b) RRcs

Figure 2.1: Matrix $\mathbf{A}_u$ under both parametrizations: the naïve (panel a) and the column-sharing (panel b). Each colour represents a set of values for the corresponding columns of $\mathbf{A}_u$. In RRn, the elements of the first (and only) column of $\mathbf{A}_1$ are different from the first column of $\mathbf{A}_2, \dots, \mathbf{A}_q$. In RRcs, the elements of the first (and only) column in $\mathbf{A}_1$ are the same as those of the first column of $\mathbf{A}_2, \dots, \mathbf{A}_q$.

2.3 Posterior sampling

In this section, we yield the estimation details of the proposed mixture prior and derive the model uncertainty quantification indexes. Initially, we provide the representation of the joint posterior distribution of the model parameters. Let us define $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1 = (\psi_{11}, \dots, \psi_{p1}), \dots, \boldsymbol{\psi}_q = (\psi_{1q}, \dots, \psi_{pq}))$, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_q)$, $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_q)$ and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)$. The parameters can be included in $\Theta = (\boldsymbol{\Sigma}, \mathbf{w}, \mathbf{A}, \mathbf{B}, \boldsymbol{\Psi}, \boldsymbol{\tau}, \boldsymbol{\Phi}, \boldsymbol{\alpha})$, and the joint posterior distribution is given by

$$\begin{aligned} p(\Theta, u \mid \mathbf{Y}) \propto\, & p(\mathbf{Y} \mid \Theta, u)\, p(\mathbf{A} \mid u)\, p(\mathbf{B} \mid \boldsymbol{\Psi}, \boldsymbol{\tau}, \boldsymbol{\Phi}, u)\, p(u \mid \mathbf{w}) \\ & \times p(\boldsymbol{\tau} \mid \boldsymbol{\alpha})\, p(\boldsymbol{\Phi} \mid \boldsymbol{\alpha})\, p(\boldsymbol{\Sigma})\, p(\mathbf{w})\, p(\boldsymbol{\alpha})\, p(\boldsymbol{\Psi}). \end{aligned}$$

The choice of the prior distributions allows for a straightforward implementation of an efficient Markov Chain Monte Carlo (MCMC) algorithm to update the parameters by sampling from the full conditional posterior distributions (see Appendix A.1 for a detailed derivation). The proposed Gibbs sampler for RRn and RRcs are outlined in Algorithm 1 and Algorithm 2, respectively, where GiG$(\cdot)$ and iG$(\cdot)$ denote the generalised inverse Gaussian and the inverse Gaussian distributions, and a $*$ superscript denotes the value of the hyper-parameters of the posterior distribution. A comprehensive description of hyperparameter specifications can be found in Appendix A.1.

The algorithm of the column-sharing approach (RRcs) performs significantly faster than the naïve (RRn) parametrization since at each iteration of the MCMC, the number of columns updated (through the posterior distribution or the prior) for $\mathbf{A}$ and $\mathbf{B}$ is R, the maximum possible rank, as opposed to $R(R+1)/2$ in the naïve approach. In addition to faster computational performance, the column-sharing case presents a better mixing around the rank estimate (see Section 2.4).

2.3.1 Variable selection

Further interest is placed on obtaining exact zeros in the coefficient matrix $\mathbf{C}$ for covariate selection, and uncertainty quantification of covariates inclusion becomes relevant. Shrinkage priors allow for parameter estimates that are very close to zero but not exactly zero, and the nature of hierarchical shrinkage priors makes common MCMC methods for sparsification that rely on cross-validation to be computationally prohibitive. However, Ray and Bhattacharya (2018) introduced the signal adaptive variable selector (SAVS), a simple algorithm for the sparsification step on the posterior mean of the parameter of interest. Employing SAVS allows to have exact zeros, but uncertainty

Algorithm 1 Gibbs sampler for RRn specification

```
Sample $u$ from the posterior distribution on a logarithmic scale through inverse transform sampling;
Sample $\mathbf{w}$ from $\mathcal{D}ir(\boldsymbol{\omega}^*)$;
Sample $\boldsymbol{\Sigma}$ from $\mathcal{IW}(\nu^*, \Upsilon^*)$;
for $s = 1$ to $R$ do
    if $s = u$ then
        Sample $\mathbf{A}_u$ by drawing vec($\mathbf{A}_u$) from $\mathcal{N}_{qu}(\boldsymbol{\mu}^*_{\mathbf{A}_u}, \boldsymbol{\Sigma}^*_{\mathbf{A}_u})$;
        Sample $\mathbf{B}_u$ by drawing vec($\mathbf{B}'_u$) from $\mathcal{N}_{pu}(\boldsymbol{\mu}^*_{\mathbf{B}_u}, \boldsymbol{\Sigma}^*_{\mathbf{B}_u})$;
    else
        Sample $\mathbf{A}_s$ and $\mathbf{B}_s$ from the prior;
    end if
    for $h = 1$ to $s$ do
        Sample $\tau_h$ from GiG($p^*_{\tau_h}, a^*_{\tau_h}, b^*_{\tau_h}$);
        Sample $\tilde{\psi}_{lh}$ from iG($a^*_{\psi_{lh}}, b^*_{\psi_{lh}}$), then set $\psi_{lh} = \tilde{\psi}^{-1}_{lh}$, for each $l = 1, \dots, p$;
        Sample $T_{lh}$ from GiG($p^*_{\phi_{lh}}, a^*_{\phi_{lh}}, b^*_{\phi_{lh}}$), then set $\phi_{lh} = T_{lh} / \sum_{i=1}^{p} T_{ih}$, for each $l = 1, \dots, p$;
        Sample $\alpha_h$ from its full conditional using a griddy Gibbs sampler (Ritter and Tanner, 1992).
    end for
end for
```

Algorithm 2 Gibbs sampler for RRcs specification

```
Sample $u$ from the posterior distribution on a logarithmic scale through inverse transform sampling;
Sample $\mathbf{w}$ from $\mathcal{D}ir(\boldsymbol{\omega}^*)$;
Sample $\boldsymbol{\Sigma}$ from $\mathcal{IW}(\nu^*, \Upsilon^*)$;
for $s = 1$ to $u$ do
    Sample $\mathbf{A}_u$ by drawing each column $\mathbf{a}_s$ from $\mathcal{N}_q(\boldsymbol{\mu}^*_{\mathbf{a}_s}, \boldsymbol{\Sigma}^*_{\mathbf{a}_s})$;
    Sample $\mathbf{B}_u$ by drawing each column $\mathbf{b}_s$ from $\mathcal{N}_p(\boldsymbol{\mu}^*_{\mathbf{b}_s}, \boldsymbol{\Sigma}^*_{\mathbf{b}_s})$;
end for
for $s = u + 1$ to $R$ do
    Sample columns $\mathbf{a}_s$ and $\mathbf{b}_s$ from the prior;
end for
for $s = 1$ to $R$ do
    Sample $\tau_s$ from GiG($p^*_{\tau_s}, a^*_{\tau_s}, b^*_{\tau_s}$);
    Sample $\tilde{\psi}_{ls}$ from iG($a^*_{\psi_{ls}}, b^*_{\psi_{ls}}$), then set $\psi_{ls} = \tilde{\psi}^{-1}_{ls}$, for each $l = 1, \dots, p$;
    Sample $T_{ls}$ from GiG($p^*_{\phi_{ls}}, a^*_{\tau_{ls}}, b^*_{\tau_{ls}}$), then set $\phi_{ls} = T_{ls} / \sum_{i=1}^{p} T_{is}$, for each $l = 1, \dots, p$;
    Sample $\alpha_s$ from its full conditional using a griddy Gibbs sampler (Ritter and Tanner, 1992).
end for
```

quantification remains uncovered. Huber et al. (2021) apply the SAVS method to sparsify every MCMC draw in the shrinkage step, thus allowing for parameter uncertainty. By following their approach to sparsify each draw, we allow for parameter uncertainty quantification. This yields the following estimate to obtain a sparse draw of $\mathbf{C}_{jk}$ at the mth iteration of the MCMC:

$$\bar{C}_{jk}^{(m)} = \text{sign}\left(C_{jk}^{(m)}\right) \|\mathbf{X}_j\|^{-2} \left(|C_{jk}^{(m)}| \, \|\mathbf{X}_j\|^2 - |C_{jk}^{(m)}|^{-2}\right)_+$$

with $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})'$ denoting the jth column of the matrix $\mathbf{X}$, $(x)_+ = \max(x, 0)$ and $\text{sign}(x) = 1$ for $x \geq 0$ and -1 otherwise. Considering an MCMC of length M iterations, by applying the SAVS at each iteration m, we obtain a collection of M sparse estimates of every element C_{jk}. Therefore, the posterior probability that the coefficient C_{jk} is not zero is the proportion of MCMC iterations such that the estimate $\bar{C}_{jk} \neq 0$. We define this proportion as the posterior inclusion probability (PIP):

$$\text{PIP}_{jk} = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\left(\bar{C}_{jk}^{(m)} \neq 0\right). \tag{2.2}$$

The posterior estimate of the jkth entry of $\mathbf{C}$ is set to 0 if $\text{PIP}_{jk} \leq 0.5$ and to the posterior mean of $\bar{C}_{jk}$ if $\text{PIP}_{jk} > 0.5$.

By definition, the PIP is a value between 0 and 1, where a PIP close to 0 indicates that the corresponding entry is not likely to be important. In contrast, a PIP close to 1 suggests that the inclusion of the entry is well-supported by the data. Hence, for both high and low PIPs, the decision about including an element is less uncertain, opposite to PIPs that lie around 0.5. Even though PIPs are, by nature, a quantification of uncertainty, their direct interpretation may not appear evident to the reader as the degree of certainty is not monotonous. The need for a straightforward and easily interpreted manner to quantify uncertainty about variable inclusion leads us to the following definition.

Definition 2.3.1 (PIP uncertainty index)**.** *Let us assume the posterior inclusion probability (PIP) as in Eq.* (2.2)*, the PIP uncertainty index, which takes values in* $[0, 1]$ *is defined as:*

$$\zeta_{jk} = 1 - 2\big|PIP_{jk} - 0.5\big|,$$

where ζ_{jk} close to 0 *(to* 1*) means low (high) uncertainty about the decision of setting the jkth entry of* $\mathbf{C}$ *to an exact* 0 *or not.*[3]

The following example illustrates the previous definition.

Example 1. *Let us consider three different entries: jk, jk^* and jk^{**}. We assume $PIP_{jk} = 0.49$, $PIP_{jk^*} = 0.94$, and $PIP_{jk^{**}} = 0.01$. Then, the point estimates are $\hat{C}_{jk} = \hat{C}_{jk^{**}} = 0$, whereas $\hat{C}_{jk^*} = M^{-1} \sum_{m=1}^{M} \bar{C}_{jk*}^{(m)}$. The associated uncertainty indices are $\zeta_{jk} = 0.98$, $\zeta_{jk^*} = 0.12$, and $\zeta_{jk^{**}} = 0.02$, meaning that the decision of setting $\hat{C}_{jk^{**}}$ to* 0 *and $\hat{C}_{jk^*}$ to the posterior mean have low uncertainty (as $\zeta_{jk^{**}} = 0.02$ and $\zeta_{jk^*} = 0.12$). Conversely, the decision of setting to* 0 *the entry $\hat{C}_{jk}$ is very uncertain (since $\zeta_{jk} = 0.98$).*

[3] The quantity ζ_{jk} is in strategy comparable to the Bernoulli variance. The linear transformation of PIP_{jk} allows equal weights for all probabilities, in contrast to different lengths in the range of values, pointing to different levels of variation in the Bernoulli variance. In this sense, ζ_{jk} facilitates the interpretability of the results. A sensitivity analysis about the choice of ζ_{jk} is included in Appendix A.2.

The element-wise PIP and the associated uncertainty index, ζ, provide information on what coefficients are nonzero and quantify the uncertainty about this statement. Instead, variable selection procedures are concerned with statistical techniques designed to identify and eliminate the subset of irrelevant covariates from the regression model. The PIP-based approach previously described can easily address this issue in univariate settings, but multivariate regressions call for the adoption of different methods as a prediction can have different impacts on each response variable. In fact, it may be possible that a covariate is irrelevant to predict a subset of the responses but exerts an influence on the remaining ones. In Appendix A.4.2, we report a variable selection method based on a single scalar quantity analogous to the group lasso of Chakraborty et al. (2019) and another one based on the PIP previously defined.

To address the limitation of the element-wise approach, we propose a novel index that assesses the relative importance of each covariate and is computationally inexpensive, as it relies on the output of the SAVS computed at every iteration of the MCMC.

Definition 2.3.2 (Relevance Index)**.** *The relevance index of the jth covariate, RI_j, with $j = 1, \dots, p$, is defined as:*

$$RI_j(k) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\big(N_{\bullet j}^{(m)} = kq\big), \qquad k = 0, \frac{1}{q}, \frac{2}{q}, \dots, 1, \tag{2.3}$$

where $N_{\bullet j}^{(m)} = \sum_{k=1}^{q} \mathbb{I}(\bar{\mathbf{C}}_{jk}^{(m)} \neq 0)$ is the number of nonzero entries in the jth row of $\bar{\mathbf{C}}$, and M is the length of the MCMC chain after burn-in, under a stationary regime. Notice that RI_j is a discrete distribution supported on $\mathcal{D} = \{0, 1/q, 2/q, \dots, 1\}$, with mass at each $k \in \mathcal{D}$ given by Eq. (2.3)*.*

The RI can be interpreted as representing the distribution (across MCMC) of the "relevance" of a covariate, as measured by the share of response variables on which the covariate exerts a significant impact (i.e., nonzero). This motivates the support being the discrete grid between 0 and 1, with step size $1/q$. Therefore, a strongly right-skewed distribution suggests that the covariate is irrelevant or relevant just to a small share of response variables, whereas a left-skewed distribution is typical of common predictors that impact all the responses. An important feature of the RI is that it easily allows for uncertainty quantification by means of the variance of the distribution. For instance, take two right-skewed distributions, RI_j and RI_k, characterised by different variances, $\sigma_j^2 > \sigma_k^2$. In this case, we have evidence in favour of considering both x_j and x_k irrelevant, but with higher uncertainty of this statement about x_j. Eventually, for sufficiently large variance, a binary statement about the irrelevance of the covariate may appear hazardous.

The approach to variable selection based on the RI allows the assessment of the relevance of each variable and provides full uncertainty quantification. In practice, one is often concerned with identifying and excluding irrelevant predictors. The shape and dispersion of the RI_j probability mass function intrinsically inform the impact exerted by the covariate in the overall model and the degree of uncertainty about this belief. Besides, when the practitioner needs to make a binary decision about variable selection, it is possible to summarise the information content of the RI to answer this question. A possible heuristic for choosing whether to include or exclude a covariate relies on the tail distribution (or survival function) of RI_j, as described in the following rule of thumb.

Definition 2.3.3 (Rule of thumb for variable selection)**.** *Let $\overline{sr} \in [0, 1]$ be the share of response* *variables on which the* j*th covariate is required to have a significant impact, and* $\overline{p} \in (0,1)$ *be the desired minimum probability. Then, a rule of thumb consists in excluding the* j*th covariate if the following statement is not satisfied:*

$$S_{RI_j}(\overline{sr}) = \mathbb{P}(RI_j > \overline{sr}) \geq \overline{p}, \tag{2.4}$$

where the probability is computed with respect to the (discrete) distribution of RI_j*.*

However, we remark that any measure for a binary decision will inevitably lose part of the information embedded in the RI. Therefore, when variable selection is concerned, we suggest to *jointly* interpret the output of the binary rule and the RI distribution. The following example illustrates the application of the rule of thumb to make a binary decision about variable selection.

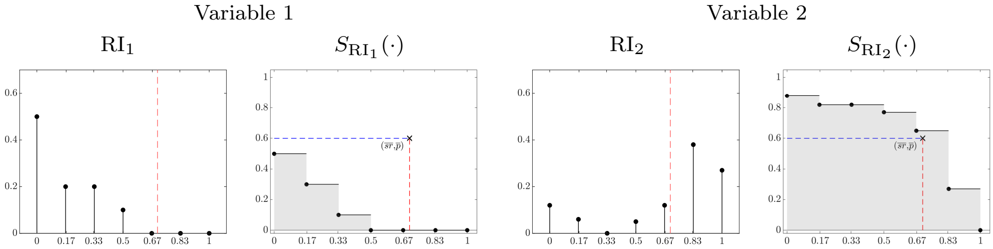


Figure 2.2: Example of RI for two fictitious covariates: variable 1 (first and second plots) and variable 2 (third and fourth). The probability mass function (first and third plots) and the survival function (second and fourth) of the respective RI are reported for each covariate. Here we used $\overline{sr} = 0.70$ (red vertical line) and $\overline{p} = 0.60$ (blue horizontal line). The area below the survival function is shaded: if the point $(\overline{sr}, \overline{p})$ is located outside the shaded area, then the covariate is to be excluded.

Example 2. *We set* $\overline{sr} = 0.70$ *and* $\overline{p} = 0.60$*; thus, we require the RI of the* j*th covariate to assign at least* 0.60 *total probability mass on or above* 0.70*. In other words, this means that not to exclude the* j*th covariate, we require the probability of being relevant for more than the* 70% *of responses to be at least* 0.60*. This example is represented in Figure 2.2, which considers two covariates, characterised by a right-skewed and a left-skewed RI. The associated survival functions are plotted together with the values of* $\overline{sr}$ *and* $\overline{p}$ *(vertical and horizontal dashed lines, respectively). The rule of thumb in Eq.* (2.4) *states that the covariate should be considered irrelevant if the point* $(\overline{sr}, \overline{p})$ *lies above the curve of the survival function (i.e., outside the shaded area), and vice versa. Therefore, in this simple example, the first covariate, which typically impacts a few response variables (as shown by the right-skewed RI, first plot), is considered irrelevant (second plot). Conversely, the other covariate has a nonzero impact on most responses (see the left-skewed RI in the third plot), and the irrelevance hypothesis is rejected for it (fourth plot). However, notice that this decision does not account for the uncertainty quantified by the RI. In this case, the variance of the RI for the second covariate is high, thus implying that the selection decision for this variable should be taken with caution. Instead, the RI for the first covariate is quite low, which suggests high confidence in the decision to exclude it from the model.*

2.4 Simulation study

We study the performance of the proposed reduced-rank model with the naïve (RRn) and the column-sharing (RRcs) priors across a range of simulation settings. Our primary objectives in conducting this simulation study are twofold: firstly, to assess the efficacy of the model in accurately estimating the rank in varied settings, including different data generating processes (DGP), and secondly, to evaluate the recovery of the coefficient matrix under distinct scenarios.

The data was generated from the multivariate linear model $\mathbf{Y} = \mathbf{X}\mathbf{C}_0 + \mathbf{E}$, where we considered correlated and uncorrelated structures on the errors and regressors of the model. The rows of $\mathbf{X}$ were independently drawn from $\mathcal{N}(0, \mathbf{\Sigma}_\mathbf{X})$, with $\mathbf{\Sigma}_\mathbf{X} = \mathbf{I}_p$ for independent regressors, and in the dependent case, the off-diagonal entries of $\mathbf{\Sigma}_\mathbf{X}$ are set to 0.5. The rows of $\mathbf{E}$ were drawn from a zero mean multivariate normal distribution; under the assumption of uncorrelated errors, the covariance matrix is diagonal with elements sampled from $\mathcal{U}(0.5, 1.75)$; for correlated errors, we consider the compound symmetry as in $\mathbf{X}$. We work with centred responses and exclude the intercept term for simplicity.

Recalling the decomposition of the matrix of coefficients $\mathbf{C}_0$ into the product of two matrices $\mathbf{A}_0$ and $\mathbf{B}_0$, the entries of both matrices are generated from the standard Gaussian, and their dimensions depend on the true rank $r_0 < R$. We consider two cases for the DGP of matrix $\mathbf{B}_0$: non-sparse DGP, where the number of nonzero rows is equal to p, and sparse DGP, where the number of nonzero rows is $p^* < p$. Furthermore, our coefficient matrix estimation method is not limited to low-rank structures. We have also tested our approach on an additional "random zeros" DGP, where a share of entries z of the matrix $\mathbf{C}$ is randomly set to zero (Figures 2.4e and 2.4f). The performance evaluation of the estimator $\hat{\mathbf{C}}$ of the coefficient matrix was conducted by considering the mean squared error, $\text{MSE} = \|\hat{\mathbf{C}} - \mathbf{C}_0\|_F^2/(pq)$, where $\|\cdot\|_F$ is the Frobenius norm.[4]

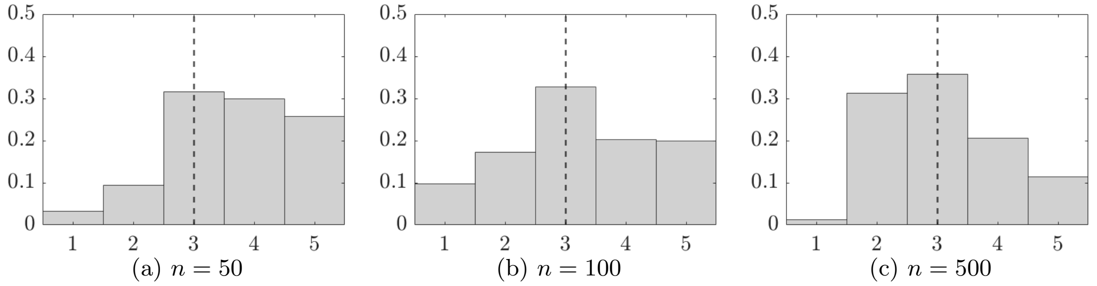
The simulation study assumes correctly specified multivariate normal errors, whether they be correlated or uncorrelated. However, in practice, errors may follow heavier-tailed distributions such as the multivariate Student-t distribution. In such cases, certain aspects of our method may be affected by extreme values, potentially leading to an increased variance or biased inference, which could significantly impact posterior concentration and variable selection. Exploring robustness to model misspecification would be a valuable direction for future work.

2.4.1 Simulation results

The RRn tends to underestimate the value of the rank across the majority of settings, particularly when the number of covariates and responses increases. As more data become available, the posterior variance of the rank parameter shrinks, resulting in the concentration of the posterior distribution. Therefore, the MCMC algorithm is further restricted in its ability to explore alternative values, resulting in a more pronounced underestimated rank. However, in all these cases, the estimated coefficient matrix $\hat{\mathbf{C}}$ is quite close to the true $\mathbf{C}$ in all settings, succeeding in identifying the entries with high magnitude, and the estimates improve as the sample size grows.

After having shown the performance of the RRn, we provide evidence of the results for the RRcs approach. The column-sharing exhibits superior performance to the naïve parametrization in terms of the MSE, convergence to the true rank and computational resources. The posterior distribution of u tends to concentrates around the true value as n increases (see Figure 2.3). However, when $\mathbf{B}$ is sparse, the posterior distribution tends to put more mass on ranks smaller than r_0, resulting in a slight underestimation of the rank as the dimensionality of (q, p) increases. A possible motivation for these results is that having zero rows in $\mathbf{B}$ implies sparsity in $\mathbf{C}$ as well, which induces our method to prefer approximate $\mathbf{C}$ with a small r than to introduce additional parameters (higher

[4] See Appendix A.3.2 for a CODA analysis for the posterior distribution of the rank and the estimates of the matrix $\mathbf{C}$.



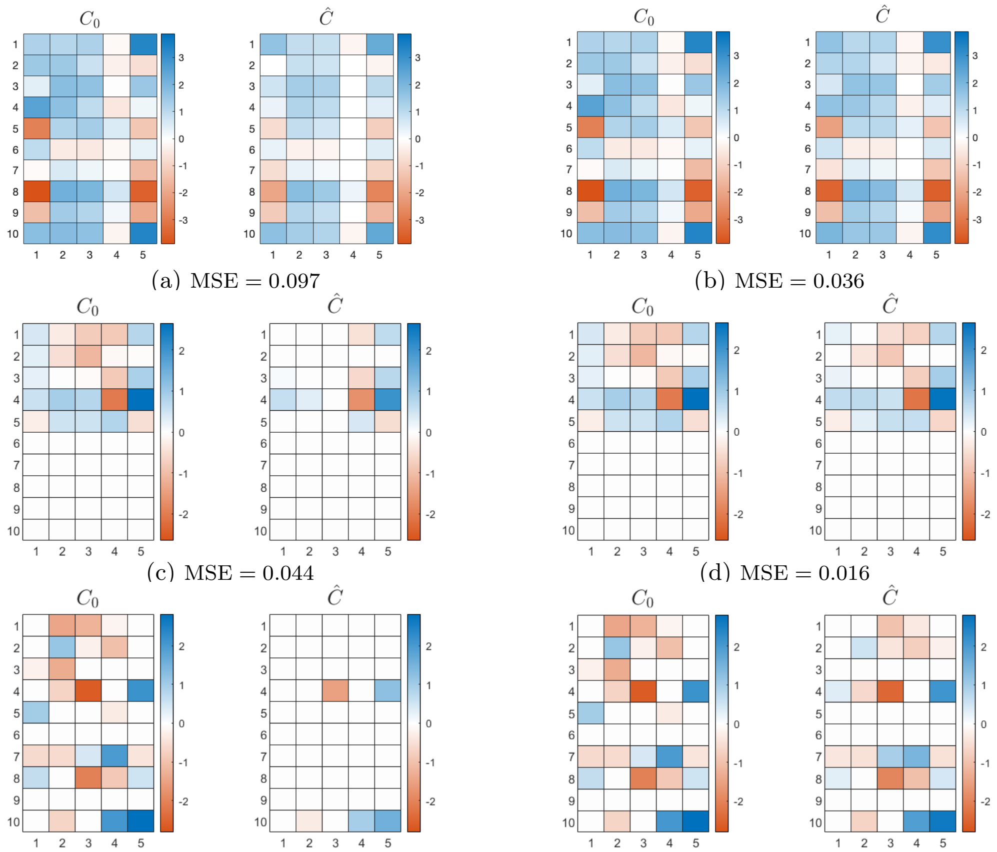
(a) $n = 50$ (b) $n = 100$ (c) $n = 500$

Figure 2.3: Posterior distribution of the rank in the non-sparse simulation setting with $(q, p) = (5, 10)$ and true rank $r_0 = 3$ (dashed line) for different sample sizes, $n \in \{50, 100, 500\}$.

r). This feature of the model can be interpreted as favouring more parsimonious parametrizations. Notice that the underestimation of r in these cases has little impact, as $\hat{\mathbf{C}}$ is nonetheless close to $\mathbf{C}_0$ (Figure 2.4). RRcs consistently achieves a better mixing of the MCMC chain compared to RRn. Notably, in both parameterizations, sparsity contributes to the improved mixing.[5]

The performance of RRcs deteriorates more rapidly for a fixed n as the number of responses q increases compared to the number of covariates p. Conversely, the performance decays slower as p increases and q remains unchanged. A change of p to p' means $(p' - p)R$ more parameters to estimate. However, when q increases to q', the number of elements in $\mathbf{A}$ and $\mathbf{\Sigma}$ is directly affected, changing by $(q' - q)R$ and $(q' - q)n$, respectively. The crucial point is that q represents the maximum rank in our context ($q \leq p$). Therefore, if q increases, it increases the number of the mixture prior components, their respective weights, and the dimension of $\mathbf{B}$. In the settings where the DGP is high-dimensional and non-sparse, both RRn and RRcs exhibit performance decline in terms of the MSE. This outcome is expected, as the estimation involves a large number of coefficients, and imposing sparsity leads to some being set to zero, thereby increasing the overall error.

[5] Summary results comparing RRcs against RRn are included in Appendix A.3



(a) MSE = 0.097

(b) MSE = 0.036

(c) MSE = 0.044

(d) MSE = 0.016

(e) MSE = 0.100

(f) MSE = 0.051

Figure 2.4: True ($\mathbf{C}_0$) and estimated ($\hat{\mathbf{C}}$) coefficient matrix. Data generated as described in Section 2.4, with $(q, p) = (5, 10)$ and $n = 100$; non-sparse matrix $\mathbf{B}$ ($p^* = p$) and $r_0 = 3$ (top); sparse matrix $\mathbf{B}$ ($p^* = 5$) and $r_0 = 3$ (middle); $z = 50\%$ of entries of $\mathbf{C}$ set to 0 (bottom). Results for the RRn (left - (a),(c),(e)) and RRcs (right - (b),(d)(f)), together with the MSE.

We emphasise that our analysis relies on a sparse estimate of $\mathbf{C}$, obtained using the SAVS method described in Section 2.3.1. This estimation approach sets some entries to exact zeros, facilitating the variable selection. This can be considered a binary classification problem, wherein the positive class encompasses the nonzero entries (representing significant coefficients). In contrast, the null entries (indicating irrelevant coefficients) belong to the negative category. Consequently, our task involves identifying the position of zero and nonzero coefficients.

To evaluate the performance of this classification task, we employ the Matthews correlation coefficient (MCC), which is a more reliable statistical measure compared to commonly used metrics such as F_1 score and accuracy (Chicco and Jurman, 2020). One notable advantage of MCC, particularly relevant to our study, is its robustness in scenarios where one class contains significantly more samples than the other, thus addressing the issue of imbalanced datasets. Our synthetic data for $\mathbf{C}_0$ repeatedly exhibits such imbalanced characteristics: in the non-sparse DGP when the majority of entries, if not all, deviate from zero, in the sparse DGP when only a few (or many) rows contain zero entries, and in the random zeros DGP when the percentage of zeros is low (or high). This varying distribution of zeros in different scenarios allows us to assess the classification performance of our method under varying levels of sparsity and imbalance.

The classification model predicts the class for each data instance, assigning a predicted label (positive or negative) to each sample. Depending on their actual class and their forecasted class,

DGP		**RRn**			**RRcs**		
		MCC	**TPR**	**FNR**	**MCC**	**TPR**	**FNR**
Standard	$p^* = 10$	—	0.84	0.16	—	0.94	0.06
Sparse	$p^* = 2$	0.00	0.00	1.00	0.81	0.70	0.30
	$p^* = 5$	0.53	0.44	0.56	0.78	0.76	0.24
	$p^* = 8$	0.13	0.08	0.92	0.54	0.68	0.32
	$p^* = 9$	0.38	0.62	0.38	0.83	0.96	0.04
Random zeros	$z = 0.20$	0.13	0.08	0.92	0.73	0.85	0.15
	$z = 0.50$	0.33	0.20	0.80	0.73	0.80	0.20
	$z = 0.80$	0.00	0.00	1.00	0.74	0.60	0.40
	$z = 0.90$	0.00	0.00	1.00	0.88	0.80	0.20

Table 2.1: Measures of association between the true coefficient matrix $\mathbf{C}_0$ and the sparse estimate $\hat{\mathbf{C}}$ in the setting $(q, p) = (5, 10)$, $n = 100$, $r_0 = 3$ in the standard and sparse scenarios. p^* represents the number of nonzero rows in $\mathbf{B}$ (and consequently in $\mathbf{C}_0$), and z is the proportion of randomly allocated zero entries in $\mathbf{C}_0$. For RRn and RRcs, we report the MCC, TPR and FNR.

every sample is categorised into one of the following cases: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). The MCC is given by:

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}} \in [-1, 1]. \tag{2.5}$$

A value of -1 indicates poor performance, while a value of 1 represents the highest level of performance. MCC produces a high-quality score only if the prediction obtained good results in all categories $(\text{TP}, \text{TN}, \text{FP}, \text{FN})$, proportionally both to the size of positive and negative elements in the dataset. Instead, the measure is undefined when any of the factors in the denominator is 0, and specific mathematical reasoning should be considered. For instance, if $\mathbf{C}_0$ comprises only nonzero (zero) entries, and they are all correctly identified in $\hat{\mathbf{C}}$, then $\text{TN} = 0$ $(\text{TP} = 0)$ and $\text{FN} = 0$ $(\text{FP} = 0)$, resulting in an undefined MCC. Nonetheless, the classifier successfully identifies all samples in this case and achieves a perfect score of 1. If all samples belong to the same class and are all incorrectly predicted, then $\text{MCC} = -1$. We are left to study the cases when mixed samples are categorised into the same class or homogeneous samples are allocated to mixed classes. In either case, the correlation coefficient can be approximated by 0 (see Chicco and Jurman, 2020, for a detailed description).

We report the MCC, the true positive rate (TPR = TP/(TP+FN)) of correctly identified nonzero entries, and the false negative rate (FNR = FN/(TP+FN)) of incorrectly identified zero entries across different simulation settings in Table 2.1. The non-negative MCCs obtained by our methods suggest a good performance in the estimation of the coefficient matrix, favouring once more the RRcs parametrization over RRn. Having attained MCCs close to 1, we show the robustness of the former method in accurately approximating the $\mathbf{C}$ matrix while incorporating sparsity in its estimation. The standard DGP illustrates the need for the corrections to the formula of the MCC when undefined since $\mathbf{C}_0$ consisted of only nonzero entries; meanwhile, the estimated $\hat{\mathbf{C}}$ had mixed values. The algorithm established 16% of the entries as 0, thus allowing for a sparse model with enhanced interpretability.

The code for the RRn and RRcs algorithms has been implemented in MATLAB 2021a, and run on a MacBook Pro M1 2020 computer with 8 GB RAM. The average computational time for running 100 iterations of the RRn for a model with dimensions $q = 5$, $p = 10$, and $n = 100$ is approximately 0.64 seconds, while for the RRcs, it is 0.15 seconds.

2.4.2 Comparison to other methods

The performance of the proposed mixture prior RRcs is comparable to other state-of-the-art methodologies in reduced-rank regression, as we showcase in this section. We evaluate our approach against the frequentist methods of Chen et al. (2013) and She and Chen (2017). The former utilises an adaptive nuclear norm penalisation approach (ANN), where the rank is estimated by the threshold of singular values. The latter proposes a robust reduced-rank regression approach (RRRR), which requires the user to choose the optimal rank, a task that is achieved through a suggested criterion.

The data was generated as outlined in Section 2.4. For each configuration, we conducted 50 replications of the experiment.

Consequently, the results presented reflect the average estimated rank, the share of replications where the rank was correctly selected, the MSE across these repetitions, and an additional error metric $\varrho(\mathbf{C}) = \|\hat{\mathbf{C}}(\hat{\mathbf{C}}'\hat{\mathbf{C}})^{+}\hat{\mathbf{C}}' - \mathbf{C}(\mathbf{C}'\mathbf{C})^{+}\mathbf{C}'\|_2$. Notably, our method achieves similar outcomes as both ANN and RRRR, particularly outperforming them when the sample size is $n = 50$ and the dimensions of q and p increase (see Table 2.2 and Table 2.3). While the hit rate for estimating the exact value of the rank tends to be low for all methods considered, our approach is able to provide an exact posterior estimate for r. Indeed, we find that with our method the true rank value tends to have substantial posterior mass. The metric $\varrho(\mathbf{C})$ of our method outperforms the one of its competitors or has comparable performance in almost all the simulated scenarios. We defer additional results to the Appendix.

					$\mathbf{\Sigma}_{ind}$			$\mathbf{\Sigma}_{corr}$	
(q,p)	r_0	**X**	**Measure**	**ANN**	**RRRR**	**RRcs**	**ANN**	**RRRR**	**RRcs**
(5,15)	3	$\mathbf{X}_{ind}$	$\hat{r}$	2.320	2.420	1.860	2.420	2.460	1.720
			$r_{\%}$	0.400	0.480	0.160	0.480	0.500	0.180
			MSE($\mathbf{C}$)	0.030	0.030	0.161	0.027	0.028	0.189
			$\varrho(\mathbf{C})$	0.745	0.709	0.984	0.674	0.670	0.988
		$\mathbf{X}_{corr}$	$\hat{r}$	2.280	2.240	1.820	2.020	1.940	1.580
			$r_{\%}$	0.340	0.300	0.180	0.180	0.140	0.200
			MSE($\mathbf{C}$)	0.043	0.047	0.251	0.046	0.051	0.173
			$\varrho(\mathbf{C})$	0.779	0.804	0.995	0.898	0.917	0.986
	5	$\mathbf{X}_{ind}$	$\hat{r}$	3.160	3.280	2.380	3.240	3.240	2.100
			$r_{\%}$	0.000	0.020	0.240	0.000	0.000	0.120
			MSE($\mathbf{C}$)	0.036	0.035	0.376	0.031	0.034	0.394
			$\varrho(\mathbf{C})$	1.000	0.984	0.380	1.000	1.000	0.419
		$\mathbf{X}_{corr}$	$\hat{r}$	2.700	2.940	2.400	2.900	2.920	2.120
			$r_{\%}$	0.000	0.000	0.300	0.000	0.020	0.180
			MSE($\mathbf{C}$)	0.072	0.063	0.470	0.063	0.065	0.463
			$\varrho(\mathbf{C})$	1.000	1.000	0.440	1.000	0.987	0.442
(5,50)	3	$\mathbf{X}_{ind}$	$\hat{r}$	0.160	5.000	1.540	0.240	5.000	1.260
			$r_{\%}$	0.000	0.000	0.100	0.020	0.000	0.060
			MSE($\mathbf{C}$)	1.010	144.533	0.148	0.667	17.328	0.183
			$\varrho(\mathbf{C})$	1.000	1.000	1.000	0.998	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	0.380	5.000	1.720	0.300	5.000	1.360
			$r_{\%}$	0.000	0.000	0.140	0.000	0.000	0.060
			MSE($\mathbf{C}$)	2.197	32.051	0.165	1.181	32.620	0.209
			$\varrho(\mathbf{C})$	1.000	1.000	1.000	1.000	1.000	1.000
	5	$\mathbf{X}_{ind}$	$\hat{r}$	0.060	5.000	2.360	0.020	5.000	2.040
			$r_{\%}$	0.000	1.000	0.280	0.000	1.000	0.120
			MSE($\mathbf{C}$)	1.849	813.710	0.309	0.917	102.402	0.298
			$\varrho(\mathbf{C})$	1.000	0.987	0.901	1.000	0.991	0.912
		$\mathbf{X}_{corr}$	$\hat{r}$	0.160	5.000	2.560	0.040	5.000	2.340
			$r_{\%}$	0.000	1.000	0.380	0.000	1.000	0.260
			MSE($\mathbf{C}$)	6.032	65.801	0.371	0.930	52.111	0.336
			$\varrho(\mathbf{C})$	1.000	0.992	0.904	1.000	0.988	0.925
(10,50)	3	$\mathbf{X}_{ind}$	$\hat{r}$	1.360	10.000	1.000	1.820	10.000	1.000
			$r_{\%}$	0.220	0.000	0.000	0.460	0.000	0.000
			MSE($\mathbf{C}$)	8.372	236.725	0.222	16.348	197.125	0.251
			$\varrho(\mathbf{C})$	0.962	1.000	1.000	0.943	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	1.260	10.000	1.040	1.600	10.000	1.000
			$r_{\%}$	0.220	0.000	0.000	0.220	0.000	0.000
			MSE($\mathbf{C}$)	36.481	13.903	0.300	3.239	43.166	0.281
			$\varrho(\mathbf{C})$	0.977	1.000	1.000	0.971	1.000	0.998
	5	$\mathbf{X}_{ind}$	$\hat{r}$	0.080	10.000	1.040	0.020	10.000	1.000
			$r_{\%}$	0.000	0.000	0.000	0.000	0.000	0.000
			MSE($\mathbf{C}$)	1.001	88.157	0.522	0.917	51.975	0.511
			$\varrho(\mathbf{C})$	1.000	1.000	0.998	1.000	1.000	0.999
		$\mathbf{X}_{corr}$	$\hat{r}$	0.480	10.000	1.140	0.500	10.000	1.020
			$r_{\%}$	0.000	0.000	0.000	0.000	0.000	0.000
			MSE($\mathbf{C}$)	1.363	64.436	0.559	2.441	30.059	0.618
			$\varrho(\mathbf{C})$	1.000	1.000	1.000	1.000	1.000	0.999

Table 2.2: Comparison of the estimated rank ($\hat{r}$), share of correctly estimated rank ($r_{\%}$), mean squared error of $\mathbf{C}$ (MSE($\mathbf{C}$)), and $\varrho(\mathbf{C}) = \|\hat{\mathbf{C}}(\hat{\mathbf{C}}'\hat{\mathbf{C}})^{+}\hat{\mathbf{C}}' - \mathbf{C}(\mathbf{C}'\mathbf{C})^{+}\mathbf{C}'\|_2$ obtained by RRcs against ANN (Chen et al., 2013) and RRRR (She and Chen, 2017) for different values of (q, p) and true rank r_0. In all settings, $n = 50$, and the DGP is sparse with $p^* = 5$ if $p = 15$, while $p^* = 10$ if $p = 50$. We present the average estimates over 50 repetitions for independent errors ($\mathbf{\Sigma}_{ind}$), correlated errors ($\mathbf{\Sigma}_{corr}$), independent regressors ($\mathbf{X}_{ind}$), and correlated regressors ($\mathbf{X}_{corr}$).

				$\boldsymbol{\Sigma}_{ind}$			$\boldsymbol{\Sigma}_{corr}$		
(q,p)	r_0	**X**	**Measure**	**ANN**	**RRRR**	**RRcs**	**ANN**	**RRRR**	**RRcs**
(5,15)	3	$\mathbf{X}_{ind}$	$\hat{r}$	2.940	2.880	2.060	2.940	2.960	1.960
			$r_{\%}$	0.940	0.880	0.120	0.940	0.960	0.160
			MSE($\mathbf{C}$)	0.023	0.026	0.799	0.021	0.023	0.675
			$\varrho(\mathbf{C})$	0.260	0.292	1.000	0.228	0.223	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	2.860	2.880	2.160	2.940	2.960	1.860
			$r_{\%}$	0.860	0.840	0.120	0.940	0.920	0.060
			MSE($\mathbf{C}$)	0.042	0.045	0.697	0.040	0.043	0.748
			$\varrho(\mathbf{C})$	0.340	0.357	1.000	0.296	0.319	1.000
	5	$\mathbf{X}_{ind}$	$\hat{r}$	4.000	4.360	2.540	3.940	4.380	2.600
			$r_{\%}$	0.000	0.380	0.360	0.000	0.420	0.360
			MSE($\mathbf{C}$)	0.058	0.037	1.594	0.051	0.033	1.624
			$\varrho(\mathbf{C})$	1.000	0.743	0.880	1.000	0.730	0.888
		$\mathbf{X}_{corr}$	$\hat{r}$	3.940	4.240	2.580	3.880	4.160	2.600
			$r_{\%}$	0.000	0.320	0.380	0.000	0.280	0.400
			MSE($\mathbf{C}$)	0.088	0.070	1.633	0.080	0.064	1.724
			$\varrho(\mathbf{C})$	1.000	0.806	0.891	1.000	0.855	0.898
(5,50)	3	$\mathbf{X}_{ind}$	$\hat{r}$	0.940	5.000	1.280	1.620	5.000	1.160
			$r_{\%}$	0.260	0.000	0.020	0.360	0.000	0.000
			MSE($\mathbf{C}$)	4.957	17.005	1.285	98.964	52.488	1.535
			$\varrho(\mathbf{C})$	0.953	1.000	1.000	0.954	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	0.800	5.000	1.860	1.620	5.000	1.400
			$r_{\%}$	0.200	0.000	0.020	0.420	0.000	0.000
			MSE($\mathbf{C}$)	15.744	155.578	1.395	303602.928	34.301	1.653
			$\varrho(\mathbf{C})$	0.969	1.000	1.000	0.973	1.000	1.000
	5	$\mathbf{X}_{ind}$	$\hat{r}$	0.000	5.000	2.600	0.160	5.000	2.520
			$r_{\%}$	0.000	1.000	0.400	0.000	1.000	0.380
			MSE($\mathbf{C}$)	5.179	228.422	2.586	5.526	580.312	2.741
			$\varrho(\mathbf{C})$	1.000	0.972	0.990	1.000	0.952	0.990
		$\mathbf{X}_{corr}$	$\hat{r}$	0.180	5.000	2.600	0.420	5.000	2.600
			$r_{\%}$	0.000	1.000	0.400	0.000	1.000	0.400
			MSE($\mathbf{C}$)	12.083	36.098	3.064	463.520	128.832	3.293
			$\varrho(\mathbf{C})$	1.000	0.962	0.989	1.000	0.978	0.989
(10,50)	3	$\mathbf{X}_{ind}$	$\hat{r}$	3.000	10.000	1.260	3.200	10.000	1.220
			$r_{\%}$	1.000	0.000	0.040	0.800	0.000	0.080
			MSE($\mathbf{C}$)	1391.891	494.411	1.570	25.856	99.000	1.711
			$\varrho(\mathbf{C})$	0.730	1.000	1.000	0.782	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	3.000	10.000	1.240	3.260	10.000	1.180
			$r_{\%}$	1.000	0.000	0.020	0.740	0.000	0.000
			MSE($\mathbf{C}$)	298.131	76.197	2.062	25.326	84.624	1.884
			$\varrho(\mathbf{C})$	0.806	1.000	1.000	0.839	1.000	1.000
	5	$\mathbf{X}_{ind}$	$\hat{r}$	2.420	10.000	1.280	4.060	10.000	1.060
			$r_{\%}$	0.480	0.000	0.000	0.500	0.000	0.000
			MSE($\mathbf{C}$)	25.933	171.279	3.548	18.649	43.215	4.064
			$\varrho(\mathbf{C})$	0.922	1.000	1.000	0.919	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	2.200	10.000	1.200	2.680	10.000	1.220
			$r_{\%}$	0.400	0.000	0.000	0.380	0.000	0.000
			MSE($\mathbf{C}$)	10.417	47.036	3.913	545.300	67.874	4.231
			$\varrho(\mathbf{C})$	0.953	1.000	1.000	0.948	1.000	1.000

Table 2.3: Comparison of the estimated rank ($\hat{r}$), share of correctly estimated rank ($r_{\%}$), mean squared error of $\mathbf{C}$ (MSE($\mathbf{C}$)), and $\varrho(\mathbf{C}) = \|\hat{\mathbf{C}}(\hat{\mathbf{C}}'\hat{\mathbf{C}})^{+}\hat{\mathbf{C}}' - \mathbf{C}(\mathbf{C}'\mathbf{C})^{+}\mathbf{C}'\|_2$ obtained by RRcs against ANN (Chen et al., 2013) and RRRR (She and Chen, 2017) for different values of (q, p) and true rank r_0. In all settings, $n = 50$, and the DGP is non-sparse with $p^* = p$. We present the average estimates over 50 repetitions for independent errors ($\boldsymbol{\Sigma}_{ind}$), correlated errors ($\boldsymbol{\Sigma}_{corr}$), independent regressors ($\mathbf{X}_{ind}$), and correlated regressors ($\mathbf{X}_{corr}$).

2.5 Applications

This section showcases the efficacy of our proposed method through its practical implementation on two actual datasets. By providing empirical demonstrations of the method, we illustrate its value and effectiveness in addressing real-life situations.

2.5.1 Chemical composition of tobacco

The dataset on the chemical composition of $n = 25$ tobacco leaf samples, taken from Anderson and Bancroft (1952), illustrates how our proposed methodology produces comparable results as those obtained in the literature with further advantages addressed subsequently.

Tobacco leaves are made up of organic and inorganic chemical constituents, and the typical interest is investigating the relationship between certain constituents. The $p = 6$ covariates are per cent nitrogen, per cent chlorine, per cent potassium, per cent phosphorus, per cent calcium, and per cent magnesium. We also include an intercept term. The $q = 3$ response variables are the rate of cigarette burn in inches per 1,000 seconds, the percentage of sugar in the leaf, and the percentage of nicotine in the leaf.

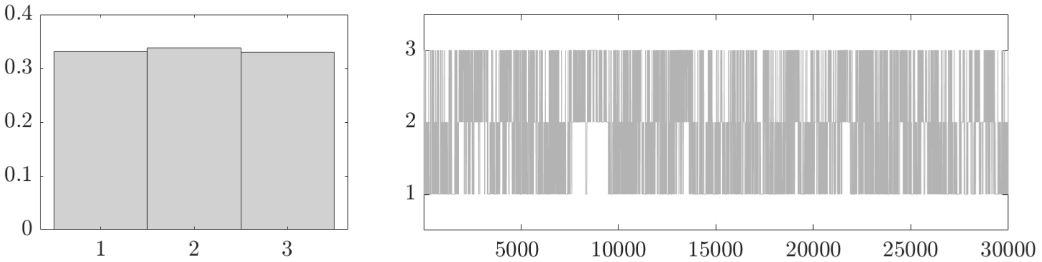
Under the column-sharing parametrization, the posterior distribution for the rank, as illustrated in Figure 2.5, shows no significant difference between any of the three values, identifying a potential uniform posterior distribution for r. Our result agrees with the findings of Izenman (2008), who uses the rank trace method. This procedure first estimates the coefficient matrix for each rank that minimises a weighted sum of squares criterion and the residual covariance matrix. Then, the rank is gradually increased, and the entries in both matrices will change significantly until the true rank is reached, where the matrices will stabilise. The change in the residual covariance matrix at each increment of r is plotted against the change in $\mathbf{C}$ in a scatterplot. Finally, the rank of $\mathbf{C}$ is assessed as the smallest rank for which the differences are close to 0. The rank-trace plot applied to the tobacco dataset shows that the rank-1 or rank-2 solutions have no discernible difference between them and the full-rank solution. The main drawback of this method is that conclusions on the effective dimensionality of the multivariate regression involve subjective judgement and visual interpretation of the rank trace plot (see Izenman, 2008, for more details).

To statistically validate our hypothesis that the posterior distribution of the rank is uniform, we conduct a Pearson's chi-squared test for goodness of fit, obtaining a p-value of 0.1595, thus implying the non-rejection of the null hypothesis. Formal statistical testing for uniformity yields a more objective result as opposed to relying purely on visual analysis. Moreover, our method requires the estimation of one model, whereas the rank trace performs the estimation of three distinct models. Besides the estimates of the rank and the coefficient matrix, we provide in conjunction uncertainty quantification as an objective means of analysis, moving away from reliance on subjective judgement.

Regarding the estimated coefficient matrix for each rank, our algorithm produces consistent estimates compared to those reported in Izenman (2008), with the aforementioned advantages.

2.5.2 COMBO-17 galaxy photometric data

The second dataset consists of a subset of a public catalogue of astronomical objects, COMBO-17 (Classifying Objects by Medium-Band Observations in 17 filters), a project of international collaboration aimed at exploring the evolution of galaxies (Wolf et al., 2004). The present dataset



(a) Posterior distribution of u. (b) MCMC chain of u.

Figure 2.5: Tobacco dataset: posterior distribution (panel a) and MCMC chain after burn-in (panel b) of the rank u for the coefficient matrix $\mathbf{C}$.

is utilised herein to illustrate the proposed variable selection procedure and the associated uncertainty.[6] The methodology serves as a reliable means of informing decision-making regarding selecting covariates, offering both suggestions and quantifying uncertainty in such selections.

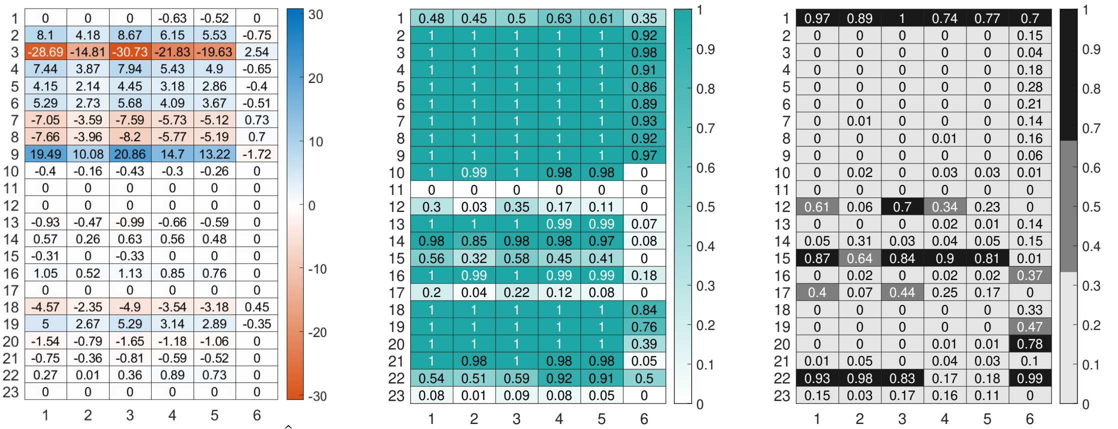
The original dataset consists of $63,501$ objects in the area of the sky named Chandra Deep Field South with brightness measurements in 17 passbands from 350 to 930 nm. We restrict the analysis to $3,438$ objects, all classified as "Galaxies" by Wolf et al. (2004), and with no missing values for any of the 65 variables. The measurement errors and five redundant variables were omitted, resulting in a total of 29 variables divided into $p = 23$ covariates and $q = 6$ responses, as done in Izenman (2008). Regarding the covariates, 10 variables correspond to the absolute magnitudes of the galaxy in 10 bands, while the remaining variables are the observed brightness in 13 bands across the range $420-915$ nm. Meanwhile, the responses are the total R-band magnitude, the aperture difference of the R-band, the central surface brightness in the R-band, two redshift estimates, and the reduced chi-squared value of the best-fitting template galaxy spectrum.

Given that the whole subset of galaxies consists of a considerable number of $n = 3,438$ observations, the level of estimation uncertainty is likely to be quite small. Therefore, motivated by the intention of highlighting the ability to quantify the uncertainty of the proposed BRECS method, we apply it to a sub-sample of $n = 500$ observations randomly selected from the entire subset of data.The rank selection, sparse estimation, and variable selection procedures are also repeated for the full sample and for other randomly chosen sub-samples of size $n = 500$ and $n = 1,500$ (see Appendix A.4.1). In line with expectations, we find that the use of larger samples reduces the uncertainty, but the key insights about the advantages of using the BRECS method are unaltered.

The posterior distribution of the rank is right-skewed and achieves the maximum (MAP) at $\hat{u} = 2$. Considering all the $3,438$ observations, the posterior distribution is more concentrated around the same maximum point (see Appendix A.4).

The effect of the covariates on the responses emphasises the importance of estimating the matrix $\mathbf{C}$. For any pair (jk), a zero entry $\hat{C}_{jk} = 0$ means that there is no association between the jth covariate and the kth response. Therefore, zero entries facilitate interpretation as the nonzero rows of $\mathbf{C}$ identify the covariates that influence at least one response. We obtain a sparse estimate of C_{jk} by applying the SAVS algorithm at each iteration of the MCMC, computing its posterior inclusion probability PIP_{jk}, and set the element to 0 if $\text{PIP}_{jk} \leq 0.5$ or to its posterior mean otherwise, as described in Section 2.3. The matrix of PIPs takes values in $[0, 1]$, and the elements of $\mathbf{C}$ with equal or less than 50% probability of inclusion are set to 0. We emphasise that even though some entries of the sparse estimate $\hat{\mathbf{C}}$ are 0, the probability of including them is not exactly 0. At first

[6] An additional forecasting exercise is presented in Appendix A.4.3.



	1	2	3	4	5	6
1	0	0	0	-0.63	-0.52	0
2	8.1	4.18	8.67	6.15	5.53	-0.75
3	-28.69	-14.81	-30.73	-21.83	-19.63	2.54
4	7.44	3.87	7.94	5.43	4.9	-0.65
5	4.15	2.14	4.45	3.18	2.86	-0.4
6	5.29	2.73	5.68	4.09	3.67	-0.51
7	-7.05	-3.59	-7.59	-5.73	-5.12	0.73
8	-7.66	-3.96	-8.2	-5.77	-5.19	0.7
9	19.49	10.08	20.86	14.7	13.22	-1.72
10	-0.4	-0.16	-0.43	-0.3	-0.26	0
11	0	0	0	0	0	0
12	0	0	0	0	0	0
13	-0.93	-0.47	-0.99	-0.66	-0.59	0
14	0.57	0.26	0.63	0.56	0.48	0
15	-0.31	0	-0.33	0	0	0
16	1.05	0.52	1.13	0.85	0.76	0
17	0	0	0	0	0	0
18	-4.57	-2.35	-4.9	-3.54	-3.18	0.45
19	5	2.67	5.29	3.14	2.89	-0.35
20	-1.54	-0.79	-1.65	-1.18	-1.06	0
21	-0.75	-0.36	-0.81	-0.59	-0.52	0
22	0.27	0.01	0.36	0.89	0.73	0
23	0	0	0	0	0	0

(a) Sparse estimate $\hat{\mathbf{C}}$.

	1	2	3	4	5	6
1	0.48	0.45	0.5	0.63	0.61	0.35
2	1	1	1	1	1	0.92
3	1	1	1	1	1	0.98
4	1	1	1	1	1	0.91
5	1	1	1	1	1	0.86
6	1	1	1	1	1	0.89
7	1	1	1	1	1	0.93
8	1	1	1	1	1	0.92
9	1	1	1	1	1	0.97
10	1	0.99	1	0.98	0.98	0
11	0	0	0	0	0	0
12	0.3	0.03	0.35	0.17	0.11	0
13	1	1	1	0.99	0.99	0.07
14	0.98	0.85	0.98	0.98	0.97	0.08
15	0.56	0.32	0.58	0.45	0.41	0
16	1	0.99	1	0.99	0.99	0.18
17	0.2	0.04	0.22	0.12	0.08	0
18	1	1	1	1	1	0.84
19	1	1	1	1	1	0.76
20	1	1	1	1	1	0.39
21	1	0.98	1	0.98	0.98	0.05
22	0.54	0.51	0.59	0.92	0.91	0.5
23	0.08	0.01	0.09	0.08	0.05	0

(b) PIP of $\mathbf{C}_{jk}$.

	1	2	3	4	5	6
1	0.97	0.89	1	0.74	0.77	0.7
2	0	0	0	0	0	0.15
3	0	0	0	0	0	0.04
4	0	0	0	0	0	0.18
5	0	0	0	0	0	0.28
6	0	0	0	0	0	0.21
7	0	0.01	0	0	0	0.14
8	0	0	0	0.01	0	0.16
9	0	0	0	0	0	0.06
10	0	0.02	0	0.03	0.03	0.01
11	0	0	0	0	0	0
12	0.61	0.06	0.7	0.34	0.23	0
13	0	0	0	0.02	0.01	0.14
14	0.05	0.31	0.03	0.04	0.05	0.15
15	0.87	0.64	0.84	0.9	0.81	0.01
16	0	0.02	0	0.02	0.02	0.37
17	0.4	0.07	0.44	0.25	0.17	0
18	0	0	0	0	0	0.33
19	0	0	0	0	0	0.47
20	0	0	0	0.01	0.01	0.78
21	0.01	0.05	0	0.04	0.03	0.1
22	0.93	0.98	0.83	0.17	0.18	0.99
23	0.15	0.03	0.17	0.16	0.11	0

(c) PIP uncertainty index ζ_{jk}.

Figure 2.6: Sparse estimate $\hat{\mathbf{C}}$ of the coefficient matrix $\mathbf{C}$ of the linear regression model with the galaxy dataset of $n = 500$ observations (panel a), the uncertainty about this estimation through the posterior inclusion probabilities (panel b), and the PIP uncertainty index, in a grey-colour scale according to low ($\zeta_{jk} \leq 1/3$), medium ($1/3 < \zeta_{jk} \leq 2/3$), or high ($\zeta_{jk} > 2/3$) uncertainty (panel c).

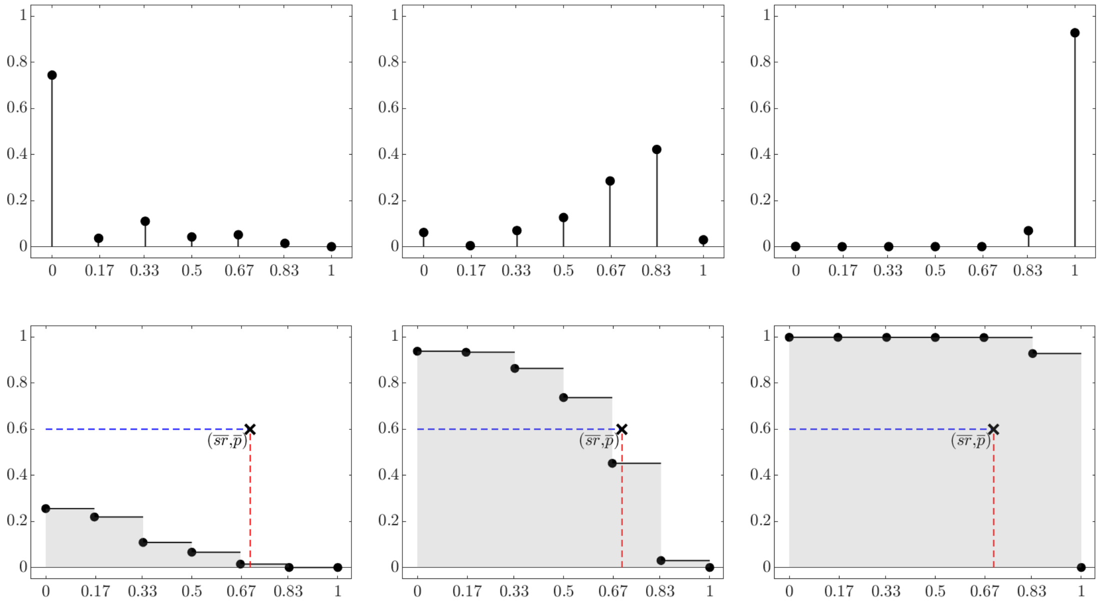


Figure 2.7: Probability mass function (top) and survival function (bottom) of RI for covariates 17 (left), 22 (centre) and 7 (right). The x-axis represents the share of nonzero elements with the probabilities on the y-axis. If we set $\overline{sr} = 0.70$ (red vertical line) and $\overline{p} = 0.60$ (blue horizontal line), then covariates 17 and 22 are to be excluded, and we include covariate 7.

inspection, the 12th covariate, the observed brightness of the galaxy in the corresponding band, is to be ruled out of the model since its coefficients are only zeros (see Figure 2.6a). However, not all of the PIPs of covariate 12 are close to 0, especially $\text{PIP}_{12,1} = 0.30$ and $\text{PIP}_{12,3} = 0.35$ (see Figure 2.6b). The estimated coefficient matrix $\hat{\mathbf{C}}$ of the galaxy dataset exposes 4 complete zero rows, thus reducing the number of covariates exerting a significant impact to 19.

Basing the decision of covariate selection solely on thresholding the PIP could lead to misleading interpretations of results. For this reason, we quantify the uncertainty about the decision of including the jkth element of $\mathbf{C}$ through the PIP uncertainty index ζ_{jk}. The PIP uncertainty

x_j	**Mode**	**Mean**	**Std**	**Q25**	**Q50**	**Q75**
1	0	0.504	0.392	0	0.667	0.833
2	1	0.986	0.055	1	1	1
3	1	0.997	0.023	1	1	1
4	1	0.985	0.047	1	1	1
5	1	0.976	0.059	1	1	1
6	1	0.982	0.052	1	1	1
7	1	0.987	0.057	1	1	1
8	1	0.985	0.063	1	1	1
9	1	0.994	0.038	1	1	1
10	0.833	0.826	0.048	0.833	0.833	0.833
11	0	0	0.007	0	0	0
12	0	0.161	0.233	0	0	0.333
13	0.833	0.843	0.049	0.833	0.833	0.833
14	0.833	0.806	0.112	0.833	0.833	0.833
15	0	0.387	0.348	0	0.333	0.667
16	0.833	0.859	0.073	0.833	0.833	0.833
17	0	0.111	0.211	0	0	0.167
18	1	0.972	0.062	1	1	1
19	1	0.96	0.071	1	1	1
20	0.833	0.897	0.083	0.833	0.833	1
21	0.833	0.831	0.068	0.833	0.833	0.833
22	0.833	0.659	0.234	0.5	0.667	0.833
23	0	0.052	0.147	0	0	0

Table 2.4: Summary statistics of the distribution of RI in the sub-sample of $n = 500$ for each covariate: mode, mean, standard deviation (Std), and the 25th, 50th and 75th quartiles (Q25, Q50, Q75). The shaded rows identify covariates with medium uncertainty ($\text{Std}(\text{RI}_j) \geq 0.20$) and high uncertainty ($\text{Std}(\text{RI}_j) \geq 0.30$). A description of each covariate is found in Appendix A.4

index provides a straightforward interpretation of the probabilities observed in the matrix of PIPs as a way to quantify uncertainty about the effect of each covariate on every response, which is particularly advantageous when addressing each response separately. Notwithstanding, the decision about irrelevant covariates in the multivariate regression model is yet to be determined since some covariates may influence only a subset of the responses. For instance, covariates 1 and 15 have an apparent influence only on responses $\{4, 5\}$ and $\{1, 3\}$, respectively (Figure 2.6). Therefore, the relevance index RI is used to assess the relative importance of each covariate. Table 2.4 reports the summary statistics of the relevance index for each covariate.

As illustrated in the top panels of Figure 2.7, there is strong evidence for the exclusion of covariate 17 and for the inclusion of covariate 7. However, the exclusion or inclusion of covariate 22 is not apparent, for there is strong variation in the distribution, and additional study should be given due consideration.[7] The *rule of thumb* in (2.4) states not to exclude the jth covariate if $S_{\text{RI}_j}(\overline{sr}) = \mathbb{P}(\text{RI}_j > \overline{sr}) \geq \overline{p}$, for appropriate values of $\overline{sr}$ and $\overline{p}$. By setting $\overline{sr} = 0.70$ and $\overline{p} = 0.60$, we are excluding covariate 17 while maintaining covariate 7 in the model (bottom panels of Figure 2.7), as inferred from the probability mass function of RI_7 and RI_{22}, accordingly. Notice that for covariate 22, the point $(0.70, 0.60)$ lies in the rejection region for inclusion (above the curve of the tail distribution), a conclusion that would not have been reached had we required a minimum probability of 0.60 for more than 40% of the responses ($\overline{sr} = 0.60$, $\overline{p} = 0.40$). Under the previous specification of $\overline{sr}$ and $\overline{p}$, the covariates considered irrelevant in the majority of responses are 1, 11, 12, 15, 17, 22, 23.

The methodology applied to the entire subset of $n = 3,438$ objects reduces the percentage of zero elements in the sparse estimate of the coefficient matrix from 28% to 14% and the number of zero rows by 3. The quantity of covariates with medium and high PIP uncertainty indices decreases accordingly, in line with stronger information provided by the data to require more variables that explain the outcomes.

[7] See Appendix A.4.2 for the results based on alternative variable selection methods.

2.6 Concluding remarks

We have proposed BRECS, a novel Bayesian approach for estimating the rank of the matrix of coefficients along with parameters in a reduced-rank regression model. Our method employs a mixture prior to rank estimation and shrinkage prior on the factor matrix resulting from the decomposition of the coefficient matrix for shrinking its irrelevant entries to 0. Furthermore, variable selection is achieved by adopting SAVS to obtain a sparse estimate of $\mathbf{C}$, in conjunction with the uncertainty about this estimation through the relevance index. By employing our method, researchers can avoid post-processing steps and obtain a quantification of uncertainty in estimating the rank, together with accurate statistical inference for the coefficient matrix.

The results of our simulation study suggest that RRcs exhibits superior performance over RRn in terms of converging more accurately to the true rank and being considerably faster in computational time. We observed that the algorithm's performance deteriorates more rapidly as the number of responses increases compared to the number of covariates. Thus, enhancing the model's scalability remains an area for future research. Overall, our method provides a reliable estimation of the coefficient matrix, even in cases where the rank is underestimated, effectively preventing over-fitting and resulting in a satisfactory approximation of the true matrix.

Our proposed approach was applied to real datasets about the chemical composition of tobacco leaves and photometric galaxy data. The obtained results are consistent with the findings presented in the literature, adding a quantification of the uncertainty about the obtained estimates. Additional domains for its applicability should be explored, including genomics (Hilafu et al., 2020) and macroeconomics (Reinsel et al., 2022). The latter explicitly suggests extending our research to the time series model and tensor regression (Billio et al., 2023; Luo and Griffin, 2024). Given the prevalence of models with incomplete response matrices in biostatistics, it is imperative to develop Bayesian approaches to address these scenarios. It is thus a promising avenue for further investigation (Mai and Alquier, 2022).

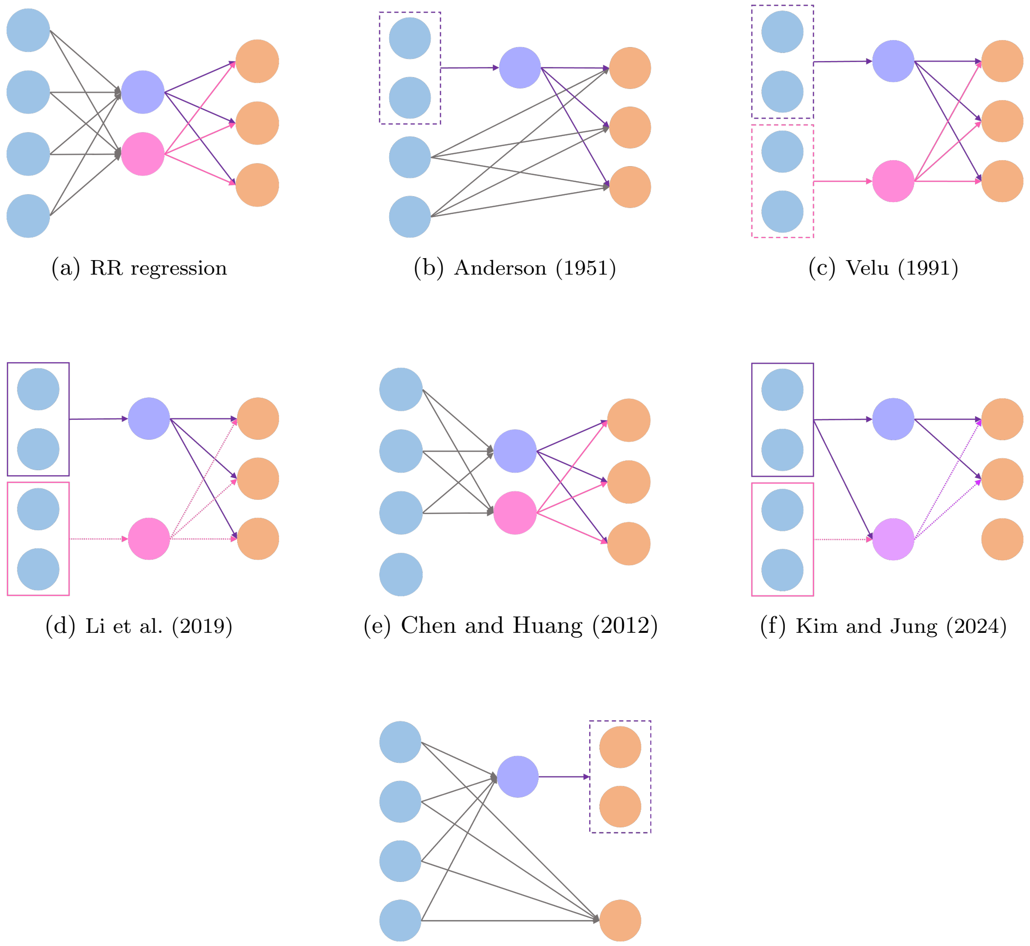
Chapter 3

Bayesian Partial Reduced-Rank Regression

3.1 Introduction

In MLR, block structures among the predictor variables are commonly addressed, particularly in the feature selection problem, where only a subset of independent variables is included as relevant predictors of the dependent variables. The reduced-rank (RR) regression model offers a more natural means of handling block structures and dimension reduction, since the covariates are related to the responses through fewer linear combinations or latent factors (Figure 3.1a). Hence, different variants of the reduced-rank regression model have been explored in the literature. For instance, Anderson (1951) examined a partitioned coefficient matrix associated with a low-rank group and a full-rank group in the covariates using a predefined grouping, thus obtaining a model where only some dependent variables express a low-rank structure while the rest maintain a full-rank relationship with the responses, and all of them enter into the regression (Figure 3.1b). This result is further elaborated upon by Velu (1991), who extended it into two sets of regressors, each of them characterised by a different low-rank structure (Figure 3.1c). Recently, Li et al. (2019) investigated an integrative reduced-rank regression model for analysing multi-view data, where each view, which is a prespecified group, consists of several predictors and has its own low-rank coefficient matrix (Figure 3.1d). On a different direction, Chen and Huang (2012) proposed a sparse reduced-rank regression that introduces row-wise sparsity in $\mathbf{C}$, enabling predictors with no association to latent factors, hence to the responses (Figure 3.1e). Kim and Jung (2024) combined the latter two approaches in a reduced-rank regression setting with multi-source data, where each predictor set is associated with a low-rank coefficient matrix while simultaneously allowing for sparsity in both covariates and responses (Figure 3.1f).

Conversely, potential groups among the responses are typically overlooked. An alternative extension of the standard RR regression model is the partially reduced-rank regression framework (Reinsel and Velu, 2006, PRR), wherein only a subset of the response variables exhibits a parsimonious relationship with the covariates via a low-rank component of the coefficient matrix, while the remaining responses are modeled through a full-rank regression to capture more complex associations (Figure 3.1g). In this scenario, the set of response variables is divided into two subsets $\mathbf{Y}_1$ and $\mathbf{Y}_2$, and the reduced-rank structure is imposed on a submatrix $\mathbf{C}_1$ of $\mathbf{C}$, driving the rela-



(a) RR regression

(b) Anderson (1951)

(c) Velu (1991)

(d) Li et al. (2019)

(e) Chen and Huang (2012)

(f) Kim and Jung (2024)

(g) Reinsel and Velu (2006)

Figure 3.1: Different reduced-rank regression models with grouping structures in an example with four covariates (blue disks) and three responses (orange disks). The central disks represent the linear combinations of the covariates, associated through dashed arrows if the association is allowed not to enter the model. Solid rectangles around the covariates indicate known group structures from the data source, while dashed rectangles represent groups assumed to be known or inferred.

tionship between $\mathbf{Y}_1$ and the covariates, $\mathbf{X}$. The reduced-rank assumption on $\mathbf{C}_1$ implies that the regression of $\mathbf{Y}_1$ on $\mathbf{X}$ is influenced by only a limited number of predictive variables constructed as linear combinations of $\mathbf{X}$.

Considering group structures within the response variables enhances our comprehension of the data in diverse fields such as macroeconomics (Reinsel and Velu, 2006) and genetics (Li et al., 2015; Luo and Chen, 2020). The PRR model, incorporating the proposed response groups, has a potential utility in multi-view data scenarios that consider two views in the responses, thereby enhancing model fit through the specialised structure in $\mathbf{C}$. Additionally, PRR adds flexibility to the model in accommodating complex relationships observed in real-world datasets, such as those found in economic contexts.

We propose the first Bayesian approach to PRR regression, where the low-rank and full-rank response groupings are unknown and directly inferred from the data. This approach considers an agnostic position about the optimal allocation of response variables and opens the possibility of using PRR models even in the absence of strong and reliable information to dictate the grouping. Reinsel and Velu (2006) fix the grouping structure in the estimation, then propose a stepwise backwards-elimination procedure to identify the optimal one. This approach is based on a sequence of likelihood ratio tests performed after estimating models with different grouping structures. Conversely, our model offers a unified framework where model selection and parameter estimation are performed jointly. Additionally, our Bayesian approach allows us to obtain a posterior distribution for the allocations and to quantify the uncertainty around it, which is not obtainable from the test-based procedure in Reinsel and Velu (2006). Moreover, we design and implement a partially collapsed Gibbs sampler that, at each iteration, first samples both the grouping structure and the (reduced) rank relying on the Laplace method, then the remaining parameters in subsequent basic steps. We call the proposed method, Bayesian partial reduced-rank (BPRR) regression and we compare its performance against well-known specifications available in the literature, such as full-rank regression and standard reduced-rank regression.

The remainder of this chapter is as follows. Section 3.2 presents the PRR framework and describes the structure of the prior distributions. Then, Section 3.3 describes in detail the challenges encountered in the design of the algorithm to perform posterior sampling, together with the proposed solutions. The performance of the algorithm is tested on synthetic data and compared to benchmark methods in linear regression in Section 3.4. Then, it is applied to macroeconomic data from the United States in Section 3.5. Section 3.6 draws the conclusions.

3.2 Partial reduced-rank regression model

Parting from the MLR model in Eq. (1.1), we assume that the response variables can be split into two different groups $\mathbf{Y}_1$ and $\mathbf{Y}_2$ of dimensions $n \times q_\gamma$ and $n \times (q - q_\gamma)$, respectively, where $q_\gamma \in \{2, \dots, q-1\}$ denotes the number of low-rank responses. Moreover, we assume that the relationship between $\mathbf{Y}_1$ and $\mathbf{X}$ admits a low-rank structure, while the regression of $\mathbf{Y}_2$ on $\mathbf{X}$ has full-rank. Under this assumption, the coefficient matrix $\mathbf{C} \in \mathbb{R}^{p\times q}$ can be partitioned as $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2]$, with $\mathbf{C}_1 \in \mathbb{R}^{p\times q_\gamma}$ having reduced rank $r = \text{rank}(\mathbf{C}_1) \leq \min(p, q_\gamma) - 1$ and $\mathbf{C}_2 \in \mathbb{R}^{p\times(q-q_\gamma)}$ with full rank $r_2 = \text{rank}(\mathbf{C}_2) = \min(p, q - q_\gamma)$.

Therefore, the model can be represented with partitioned matrices as:

$$[\mathbf{Y}_1,\, \mathbf{Y}_2] = \mathbf{X}\,[\mathbf{C}_1,\, \mathbf{C}_2] + [\mathbf{E}_1,\, \mathbf{E}_2]\,. \tag{3.1}$$

Notice that each response vector $\mathbf{y}_i$, $i = 1, \dots, n$, is of the form $\mathbf{y}_i = (y_{i,1}, \dots, y_{i,q_\gamma}, y_{i,q_\gamma+1}, \dots, y_{i,q})'$ $\in \mathbb{R}^q$. We write $\mathbf{e}_i = (\mathbf{e}_{1i}', \mathbf{e}_{2i}')$ with $\mathbf{e}_{1i} = (e_{i,1}, \dots, e_{i,q_\gamma})'$ and $\mathbf{e}_{2i} = (e_{i,q_\gamma+1}, \dots, e_{i,q})'$, and assume $\mathbf{e}_{(i)}$ is normally distributed with mean zero and a partitioned covariance matrix $\mathbf{\Sigma} = \text{cov}(\mathbf{e}_i)$ such that $\mathbf{\Sigma}_{11} = \text{cov}(\mathbf{e}_{1i})$, $\mathbf{\Sigma}_{22} = \text{cov}(\mathbf{e}_{2i})$, and $\mathbf{\Sigma}_{12} = \text{cov}(\mathbf{e}_{1i}, \mathbf{e}_{2i})$.

3.2.1 Prior specifications

The model in Eq. (3.1) classifies the response variable into two groups. Differently from Reinsel and Velu (2006), we assume the grouping structure to be unknown and aim at inferring it from the data. Therefore, we introduce a binary vector $\boldsymbol{\gamma} \in \{0, 1\}^q$ to categorise the responses into the low-rank and the full-rank groups. As we lack any prior information regarding the criteria for this classification, we assume that each element γ_j, $j = 1, \dots, q$, follows independently a Bernoulli prior distribution with probability ρ of being assigned to the low-rank group. Consequently, the joint prior distribution on $\boldsymbol{\gamma}$ is:

$$p(\boldsymbol{\gamma} \mid \rho) = \left[\prod_{j=1}^{q} \mathrm{Bern}(\gamma_j \mid \rho) \right] \mathbb{I}(1 < q_{\boldsymbol{\gamma}} < q) \,, \tag{3.2}$$

where $q_{\boldsymbol{\gamma}} = \sum_{j=1}^{q} \gamma_j$, and $\rho \in (0,1)$ is the prior probability of being assigned to the low-rank group. The constraint imposed by the indicator function in Eq. (3.2) allows for the existence of the low-rank group and, thus, of a PRR model. In fact, if $q_\gamma = 1$, $\mathbf{Y}_1$ comprises only one response, making $\mathbf{C}$ full-rank. Conversely, when $q_\gamma = q$, all responses are part of the low-rank group, which collapses into the standard RR model. Additionally, we employ a hierarchical prior structure, where ρ is assigned a Beta prior distribution, $\rho \sim \mathcal{B}e(\rho \mid \underline{a}_\rho, \underline{b}_\rho)$, resulting in the Beta-Binomial prior (Scott and Berger, 2010).

Remark 1. Reinsel and Velu (2006) fix the grouping structure in the estimation, then propose a stepwise backwards-elimination procedure to identify the optimal one. This approach is based on a sequence of likelihood ratio tests performed after estimating models with different grouping structures. Conversely, our model offers a unified framework where model selection and parameter estimation are performed jointly. Moreover, the posterior distribution for the allocation vector $\boldsymbol{\gamma}$ allows us to compute a point estimate of the grouping structure and to quantify the uncertainty around it, which is not obtainable from the test-based procedure in Reinsel and Velu (2006).

Moving to the coefficients matrix, $\mathbf{C}$, we can provide a specification for both the low and the full rank matrix. In particular, the submatrix of coefficients $\mathbf{C}_1$ is assumed to have reduced rank $r \leq r_{\max} = \min(p, q_\gamma) - 1$, which depends on the binary parameter $\boldsymbol{\gamma}$. Therefore, conditional on q_γ (hence on $\boldsymbol{\gamma}$), we assume an uninformative uniform prior distribution for r over the discrete set $\{1, \ldots, r_{\max}\}$, that is $r \mid \boldsymbol{\gamma} \sim \mathcal{U}(r \mid \{1, \ldots, r_{\max}\})$.

Given that $\mathbf{C}_1$ is a low-rank matrix, we can express it as the product of two full-rank matrices $\mathbf{A} \in \mathbb{R}^{q_\gamma \times r}$ and $\mathbf{B} \in \mathbb{R}^{p \times r}$, such that $\mathbf{C}_1 = \mathbf{B}\mathbf{A}'$. This decomposition is not unique, since for any orthogonal $r \times r$ matrix $\mathbf{P}$, we have that $\mathbf{C}_1 = (\mathbf{B}\mathbf{P})(\mathbf{P}'\mathbf{A}')$. To achieve a unique decomposition of $\mathbf{C}_1$, we follow Geweke (1996) and impose an identifying restriction by assuming the first r rows of $\mathbf{A}$ are the identity matrix $\mathbf{I}_r$, that is

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}_r \\ \mathbf{F} \end{bmatrix}, \tag{3.3}$$

where $\mathbf{F}$ is a $(q_\gamma - r) \times r$. Denoting with $\mathrm{vec}(\cdot)$ the vectorisation operator, we assume a multivariate Gaussian prior distribution on $\boldsymbol{\alpha}_F = \mathrm{vec}(\mathbf{F}')$, that is

$$\boldsymbol{\alpha}_F \mid \boldsymbol{\gamma}, r \sim \mathcal{N}_{(q_\gamma - r)r}(\boldsymbol{\alpha}_F \mid \mathbf{0}, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}),$$

where $\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}} = \underline{a}\, \mathbf{I}_{(q_\gamma - r)r}$, for fixed $\underline{a} > 0$. Similarly, defining $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B})$ and $\boldsymbol{\delta} = \mathrm{vec}(\mathbf{C}_2)$, we assume the multivariate Gaussian prior distributions:

$$\begin{aligned} \boldsymbol{\beta} \mid \boldsymbol{\gamma}, r &\sim \mathcal{N}_{pr}(\boldsymbol{\beta} \mid \mathbf{0}, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}), \\ \boldsymbol{\delta} \mid \boldsymbol{\gamma} &\sim \mathcal{N}_{p(q - q_\gamma)}(\boldsymbol{\delta} \mid \mathbf{0}, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}), \end{aligned}$$

where $\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \underline{b}\, \mathbf{I}_{pr}$, and $\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}} = \underline{d}\, \mathbf{I}_{p(q - q_\gamma)}$, for fixed $\underline{b}$, $\underline{d} > 0$. Lastly, we adhere to the conventional practice and assign a conjugate inverse Wishart prior to $\boldsymbol{\Sigma}$, that is $\boldsymbol{\Sigma} \sim \mathcal{IW}_q(\boldsymbol{\Sigma} \mid \underline{\nu}, \underline{\boldsymbol{\Psi}})$, with $\underline{\nu}$

and $\underline{\boldsymbol{\Psi}}$ being the fixed degrees of freedom and scale matrix.

3.3 Posterior sampling

In this section, we design an MCMC algorithm to draw samples from the joint posterior distribution $p(\mathbf{A}, \mathbf{B}, \mathbf{C}_2, \boldsymbol{\Sigma}, r, \boldsymbol{\gamma}, \rho \mid \mathbf{Y})$. The most critical issue to tackle is that the dimensions of the matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}_2$ depend on the states of $\boldsymbol{\gamma}$ and r, which implies that the dimension of the parameter space may change across the iterations of the MCMC algorithm. Consequently, the traditional Gibbs sampler is invalid in this setting, whereas a reversible jump MCMC, although theoretically feasible, poses significant challenges in terms of implementation and proper execution (Robert and Casella, 1999).[1] To address this challenge, we implement a partially collapsed Gibbs sampler (PCG, see van Dyk and Park, 2008), which generalises the composition of the conditional distributions in Gibbs samplers, relying on three basic tools: marginalization, permutation, and trimming. Specifically, we avoid the need for transdimensional samplers by drawing $(\boldsymbol{\gamma}, r)$ from a joint distribution marginalised over the parameters $(\mathbf{A}, \mathbf{B}, \mathbf{C}_2)$ whose size depends on $(\boldsymbol{\gamma}, r)$. Subsequently, we sample $(\mathbf{A}, \mathbf{B}, \mathbf{C}_2)$ conditionally on the updated values of $(\boldsymbol{\gamma}, r)$. The entire sampling process is summarised in Algorithm 3.

In the remainder of this section, we describe the procedures adopted to integrate out $(\mathbf{A}, \mathbf{B}, \mathbf{C}_2)$ from the likelihood, then we explain the main computational challenges and the proposed solutions. The first problem arises in Step 1, the most computational intensive step, where sampling $\boldsymbol{\gamma}$ entails the exploration of a 2^q dimensional parameter space and the computation of analytically intractable integrals, a limitation also encountered in Step 2. The second issue concerns the dimensions of matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}_2$ in practically implementing Steps 3 and 4, since their size depends on the states of $(\boldsymbol{\gamma}, r)$, producing a potential incompatibility between the dimension of the parameters at the previous MCMC iteration and the current one corresponding to the updated $(\boldsymbol{\gamma}, r)$.

Algorithm 3 PCG for Bayesian PRR model

```
1: Sample $\boldsymbol{\gamma}$ from $p(\boldsymbol{\gamma} \mid \mathbf{Y}, \boldsymbol{\Sigma}, \rho)$.
2: Sample $r$ from $p(r \mid \boldsymbol{\gamma}, \mathbf{Y}, \boldsymbol{\Sigma})$.
3: Sample $\boldsymbol{\delta} = \mathrm{vec}(\mathbf{C}_2)$ from $p(\boldsymbol{\delta} \mid \mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \mathbf{A}, \mathbf{B}) = \mathcal{N}_{p(q-q_\gamma)}(\overline{\boldsymbol{\mu}}_\delta, \overline{\boldsymbol{\Sigma}}_\delta)$.
4: Sample $\boldsymbol{\alpha}_{\mathbf{F}} = \mathrm{vec}(\mathbf{F}')$ from $p(\boldsymbol{\alpha}_F \mid \mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r, \mathbf{B}, \mathbf{C}_2) = \mathcal{N}_{(q_\gamma - r)r}(\overline{\boldsymbol{\mu}}_\alpha, \overline{\boldsymbol{\Sigma}}_\alpha)$,
   then set $\mathbf{A} = [\mathbf{I}_r, \mathbf{F}']'$.
5: Sample $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B})$ from $p(\boldsymbol{\beta} \mid \mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r, \mathbf{A}, \mathbf{C}_2) = \mathcal{N}_{pr}(\overline{\boldsymbol{\mu}}_\beta, \overline{\boldsymbol{\Sigma}}_\beta)$.
6: Sample $\boldsymbol{\Sigma}$ from $p(\boldsymbol{\Sigma} \mid \mathbf{Y}, \boldsymbol{\gamma}, \mathbf{A}, \mathbf{B}, \mathbf{C}_2) = \mathcal{IW}_q(\overline{\nu}, \overline{\boldsymbol{\Psi}})$.
7: Sample $\rho$ from $p(\rho \mid \boldsymbol{\gamma}) = \mathcal{B}e(\overline{a}_\rho, \overline{b}_\rho)$.
```

Remark 2. The proposed BPRR model can be extended to the case of non-Gaussian noise while keeping its main features. In particular, the BPRR can be applied in the generalized linear model

[1] Implementing a reversible jump MCMC in our setting would require the definition of a cross-model move involving the allocation vector $\boldsymbol{\gamma}$ and the rank r since both have an impact on the size of the parameter space. This move requires defining (a) the mapping function and (b) the proposal distribution for an auxiliary vector, w, which are highly arbitrary. The mapping function expresses a functional relationship between the parameters of two different models (the current and the proposed one), implying that a good choice for this function is likely to improve the sampler's performance in terms of between-model acceptance rates and chain mixing. However, the major difficulty is that good relationships can be hard to define even in the simplest model settings. In general cases like our framework, where two models are characterised by different binary vector $\boldsymbol{\gamma}$ and/or integer variable r, parameters between models may not be obviously comparable, thus making the definition of a relationship between them even harder to define. Finally, there are no easy-to-use criteria for choosing the proposal distributions for the auxiliary vectors w.

(GLM) framework by choosing an appropriate link function to responses from the natural exponential family. In several cases, it is possible to design a suitable data augmentation scheme to obtain a full conditional posterior for the coefficients of the same type as the Gaussian BPRR model (i.e., preserving the conjugacy of the prior). For example, regression models using a binomial likelihood can leverage the Pólya-gamma data augmentation of Polson et al. (2013) to obtain conditional conjugacy for the coefficients of the (latent) linear regression. Instead, for Poisson and Gamma likelihoods, the full conditional distribution for the coefficients is not conjugate, thus requiring the use of Metropolis-Hastings algorithms.

3.3.1 Preliminaries

Let us define the likelihood function as:

$$f(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \mathbf{C}_2, \boldsymbol{\Sigma}, r, \boldsymbol{\gamma}) = \frac{1}{(2\pi)^{nq/2} \, |\boldsymbol{\Sigma}|^{n/2}} \exp\left\{ -\frac{1}{2} \left\| (\mathbf{Y} - \mathbf{X}\mathbf{C})\boldsymbol{\Sigma}^{-1/2} \right\|_F^2 \right\},$$

where $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2]$, $\mathbf{C}_1 = \mathbf{B}\mathbf{A}'$, and $\|\cdot\|_F$ denotes the Frobenius norm. The first step involves explicitly expressing the likelihood in terms of $\mathbf{C}_1$ and $\mathbf{C}_2$. Therefore, we exploit the partitioning of $\mathbf{C}$ and rewrite the model in Eq. (3.1) equivalently as

$$\mathbf{Y} = \mathbf{X}\mathbf{C}_1\mathbf{V}_1 + \mathbf{X}\mathbf{C}_2\mathbf{V}_2 + \mathbf{E}, \tag{3.4}$$

where $\mathbf{V}_1 = \left[\mathbf{I}_{q_\gamma}, \mathbf{0}_{q_\gamma \times (q - q_\gamma)}\right]$, and $\mathbf{V}_2 = \left[\mathbf{0}_{(q-q_\gamma)\times q_\gamma}, \mathbf{I}_{q-q_\gamma}\right]$. By vectorising Eq. (3.4), we obtain

$$\mathbf{y} = \mathbf{U}_1\mathbf{c}_1 + \mathbf{U}_2\boldsymbol{\delta} + \mathbf{e}, \tag{3.5}$$

where $\mathbf{y} = \mathrm{vec}(\mathbf{Y})$, $\mathbf{c}_1 = \mathrm{vec}(\mathbf{C}_1)$, $\boldsymbol{\delta} = \mathrm{vec}(\mathbf{C}_2)$, $\mathbf{e} = \mathrm{vec}(\mathbf{E})$, and $\mathbf{U}_i = \mathbf{V}_i' \otimes \mathbf{X}$, for each $i = 1, 2$. It follows that $\mathbf{y} \mid \mathbf{A}, \mathbf{B}, \mathbf{C}_2, \boldsymbol{\Sigma}, r, \boldsymbol{\gamma} \sim \mathcal{N}_{nq}(\mathbf{y} \mid \mathbf{U}_1\mathbf{c}_1 + \mathbf{U}_2\boldsymbol{\delta}, \tilde{\boldsymbol{\Sigma}})$, where $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} \otimes \mathbf{I}_n$.

Under the vectorised model in Eq. (3.5), the likelihood can be marginalized over $\mathbf{C}_2$ analytically to obtain

$$\begin{aligned} f(\mathbf{y} \mid \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r) &= \int f(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \mathbf{C}_2, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r)\, p(\mathbf{C}_2 \mid \boldsymbol{\gamma})\, \mathrm{d}\mathbf{C}_2 \\ &= \int \mathcal{N}_{nq}(\mathbf{y} \mid \mathbf{U}_1\mathbf{c}_1 + \mathbf{U}_2\boldsymbol{\delta}, \tilde{\boldsymbol{\Sigma}})\, \mathcal{N}_{p(q-q_\gamma)}(\boldsymbol{\delta} \mid \mathbf{0}, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}})\, \mathrm{d}\boldsymbol{\delta} \\ &= \mathcal{N}_{nq}(\mathbf{y} \mid \mathbf{U}_1\mathbf{c}_1,\, \tilde{\boldsymbol{\Sigma}} + \mathbf{U}_2\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}\mathbf{U}_2'). \end{aligned} \tag{3.6}$$

This distribution represents the starting point in the design of the proposed PCG sampler.

3.3.2 Sampling the response allocation, $\boldsymbol{\gamma}$, and rank, r

Starting from Eq. (3.6), we are left with the task of marginalising $\mathbf{A}$ and $\mathbf{B}$. As analytical integration is unfeasible, we obtain an approximation to the (marginal) posterior of $\boldsymbol{\gamma}$ via the Laplace method, which provides a trade-off between computational speed and accuracy (e.g., see Tierney and Kadane, 1986; Tierney et al., 1989; Kass and Raftery, 1995, among others). Then, a sample from the approximate posterior is obtained through the Metropolized Shotgun Stochastic Search (MSSS) algorithm (Hans et al., 2007). Our strategy is similar in spirit to Yang et al. (2022), which is concerned with rank estimation in a RR model.

The posterior of $\boldsymbol{\gamma}$ given by Bayes' theorem after integrating out $\mathbf{A}, \mathbf{B}$, $\mathbf{C}_2$ and r is

$$p(\boldsymbol{\gamma} \mid \mathbf{Y}, \boldsymbol{\Sigma}, \rho) = \frac{f_\gamma(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma} \mid \rho)}{\sum_{\boldsymbol{\gamma}^\dagger \in \{0,1\}^q} f_\gamma(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}^\dagger) p(\boldsymbol{\gamma}^\dagger \mid \rho)},$$

where $f_\gamma(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma})$ is obtained from Eq. (3.7) by marginalising over $\mathbf{A}$, $\mathbf{B}$, and r, that is

$$\begin{aligned} f_\gamma(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}) &= \sum_{r=1}^{r_{\max}} \frac{1}{r_{\max}} \iint f(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r)\, p(\mathbf{A}, \mathbf{B} \mid r, \boldsymbol{\gamma})\, \mathrm{d}\mathbf{A}\, \mathrm{d}\mathbf{B} \qquad (3.7) \\ &= \sum_{r=1}^{r_{\max}} \frac{1}{r_{\max}} f_r(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r). \end{aligned}$$

Note that the integration with respect to r is performed analytically since the latter is a discrete parameter with finite support. Conversely, we use the Laplace method (Kass and Raftery, 1995) to approximate the analytically intractable integration of $\mathbf{A}, \mathbf{B}$, and ignore constant order terms to obtain

$$\log f_r(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r) \approx \log f(\mathbf{Y} \mid \hat{\mathbf{A}}, \hat{\mathbf{B}}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r) - \frac{1}{2}(pr + (q_\gamma - r)r) \log n, \qquad (3.8)$$

where $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are the maximum likelihood estimators (MLEs) of $\mathbf{A}$ and $\mathbf{B}$, given r and $\boldsymbol{\gamma}$. For situations in which the sample size is moderate, it produces accurate answers required for the estimation tasks involved in the simulation study of Section 3.4. The relative error is of order $O(n^{-1})$ (Kass and Raftery, 1995), and the method provides adequate approximations for well-behaved problems where the likelihood function is approximately normal, as is assumed in our framework. Since the analytical form of the posterior distribution is intractable, neither the Hessian of the log-posterior nor the maximum a posteriori can be obtained in closed form. Furthermore, the prior assumptions are not strongly informative, rendering the MLE a suitable point for evaluation (Schwarz, 1978). The main difficulty relies on the computation of these MLEs, although Reinsel et al. (2022) make available these estimators for i.i.d. Gaussian errors, our setting differs from this baseline in two main aspects. First, the Gaussian density in Eq. (3.6) has a covariance matrix $\boldsymbol{\Sigma}_\mathbf{y} = \tilde{\boldsymbol{\Sigma}} + \mathbf{U}_2 \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}} \mathbf{U}_2'$ that incorporates heteroscedastic errors through the dependency of $\mathbf{U}_2$ on $\mathbf{V}_2$. As a consequence, the block of the covariance matrix corresponding to $\mathbf{C}_2$ introduces a different variance for each observation. Second, we are imposing an identification restriction on the matrix $\mathbf{A}$ and an additional restriction on the vectorised linear model's coefficient via the binary matrix $\mathbf{V}_1$.

The first restriction requires the first r rows of $\mathbf{A}$ to be the identity matrix. The second restriction pertains to the representation of $\mathbf{Y}$ in Eq. (3.4) as the sum of a low-rank component and its full-rank counterpart, which clearly demands the introduction of zero factors through $\mathbf{V}_1$ and $\mathbf{V}_2$ to accommodate the desired structure. Consequently, an alternative procedure to compute the MLEs is required.

Hansen (2002) proposed an ML estimation technique for a general class of reduced rank regression models (called GRRR), including models with a generic structure of the covariance matrix and potential restrictions on the coefficient matrices. We exploit the GRRR setting to accommodate the heteroscedasticity and the aforementioned restrictions in the computation of the MLEs.

The GRRR problem considers the regression given by the vectorised model of Eq. (3.6)

$$\mathbf{y}_i = \mathbf{V}_1' \mathbf{A} \mathbf{B}' \mathbf{x}_i + \tilde{\mathbf{e}}_i,$$

where $\tilde{\mathbf{e}}_i$ is the ith column of $\tilde{\mathbf{E}} \in \mathbb{R}^{q\times n}$ and $\text{vec}(\tilde{\mathbf{E}}) \sim \mathcal{N}_{nq}(\mathbf{0}, \mathbf{\Sigma_y})$, subject to the restriction

$$\text{vec}(\mathbf{V}_1'\mathbf{A}) = \mathbf{G}\psi + \mathbf{g},$$

where ψ is the true vector of parameters to be estimated, $\mathbf{G}$ is a binary $qr \times r(q_\gamma - r)$ matrix and $\mathbf{g}$ is the qr-dimensional binary vector of restrictions (see Appendix B.2 for details). Then, the MLEs following the GRRR method are obtained as

$$\hat{\boldsymbol{\alpha}}_{\mathbf{V}_1} = \text{vec}(\mathbf{V}_1'\mathbf{A}) = \mathbf{G}(\mathbf{G}'\mathbf{M_B}\mathbf{G})^{-1}\mathbf{G}'(\mathbf{n_B} - \mathbf{M_B}\mathbf{g}) + \mathbf{g}, \tag{3.9}$$

$$\hat{\boldsymbol{\beta}} = \text{vec}(\mathbf{B}) = \mathbf{M}_\mathbf{A}^{-1}\mathbf{n_A}, \tag{3.10}$$

where $\mathbf{M_B} = (\mathbf{XB} \otimes \mathbf{I}_q)'\tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1}(\mathbf{XB} \otimes \mathbf{I}_q)$, $\mathbf{n_B} = (\mathbf{XB} \otimes \mathbf{I}_q)'\tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1}\,\text{vec}(\mathbf{Y}')$, $\mathbf{M_A} = \mathbf{K}_{p,r}'(\mathbf{X} \otimes \mathbf{V}_1'\mathbf{A})'\tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1}(\mathbf{X} \otimes \mathbf{V}_1'\mathbf{A})\mathbf{K}_{p,r}$, $\mathbf{n_A} = \mathbf{K}_{p,r}'(\mathbf{X} \otimes \mathbf{V}_1'\mathbf{A})'\tilde{\mathbf{\Sigma}}_\mathbf{y}^{-1}\,\text{vec}(\mathbf{Y}')$, $\tilde{\mathbf{\Sigma}}_\mathbf{y} = \mathbf{K}_{n,q}\mathbf{\Sigma_y}\mathbf{K}_{n,q}'$, and $\mathbf{K}_{m,n}$ is the $mn \times mn$ commutation matrix, which transforms the vectorisation of a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ into the vectorisation of its transpose, such that $\mathbf{K}_{m,n}\,\text{vec}(\mathbf{M}) = \text{vec}(\mathbf{M}')$.

Noticing that the expressions in Eq. (3.9) and (3.10) depend on each other, the practical implementation of the GRRR method is done in a recursive algorithm. After a random initialization of the parameter values, $\boldsymbol{\alpha}_{\mathbf{V}_1}$ and $\boldsymbol{\beta}$ are iteratively updated until convergence. Once a solution $\hat{\boldsymbol{\alpha}}_{\mathbf{V}_1}, \hat{\boldsymbol{\beta}}$ is obtained, it suffices to transform the vectorised MLEs back to their matrix forms $\mathbf{V}_1\hat{\mathbf{A}}'$ and $\hat{\mathbf{B}}$ to obtain the MLEs of the low-rank coefficient matrix as $\hat{\mathbf{C}}_1 = \hat{\mathbf{B}}(\mathbf{V}_1')^+\mathbf{V}_1'\hat{\mathbf{A}} = \hat{\mathbf{B}}\hat{\mathbf{A}}'$, where $\mathbf{M}^+$ refers to the Moore-Penrose pseudoinverse of $\mathbf{M}$.[2]

Consequently, we obtain the approximation

$$f_\gamma(\mathbf{Y} \mid \mathbf{\Sigma}, \boldsymbol{\gamma}) \approx \sum_{r=1}^{r_{\max}} \frac{1}{r_{\max}} \tilde{f}_r(\mathbf{Y} \mid \mathbf{\Sigma}, \boldsymbol{\gamma}, r) \equiv \tilde{f}_\gamma(\mathbf{Y} \mid \mathbf{\Sigma}, \boldsymbol{\gamma}), \tag{3.11}$$

where $\tilde{f}_r(\mathbf{Y} \mid \mathbf{\Sigma}, \boldsymbol{\gamma}, r)$ is the Laplace approximation of $f_r(\mathbf{Y} \mid \mathbf{\Sigma}, \boldsymbol{\gamma}, r)$ obtained from Eq. (3.8) to the integral in Eq. (3.7). Therefore, the posterior distribution of $\boldsymbol{\gamma}$ is approximated by

$$\tilde{p}(\boldsymbol{\gamma} \mid \mathbf{Y}, \mathbf{\Sigma}, \rho) = \frac{\tilde{f}_\gamma(\mathbf{Y} \mid \mathbf{\Sigma}, \boldsymbol{\gamma})p(\boldsymbol{\gamma} \mid \rho)}{\sum_{\boldsymbol{\gamma}^\dagger \in \{0,1\}^q} \tilde{f}_\gamma(\mathbf{Y} \mid \mathbf{\Sigma}, \boldsymbol{\gamma}^\dagger)p(\boldsymbol{\gamma}^\dagger \mid \rho)}.$$

The performance of the (approximated) posterior for $\boldsymbol{\gamma}$ via a Laplace approximation yields nice results (see the simulation study in Section 3.4). However, a deeper investigation of the theoretical properties guaranteeing the performance of the estimator based on this approximation is left for future work.

Given that, $\boldsymbol{\gamma}$ is a q-dimensional binary vector, the collection of all possible configurations for the response allocation encompasses 2^q distinct elements, which quickly becomes exceedingly large even for moderate q. This calls for the adoption of an approximate method to sample $\boldsymbol{\gamma}$. Given the discreteness of the support of $\boldsymbol{\gamma}$, we use a Metropolized Shotgun Stochastic Search (MSSS) algorithm proposed by Hans et al. (2007). The MSSS approach explores regions of the high-dimensional parameter space by examining a selection of neighbours of the current iteration's $\boldsymbol{\gamma}$ and rapidly identifying those with high posterior probability. Defining the set of neighbours to contain only a subset of all the possible values of $\boldsymbol{\gamma}$ allows for a trade-off between the space exploration speed and the computational time. Similar to Yang et al. (2022), we take the neighbourhood

[2] $\mathbf{V}_1$ is not an invertible matrix given that its dimensionality is $q_\gamma \times q$, with $q_\gamma < q$.

to be every binary vector that is a one-variable change to the current allocation $\boldsymbol{\gamma}$ and at the same time complies with the existence of a low-rank group. For example, if $\boldsymbol{\gamma} = (1, 0, 1, 0)$, then $\mathrm{nbd}(\boldsymbol{\gamma}) = \{(1, 1, 1, 0), (1, 0, 1, 1)\}$, but $(0, 1, 1, 0)$, $(0, 0, 1, 0)$ and $(1, 0, 0, 0)$ are not a neighbours.[3] This restriction improves computational efficiency compared to an unrestricted neighbourhood comprising all elements while allowing the (reduced) exploration of the space (see Appendix B.3 for a detailed description of MSSS). We define a proposal distribution by

$$g(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^{(m)}) \propto \tilde{p}(\boldsymbol{\gamma} \mid \mathbf{Y}, \boldsymbol{\Sigma}, \rho)\, \mathbb{I}\left(\boldsymbol{\gamma} \in \mathrm{nbd}(\boldsymbol{\gamma}^{(m)})\right),$$

where $\boldsymbol{\gamma}^{(m)}$ is the value of $\boldsymbol{\gamma}$ at the mth iteration of the MCMC. Summarising, the first step of the proposed PCG sampler generates a draw from the marginal posterior $p(\boldsymbol{\gamma} \mid \mathbf{Y}, \boldsymbol{\Sigma}, \rho)$ with the following procedure:

1. Generate $\boldsymbol{\gamma}^*$ from $g(\boldsymbol{\gamma} \mid \boldsymbol{\gamma}^{(m)})$.

2. Accept $\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^*$ with probability

$$\rho_\gamma = \min\left\{1, \frac{\sum_{\boldsymbol{\gamma} \in \mathrm{nbd}(\boldsymbol{\gamma}^{(m)})} \tilde{f}_\gamma(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma} \mid \rho)}{\sum_{\boldsymbol{\gamma}^\dagger \in \mathrm{nbd}(\boldsymbol{\gamma}^*)} \tilde{f}_\gamma(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}^\dagger) p(\boldsymbol{\gamma}^\dagger \mid \rho)}\right\},$$

and otherwise, set $\boldsymbol{\gamma}^{(m+1)} = \boldsymbol{\gamma}^{(m)}$.

Similarly to $\boldsymbol{\gamma}$, the conditional posterior of r is approximated through the Laplace method. Specifically, we compute the approximated posterior

$$\tilde{p}(r \mid \mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}) = \frac{\tilde{f}_r(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r) p(r \mid \boldsymbol{\gamma})}{\sum_{r^\dagger = 1}^{r_{\max}} \tilde{f}_r(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r^\dagger) p(r^\dagger \mid \boldsymbol{\gamma})}, \tag{3.12}$$

where $\tilde{f}_r(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}, r) p(r \mid \boldsymbol{\gamma})$ is the same as in Eq. (3.8). Then, a new value of r is sampled from the discrete distribution on $\{1, \ldots, r_{\max}\}$ with the probabilities given in Eq. (3.12).

3.3.3 Sampling the matrices A and C$_2$

The proposed PCG sampler introduces a particular challenge related to the dimensions of matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}_2$ in Steps 3 and 4 of Algorithm 3. Suppose that at the end of iteration m, we have generated $(\boldsymbol{\gamma}^{(m)}, r^{(m)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}, \mathbf{C}_2^{(m)})$. Then, at iteration $m + 1$, we obtain new values $(\boldsymbol{\gamma}^{(m+1)}, r^{(m+1)})$ in Steps 1 and 2. Afterwards, we shall update $\mathbf{C}_2$ by sampling $\boldsymbol{\delta}^{(m+1)}$ conditioned on the value $\boldsymbol{\gamma}^{(m+1)}$ just generated. To this aim, notice that $\mathbf{C}_2^{(m+1)}$ should be of "new" dimension $p \times (q - q_{\boldsymbol{\gamma}^{(m+1)}})$, and its posterior distribution depends on the matrix $\mathbf{C}_1$, which should have "new" dimension $p \times q_{\boldsymbol{\gamma}^{(m+1)}}$. However, the matrix $\mathbf{C}_1^{(m)}$ available at this step is of dimension $p \times q_{\boldsymbol{\gamma}^{(m)}}$. An analogous issue is encountered in Step 4 when sampling $\mathbf{A}^{(m+1)}$, as the posterior of the latter parameter would require a matrix $\mathbf{B}$ with "new" dimension $p \times r^{(m+1)}$, whereas the available matrix $\mathbf{B}^{(m)}$ has $r^{(m)}$ columns.

It is worth emphasising that these dimensionality issues stem from considering $(\boldsymbol{\gamma}, r)$ as parameters to be estimated, thus varying quantities across the MCMC iterations. Moreover, changing the order of the Gibbs steps (while keeping the PCG sampler) would not circumvent the problem.

[3] The element $(0, 1, 1, 0)$ is not a neighbour because two variables changed; $(0, 0, 1, 0)$ and $(1, 0, 0, 0)$ are not neighbours because they violate the constraint $q_\gamma \in \{1, \ldots, q - 1\}$.

A possible way out of the issue in Step 3 can be found by recalling the decomposition of the coefficient matrix in Eq. (3.4), that is $\mathbf{C}^{(m)} = \left[\mathbf{C}_1^{(m)}, \mathbf{C}_2^{(m)}\right]$. Importantly, for any iteration $m = 1, \dots, M$, this matrix of coefficients has a fixed dimension $p \times q$, whereas the number of columns of $\mathbf{C}_1^{(m)}$ and $\mathbf{C}_2^{(m)}$ are possibly changing across iterations in consequence of varying $\boldsymbol{\gamma}^{(m)}$. Hence, to sample $\boldsymbol{\delta}^{(m+1)}$ conditioned on $\boldsymbol{\gamma}^{(m+1)}, \mathbf{A}^{(m)}, \mathbf{B}^{(m)}$, we construct an auxiliary matrix $\mathbf{C}_{1*}^{(m)}$ that is consistent with the newly sampled value of $\boldsymbol{\gamma}^{(m+1)}$, formed by selecting the first $q_{\boldsymbol{\gamma}^{(m+1)}}$ columns of the available $\mathbf{C}^{(m)}$:

$$\mathbf{C}_{1*}^{(m)} = \left[\mathbf{C}_{\bullet 1}^{(m)}, \dots, \mathbf{C}_{\bullet q_{\boldsymbol{\gamma}^{(m+1)}}}^{(m)}\right],$$

where $\mathbf{C}_{\bullet j}^{(m)}$ is the jth column of matrix $\mathbf{C}^{(m)}$. At iteration $m+1$, the auxiliary matrix $\mathbf{C}_{1*}^{(m)}$ is used to compute the updated parameters of the posterior distribution of $\boldsymbol{\delta}^{(m+1)}$.

To address the dimensionality inconsistency in Step 4, let us recall the restriction in Eq. (3.3); then, substituting $\mathbf{A}^{(m)} = \left[\mathbf{I}_{r^{(m)}}, \mathbf{F}^{(m)\prime}\right]'$ and $\mathbf{B}^{(m)}$ in the low-rank matrix representation yields

$$\mathbf{C}_1^{(m)} = \left[\mathbf{B}^{(m)}, \mathbf{B}^{(m)}\mathbf{F}^{(m)\prime}\right]. \tag{3.13}$$

Based on Eq. (3.13), it is evident that the first $r^{(m)}$ columns of $\mathbf{C}_1^{(m)}$ coincide with the matrix $\mathbf{B}^{(m)}$. Consequently, to update $\mathbf{A}^{(m+1)}$, we define the auxiliary matrix $\mathbf{B}_*^{(m)}$ as:

$$\mathbf{B}_*^{(m)} = \left[\mathbf{C}_{\bullet 1}^{(m)}, \dots, \mathbf{C}_{\bullet r^{(m+1)}}^{(m)}\right].$$

It is important to remark that the auxiliary matrices $\mathbf{C}_{1*}^{(m)}$ and $\mathbf{B}_*^{(m)}$ have appropriate dimensions, corresponding to the updated values $\boldsymbol{\gamma}^{(m+1)}$ and $r^{(m+1)}$, and contain elements already available at iteration $m+1$, that is $\mathbf{C}^{(m)}$, $\mathbf{B}^{(m)}, \mathbf{F}^{(m)}$.

In more detail, in Step 3, the full-rank coefficient matrix is sampled in vectorised form. Denoting $\mathbf{c}_{1*} = \operatorname{vec}(\mathbf{C}_{1*})$, the posterior distribution of $\boldsymbol{\delta}$ is proportional to the multivariate Gaussian distribution $p(\boldsymbol{\delta} \mid \mathbf{Y}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \mathbf{c}_{1*}) \propto p(\boldsymbol{\delta})\, p(\mathbf{y} \mid \boldsymbol{\Sigma}, \boldsymbol{\delta}, \mathbf{c}_{1*}) \sim \mathcal{N}_{p(q-q_\gamma)}(\boldsymbol{\delta} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{\delta}}, \overline{\boldsymbol{\Sigma}}_\delta)$ with mean $\overline{\boldsymbol{\mu}}_{\boldsymbol{\delta}} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}} \mathbf{U}_2' \tilde{\boldsymbol{\Sigma}}^{-1}(\mathbf{y} - \mathbf{U}_1\mathbf{c}_{1*})$ and covariance matrix $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}} = (\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}^{-1} + \mathbf{U}_2'\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{U}_2)^{-1}$.

The update of $\mathbf{A} = [\mathbf{I}_r, \mathbf{F}']'$ given $(\mathbf{Y}, \boldsymbol{\gamma}, r, \boldsymbol{\Sigma}, \mathbf{B}, \mathbf{C}_2)$ is performed by sampling $\boldsymbol{\alpha}_{\mathbf{F}} = \operatorname{vec}(\mathbf{F}')$. The posterior distribution of $\boldsymbol{\alpha}_{\mathbf{F}}$ is proportional to the multivariate Gaussian distribution $p(\boldsymbol{\alpha}_{\mathbf{F}} \mid$ $\mathbf{Y},$

$\boldsymbol{\Sigma}, \boldsymbol{\gamma}, r, \mathbf{C}_2, \mathbf{B}_*) \propto p(\boldsymbol{\alpha}_{\mathbf{F}} \mid \boldsymbol{\gamma}, r)\, p(\mathbf{Y} \mid \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{B}_*, \mathbf{C}_2) \sim \mathcal{N}_{(q_\gamma - r)r}(\boldsymbol{\alpha}_{\mathbf{F}} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}})$ with mean $\overline{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} =$ $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}\big(\mathbf{m}_J$

$- \mathbf{H}_{[J,J]}\mathbf{v}\big)$ and covariance matrix $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}} = \left(\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{-1} + \mathbf{H}_{[J,J]}\right)^{-1}$, where $\mathbf{v} = \operatorname{vec}(\mathbf{I}_r)$, $\mathbf{m} = \mathbf{M}_{\boldsymbol{\alpha}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{y}}_2$, $\mathbf{H} = \mathbf{M}_{\boldsymbol{\alpha}}'\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{M}_{\boldsymbol{\alpha}}$, $\tilde{\mathbf{y}}_2 = \mathbf{y} - \mathbf{U}_2\boldsymbol{\delta}$, and $\mathbf{M}_{\boldsymbol{\alpha}} = \mathbf{U}_1(\mathbf{I}_{q_\gamma} \otimes \mathbf{B})$. Moreover, $\mathbf{H}_{[J,J]}$ indicates the Jth row and the Jth column in $\mathbf{H}$ for the sequence $J = \left\{r^2+1, r^2+2, \dots, q_\gamma r\right\}$.

3.3.4 Sampling the other parameters

Regarding Steps (5)–(7) of Algorithm 3, the conditional posterior distribution of $\boldsymbol{\beta}$, given $(\mathbf{Y}, \boldsymbol{\gamma}, r,$ $\boldsymbol{\Sigma}, \mathbf{A}, \mathbf{C}_2)$, is proportional to the multivariate Gaussian distribution $\mathcal{N}_{pr}(\boldsymbol{\beta} \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}})$, where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} =$ $(\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{-1} + \mathbf{M}_{\boldsymbol{\beta}}'\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{M}_{\boldsymbol{\beta}})^{-1}$ and $\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\mathbf{M}_{\boldsymbol{\beta}}'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{y}}_2$, with $\mathbf{M}_{\boldsymbol{\beta}} = \mathbf{U}_1(\mathbf{A} \otimes \mathbf{I}_p)$.

The conditional posterior of the innovation covariance matrix $\boldsymbol{\Sigma}$, given $(\mathbf{A}, \mathbf{B}, \mathbf{C}_2, \mathbf{Y})$, is the inverse Wishart $\mathcal{IW}_q(\boldsymbol{\Sigma} \mid \overline{\nu}, \overline{\boldsymbol{\Psi}})$, where $\overline{\nu} = \underline{\nu} + n$ and $\overline{\boldsymbol{\Psi}} = \underline{\boldsymbol{\Psi}} + (\mathbf{Y} - \mathbf{X}\mathbf{C})'(\mathbf{Y} - \mathbf{X}\mathbf{C})$. Finally, the posterior distribution of ρ, the probability of a response variable belonging to the low-rank group, is the Beta distribution $\mathcal{B}e(\rho \mid \overline{a}_\rho, \overline{b}_\rho)$, with $\overline{a}_\rho = \underline{a}_\rho + q_\gamma$, and $\overline{b}_\rho = \underline{b}_\rho + q - q_\gamma$. A detailed derivation of the posterior distributions is available in Appendix B.1.

3.4 Simulation study

This section is devoted to examining the proposed model's performance in terms of the ability to recover the true group allocation, that is, the classification of variables into the low- and full-rank groups in different simulation studies. Next, the performance of the sampler is tested about rank estimation and the recovery of the overall coefficient matrix, $\mathbf{C}$.

The data was generated from the multivariate linear model $\mathbf{Y}_0 = \mathbf{X}\mathbf{C}_0 + \mathbf{E}_0$. The rows of $\mathbf{X}$ were independently drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ and the rows of $\mathbf{E}_0$ were drawn from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$, where the covariance matrix $\boldsymbol{\Sigma}_0$ is diagonal with elements sampled from $\mathcal{U}(0.5, 1.75)$. We work with centred responses and exclude the intercept term for simplicity. To generate the coefficient matrix $\mathbf{C}_0$, we first recall its partition into the low-rank and full-rank matrices $\mathbf{C}_1 = \mathbf{B}\mathbf{A}'$ and $\mathbf{C}_2$, then draw each free entry of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}_2$ from a standard Gaussian. Notice that the dimensions of these matrices depend on the fixed number of responses in the low-rank group, $q_\gamma < q$, and the true rank, $r \leq r_{\max}$.

The allocation of the response variables to the reduced-rank group is randomly selected given q_γ, and represented by the binary vector $\boldsymbol{\gamma}$. The columns of the matrix $\mathbf{Y}_0$ are then permuted following the allocation imposed by $\boldsymbol{\gamma}$. The response matrix so generated, $\mathbf{Y}$, need not necessarily be partitioned as $\mathbf{Y}_0 = [\mathbf{Y}_1, \mathbf{Y}_2]$, which is the representation postulated by our BPRR model. For example, if $q = 5$ and $q_\gamma = 3$, the data generating process (DGP) initially comprises the partitioned response matrix $\mathbf{Y}_0 = [\mathbf{Y}_1, \mathbf{Y}_2]$, with $\mathbf{Y}_1 = [\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \mathbf{y}_{(3)}] \in \mathbb{R}^{n\times 3}$ and $\mathbf{Y}_2 = [\mathbf{y}_{(4)}, \mathbf{y}_{(5)}] \in \mathbb{R}^{n\times 2}$, where $\mathbf{y}_{(j)}$ represents the vector of all observations for the jth response. However, after a random $\boldsymbol{\gamma}$ has been generated, say $\boldsymbol{\gamma}_0 = (0, 1, 0, 1, 1)$, then the final generated response matrix that is fed into our model is $\mathbf{Y} = [\mathbf{y}_{(4)}, \mathbf{y}_{(1)}, \mathbf{y}_{(5)}, \mathbf{y}_{(2)}, \mathbf{y}_{(3)}]$, aiming to reorder the responses to their true form, $\mathbf{Y}_0$, if $\hat{\boldsymbol{\gamma}}$ is estimated correctly.

The hyperparameters are set to consider noninformative priors, specifically $\underline{a}_\rho = \underline{b}_\rho = 1$, $\underline{a} = \underline{b} = \underline{d} = 0.5$, $\underline{\nu} = q + 1$, and $\underline{\boldsymbol{\Psi}} = \mathbf{I}_q$. The starting value of $\boldsymbol{\Sigma}$ is the identity matrix, while the coefficient matrix and the response allocation vector are initialised at random.

We consider different simulation settings with varying dimensionality, number of low-rank responses and true rank. Our method (BPRR) is compared with the following competitors: full-rank (FR), *full* low-rank (RR), and *pre-specified allocation* partial low-rank (PRR*). The first one is a standard linear regression model, where no low-rank structure is assumed. The full low-rank concerns a usual reduced-rank regression without an imposed partition. The last competitor is a partially reduced-rank regression model in which the low-rank group is fixed at random, a feature that serves two purposes. First, it accommodates scenarios where the researcher might have prior knowledge about the grouping structure, enabling the estimation procedure of the partial reduced-rank model to be conducted with a constraint by the imposed $\boldsymbol{\gamma}$. Second, it allows us to examine whether the automatic model selection in BPRR offers advantages over a random grouping choice.

The performance of the estimator $\hat{\mathbf{C}}$ of the coefficient matrix is evaluated using the mean squared error, $\text{MSE} = \|\hat{\mathbf{C}} - \mathbf{C}_0\|_{\text{F}}^2/(pq)$, where $\hat{\mathbf{C}}$ is the posterior mean of the predicted coefficients in their original ordering, warranting a fair comparison if the estimated allocation vector, $\hat{\boldsymbol{\gamma}}$, is not the correct grouping. The point estimates of the binary vector and the rank ($\hat{\boldsymbol{\gamma}}$ and $\hat{r}$, respectively) are their corresponding maximum a posteriori.

					BPRR metrics				MSE $\times 10^2$				
p	q	q_γ	r	n	$\hat{q}_\gamma$	$\hat{r}$	Accuracy	F_1 score	BPRR	FR	RR	PRR*	Oracle
5	5	3	1	20	3.240	1.740	0.584	0.659	11.831	10.369	13.350	15.275	8.566
5	6	3	1	20	3.650	2.030	0.612	0.653	10.218	8.900	12.420	13.712	8.227
5	6	3	1	50	3.950	2.120	0.698	0.743	3.434	2.646	4.136	6.354	2.510
5	6	5	2	50	4.790	2.090	0.862	0.913	2.646	3.252	3.480	9.673	2.353
5	6	5	2	100	4.910	2.090	0.925	0.953	1.202	1.277	1.580	5.366	1.108
5	7	3	1	20	3.770	2.090	0.576	0.560	11.356	10.782	16.966	15.689	9.302
5	7	3	1	50	4.650	2.610	0.610	0.657	3.182	2.673	5.172	5.900	2.972
5	7	5	2	50	5.320	2.500	0.786	0.849	3.503	3.186	4.632	7.062	2.936
5	8	6	2	20	5.100	2.120	0.713	0.777	14.684	16.345	17.112	21.816	14.467
10	5	3	1	20	3.400	1.910	0.636	0.714	14.925	16.998	18.838	23.548	10.309
10	5	3	1	100	3.750	1.920	0.766	0.831	1.948	1.297	1.953	8.659	1.712
10	6	3	1	20	3.490	2.100	0.622	0.656	16.145	16.182	19.166	23.283	11.830
10	6	5	2	20	4.600	2.310	0.797	0.866	30.161	38.745	29.880	43.679	26.247
10	6	5	2	50	4.980	2.250	0.893	0.936	4.892	5.561	5.573	10.051	4.563
10	7	3	1	20	3.620	2.280	0.591	0.580	15.644	14.609	17.508	20.657	10.661
10	7	5	2	20	4.870	2.650	0.696	0.769	23.839	26.104	25.028	38.296	21.142
10	8	3	1	20	3.030	1.810	0.664	0.561	18.609	14.082	16.886	23.783	12.490
10	8	6	2	20	5.290	2.590	0.696	0.765	22.094	26.685	23.590	35.390	19.869
20	5	3	1	20	2.610	1.480	0.570	0.610	32.304	30.856	36.725	39.718	24.326
20	5	3	1	50	3.920	2.390	0.608	0.719	6.866	4.820	8.097	14.118	5.980
20	6	3	1	20	2.410	1.290	0.595	0.545	32.099	28.815	35.023	38.934	23.014
20	6	5	2	20	2.530	1.310	0.528	0.609	56.350	54.898	54.315	58.261	44.784
20	7	3	1	20	2.200	1.070	0.609	0.463	31.839	27.640	34.408	37.256	22.802
20	7	5	2	20	2.350	1.220	0.521	0.535	57.864	56.544	59.930	67.723	51.176
50	5	3	1	50	3.490	2.140	0.570	0.661	23.463	25.552	32.526	30.296	16.145

Table 3.1: Average MSE ($\times 10^2$) over 100 replicates of the listed simulation settings in columns 1 to 5, for each model: Bayesian partial reduced-rank regression (BPRR), *full* low-rank regression (FR), reduced-rank regression (RR) and *pre-specified allocation* partial low-rank regression (PRR*). For each combination of parameters, columns 6 to 9 provide the average estimates of the number of low-rank responses, the rank of $\mathbf{C}_1$, the accuracy and the F_1 score.

Table 3.1 summarises the simulation results by providing the average MSE over 100 independent experiments of each scenario. Our method predominantly achieves the minimum mean squared error, and increasing the number of observations results in a smaller error. The former result is visually explored in Fig. 3.2, which examines the similarity of the coefficient matrix estimated by each of the four models with the true ordering. BPRR approximates $\mathbf{C}_0$ more accurately than its competitors, and the automatic grouping choice eliminates the need for a pre-processing step to select a potentially incorrect $\boldsymbol{\gamma}$, which could result in an inferior estimation as in this case. Additionally, we present the average estimated number of low-rank response variables with their corresponding rank estimate. As a measure of classification for the allocation parameter, we employ the accuracy and the F_1 score, both of which range from 0 to 1. A higher value indicates a more accurate classification of the responses.

We observe that misspecifying a PRR regression model, either by choosing an incorrect regression model or by inaccurately determining the reduced-rank grouping, can lead to substantive performance loss, as demonstrated earlier. In nearly half of the simulation scenarios reported in Table 3.1, the FR model achieves a lower error than BPRR, an outcome expected due to the additional parameters introduced in the regression. In contrast, the RR model consistently yields a higher error relative to our proposed method, likely due to its restrictive assumption that all response variables share a common low-rank structure, which can lead to underfitting when this structure is only partially present. Hence, BPRR offers a compromise between the full-rank model and the reduced-rank choice, balancing flexibility and parsimony. Furthermore, BPRR possesses

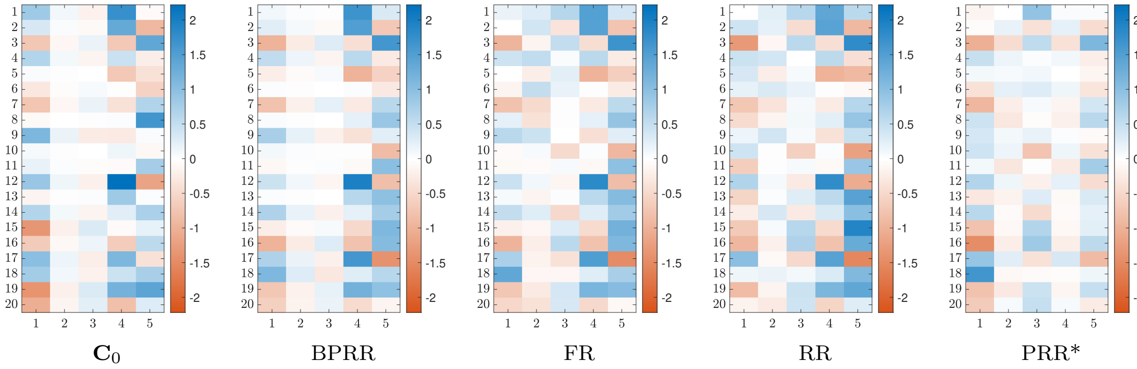


Figure 3.2: True coefficient matrix (first left) and estimated $\mathbf{C}$ matrix by each model in the simulation scenario where $p = 20$, $q = 5$, $q_\gamma = 3$, $r = 1$ and $n = 20$.

the added advantage of estimating the response groups, which avoids the need to specify a possibly incorrect classification that may hinder the performance of the model, as in PRR*.

The code for the BPRR algorithm has been implemented in MATLAB 2021a, and run on a MacBook Pro M1 2020 computer with 8 GB RAM. The average computational time for running 100 iterations of a model with dimensions $q = 5$, $p = 5$, and $n = 100$ is approximately 365 seconds.

3.4.1 Convergence diagnostic analysis

We assess the computational performance of the proposed MCMC algorithm by means of a convergence diagnostic analysis (CODA). Specifically, for each parameter under investigation, we consider Geweke's convergence diagnostic, Heidelberger and Welch's stationarity and half-width tests (see Geweke, 1992; Heidelberger and Welch, 1983, for further details). Since the coefficient matrix $\mathbf{C}$ has pq entries, Table 3.2 reports the share of entries that pass each of the aforementioned tests (i.e., p-value > 0.05; ratio < 0.10). The convergence diagnostics of the rank of the low-rank matrix $\mathbf{C}_1$ are equally included in Table 3.2. The results are satisfactory and suggest convergence of the chains for both parameters.[4]

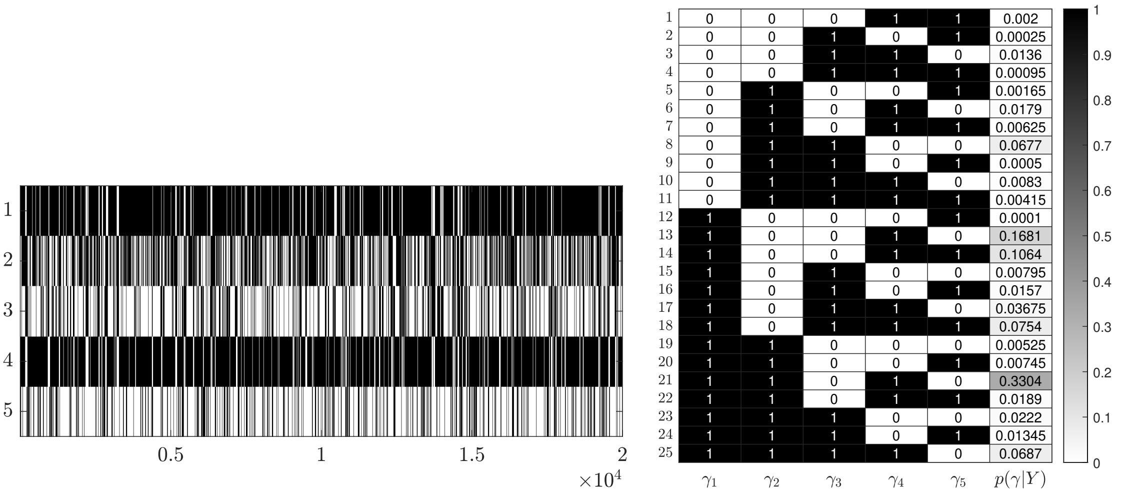
					Geweke test		HW Stationarity test		HW Half-width test	
p	q	q_γ	r	n	r (p-value)	$\mathbf{C}$ (share)	r (p-value)	$\mathbf{C}$ (share)	r (ratio)	$\mathbf{C}$ (share)
5	5	3	1	20	0.814	0.920	0.376	1.000	0.011	0.800
5	5	3	1	40	0.450	0.800	0.072	0.680	0.018	0.882
5	8	3	1	20	0.193	0.750	0.216	1.000	0.037	0.900
5	8	6	2	20	0.019	0.975	0.223	1.000	0.008	0.950
5	8	6	4	20	0.203	0.975	0.290	0.975	0.003	0.949
10	5	3	1	20	0.275	0.920	0.360	0.980	0.016	0.918
10	5	3	1	40	0.795	1.000	0.306	0.980	0.015	1.000
10	8	3	1	20	0.486	0.825	0.218	0.963	0.132	0.870
10	8	6	2	20	0.819	0.975	0.550	1.000	0.033	0.813
10	8	6	4	20	0.268	0.913	0.148	0.988	0.104	0.848
20	5	3	1	20	0.011	0.830	0.433	0.990	0.004	0.909
20	5	3	1	40	0.072	0.620	0.234	1.000	0.017	0.930

Table 3.2: Convergence diagnostics of the rank and the coefficient matrix $\mathbf{C}$: Geweke and the Heidelberger and Welch's (HW) stationarity tests p-values of the rank (> 0.05), and the rank's HW half-width test ratio (< 0.10). Share of entries of the coefficient matrix $\mathbf{C}$ that pass the Geweke test, HW stationarity and half-width tests (in column).

Instead, as pertains to the binary allocation vector $\boldsymbol{\gamma}$, similar tests are not available. Therefore,

[4] We performed the CODA analysis also on other independent runs of the algorithm, finding analogous results.

we rely on the visual inspection of the trace plot and posterior distribution in Fig. 3.3 to assess the convergence and mixing of the chain. It is worth emphasising that we also performed independent runs of the algorithm with random initialisation and obtained similar plots. Therefore, the results in Fig. 3.3 suggest a good exploration of the space of configurations and a good acceptance rate of the MSSS step. In particular, the posterior distribution assigns positive mass on several configurations while displaying a unique maximum at $\hat{\boldsymbol{\gamma}} = (1, 1, 0, 1, 0)$, which corresponds to the true allocation (see Appendix B.4 for further results).



	γ_1	γ_2	γ_3	γ_4	γ_5	$p(\gamma\|Y)$
1	0	0	0	1	1	0.002
2	0	0	1	0	1	0.00025
3	0	0	1	1	0	0.0136
4	0	0	1	1	1	0.00095
5	0	1	0	0	1	0.00165
6	0	1	0	1	0	0.0179
7	0	1	0	1	1	0.00625
8	0	1	1	0	0	0.0677
9	0	1	1	0	1	0.0005
10	0	1	1	1	0	0.0083
11	0	1	1	1	1	0.00415
12	1	0	0	0	1	0.0001
13	1	0	0	1	0	0.1681
14	1	0	0	1	1	0.1064
15	1	0	1	0	0	0.00795
16	1	0	1	0	1	0.0157
17	1	0	1	1	0	0.03675
18	1	0	1	1	1	0.0754
19	1	1	0	0	0	0.00525
20	1	1	0	0	1	0.00745
21	1	1	0	1	0	0.3304
22	1	1	0	1	1	0.0189
23	1	1	1	0	0	0.0222
24	1	1	1	0	1	0.01345
25	1	1	1	1	0	0.0687

Figure 3.3: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario where $p = 20$, $q = 5$, $q_\gamma = 3$, $r = 1$ and $n = 20$.

3.4.2 Forecasting exercise

As a further evaluation step, we conduct a forecasting exercise to evaluate the predictive performance of our proposed method. The dimensions of the artificially generated data for this purpose were $p = 10$, $q = 5$, $q_\gamma = 3$, $r = 1$. Two different sample sizes were considered, $n = 60$ and $n = 100$. We make a one-step-ahead prediction for n_{test} observations using our model and the three competitors: full-rank (FR), *full* low-rank (RR), and *pre-specified allocation* partial low-rank (PRR*). A fitted matrix of responses, $\hat{\mathbf{Y}}$, was constructed with all n_{test} predictions, and compared against the true values using the mean squared error (MSE) and the mean absolute error (MAE), defined as

$$\text{MSE} = \sum_{i=1}^{n_{\text{test}}} \sum_{j=1}^{q} (\hat{y}_{ij} - y_{ij})^2/(n_{\text{test}} q), \text{ and MAE} = \sum_{i=1}^{n_{\text{test}}} \sum_{j=1}^{q} |\hat{y}_{ij} - y_{ij}|/(n_{\text{test}} q). \tag{3.14}$$

The forecast error metrics of the models are presented in Table 3.3, demonstrating that our model achieves superior predictive performance compared to the alternatives.

(n, n_{test})		BPRR	FR	RR	PRR*
$(60, 20)$	MSE	1.405	1.501	1.535	2.059
	MAE	0.903	0.940	0.963	1.043
$(100, 40)$	MSE	1.467	1.529	1.456	2.797
	MAE	0.941	0.959	0.945	1.100

Table 3.3: Mean squared error (MSE) and mean absolute error (MAE) of the true responses versus the fitted values through a rolling forecast with the models BPRR, FR, RR, and PRR*.

3.5 Application

This section aims to demonstrate the usefulness of our method when applied to real-world data. We consider quarterly macroeconomic data for the United States from 2014Q1 to 2023Q4, which were retrieved from FRED, Federal Reserve Bank of St. Louis, and the OECD, Organisation for Economic Co-operation and Development (see Appendix B.5 for details).

The $q = 5$ responses are the index of industrial production (y_1), personal consumption of food and drinks (y_2), unemployment rate (y_3), volume index of imports of goods and services (y_4), and volume index of exports of goods and services (y_5). The $p = 5$ covariates are civilian labour force level (x_1), median weekly earnings (x_2), price index of imports of goods and services (x_3), price index of exports of goods and services (x_4), and price index of final consumption expenditure (x_5). All variables were standardised before conducting the analysis.

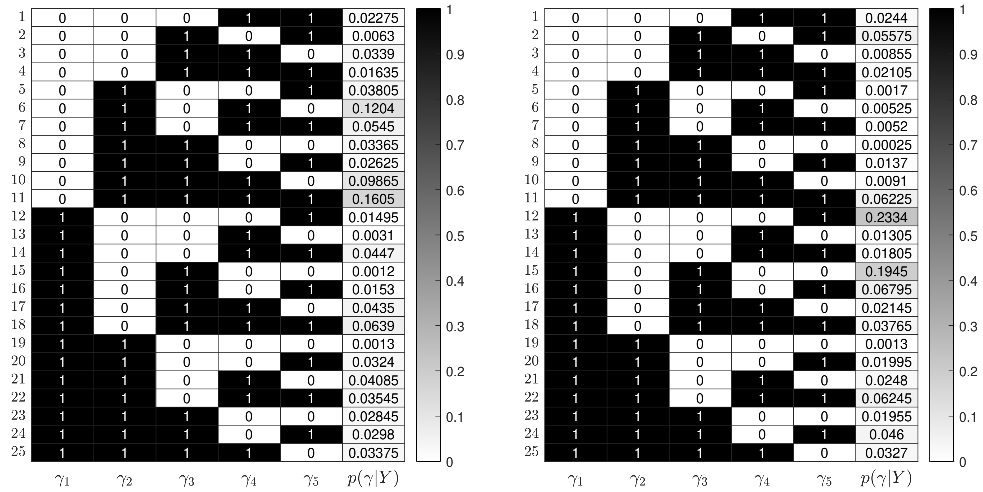
To account for possible temporal dependence, we introduce time variation in the innovation covariance, by assuming $\mathbf{e}_i \sim \mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma}_i)$, where we assume the decomposition $\mathbf{\Sigma}_i = \mathbf{W}^{-1}\mathbf{D}_i\mathbf{W}^{-1\prime}$ with $\mathbf{W}$ being a lower triangular matrix with ones on the diagonal and $\mathbf{D}_i = \operatorname{diag}\big(\exp(h_{1i}), \ldots, \exp(h_{qi})\big)$ is a diagonal matrix of variances (e.g., see Carriero et al., 2019). Each log-variance is assumed to follow a random walk process

$$h_{ji} = h_{ji-1} + \varepsilon_{ji}, \qquad \varepsilon_{ji} \sim \mathcal{N}(0, \sigma_j^2), \qquad j = 1, \ldots, q.$$

The model specification is completed by assuming a Gaussian prior for the free entries of $\mathbf{W}$, denoted $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \underline{\Omega})$, a conjugate prior for the variance, $\sigma_j^2 \sim \mathcal{IG}(\underline{a}_\sigma, \underline{b}_\sigma)$, and the initial point, $h_{j0} \sim \mathcal{N}(0, \underline{\varsigma}_j^2)$.

We investigate whether the pre-COVID period and the years following the outbreak of the COVID pandemic have similar drivers. Therefore, we consider two sub-periods by splitting the sample into the pre-COVID (2014Q1 to 2018Q4) and the (post)-COVID (2019Q1 to 2023Q4) periods, where the latter includes the outbreak of the pandemic and the subsequent recovery. Each period consists of $n = 20$ quarterly observations. Our interest lies, in particular, in investigating whether and how the estimation of the response allocation vector changes over time and quantifying the associated uncertainty.

In the first period, the estimated allocation is $\hat{\boldsymbol{\gamma}} = (0, 1, 1, 1, 1)$. However, upon inspecting the posterior distribution of $\boldsymbol{\gamma}$ in Fig. 3.4, it is evident that this result is highly uncertain. The allocation has posterior probability of 0.16, followed closely by $(0, 1, 0, 1, 0)$ at 0.12, and $(0, 1, 1, 1, 0)$ at 0.10. In contrast, the posterior distribution of $\boldsymbol{\gamma}$ for the period including the COVID pandemic exhibits a mode at $\hat{\boldsymbol{\gamma}} = (1, 0, 0, 0, 1)$. These results suggest a significant shift in the low-rank structure of the coefficient matrix, which moved from an almost reduced rank to an almost full rank structure. It is worth emphasising that the two periods differ not only in the point estimate


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

	γ_1	γ_2	γ_3	γ_4	γ_5	$p(\gamma\|Y)$
1	0	0	0	1	1	0.02275
2	0	0	1	0	1	0.0063
3	0	0	1	1	0	0.0339
4	0	0	1	1	1	0.01635
5	0	1	0	0	1	0.03805
6	0	1	0	1	0	0.1204
7	0	1	0	1	1	0.0545
8	0	1	1	0	0	0.03365
9	0	1	1	0	1	0.02625
10	0	1	1	1	0	0.09865
11	0	1	1	1	1	0.1605
12	1	0	0	0	1	0.01495
13	1	0	0	1	0	0.0031
14	1	0	0	1	1	0.0447
15	1	0	1	0	0	0.0012
16	1	0	1	0	1	0.0153
17	1	0	1	1	0	0.0435
18	1	0	1	1	1	0.0639
19	1	1	0	0	0	0.0013
20	1	1	0	0	1	0.0324
21	1	1	0	1	0	0.04085
22	1	1	0	1	1	0.03545
23	1	1	1	0	0	0.02845
24	1	1	1	0	1	0.0298
25	1	1	1	1	0	0.03375

	γ_1	γ_2	γ_3	γ_4	γ_5	$p(\gamma\|Y)$
1	0	0	0	1	1	0.0244
2	0	0	1	0	1	0.05575
3	0	0	1	1	0	0.00855
4	0	0	1	1	1	0.02105
5	0	1	0	0	1	0.0017
6	0	1	0	1	0	0.00525
7	0	1	0	1	1	0.0052
8	0	1	1	0	0	0.00025
9	0	1	1	0	1	0.0137
10	0	1	1	1	0	0.0091
11	0	1	1	1	1	0.06225
12	1	0	0	0	1	0.2334
13	1	0	0	1	0	0.01305
14	1	0	0	1	1	0.01805
15	1	0	1	0	0	0.1945
16	1	0	1	0	1	0.06795
17	1	0	1	1	0	0.02145
18	1	0	1	1	1	0.03765
19	1	1	0	0	0	0.0013
20	1	1	0	0	1	0.01995
21	1	1	0	1	0	0.0248
22	1	1	0	1	1	0.06245
23	1	1	1	0	0	0.01955
24	1	1	1	0	1	0.046
25	1	1	1	1	0	0.0327

Figure 3.4: Posterior distribution of the allocation vector, $\boldsymbol{\gamma}$, for the period 2014Q1-2018Q4 (left) and 2019Q1-2023Q4 (right).

of the allocation vector, $\hat{\boldsymbol{\gamma}}$, but also in the uncertainty about the estimate, which is higher in the pre-COVID period. The proposed method allows us to uncover both findings directly from the data, as opposed to the traditional PRR model with *a-priori fixed* allocation.

The uncertainty regarding parameter estimates for the 2014-2018 period is less pronounced in the rank's posterior distribution, which shows high probabilities for 1 and 2, with a clear single mode at 1 (see Fig. 3.5). Conversely, in the second period, we obtain a solid conclusion for a rank 1 model.

Overall, these findings suggest that in the 2014-2018 period the relationship between the responses and covariates is rather simple, as most responses are allocated to the low-rank group with an estimated rank of 1. However, the high uncertainty with posterior mass concentrated on models with large low-rank groups indicates the lack of a clear grouping and suggests that the most relevant feature to account for in this period is the overall reduced-rank structure of the coefficient matrix, which yields a total rank of 2 under the model with the maximum a posteriori $\boldsymbol{\gamma}$. A standard reduced-rank regression estimated a rank 3 coefficient matrix. Conversely, the 2019-2023 period features a small low-rank group with $\hat{r} = 1$ and a three-dimensional full-rank group, which is indicative of a significantly more complex structure of the relationship. The rank obtained under the traditional reduced-rank setting is again 3, failing to capture any increase in the complexity of the data from the previous period to the current one.

A possible reason for this change in the uncertainty of the posterior distribution for the grouping structure is the shift brought by COVID. Prior to the outbreak of the pandemic, the data were characterised by a large group with a simple structure, as pointed by $q_\gamma = 4$ and a low rank with small r across all models with high posterior probability. The uncertainty highlights the prominence of the reduced-rank feature over the qualitative information about which variables should be included in the group. Second, the outbreak of the pandemic could also explain the change towards more complex relationships, since in 2019-2023 we find $q_\gamma = 2$ with high probability (more than 45%).

Pertaining to the regression coefficients, in Fig. 3.6 we find evidence of a change in the relationship structure between the two periods. In the first period, median weekly earnings (x_2), and the price index of exports of goods and services (x_4) appear to have a negligible impact on

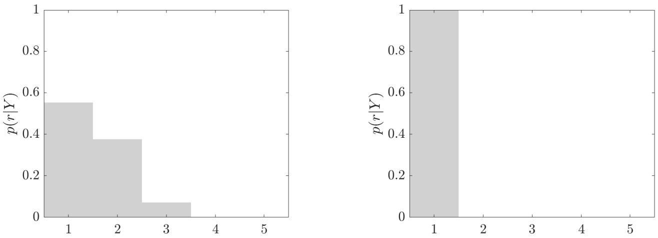


Figure 3.5: Posterior distribution of the rank, r, for the period 2014Q1-2018Q4 (left) and 2019Q1-2023Q4 (right).

explaining the responses of the low-rank group (y_2 to y_5). This pattern changes in the subsequent quarters, where the index of industrial production (y_1) and the volume index of exports of goods and services (y_5) exhibit a relationship with the covariates that are effectively captured by a rank-1 coefficient matrix. Furthermore, the weak signal of the covariates in the first period has strengthened in the second, and this estimation is more accurate in terms of the MSE compared to the other reduced-rank models presented in Section 3.4.[5]

The variation between the two periods and the associated uncertainty involved suggest a research direction concerning the incorporation of time-varying parameters within a time-series framework, which could potentially facilitate the identification of structural breaks.

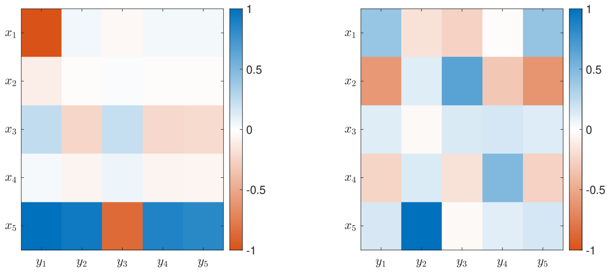


Figure 3.6: Posterior mean of the coefficient matrix, $\mathbf{C}$, for the period 2014Q1-2018Q4 (left) and 2019Q1-2023Q4 (right). Responses are labelled on the horizontal axis, and covariates on the vertical axis.

3.6 Concluding remarks

We have proposed a novel Bayesian approach to inference for a partial reduced rank regression model (BPRR). To circumvent the need for transdimensional samplers, we rely on a partially collapsed Gibbs sampler, where the allocation vector and the rank parameters are drawn from their joint distribution marginalised over the coefficient matrix. Then, a Metropolis-Hasting step with local exploration is used to draw the allocation vector, reducing the unfeasible exploring of the entire space of configurations to a computationally manageable local search.

The simulation study has highlighted the good performance of the model and the proposed partially collapsed Gibbs sampler algorithm. BPRR outperforms its competitors regarding the MSE and effectively estimates the allocation vector, the rank of the reduced-rank matrix, and the regression coefficients. The MCMC convergence diagnostics support the efficacy of the algorithm. Our approach's usefulness has also been demonstrated in real macroeconomic data, showing a

[5] $\text{MSE}_{\text{BPRR}} = 0.052$, $\text{MSE}_{\text{FR}} = 0.051$, $\text{MSE}_{\text{RR}} = 0.065$, $\text{MSE}_{\text{PRR}^*} = 0.062$.

significant shift in both the point estimates and posterior uncertainty about the allocation vector and the rank parameters since the outbreak of the COVID pandemic.

The proposed approach can be extended in several future directions. For instance, when dealing with time series data, it would be interesting to allow the allocation vector to vary over time, that is, γ_t. Another point worth exploring is the design of computational tools to speed up the inferential algorithm when sampling the allocation vector (e.g., Geels et al., 2023).

Chapter 4

Markov-Switching Partial Reduced-Rank Regression

4.1 Introduction

A particular form of cluster in multivariate data relates to the structure of the response variables in their relation to the covariates, as expressed by the regression coefficient matrix. The reduced-rank (RR) regression model (Anderson, 1951; Izenman, 1975; Reinsel et al., 2022) imposes a lower-rank constraint on the coefficient matrix, achieving a smaller number of relevant linear combinations of the predictor variables that explain the variation in all the responses. In contrast to traditional RR regression and its variants, commonly exploring covariate clustering (e.g., see Anderson, 1951; Velu, 1991; Li et al., 2019; Kim and Jung, 2024), we consider a Partial RR (PRR) regression model (Reinsel and Velu, 2006; Pintado et al., 2025), where the reduced-rank structure applies to only a subset of the response variables. The PRR framework first partitions the observed response variables into two clusters differing in the complexity of the relationship with the covariates. Specifically, a subset of the responses is assumed to be driven by a limited number of linear combinations of the covariates. In contrast, the other subset is characterised by a more complex relationship to the predictors through a full-rank coefficient matrix. Hence, the response vector and the coefficient matrix are partitioned into low- and full-rank groups. Rather than relying on an a priori fixed grouping structure, we treat the group memberships and rank as unknown parameters to be estimated in a Bayesian approach.

We generalise the PRRR framework in two directions. First, we replace the full-rank linear part of the model with a more flexible nonparametric term using a Gaussian process. Second, we introduce a Markov-switching process to allow for a time-varying and persistent clustering structure. This choice allows clustering of the response variables regarding the degree of complexity of their relationship with the covariates: one group retains a simple linear, low-rank specification, whereas the other considers a flexible nonparametric regression.

Gaussian processes (GP) define a prior distributions over functions, a useful feature for flexible nonparametric regression (see Neal, 1999; Murphy, 2012; Rasmussen and Williams, 2005), while its specification requires a mean and a kernel function over the input space. The function values are jointly Gaussian, and for two different input points, they will be similar if the inputs are considered to be similar by the kernel. In the context of time series, Bonnerjee et al. (2024) constructed a

Gaussian approximation for non-stationary time series, and show theoretical results for change-point detection and simultaneous inference in the presence of non-stationary errors. Cunningham et al. (2012) described a model for time series with multiple time markings in a multidimensional time-marked GP.

Nonparametric regression techniques to estimate unknown functional relations in the conditional mean have gained considerable attention in the statistics literature, particularly in improving their robustness to outliers (e.g., see Čížek and Sadıkoğlu, 2020; Salibian-Barrera, 2023). These models offer substantial flexibility, as they do not assume any particular form of the regression function, exhibit robustness to model misspecification, and can scale well in high dimensions. Nonetheless, they are also difficult to interpret, and shrinkage is often achieved through forcing the underlying response surface towards the space of smoothly varying functions.

However, when multivariate data observed at multiple time points are available, it is worthwhile to incorporate the time structure into the clustering result to understand how objects move between the clusters over time. A naïve approach to clustering multivariate time-series data is to apply a clustering technique independently at each time step. Nevertheless, interpreting the resulting clustering structure across time would be particularly difficult and hampered by the inconsistent cluster structures over time. Conversely, specifying a time-varying and dependent clustering would allow to address this issue as the structure itself would be considered persistent over time (e.g., see Maruotti and Punzo, 2017). Moreover, this approach could help relate changes in the clustering structure to specific events.

Several contributions in the literature adopt a dynamic clustering perspective. For instance, Creal et al. (2014) developed a model for credit ratings based on market data, with the main objective of classifying firms into different rating categories over time. They, therefore, allowed for transitions across clusters (dynamic clustering) while the parameters in their underlying mixture model were kept constant. In addition to clustering time-dependent data, Paci and Finazzi (2018) constructed a model that includes a spatial component to identify dynamic clusters in spatiotemporal data and applied it to assess air quality trends in Europe. Corneli et al. (2018) clustered the vertices in a dynamic network model while detecting change points in the intensities of the interactions over a continuous time interval. They relied on a non-homogeneous Poisson point process with intensity functions that depend on the hidden node clusters. These works focus primarily on how observations move between clusters over time. In contrast, our approach focuses not on clustering individual units, but on dynamically identifying the structural grouping of the response variables themselves, based on the complexity of their relationship with the covariates. Therefore, we adapt the PRR model to the Hidden Markov regression models (HMRMs) framework by introducing a Markov-switching process that drives the changes in the grouping structure and model parameters over time, accounting for time variation and persistence. HMRMs are state of the art in analysing time-dependent data (Hamilton, 1990; Frühwirth-Schnatter, 2006). Under the HMRM framework, the benchmark natural model for coping with continuous response vectors is represented by the HMRM based on a state-specific Gaussian distribution for the error term, which is referred to as Gaussian HMRM herein. Our proposed model that merges the partial reduced-rank specification with the Markov-Switching structure is called Markov-switching Bayesian Partial Reduced-Rank (MS-PRR) regression. Given the number of states, we design an MCMC algorithm that first samples the path of the hidden Markov chain and then estimates a different clustering structure of the response variables in each state. Conditionally on the states and grouping, all the remaining parameters are sampled from their full conditional distributions.

In a simulation study, the proposed model is tested on various synthetic datasets to investigate its performance in recovering the change points between different regimes and the associated clustering structure. The results show the effectiveness of the method in capturing the regime switches, as well as the underlying grouping structure of the response variables and the associated low-rank. Finally, we apply the model to a real macroeconomic dataset from the United States, finding evidence of shifts in the response grouping at certain periods corresponding to the early 2000s recession, the 2008 economic crisis and the COVID pandemic.

The remainder of the chapter is as follows. Section 4.2 introduces and describes the proposed MS-PRR model. Then, Section 4.3 presents the prior structure and provides details about sampling from the posterior distribution using an MCMC algorithm. Section 4.4 introduces time-varying volatility. To validate the algorithm, Section 4.5 tests its performance on synthetic data, whereas Section 4.6 provides an application to a real-world time series dataset. Section 4.7 draws the concluding remarks.

4.2 Markov-switching partial reduced-rank regression model

For each time point $t = 1, \dots, T$, let $\mathbf{y}_t \in \mathbb{R}^q$ be the vector of response variables, $\mathbf{x}_t \in \mathbb{R}^p$ the vector of explanatory variables, $\mathbf{C} \in \mathbb{R}^{p\times q}$ the matrix of regression coefficients, and $\mathbf{e}_t$ the noise vector associated to each observation t. The multivariate linear regression model is defined as

$$\mathbf{y}_t = \mathbf{C}'\mathbf{x}_t + \mathbf{e}_t, \qquad \mathbf{e}_t \sim \mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma}).$$

We introduce in the model a Markov-switching structure to simulate the switches available in the real data in each observation t, assuming the periods vary between K different states driven by a hidden Markov chain with transition matrix $\boldsymbol{\xi}$. Hence, the parameters are indexed by the state $k \in \{1, \dots, K\}$, and $\mathbf{s} = (s_1, s_2, \dots, s_T)$ are the latent states at each of the T observations.

Following the definitions in the previous chapters, we assume that each response variable can be split into two groups of dimensions q_{γ,s_t} and $q - q_{\gamma,s_t}$, where $q_{\gamma,s_t} \in \{2, \dots, q-1\}$. The first q_{γ,s_t} elements of $\mathbf{y}_t$ correspond to the first group, which admits a low-rank structure in its regression on $\mathbf{x}_t$. The relationship between the following $q - q_{\gamma,s_t}$ entries and $\mathbf{x}_t$ is assumed to have full rank. Thus, each vector of responses is described as $\mathbf{y}_t = (y_{t,1}, \dots, y_{t,q_{\gamma,s_t}}, y_{t,q_{\gamma,s_t}+1}, \dots, y_{t,q})' \in \mathbb{R}^q$, and an analogous structure follows for the error terms $\mathbf{e}_t$. Under this assumption, the matrix of coefficients has a low-rank and a full-rank component. Thus, the notation of the matrix of regression coefficients is determined by a low-rank component defined by a matrix $\mathbf{C}_{s_t} \in \mathbb{R}^{p\times q_{\gamma,s_t}}$ with reduced rank $r_{s_t} = \operatorname{rank}(\mathbf{C}_{s_t}) \le \min(p, q_{\gamma,s_t}) - 1$, and a flexible, more complex, component modelled nonparametrically by $\mathbf{f}_{s_t}(\mathbf{x}_t) \in \mathbb{R}^{q-q_{\gamma,s_t}}$. Therefore, the model can be represented as follows:

$$\mathbf{y}_t = \mathbf{V}'_{1s_t}\mathbf{C}'_{s_t}\mathbf{x}_t + \mathbf{V}'_{2s_t}\mathbf{f}_{s_t}(\mathbf{x}_t) + \mathbf{e}_t, \qquad \mathbf{e}_t \sim \mathcal{N}_q(\mathbf{0}, \mathbf{\Sigma}), \tag{4.1}$$

where $\mathbf{V}_{1s_t} = \left[\mathbf{I}_{q_{\gamma,s_t}}, \mathbf{0}_{q_{\gamma,s_t}\times(q-q_{\gamma,s_t})}\right] \in \mathbb{R}^{q_{\gamma,s_t}\times q}$, and $\mathbf{V}_{2s_t} = \left[\mathbf{0}_{(q-q_{\gamma,s_t})\times q_{\gamma,s_t}}, \mathbf{I}_{q-q_{\gamma,s_t}}\right] \in \mathbb{R}^{(q-q_{\gamma,s_t})\times q}$.

4.2.1 Prior specifications

As done in Chapter 3, we introduce a binary vector $\boldsymbol{\gamma}_k \in \{0,1\}^q$ for each state $k = 1, \dots, K$ to categorise the responses into the low-rank group and the flexible component. As we lack any prior information regarding the criteria for this classification, we assume that each element $\gamma_{j,k}$

($j = 1, \dots, q$) in each state $k = 1, \dots, K$ follows an independent Bernoulli prior distribution with probability ρ_k of being assigned to the low-rank group. Consequently, the joint prior distribution on $\boldsymbol{\gamma}_k$ is

$$p(\boldsymbol{\gamma}_k \mid \rho_k) = \left[\prod_{j=1}^{q} \text{Bern}(\gamma_{j,k} \mid \rho_k) \right] \mathbb{I}(1 < q_{\boldsymbol{\gamma},k} < q) \, , \tag{4.2}$$

where $q_{\boldsymbol{\gamma},k} = \sum_{j=1}^{q} \gamma_{j,k}$, and $\rho_k \in (0,1)$ is the prior probability of being assigned to the low-rank group. The constraint imposed by the indicator function in Eq. (4.2) allows for the existence of the low-rank group and, thus, of a PRR model. Additionally, ρ_k is assigned a Beta prior distribution, $\rho_k \sim \mathcal{B}e(\rho_k \mid \underline{a}_\rho, \underline{b}_\rho)$.

The matrix of coefficients $\mathbf{C}_k$ is assumed to have reduced rank $r_k \leq R_k = \min(p, q_{\boldsymbol{\gamma},k}) - 1$, which depends on the binary parameter $\boldsymbol{\gamma}_k$. Therefore, conditional on $q_{\boldsymbol{\gamma},k}$ (hence on $\boldsymbol{\gamma}_k$), we assume an non-informative uniform prior distribution for r_k over the discrete set $\{1, \dots, R_k\}$, that is $r_k \mid \boldsymbol{\gamma}_k \sim \mathcal{U}(r_k \mid \{1, \dots, R_k\})$.

Given that $\mathbf{C}_k$ is a low-rank matrix, we can express it as the product of two full-rank matrices $\mathbf{A}_k \in \mathbb{R}^{q_{\gamma,k} \times r_k}$ and $\mathbf{B}_k \in \mathbb{R}^{p \times r_k}$, such that $\mathbf{C}_k = \mathbf{B}_k \mathbf{A}_k'$. This decomposition is not unique, since for any orthogonal $r_k \times r_k$ matrix $\mathbf{P}$, we have that $\mathbf{C}_k = (\mathbf{B}_k \mathbf{P})(\mathbf{P}' \mathbf{A}_k')$. To achieve a unique decomposition of $\mathbf{C}_k$, we follow Geweke (1996) and impose an identifying restriction by assuming the first r_k rows of $\mathbf{A}_k$ are equal to the identity matrix $\mathbf{I}_{r_k}$, that is

$$\mathbf{A}_k = \begin{bmatrix} \mathbf{I}_{r_k} \\ \mathbf{A}_{0k} \end{bmatrix} ,$$

where $\mathbf{A}_{0k}$ is of dimension $(q_{\gamma,k} - r_k) \times r_k$. Denoting with $\text{vec}(\cdot)$ the vectorisation operator, we assume a multivariate Gaussian prior distribution on $\boldsymbol{\alpha}_k = \text{vec}(\mathbf{A}_{0k}')$, that is

$$\boldsymbol{\alpha}_k \mid \boldsymbol{\gamma}_k, r_k \sim \mathcal{N}_{(q_{\gamma,k} - r_k) r_k}(\boldsymbol{\alpha}_k \mid \mathbf{0}, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha},k}),$$

where $\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha},k} = \underline{a}\, \mathbf{I}_{(q_{\gamma,k} - r_k) r_k}$, for fixed $\underline{a} > 0$. Similarly, defining $\boldsymbol{\beta}_k = \text{vec}(\mathbf{B}_k)$, we assume a multivariate Gaussian prior distribution:

$$\boldsymbol{\beta}_k \mid \boldsymbol{\gamma}_k, r_k \sim \mathcal{N}_{p r_k}(\boldsymbol{\beta}_k \mid \mathbf{0}, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta},k}),$$

where $\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta},k} = \underline{b}\, \mathbf{I}_{p r_k}$. We adhere to the conventional practice and assign a conjugate inverse Wishart prior distribution to $\boldsymbol{\Sigma}$, that is $\boldsymbol{\Sigma} \sim \mathcal{IW}_q(\boldsymbol{\Sigma} \mid \underline{\nu}, \underline{\boldsymbol{\Psi}})$, with $\underline{\nu}$ and $\underline{\boldsymbol{\Psi}}$ being the fixed degrees of freedom and scale matrix, respectively. In the simulation experiment and real data application, we will relax this assumption by setting a stochastic volatility process. Lastly, we assign a Dirichlet prior distribution to each row of the transition matrix, $\boldsymbol{\Xi}$, for the latent Markov chain, that is $\boldsymbol{\xi}_k \sim \mathcal{D}ir(\underline{\mathbf{d}})$ for each $k = 1, \dots, K$, with $\underline{\mathbf{d}} \in \mathbb{R}_+^K$.

Moving to the other component, we adopt a Gaussian process prior for the function $f_{k,j} : \mathbb{R}^p \to \mathbb{R}$, that is $f_{k,j}(\mathbf{x}) \sim \mathcal{GP}\big(\underline{m}_k(\mathbf{x}), \underline{\boldsymbol{\Omega}}_k(\mathbf{x}, \mathbf{x}^*)\big)$, $j = 1, \dots, q - q_{\gamma,k}$. Specifically, we assume zero mean $\underline{m}_k(\mathbf{x}) = \mathbf{0}$ and covariance function from the Matérn class given by

$$\Omega(\mathbf{x}, \mathbf{x}^*) = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}\, \|\mathbf{x} - \mathbf{x}^*\|_2}{\zeta} \right)^{\nu} H_\nu \left(\frac{\sqrt{2\nu}\, \|\mathbf{x} - \mathbf{x}^*\|_2}{\zeta} \right) ,$$

where $\Gamma(\cdot)$ is the Gamma function, $H_\nu(\cdot)$ is a modified Bessel function, and ν, ζ and σ_f^2 are positive parameters. We follow the conventional practice of fixing ν, which controls the smoothness level of the function, as a half-integer, adhering to the popular choice $\nu = 3/2$ (Rasmussen and Williams, 2005, ch.4). This corresponds to the finite-dimensional prior $\mathbf{f}_{k,j} \sim \mathcal{N}_{T_k}(\mathbf{0}, \underline{\mathbf{\Omega}}_k)$, where the prior covariance matrix has generic element

$$\underline{\mathbf{\Omega}}_{k;i,l} = \sigma_f^2 \left(1 + \frac{\sqrt{3}\,\|\mathbf{x}_i - \mathbf{x}_l\|_2}{\zeta}\right) \exp\left(-\frac{\sqrt{3}\,\|\mathbf{x}_i - \mathbf{x}_l\|_2}{\zeta}\right), \tag{4.3}$$

for time points i and l such that $s_i = s_l = k$.

The hyperparameters, ζ and σ_f^2, of the Gaussian process prior control the behaviour of the function $f_{k,j}$. The signal variance σ_f^2 defines the amplitude of the function, thus the variation of the function values from the mean. The length scale ζ determines how quickly the process varies as a function of the input points, affecting the smoothness of the function. Hence, setting ζ too large yields a model which might miss higher frequency information, whereas a ζ set too small leads to overfitting. We assign independent Gamma priors to the two hyperparameters, as follows:

$$\sigma_f^2 \sim \mathcal{G}a\big(\sigma_f^2 \mid \underline{a}_{\sigma_f}, \underline{b}_{\sigma_f}\big), \qquad \zeta \sim \mathcal{G}a\big(\zeta \mid \underline{a}_{\zeta}, \underline{b}_{\zeta}\big),$$

where $\mathcal{G}a(\cdot)$ denotes the Gamma distribution, and $\underline{a}_{\sigma_f}, \underline{b}_{\sigma_f}, \underline{a}_{\zeta}, \underline{b}_{\zeta}$ are fixed positive values.

4.2.2 Definition of the likelihood function

Before moving to the posterior distributions, we clarify the notation and other derivations of importance for defining the likelihood function.

Let us begin by defining $\mathbf{Y} \in \mathbb{R}^{T \times q}$ as the matrix with row t as $\mathbf{y}_t'$. Consider now a generic state k, and denote by $\mathcal{T}_k$ the set of time indices assigned to the kth state, that is $\mathcal{T}_k = \{t \in \{1, \ldots, T\} : s_t = k\}$, and has cardinality $|\mathcal{T}_k| = T_k$. Let $\tilde{\mathbf{Y}}_k \in \mathbb{R}^{T_k \times q}$ denote the submatrix of $\mathbf{Y}$ containing only the rows with time index $t \in \mathcal{T}_k$. Similarly, $\tilde{\mathbf{X}}_k \in \mathbb{R}^{T_k \times p}$ denotes the matrix with row $t \in \mathcal{T}_k$ as $\mathbf{x}_t'$. Let $\tilde{\mathbf{y}}_k \in \mathbb{R}^{qT_k}$ denote the vector stacking the observations $\mathbf{y}_t$ for all the periods when $s_t = k$, that is $\tilde{\mathbf{y}}_k = \mathrm{vec}(\tilde{\mathbf{Y}}_k)$, and define analogously $\tilde{\mathbf{e}}_k \in \mathbb{R}^{qT_k}$.

Define the index set collecting the indices of all response variables allocated to the flexible group in state k as $Q_k = \{j \in \{1, \ldots, q\} : \gamma_{j,k} = 0\}$ with cardinality $q - q_{\gamma,k}$. Let $\mathbf{f} = \{f_{k,j}(\mathbf{x}_t) \in \mathbb{R} : k = 1, \ldots, K,\ j \in Q_k,\ t \in \mathcal{T}_k\}$ denote the collection of functions $f_{k,j}$ evaluated at each observed covariate. Meanwhile, $\mathbf{f}_k = \mathbf{f}_k(\mathbf{x}_t) = (f_{k,1}(\mathbf{x}_t), \ldots, f_{k,q-q_{\gamma,k}}(\mathbf{x}_t))' \in \mathbb{R}^{q-q_{\gamma,k}}$. Finally, $\tilde{\mathbf{f}}_k = \{f_{k,j}(\mathbf{x}_t) \in \mathbb{R} : s_t = k,\ j \in Q_k,\ t \in \mathcal{T}_k\}$ is the vector in $\mathbb{R}^{(q-q_{\gamma,k})T_k}$ stacking the values of $f_{k,j}$ evaluated at the observed covariate values, for all periods when $s_t = k$ and all responses belonging to the flexible group $j \in Q_k$.

The conditional likelihood can be rewritten as follows

$$p(\mathbf{Y} \mid \mathbf{s}, \mathbf{\Sigma}, \mathbf{A}, \mathbf{B}, \mathbf{f}) = \prod_{k=1}^{K} \prod_{t \in \mathcal{T}_k} p(\mathbf{y}_t \mid s_t = k, \mathbf{\Sigma}, \mathbf{A}_k, \mathbf{B}_k, \mathbf{f}_k).$$

Parting from Eq. (4.1), a vectorised model for all time periods where $s_t = k$ is given by

$$\tilde{\mathbf{y}}_k = \mathbf{U}_{1k}\mathbf{c}_k + \mathbf{U}_{2k}\tilde{\mathbf{f}}_k + \tilde{\mathbf{e}}_k, \qquad \tilde{\mathbf{e}}_k \sim \mathcal{N}_{qT_k}(\mathbf{0}, \tilde{\mathbf{\Sigma}}_k), \tag{4.4}$$

where $\mathbf{U}_{1k} = \mathbf{V}_{1k}' \otimes \tilde{\mathbf{X}}_k$, $\mathbf{c}_k = \mathrm{vec}(\mathbf{C}_k)$, $\mathbf{U}_{2k} = \mathbf{V}_{2k}' \otimes \mathbf{I}_{T_k}$ and $\tilde{\mathbf{\Sigma}}_k = \mathbf{\Sigma} \otimes \mathbf{I}_{T_k}$. Thus, it follows

$\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \mathbf{A}_k, \mathbf{B}_k, \tilde{\mathbf{f}}_k \sim \mathcal{N}_{qT_k}(\mathbf{U}_{1k}\mathbf{c}_k + \mathbf{U}_{2k}\tilde{\mathbf{f}}_k,\ \tilde{\boldsymbol{\Sigma}}_k)$.

Recall that the prior on the scalar-valued functions $f_{k,j}(\mathbf{x}) \sim \mathcal{GP}(\underline{m}_k(\mathbf{x}), \underline{\boldsymbol{\Omega}}_k(\mathbf{x}, \mathbf{x}^*))$ corresponds to the finite-dimensional prior $\mathbf{f}_{k,j} \sim \mathcal{N}_{T_k}(\mathbf{0}, \underline{\boldsymbol{\Omega}}_k)$, where the prior covariance matrix has generic element $\underline{\boldsymbol{\Omega}}_{k;i,l}$ defined in Eq. (4.3). This implies that the vector $\tilde{\mathbf{f}}_k$ has the prior distribution

$$\tilde{\mathbf{f}}_k \sim \mathcal{N}_{(q-q_{\gamma,k})T_k}(\mathbf{0}, \overline{\boldsymbol{\Omega}}_k),$$

where $\overline{\boldsymbol{\Omega}}_k = \mathbf{I}_{q-q_{\gamma,k}} \otimes \underline{\boldsymbol{\Omega}}_k$.

The likelihood is marginalized over $\tilde{\mathbf{f}}_k$ analytically to obtain

$$\begin{aligned} p(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\gamma}_k, r_k) &= \int p(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \mathbf{A}_k, \mathbf{B}_k, \tilde{\mathbf{f}}_k, \boldsymbol{\gamma}_k, r_k)\, p(\tilde{\mathbf{f}}_k \mid \mathbf{s}, \boldsymbol{\gamma}_k)\, d\tilde{\mathbf{f}}_k \\ &= \int \mathcal{N}_{qT_k}(\tilde{\mathbf{y}}_k \mid \mathbf{U}_{1k}\mathbf{c}_k + \mathbf{U}_{2k}\tilde{\mathbf{f}}_k,\ \tilde{\boldsymbol{\Sigma}}_k)\, \mathcal{N}_{(q-q_{\gamma,k})T_k}(\tilde{\mathbf{f}}_k \mid \mathbf{0}, \overline{\boldsymbol{\Omega}}_k)\, d\tilde{\mathbf{f}}_k \\ &= \mathcal{N}_{qT_k}(\tilde{\mathbf{y}}_k \mid \mathbf{U}_{1k}\mathbf{c}_k, \boldsymbol{\Sigma}_{\mathbf{y}k}), \end{aligned} \tag{4.5}$$

where $\boldsymbol{\Sigma}_{\mathbf{y}k} = \tilde{\boldsymbol{\Sigma}}_k + \mathbf{U}_{2k}\overline{\boldsymbol{\Omega}}_k\mathbf{U}_{2k}'$.

4.3 Posterior sampling

In this section, we design an MCMC algorithm to draw samples from the joint posterior distribution $p(\mathbf{s}, \mathbf{A}, \mathbf{B}, \mathbf{f}, \boldsymbol{\Sigma}, \mathbf{r}, \boldsymbol{\gamma}, \rho \mid \mathbf{Y})$.

The traditional Gibbs approach that samples from the full conditional distributions of the parameters is invalid in our setting, given that the dimensions of $\mathbf{A}_k$, $\mathbf{B}_k$, and $\mathbf{f}_k$, for each $k = 1, \ldots, K$, depend on the states of $\boldsymbol{\gamma}_k$ and r_k. Thus, the parameter space dimension may change across iterations of the MCMC algorithm, and the reversible jump MCMC (Robert and Casella, 1999) can deal with such dimensionality issues. Nonetheless, the practical implementation of the algorithm would require the definition of cross-model moves, which are hard to define and execute properly in complex settings like ours.

To overcome this challenge, we adapt the sampling strategy proposed in Chapter 3 (Pintado et al., 2025) to the present framework, and implement a partially collapsed Gibbs sampler (PCG, see van Dyk and Park, 2008). In particular, for each state k, we draw $(\boldsymbol{\gamma}_k, r_k)$ from a joint distribution marginalised over the parameters $(\mathbf{A}_k, \mathbf{B}_k, \mathbf{f}_k)$ whose sizes depend on $(\boldsymbol{\gamma}_k, r_k)$, which obviates the need of transdimensional samplers, similar to Yang et al. (2022). Afterwards, $(\mathbf{A}_k, \mathbf{B}_k, \mathbf{f}_k)$ is sampled conditionally on the updated values of $(\boldsymbol{\gamma}_k, r_k)$. The entire sampling process is summarised in Algorithm 4.

4.3.1 Sampling the state-specific allocations and rank

We aim at sampling $\boldsymbol{\gamma}_k$ from the conditional posterior marginalised over $(r_k, \mathbf{A}_k, \mathbf{B}_k, \mathbf{f}_k)$, that is

$$p(\boldsymbol{\gamma}_k \mid \tilde{\mathbf{y}}_k, \mathbf{s}, \boldsymbol{\Sigma}, \rho_k) = \frac{p_{\boldsymbol{\gamma}_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k) p(\boldsymbol{\gamma}_k \mid \rho_k)}{\sum_{\boldsymbol{\gamma}_k^\dagger \in \{0,1\}^q} p_{\boldsymbol{\gamma}_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k^\dagger) p(\boldsymbol{\gamma}_k^\dagger \mid \rho_k)},$$

Algorithm 4 PCG for Bayesian PRR model

```
for $k = 1, \dots, K$ do
    Sample $\boldsymbol{\gamma}_k$ from $p(\boldsymbol{\gamma}_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \rho_k)$.
    Sample $r_k$ from $p(r_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k)$.
    Sample $\tilde{\mathbf{f}}_k$ from $p(\tilde{\mathbf{f}}_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, \mathbf{A}_k, \mathbf{B}_k) = \mathcal{N}_{(q-q_{\gamma,k})T_k}(\overline{\boldsymbol{\mu}}_{\mathbf{f},k}, \overline{\boldsymbol{\Sigma}}_{\mathbf{f},k})$.
    Sample $\boldsymbol{\alpha}_k = \mathrm{vec}(\mathbf{A}'_{0k})$ from $p(\boldsymbol{\alpha}_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k, \mathbf{B}_k, \mathbf{f}_k) = \mathcal{N}_{(q_{\gamma,k}-r_k)r_k}(\overline{\boldsymbol{\mu}}_{\alpha,k}, \overline{\boldsymbol{\Sigma}}_{\alpha,k})$,
    then set $\mathbf{A}_k = [\mathbf{I}_{r_k}, \mathbf{A}'_{0k}]'$.
    Sample $\boldsymbol{\beta}_k = \mathrm{vec}(\mathbf{B}_k)$ from $p(\boldsymbol{\beta}_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k, \mathbf{A}_k, \mathbf{f}_k) = \mathcal{N}_{pr_k}(\overline{\boldsymbol{\mu}}_{\beta,k}, \overline{\boldsymbol{\Sigma}}_{\beta,k})$.
    Sample $\rho_k$ from $p(\rho_k \mid \boldsymbol{\gamma}_k) = \mathcal{B}e(\overline{a}_\rho, \overline{b}_\rho)$.
    Sample $\boldsymbol{\xi}_k$ from $p(\boldsymbol{\xi}_k \mid \mathbf{s}) = \mathcal{D}ir(\overline{\mathbf{d}}_\xi)$.
end for
Sample $\zeta$ from $p(\zeta \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\gamma}, \sigma_f^2)$.
Sample $\sigma_f^2$ from $p(\sigma_f^2 \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\gamma}, \zeta)$.
Sample $\mathbf{s}$ from $p(\mathbf{s} \mid \mathbf{Y}, \boldsymbol{\gamma}, \mathbf{A}, \mathbf{B}, \mathbf{f}, \boldsymbol{\Sigma}, \boldsymbol{\Xi})$.
Sample $\boldsymbol{\Sigma}$ from $p(\boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\gamma}, \mathbf{A}, \mathbf{B}, \mathbf{f}) = \mathcal{IW}_q(\overline{\nu}, \overline{\boldsymbol{\Psi}})$.
```

where $p_{\gamma_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k)$ is defined as

$$\begin{aligned} p_{\gamma_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k) &= \sum_{r_k=1}^{R_k} \frac{1}{R_k} \iint p(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\gamma}_k, r_k)\, p(\mathbf{A}_k, \mathbf{B}_k \mid \boldsymbol{\gamma}_k, r_k)\, \mathrm{d}\mathbf{A}_k\, \mathrm{d}\mathbf{B}_k \qquad (4.6) \\ &= \sum_{r_k=1}^{R_k} \frac{1}{R_k} p_{r_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k). \end{aligned}$$

We use the Laplace method (Raftery, 1995) to approximate the analytically intractable integration of $\mathbf{A}_k, \mathbf{B}_k$ to obtain

$$\log p_{r_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k) \approx \log p(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \hat{\mathbf{A}}_k, \hat{\mathbf{B}}_k, \boldsymbol{\gamma}_k, r_k) - \frac{1}{2}\big(pr_k + (q_{\gamma,k} - r_k)r_k\big) \log T_k, \qquad (4.7)$$

where $\hat{\mathbf{A}}_k$ and $\hat{\mathbf{B}}_k$ are the maximum likelihood estimators (MLEs) of $\mathbf{A}_k$ and $\mathbf{B}_k$, given $(\boldsymbol{\gamma}_k, r_k)$. Deriving these MLEs involves non-trivial steps arising from two main respects. First, the covariance matrix of the Gaussian density in Eq. (4.5) incorporates heteroscedastic errors through the dependence of $\mathbf{U}_{2k}$ on $\mathbf{V}_{2k}$. Second, two structural restrictions are imposed on the model: the identification restriction on the matrix $\mathbf{A}_k$, and the representation of the response vector as the sum of the low-rank and full-rank components in Eq. (4.1) through the binary matrices $\mathbf{V}_{1k}$ and $\mathbf{V}_{2k}$. We exploit the general class of reduced-rank regression models (GRRR) studied in Hansen (2002), where a ML estimation technique able to accommodate our setting is proposed.

The GRRR problem considers the regression for all periods when $s_t = k$

$$\mathbf{y}_t = \mathbf{V}'_{1k}\mathbf{A}_k\mathbf{B}'_k\mathbf{x}'_t + \bar{\mathbf{e}}_t,$$

for $t \in \mathcal{T}_k$, where $\bar{\mathbf{e}}_t$ is column t of $\bar{\mathbf{E}} \in \mathbb{R}^{q \times T_k}$, and $\mathrm{vec}(\bar{\mathbf{E}}) \sim \mathcal{N}_{qT_k}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{y}k})$, subject to the restriction

$$\mathrm{vec}(\mathbf{V}'_{1k}\mathbf{A}_k) = \mathbf{G}_k\, \mathrm{vec}(\mathbf{A}_{0k}) + \mathbf{g}_k,$$

where $\mathbf{G}_k$ is a binary matrix of dimension $qr_k \times r_k(q_{\gamma,k} - r_k)$ and $\mathbf{g}_k$ is the qr_k-dimensional vector of restrictions. In detail, $\mathbf{g}_k$ has entries with ones in the index set $\{(\ell - 1)(q + 1) + 1,\ \ \text{for each } \ell =$

$1, 2, \ldots, r_k\}$, and zeros in the remaining entries. The matrix $\mathbf{G}_k$ is a block matrix defined as

$$\mathbf{G}_k = [\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_{r_k}],$$
$$\mathbf{G}_{k\ell} = [\mathbf{0}_{(q_{\gamma,k}-r_k)\times(q(\ell-1)+r_k)}, \mathbf{I}_{q_{\gamma,k}-r_k}, \mathbf{0}_{(q_{\gamma,k}-r_k)\times(q(r_k-\ell+1)-q_{\gamma,k})}]', \qquad \ell = 1, 2, \ldots, r_k.$$

Notice that there are no constraints on the matrix $\mathbf{B}_k$. Then, the MLEs of $\boldsymbol{\alpha}_{\mathbf{V}_{1k}} = \mathrm{vec}(\mathbf{V}_{1k}\mathbf{A}_k)$ and $\boldsymbol{\beta}_k = \mathrm{vec}(\mathbf{B}_k)$ following the GRRR method are obtained as

$$\hat{\boldsymbol{\alpha}}_{\mathbf{V}_{1k}} = \mathbf{G}_k(\mathbf{G}_k'\mathbf{M}_{\mathbf{B}_k}\mathbf{G}_k)^{-1}\mathbf{G}_k'(\mathbf{n}_{\mathbf{B}_k} - \mathbf{M}_{\mathbf{B}_k}\mathbf{g}_k) + \mathbf{g}_k, \tag{4.8}$$
$$\hat{\boldsymbol{\beta}}_k = \mathbf{M}_{\mathbf{A}_k}^{-1}\mathbf{n}_{\mathbf{A}_k}, \tag{4.9}$$

where $\mathbf{M}_{\mathbf{B}_k} = (\tilde{\mathbf{X}}_k\mathbf{B}_k\otimes\mathbf{I}_q)'\tilde{\mathbf{\Sigma}}_{\mathbf{y}k}^{-1}(\tilde{\mathbf{X}}_k\mathbf{B}_k\otimes\mathbf{I}_q)$, $\mathbf{n}_{\mathbf{B}_k} = (\tilde{\mathbf{X}}_k\mathbf{B}_k\otimes\mathbf{I}_q)'\tilde{\mathbf{\Sigma}}_{\mathbf{y}k}^{-1}\,\mathrm{vec}(\tilde{\mathbf{Y}}_k')$, $\mathbf{M}_{\mathbf{A}_k} = \mathbf{K}_{p,r_k}'(\tilde{\mathbf{X}}_k\otimes \mathbf{V}_{1k}'\mathbf{A}_k)'\tilde{\mathbf{\Sigma}}_{\mathbf{y}k}^{-1}(\tilde{\mathbf{X}}_k\otimes\mathbf{V}_{1k}'\mathbf{A}_k)\mathbf{K}_{p,r_k}$, $\mathbf{n}_{\mathbf{A}_k} = \mathbf{K}_{p,r_k}'(\tilde{\mathbf{X}}_k\otimes\mathbf{V}_{1k}'\mathbf{A}_k)'\tilde{\mathbf{\Sigma}}_{\mathbf{y}k}^{-1}\,\mathrm{vec}(\tilde{\mathbf{Y}}_k')$, $\tilde{\mathbf{\Sigma}}_{\mathbf{y}k} = \mathbf{K}_{T_k,q}\mathbf{\Sigma}_{\mathbf{y}k}\mathbf{K}_{T_k,q}'$, and $\mathbf{K}_{m,n}$ is the $mn \times mn$ commutation matrix, which transforms the vectorisation of a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ into the vectorisation of its transpose, such that $\mathbf{K}_{m,n}\,\mathrm{vec}(\mathbf{M}) = \mathrm{vec}(\mathbf{M}')$.

The expressions in Eq. (4.8) and (4.9) depend on each other. Consequently, the practical implementation of the GRRR method is done in a recursive algorithm, updating $\boldsymbol{\alpha}_{\mathbf{V}_{1k}}$ and $\boldsymbol{\beta}_k$ iteratively until convergence. Once a solution $(\hat{\boldsymbol{\alpha}}_{\mathbf{V}_{1k}}, \hat{\boldsymbol{\beta}}_k)$ is obtained, it suffices to transform the vectorised MLEs back to their matrix forms $\mathbf{V}_{1k}\hat{\mathbf{A}}_k'$ and $\hat{\mathbf{B}}_k$ to obtain the MLEs of the low-rank coefficient matrix as $\hat{\mathbf{C}}_k = \hat{\mathbf{B}}_k(\mathbf{V}_{1k}')^{+}\mathbf{V}_{1k}'\hat{\mathbf{A}}_k = \hat{\mathbf{B}}_k\hat{\mathbf{A}}_k'$, where $\mathbf{M}^{+}$ denotes the Moore-Penrose pseudoinverse of $\mathbf{M}$.[1]

Therefore, we obtain the approximation

$$p_{\gamma_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \mathbf{\Sigma}, \boldsymbol{\gamma}_k) \approx \sum_{r_k=1}^{R_k} \frac{1}{R_k}\tilde{p}_{r_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \mathbf{\Sigma}, \boldsymbol{\gamma}_k, r_k) \equiv \tilde{p}_{\gamma_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \mathbf{\Sigma}, \boldsymbol{\gamma}_k),$$

where $\tilde{p}_{r_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \mathbf{\Sigma}, \boldsymbol{\gamma}_k, r_k)$ is the Laplace approximation of $p_{r_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \mathbf{\Sigma}, \boldsymbol{\gamma}_k, r_k)$ obtained from Eq. (4.7) to the integral in Eq. (4.6). Therefore, the (partially marginal) posterior distribution of $\boldsymbol{\gamma}_k$ is approximated by

$$\tilde{p}(\boldsymbol{\gamma}_k \mid \tilde{\mathbf{y}}_k, \mathbf{s}, \mathbf{\Sigma}, \rho_k) = \frac{\tilde{p}_{\gamma_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \mathbf{\Sigma}, \boldsymbol{\gamma}_k)p(\boldsymbol{\gamma}_k \mid \rho_k)}{\sum_{\boldsymbol{\gamma}_k^{\dagger}\in\{0,1\}^q}\tilde{p}_{\gamma_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \mathbf{\Sigma}, \boldsymbol{\gamma}_k^{\dagger})p(\boldsymbol{\gamma}_k^{\dagger} \mid \rho_k)}.$$

Provided that $\boldsymbol{\gamma}_k$ is a q-dimensional binary vector, the total number of possible configurations for the response allocation is 2^q, which computational challenge even for moderate values of q. To circumvent this problem, we employ an approximate sampling strategy for $\boldsymbol{\gamma}_k$, namely the Metropolized Shotgun Stochastic Search (MSSS) algorithm introduced by Hans et al. (2007). The MSSS algorithm efficiently explores the high-dimensional parameter space by evaluating a subset of neighbouring configurations of the current state of $\boldsymbol{\gamma}_k$. Restricting the neighbourhood to a manageable set introduces a balance between exploration of the parameter space and computational feasibility. Following the approach of Yang et al. (2022), we define the neighbourhood of $\boldsymbol{\gamma}_k$ to consist of all binary vectors that differ from $\boldsymbol{\gamma}_k$ in exactly one component, while still satisfying the existence of a low-rank group. This neighbourhood restriction improves computational efficiency compared to considering all possible configurations, while still enabling an effective (albeit reduced) exploration of the space (see Appendix B.3 for a detailed description of MSSS). A proposal

[1] $\mathbf{V}_{1k}$ is not invertible since its dimensionality is $q_{\gamma,k} \times q$, with $q_{\gamma,k} < q$.

distribution is defined by

$$g(\boldsymbol{\gamma}_k \mid \boldsymbol{\gamma}_k^{(m)}) \propto \tilde{p}(\boldsymbol{\gamma}_k \mid \tilde{\mathbf{y}}_k, \mathbf{s}, \boldsymbol{\Sigma}, \rho_k) \, \mathbb{I}\left(\boldsymbol{\gamma}_k \in \mathrm{nbd}(\boldsymbol{\gamma}_k^{(m)})\right),$$

where $\boldsymbol{\gamma}_k^{(m)}$ denotes the value of $\boldsymbol{\gamma}_k$ at the mth iteration of the MCMC. Hence, for the kth state of the hidden Markov chain, the proposed sampling scheme draws the allocation vector from the marginal posterior $p(\boldsymbol{\gamma}_k \mid \tilde{\mathbf{y}}_k, \mathbf{s}, \boldsymbol{\Sigma}, \rho_k)$ following the steps:

1. Generate $\boldsymbol{\gamma}_k^*$ from $g(\boldsymbol{\gamma}_k \mid \boldsymbol{\gamma}_k^{(m)})$.
2. Accept $\boldsymbol{\gamma}_k^{(m+1)} = \boldsymbol{\gamma}_k^*$ with probability

$$\rho_{\boldsymbol{\gamma}_k} = \min\left\{1, \frac{\sum_{\boldsymbol{\gamma}_k \in \mathrm{nbd}(\boldsymbol{\gamma}_k^{(m)})} \tilde{p}_\gamma(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k) p(\boldsymbol{\gamma}_k \mid \rho_k)}{\sum_{\boldsymbol{\gamma}_k^\dagger \in \mathrm{nbd}(\boldsymbol{\gamma}_k^*)} \tilde{p}_\gamma(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k^\dagger) p(\boldsymbol{\gamma}_k^\dagger \mid \rho_k)}\right\},$$

   and otherwise, set $\boldsymbol{\gamma}_k^{(m+1)} = \boldsymbol{\gamma}_k^{(m)}$.

The conditional posterior of r_k, marginalised over $(\mathbf{A}_k, \mathbf{B}_k, \mathbf{f}_k)$ is approximated through the Laplace method, following the same steps described for $\boldsymbol{\gamma}_k$. Specifically, we compute the approximated (partially marginal) posterior

$$\tilde{p}(r_k \mid \tilde{\mathbf{y}}_k, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k) = \frac{\tilde{p}_{r_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k) p(r_k \mid \boldsymbol{\gamma}_k)}{\sum_{r_k^\dagger = 1}^{R_k} \tilde{p}_{r_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k^\dagger) p(r_k^\dagger \mid \boldsymbol{\gamma}_k)}, \tag{4.10}$$

where $\tilde{p}_{r_k}(\tilde{\mathbf{y}}_k \mid \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k) p(r_k \mid \boldsymbol{\gamma}_k)$ is obtained in Eq. (4.7). Then, a new value of r_k is sampled from the discrete distribution on $\{1, \ldots, R_k\}$ with the probabilities given in Eq. (4.10).

This process is iterated for each state $k = 1, \ldots, K$ to sample the collections $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K\}$ and $\mathbf{r} = \{r_1, \ldots, r_K\}$.

4.3.2 Sampling the nonparametric term

Using the function space approach, we look for the posterior distribution of $\tilde{\mathbf{f}}_k$. The distribution of the vectorised model in Eq. (4.4) implies that the vector collecting all the partial residuals from observations allocated to state k, $\overline{\mathbf{y}}_{1k} = \tilde{\mathbf{y}}_k - \mathbf{U}_{1k}\mathbf{c}_k$, is distributed as

$$\overline{\mathbf{y}}_{1k} \mid \boldsymbol{\Sigma}, \tilde{\mathbf{f}}_k \sim \mathcal{N}_{qT_k}(\mathbf{U}_{2k}\tilde{\mathbf{f}}_k, \tilde{\boldsymbol{\Sigma}}_k). \tag{4.11}$$

Notice that the dependence on $(\mathbf{s}, \mathbf{A}_k, \mathbf{B}_k)$ is implicitly accounted for in the definition of $\overline{\mathbf{y}}_{1k}$. In this respect, we face a particular challenge in the proposed PCG sampler associated with the dimensions of $(\mathbf{A}_k, \mathbf{B}_k, \mathbf{f}_k)$. The dimensions of these parameters do not necessarily agree between subsequent iterations of the MCMC, given their dependence on $(\boldsymbol{\gamma}_k, r_k)$, which may not maintain the same values. Accordingly, we overcome this issue in the current Step 4 of Algorithm 4 by defining an auxiliary matrix $\mathbf{C}_k^*$ made of the first (newly sampled value) $q_{\gamma,k}$ columns of the matrix $\mathbf{C}_k$ updated in the previous MCMC iteration (Yang et al., 2022; Pintado et al., 2025). Hence, the vector $\overline{\mathbf{y}}_{1k}^* = \tilde{\mathbf{y}}_k - \mathbf{U}_{1k}\mathbf{c}_k^*$, where $\mathbf{c}_k^* = \mathrm{vec}(\mathbf{C}_k^*)$ maintains the distribution in Eq. (4.11), and we refer to this expression for the likelihood of $\tilde{\mathbf{f}}_k$.

Therefore, the posterior distribution of $\tilde{\mathbf{f}}_k$ can be obtained by applying Bayes' rule, resulting in

$$p(\tilde{\mathbf{f}}_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \mathbf{A}_k, \mathbf{B}_k, \boldsymbol{\gamma}_k) \propto \mathcal{N}_{(q-q_{\gamma,k})T_k}(\tilde{\mathbf{f}}_k \mid \mathbf{0}, \overline{\boldsymbol{\Omega}}_k)\mathcal{N}_{qT_k}(\overline{\mathbf{y}}_{1k}^* \mid \mathbf{U}_{2k}\tilde{\mathbf{f}}_k, \tilde{\boldsymbol{\Sigma}}_k)$$

$$\propto \mathcal{N}_{(q-q_{\gamma,k})T_k}(\tilde{\mathbf{f}}_k \mid \overline{\boldsymbol{\mu}}_{\mathbf{f},k}, \overline{\boldsymbol{\Sigma}}_{\mathbf{f},k}),$$

with $\overline{\boldsymbol{\Sigma}}_{\mathbf{f},k} = \left(\overline{\boldsymbol{\Omega}}_k^{-1} + \mathbf{U}'_{2k}\tilde{\boldsymbol{\Sigma}}_k^{-1}\mathbf{U}_{2k}\right)^{-1}$ and $\overline{\boldsymbol{\mu}}_{\mathbf{f},k} = \overline{\boldsymbol{\Sigma}}_{\mathbf{f},k}\mathbf{U}'_{2k}\tilde{\boldsymbol{\Sigma}}_k^{-1}\overline{\mathbf{y}}^*_{1k}$.

We approximate the posterior full conditional distributions of the hyperparameters ζ and σ_f^2 using a Griddy Gibbs approach (Ritter and Tanner, 1992; Tanner and and, 1987), which discretises the space over a fixed grid of length d_u for each parameter. First, let us define the vector $\tilde{\mathbf{y}}_{j,k} = \{y_{j,t} : t \in \mathcal{T}_k\} = (y_{j,t_{1k}}, \ldots, y_{j,t_{T_k}})' \in \mathbb{R}^{T_k}$, with $j \in Q_k$ and $k = 1, \ldots, K$. Then, recall from Eq. (4.3) that the kernel $\underline{\boldsymbol{\Omega}}_k \in \mathbb{R}^{T_k \times T_k}$ depends on the values of the hyperparmeters, and we write $\underline{\boldsymbol{\Omega}}_k(\sigma_f^2, \zeta)$ to emphasise this dependence. The full conditional distributions are

$$\zeta \mid \bullet \propto p(\zeta) \times \prod_{k=1}^{K} \prod_{j \in Q_k} \left|\underline{\boldsymbol{\Omega}}_k(\sigma_f^2, \zeta)\right|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\tilde{\mathbf{y}}'_{j,k}\underline{\boldsymbol{\Omega}}_k(\sigma_f^2, \zeta)^{-1}\tilde{\mathbf{y}}_{j,k}\right\},$$

$$\sigma_f^2 \mid \bullet \propto p(\sigma_f^2) \times \prod_{k=1}^{K} \prod_{j \in Q_k} \left|\underline{\boldsymbol{\Omega}}_k(\sigma_f^2, \zeta)\right|^{-1/2} \exp\left\{ -\frac{1}{2}\tilde{\mathbf{y}}'_{j,k}\underline{\boldsymbol{\Omega}}_k(\sigma_f^2, \zeta)^{-1}\tilde{\mathbf{y}}_{j,k}\right\}.$$

We use a Griddy Gibbs to sample from each of these distributions setting two unidimensional grids of 100 points.

Remark 3. The vector $\tilde{\mathbf{f}}_k$ contains the GP component of the regression for all the $q - q_{\gamma,k}$ variables allocated to the flexible group. Therefore, by drawing $\tilde{\mathbf{f}}_k$ we are jointly sampling all the vectors $\mathbf{f}_{k,1}, \ldots, \mathbf{f}_{k,q-q_{\gamma,k}}$. Since the posterior values of the hyperparameters involve computations with several sparse matrices, sampling from the $(q - q_{\gamma,k})T_k$-dimensional Gaussian can be made efficiently.

4.3.3 Sampling the latent states

An efficient approach to sampling the latent state chain in Markov switching models is multi-move sampling, which involves drawing the entire sequence of states, $\mathbf{s} = (s_1, \ldots, s_T)$, jointly from its conditional posterior distribution. We employ a forward-filtering-backward sampling algorithm (e.g., see Frühwirth-Schnatter, 2006, chap. 11) to sample a path of the hidden Markov chain. The steps in detail are described in Appendix C.1.

Remark 4. Markov-switching models are known to suffer from a label-switching problem, an identification issue due to invariance to permutations of the regime labels. A standard approach to address this challenge is to impose identification constraints. We perform online identification of the states by assigning labels in correspondence to the number of observations allocated to each regime, with $k = 1$ assigned to the regime with the largest observation count, and so on.

4.3.4 Sampling the other parameters

Regarding the sampling of the other parameters, e.g., Steps (5)-(8) of Algorithm 4, we describe them in the following paragraphs.

The update of $\mathbf{A}_k = [\mathbf{I}_{r_k}, \mathbf{A}'_{0k}]'$ given $(\mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k, \mathbf{B}_k, \mathbf{f}_k)$ is performed by sampling $\boldsymbol{\alpha}_k = \text{vec}(\mathbf{A}'_{0k})$. We define the matrix $\mathbf{B}^*_k$ following the same strategy as in Section 4.3.2 to avoid dimensional incompatibility. Recalling the identification constraint in $\mathbf{A}_k$ and the factorisation $\mathbf{C}_k = \mathbf{B}_k\mathbf{A}'_k$, one obtains $\mathbf{C}_k = [\mathbf{B}_k, \mathbf{B}_k\mathbf{A}'_{0k}]$. Thus, we select the first r_k columns of $\mathbf{C}_k$, using the newly sampled value of the rank in state k at the current MCMC iteration to construct $\mathbf{B}^*_k$.

Therefore, the posterior distribution of $\boldsymbol{\alpha}_k$ is proportional to the multivariate Gaussian distribution

$$\begin{aligned} p(\boldsymbol{\alpha}_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k, \mathbf{B}_k^*, \mathbf{f}_k) &\propto p(\boldsymbol{\alpha}_k \mid \boldsymbol{\gamma}_k, r_k)\, p(\mathbf{Y} \mid \mathbf{s}, \boldsymbol{\Sigma}, \mathbf{A}_k, \mathbf{B}_{k*}, \mathbf{f}_k) \\ &\propto \mathcal{N}_{(q_{\gamma,k}-r_k)r_k}(\boldsymbol{\alpha}_k \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{\alpha},k}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha},k}) \end{aligned}$$

with mean $\overline{\boldsymbol{\mu}}_{\boldsymbol{\alpha},k} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha},k}\big(\mathbf{m}_{[J_k]} - \mathbf{H}_{[J_k,J_k]}\mathbf{v}_k\big)$ and covariance matrix $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha},k} = \big(\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha},k}^{-1} + \mathbf{H}_{[J_k,J_k]}\big)^{-1}$, where $\mathbf{v}_k = \mathrm{vec}(\mathbf{I}_{r_k})$, $\mathbf{m}_k = \mathbf{M}'_{\boldsymbol{\alpha},k}\tilde{\boldsymbol{\Sigma}}_k^{-1}\overline{\mathbf{y}}_{2k}$, $\mathbf{H}_k = \mathbf{M}'_{\boldsymbol{\alpha},k}\tilde{\boldsymbol{\Sigma}}_k^{-1}\mathbf{M}_{\boldsymbol{\alpha},k}$, $\overline{\mathbf{y}}_{2k} = \tilde{\mathbf{y}}_k - \mathbf{U}_{2k}\tilde{\mathbf{f}}_k$, and $\mathbf{M}_{\boldsymbol{\alpha},k} = \mathbf{U}_{1k}(\mathbf{I}_{q_{\gamma,k}} \otimes \mathbf{B}_k^*)$. Moreover, $\mathbf{H}_{[J_k,J_k]}$ indicates the Jth row and the Jth column in $\mathbf{H}_k$ for the sequence $J_k = \big\{r_k^2+1, r_k^2+2, \dots, q_{\gamma,k}r_k\big\}$.

The conditional posterior distribution of $\boldsymbol{\beta}_k$, given $(\mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k, \mathbf{A}_k, \mathbf{f}_k)$, is proportional to the multivariate Gaussian distribution $\mathcal{N}_{pr_r}(\boldsymbol{\beta}_k \mid \overline{\boldsymbol{\mu}}_{\boldsymbol{\beta},k}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta},k})$, where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta},k} = (\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta},k}^{-1} + \mathbf{M}'_{\boldsymbol{\beta},k}\tilde{\boldsymbol{\Sigma}}_k^{-1}\mathbf{M}_{\boldsymbol{\beta},k})^{-1}$ and $\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta},k} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta},k}\mathbf{M}'_{\boldsymbol{\beta},k}\tilde{\boldsymbol{\Sigma}}_k^{-1}\overline{\mathbf{y}}_{2k}$, with $\mathbf{M}_{\boldsymbol{\beta},k} = \mathbf{U}_{1k}(\mathbf{A}_k \otimes \mathbf{I}_p)$.

The posterior distribution of ρ_k, the probability of a response variable belonging to the low-rank group in state k, is the Beta distribution $\mathcal{B}e(\rho_k \mid \overline{a}_\rho, \overline{b}_\rho)$, with $\overline{a}_\rho = \underline{a}_\rho + q_{\gamma,k}$, and $\overline{b}_\rho = \underline{b}_\rho + q - q_{\gamma,k}$.

Assuming that the initial distribution of $\mathbf{s}$ is independent of $\boldsymbol{\Xi}$, the rows $\boldsymbol{\xi}_k$ of $\boldsymbol{\Xi}$ are independent a posteriori, and are drawn from $\boldsymbol{\xi}_k \sim \mathcal{D}ir(\overline{\mathbf{d}})$, where $\overline{\mathbf{d}} = (d_{k1} + N_{k1}(\mathbf{s}), \dots, d_{kK} + N_{kK}(\mathbf{s}))$ and $N_{kl}(\mathbf{s}) = |\{s_{t-1}=k, s_t=l\}|$, with $|\cdot|$ representing the cardinality of a set. N_{kl} counts the numbers of transitions from state k to state l for the current draw of $\mathbf{s}$. (Frühwirth-Schnatter, 2006, ch. 11.5)

Finally, the conditional posterior of the innovation covariance matrix $\boldsymbol{\Sigma}$, given $(\mathbf{Y}, \mathbf{A}, \mathbf{B}, \mathbf{f})$, is the inverse Wishart $\mathcal{IW}_q(\boldsymbol{\Sigma} \mid \overline{\nu}, \overline{\boldsymbol{\Psi}})$, where $\overline{\nu} = \underline{\nu} + T$ and $\overline{\boldsymbol{\Psi}} = \underline{\boldsymbol{\Psi}} + (\mathbf{Y} - \mathbf{M})'(\mathbf{Y} - \mathbf{M})$, letting $\mathbf{M}$ denote the $T \times q$ matrix collecting all the mean values of $\mathbf{Y}$ at every time point.

4.4 Model with SV

Given the model in the previous section, we introduce time variation in the covariance matrix by assuming the decomposition $\boldsymbol{\Sigma}_t = \mathbf{W}^{-1}\mathbf{D}_t\mathbf{W}^{-1\prime}$, where $\mathbf{W}$ is a lower triangular matrix with ones on the diagonal and $\mathbf{D}_t = \mathrm{diag}\big(\exp(h_{1t}), \dots, \exp(h_{qt})\big)$ with

$$h_{jt} = h_{jt-1} + \varepsilon_{jt}, \qquad \varepsilon_{jt} \sim \mathcal{N}(0, \sigma_j^2), \qquad j = 1, \dots, q.$$

The free entries of $\mathbf{W}$, denoted $\mathbf{w}$, are assigned a Gaussian prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \underline{\Omega}_w)$ and have a Gaussian posterior. Let us first rewrite the model as

$$\mathbf{W}\underbrace{(\mathbf{y}_t - \mathbf{V}'_{1s_t}\mathbf{C}'_{s_t}\mathbf{x}_t - \mathbf{V}'_{2s_t}\mathbf{f}_{s_t}(\mathbf{x}_t))}_{\check{\mathbf{y}}_t} = \mathbf{D}_t^{1/2}\mathbf{e}_t.$$

Owing to the lower triangular structure of $\mathbf{W}$, this system consists of a set of $j = 2, \dots, q$ equations, with equation j having as dependent variable $\check{y}_{jt}$ and as independent variables $-\check{y}_{\ell t}$, with $\ell = 1, \dots, j-1$, with coefficients $w_{j\ell}$. Therefore, multiplying the jth equation by $d_{jt}^{1/2}$ removes the heteroskedasticity associated with stochastic volatility. Replicating this approach separately for each transformed equation j leads to a Gaussian posterior distribution for the vector of coefficients of the jth equation, $\mathbf{w}_j = (w_{j1}, \dots, w_{jj-1})$'. We refer to Cogley and Sargent (2005) for further details.

The path of the stochastic volatility, $\mathbf{h}_j = (h_{j1}, \dots, h_{jT})'$, for each $j = 1, \dots, q$, is drawn jointly from the approximate posterior distribution obtained using the mixture approach of Omori et al.

(2007). It is well-suited to our context because the mixture-of-normals approximation allows for efficient and accurate update using a standard Gibbs sampler, while also accounting for leverage, an important feature for capturing the dynamics observed in real-data applications. Finally, assuming $\sigma_j^2 \sim \mathcal{IG}(\underline{a}_\sigma, \underline{b}_\sigma)$, its posterior full conditional distribution is conjugate $\sigma_j^2 | \mathbf{h}_j \sim \mathcal{IG}(\overline{a}_\sigma, \overline{b}_\sigma)$, where $\overline{a}_\sigma = \underline{a}_\sigma + n$ and $\overline{b}_\sigma = \underline{b}_\sigma + \sum_{t=1}^T (h_{jt} - h_{jt-1})^2$. The partially collapsed Gibbs sampler for our model with stochastic volatility is summarised in Algorithm 5.

Algorithm 5 PCG for Bayesian MS-PRR-SV model

```
for $k = 1, \dots, K$ do
    Sample $\boldsymbol{\gamma}_k$ from $p(\boldsymbol{\gamma}_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \rho_k)$.
    Sample $r_k$ from $p(r_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k)$.
    Sample $\overline{\mathbf{f}}_k$ from $p(\tilde{\mathbf{f}}_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, \mathbf{A}_k, \mathbf{B}_k) = \mathcal{N}_{(q-q_{\gamma,k})T_k}(\overline{\boldsymbol{\mu}}_{\mathbf{f},k}, \overline{\boldsymbol{\Sigma}}_{\mathbf{f},k})$.
    Sample $\boldsymbol{\alpha}_k = \mathrm{vec}(\mathbf{A}_{0k}')$ from $p(\boldsymbol{\alpha}_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k, \mathbf{B}_k, \mathbf{f}_k) = \mathcal{N}_{(q_{\gamma,k}-r_k)r_k}(\overline{\boldsymbol{\mu}}_{\alpha,k}, \overline{\boldsymbol{\Sigma}}_{\alpha,k})$,
    then set $\mathbf{A}_k = [\mathbf{I}_{r_k}, \mathbf{A}_{0k}']'$.
    Sample $\boldsymbol{\beta}_k = \mathrm{vec}(\mathbf{B}_k)$ from $p(\boldsymbol{\beta}_k \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}_k, r_k, \mathbf{A}_k, \mathbf{f}_k) = \mathcal{N}_{pr_k}(\overline{\boldsymbol{\mu}}_{\beta,k}, \overline{\boldsymbol{\Sigma}}_{\beta,k})$.
    Sample $\rho_k$ from $p(\rho_k \mid \boldsymbol{\gamma}_k) = \mathcal{B}e(\overline{a}_\rho, \overline{b}_\rho)$.
    Sample $\boldsymbol{\xi}_k$ from $p(\boldsymbol{\xi}_k \mid \mathbf{s}) = \mathcal{D}ir(\overline{\mathbf{d}}_\xi)$.
end for
Sample $\zeta$ from $p(\zeta \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\gamma}, \sigma_f^2)$.
Sample $\sigma_f^2$ from $p(\sigma_f^2 \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\gamma}, \zeta)$.
Sample $\mathbf{s}$ from $p(\mathbf{s} \mid \mathbf{Y}, \boldsymbol{\gamma}, \mathbf{A}, \mathbf{B}, \mathbf{f}, \boldsymbol{\Sigma}, \boldsymbol{\Xi})$.
for $j = 1, \dots, q$ do
    Sample $\mathbf{h}_j$ from $p(\mathbf{h}_j \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\gamma}, \mathbf{A}, \mathbf{B}, \mathbf{f}, \mathbf{W}, \sigma_j^2)$,
    using the auxiliary mixture sampler of Omori et al. (2007).
    Sample $\sigma_j^2$ from $p(\sigma_j^2 \mid \mathbf{h}_j) = \mathcal{IG}(\overline{b}_{\sigma,j}, \overline{b}_{\sigma,j})$.
end for
Sample $\mathbf{w}$ from $p(\mathbf{w} \mid \mathbf{Y}, \mathbf{s}, \boldsymbol{\gamma}, \mathbf{A}, \mathbf{B}, \mathbf{f}, \mathbf{h}) = \mathcal{N}_{q(q-1)/2}(\overline{\boldsymbol{\mu}}_w, \overline{\boldsymbol{\Sigma}}_w)$,
then set $\boldsymbol{\Sigma} = \{\mathbf{W}^{-1}\mathbf{D}_t\mathbf{W}^{-1\prime}\}_{t=1}^T$.
```

4.5 Simulation study

In this section, we evaluate the performance of the proposed model with respect to its ability to recover the latent state chain, and the corresponding true group allocations of the response variables into the low-rank and the flexible classifications within each state. Then, we assess the effectiveness of the sampling procedure in estimating the rank and recovering the matrix of mean values of $\mathbf{Y}$, denoted by $\mathbf{M}$.

All simulation settings use the following configuration: $K = 2$, $p = q = 5$, $\mathbf{q}_\gamma = (4, 2)$, $\mathbf{r} = (2, 1)$, and $n = 100$. The data generating process (DGP) is described in detail as follows. The rows of the design matrix $\mathbf{X}$ were sampled independently from $\mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$. The error terms, $\mathbf{e}_t$, were drawn from a multivariate normal distribution with a diagonal covariance matrix whose entries were uniformly chosen over the interval $(0.1, 1)$. The entries of the low-rank matrices $\mathbf{C}_k$ were sampled uniformly from $(-3, -1.5) \cup (1.5, 3)$. To enforce the reduced-rank structure, the smallest $p - r_k$ singular values of the matrix were set to zero. The nonparametric term was generated as a sine wave with amplitude 2, and the allocation of the response variables to the reduced-rank group was randomly assigned, given $q_{\gamma,k}$.

We consider three distinct settings, which primarily differ in the composition of the hidden Markov chain and the treatment of volatility in the model. Specifically:

1. $\mathbf{s}$ is generated with a single regime switch at the midpoint of the time series, such that $s_1, \ldots, s_{50} = 1$ and $s_{51}, \ldots, s_{100} = 2$. Estimation is performed under the assumption of constant volatility.

2. $\mathbf{s}$ is generated with random switches between the two regimes, and the model again assumes constant volatility.

3. $\mathbf{s}$ is generated with random switching, but estimation is conducted under a model that allows for stochastic volatility.

We run the MS-PRR for 11000 MCMC iterations after discarding an additional 1000 iterations as burnin, and the hyperparameters are set to reflect noninformative priors. In detail, $\underline{a}_\rho = \underline{b}_\rho = 1$, $\underline{a} = \underline{b} = 2$, $\underline{\nu} = q + 1$, $\underline{\mathbf{\Psi}} = \mathbf{I}_q$, and $\underline{\mathbf{d}}$ is a K-dimensional vector of ones.

The point estimates of the binary allocation vectors $\hat{\boldsymbol{\gamma}}$, the ranks $\hat{\mathbf{r}}$, and the latent states $\hat{\mathbf{s}}$ are obtained as their respective maximum a posteriori (MAP) estimates. The performance of the estimator $\widehat{\mathbf{M}}$ of the mean response matrix is assessed using the mean squared error (MSE), defined as $\text{MSE} = \|\widehat{\mathbf{M}} - \mathbf{M}\|_{\text{F}}^2/(Tq)$, where $\widehat{\mathbf{M}}$ is the posterior average of the predicted response matrix. We compute the mean absolute error (MAE), $\text{MAE} = \sum_{t=1}^{T} |\hat{\mathbf{s}}_t - \mathbf{s}_t|/T$, as a measure of distance between the estimated chain and the true latent states. To evaluate the classification performance of the allocation parameter, we use accuracy and the F_1 score as metrics, both ranging from 0 to 1. Higher values indicate better agreement between the estimated and true response groupings, with 1 indicating a perfect classification.

Table 4.1 summarises the simulation results by providing the average MSE and MAE over 20 independent experiments of each setting. For each state $k = 1, 2$, the average of the estimated number of low-rank responses, $\hat{q}_{\gamma,k}$, the estimated rank, $\hat{r}_k$, the accuracy and the F_1 score in each setting are provided. The proposed method accurately recovers the true latent state sequence, as indicated by the low mean absolute error. The allocation vectors are also accurately estimated, with accuracy and F_1 scores consistently attaining values close to 1. Additionally, the estimated ranks closely match the true ranks on average, demonstrating the effectiveness of the approach in recovering the underlying model structure.

	Setting					
	1		2		3	
MSE	0.530		0.468		0.427	
MAE	0.098		0.003		0.002	
k	1	2	1	2	1	2
$\hat{q}_{\gamma,k}$	4.000	2.300	4.000	2.200	4.000	2.000
$\hat{r}_k$	1.900	1.150	2.100	1.000	2.100	1.000
Accuracy	0.960	0.900	1.000	0.960	1.000	1.000
F_1 score	0.975	0.917	1.000	0.960	1.000	1.000

Table 4.1: Average MSE and MAE over 20 replicates of three simulation settings, where $K = 2$, $p = q = 5$, $\mathbf{q}_\gamma = (4, 2)$, $\mathbf{r} = (2, 1)$, and $n = 100$. For each state $k = 1, 2$, the average of the estimated number of low-rank responses, $\hat{q}_{\gamma,k}$, the estimated rank, $\hat{r}_k$, the accuracy and the F_1 score in each setting are provided.

The results discussed above are further corroborated by visual inspection of the trace plots from one run of each simulation setting. As shown in Figure 4.1, the hidden Markov chain is accurately recovered across iterations, with a distinct separation between states, even in the presence of random switching. The model also successfully identifies the true response groupings and their associated ranks (Figures 4.2, 4.3 and 4.4). In setting 3, which involves a model where the conditional variance process is misspecified, Figure 4.5 showcases both the true and estimated mean matrices, demonstrating that the proposed method maintains a high level of accuracy. Additionally, the trace plots of the Gaussian process hyperparameters suggest good mixing properties. Overall, these results support the robustness and effectiveness of the proposed approach across varying conditions.

The code for the MS-PRR algorithm has been implemented in MATLAB 2021a, and run on a MacBook Pro M1 2020 computer with 8 GB RAM. The average computational time for running 100 iterations of a model with the dimensions specified in this Section ($K = 2$, $q = 5$, $p = 5$, and $n = 100$) is approximately 362 seconds.

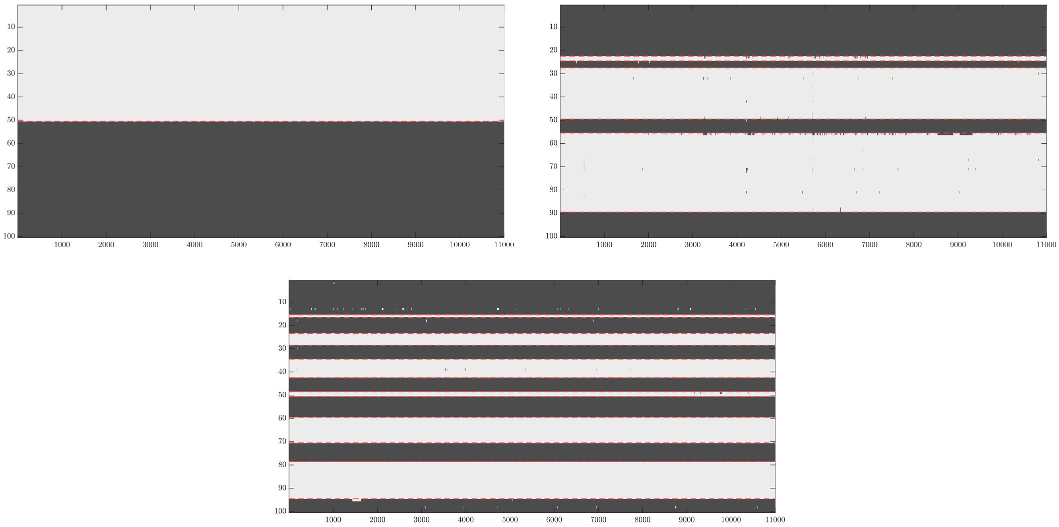


Figure 4.1: Trace plots of $\mathbf{s}$ across 11000 MCMC iterations for the different simulation scenarios: setting 1 (top left), setting 2 (top right) and setting 3 (bottom). Each colour represents one state, while a dashed horizontal red line indicates the time points where a true regime switch occurs.

4.6 Application

In the previous chapter, we demonstrated the practical applicability of the BPRR method by analysing quarterly macroeconomic data from the United States covering the years from 2014 to 2023. The dataset was partitioned into two distinct periods, pre-COVID (2014Q1–2018Q4) and post-COVID (2019Q1–2023Q4), to explore whether the years before and after the pandemic outbreak exhibit similar drivers and structure. The results indicated a substantial shift in the low-rank structure of the coefficient matrix, transitioning from a nearly fully reduced-rank form in the pre-COVID era to a structure approaching full rank in the subsequent period. Moreover, a change in the regression structure was observed over time.

Building upon these findings, we now apply the MS-PRR model by extending the temporal coverage to include data from 2000Q1 to 2023Q4. The aim is to investigate how the response allocation vector evolves over time and to quantify the associated uncertainty. Specifically, we seek to capture temporal changes in the regression structure as regime switches in a hidden Markov

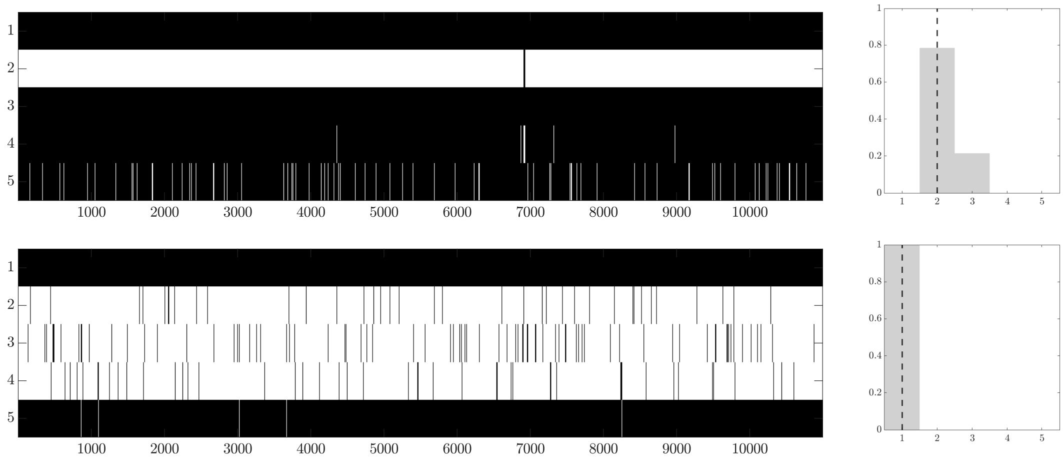


Figure 4.2: Trace plots of $\boldsymbol{\gamma}$ and posterior distribution of $\mathbf{r}$ across 11000 MCMC iterations for setting 1. The top row is state 1, where the true allocation vector is $\boldsymbol{\gamma}_1 = (1, 0, 1, 1, 1)$. The second row is state 2, where the true allocation vector is $\boldsymbol{\gamma}_1 = (1, 0, 0, 0, 1)$. The true values of the rank are indicated with dashed vertical lines.

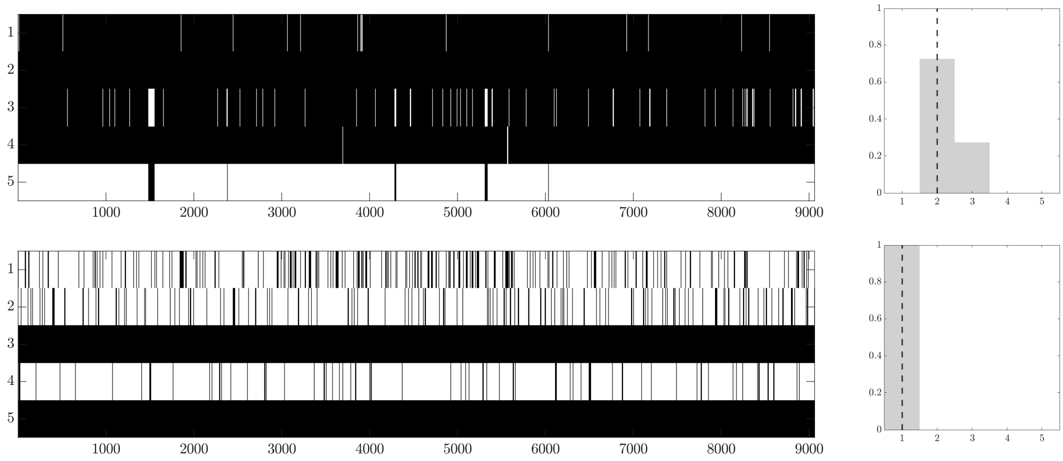


Figure 4.3: Trace plots of $\boldsymbol{\gamma}$ and posterior distribution of $\mathbf{r}$ across 11000 MCMC iterations for setting 2. The top row is state 1, where the true allocation vector is $\boldsymbol{\gamma}_1 = (1, 1, 1, 1, 0)$. The second row is state 2, where the true allocation vector is $\boldsymbol{\gamma}_1 = (0, 0, 1, 0, 1)$. The true values of the rank are indicated with dashed vertical lines.

chain. We set the number of regimes to $K = 2$ for interpretability and motivated by the structural differences identified between the pre- and post-COVID periods in Chapter 3. To account for time dependence in the data, we further introduce stochastic volatility in the innovation covariance structure, as described in Section 4.4.

Figure 4.6 reveals that our method has identified three main regime switches, depicted by the distinct clusters of horizontal dark lines. These transitions align with three major economic downturns during the analysed years: the early 2000s recession, the 2008 financial crisis, and the COVID recession.

The estimated allocations are $\hat{\boldsymbol{\gamma}}_1 = (0, 1, 0, 1, 1)$ and $\hat{\boldsymbol{\gamma}}_2 = (0, 0, 0, 1, 1)$, each of them in correspondence to a rank-1 coefficient matrix for the low-rank group. Notice in Figure 4.7 that the posterior distribution of $\boldsymbol{\gamma}$ differs across states not only in the mode, but also in terms of the

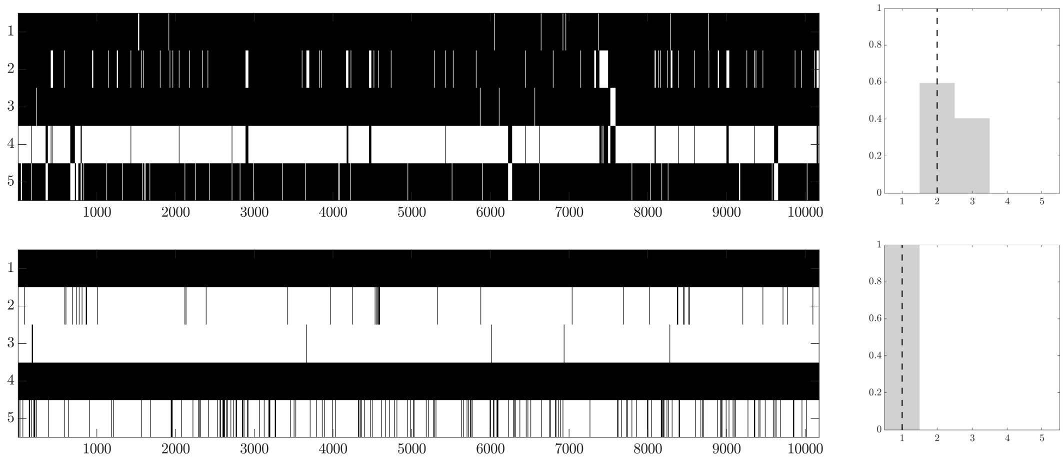


Figure 4.4: Trace plots of γ and posterior distribution of $\mathbf{r}$ across 11000 MCMC iterations for setting 3. The top row is state 1, where the true allocation vector is $\gamma_1 = (1, 1, 1, 0, 1)$. The second row is state 2, where the true allocation vector is $\gamma_1 = (1, 0, 0, 1, 0)$. The true values of the rank are indicated with dashed vertical lines.

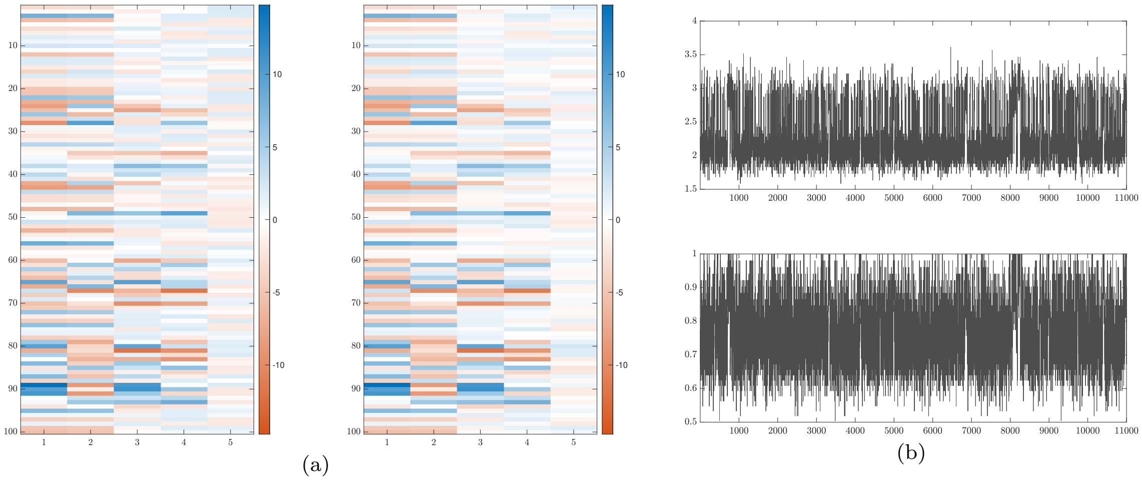


Figure 4.5: Panel (a): true mean matrix (left) and estimated mean matrix (right) in setting 3. Panel (b): trace plots of the Gaussian process hyperparameters σ_f^2 (top) and ζ (bottom) in setting 3.

number of candidate vectors (rows) having positive posterior probability throughout the MCMC. Notably, the second state exhibits greater variability in the allocation structure, with its posterior mass distributed across multiple candidate vectors, indicating greater uncertainty in the grouping structure. In contrast, the first state has most of the posterior mass concentrated on a single vector, suggesting a more stable partition. This behaviour is motivated by the fact that the second state corresponds to three short-lived periods of financial instability. Hence, a small number of observations combined with an uninformative uniform prior on the allocation vector resulted in higher posterior uncertainty. Meanwhile, the first state includes the majority of the observations, thus obtaining greater posterior concentration.

The first state has a higher number of responses allocated to the low-rank group than the second state. We emphasise that this finding agrees with the results from the previous chapter, where the observations belonging to the period before the outbreak of the COVID pandemic

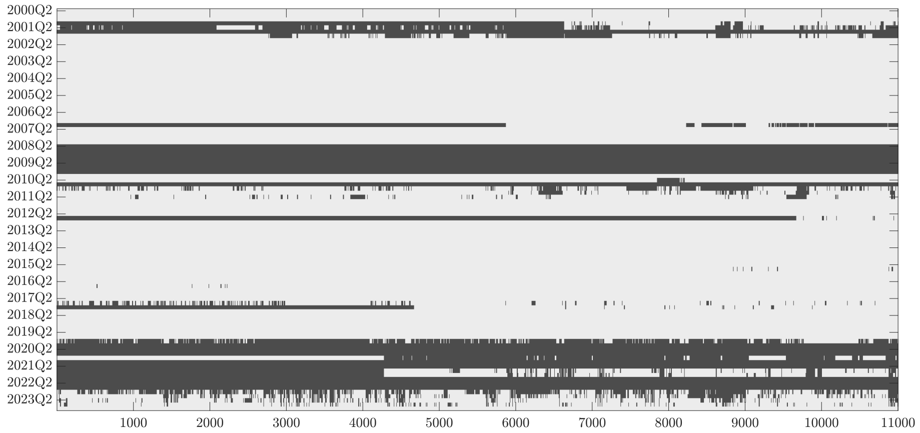
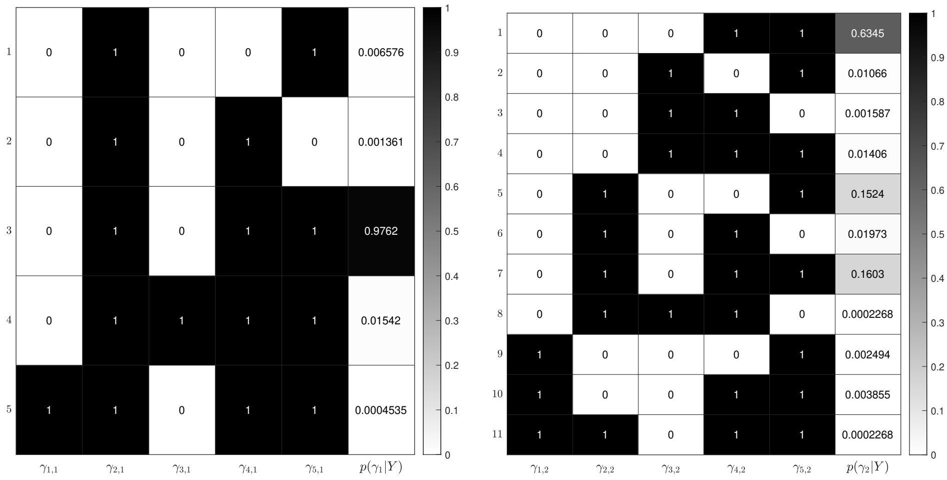


Figure 4.6: Trace plot of $\mathbf{s}$, where a light colour represents state 1 and a darker colour corresponds to state 2.



	$\gamma_{1,1}$	$\gamma_{2,1}$	$\gamma_{3,1}$	$\gamma_{4,1}$	$\gamma_{5,1}$	$p(\gamma_1\|Y)$
1	0	1	0	0	1	0.006576
2	0	1	0	1	0	0.001361
3	0	1	0	1	1	0.9762
4	0	1	1	1	1	0.01542
5	1	1	0	1	1	0.0004535

	$\gamma_{1,2}$	$\gamma_{2,2}$	$\gamma_{3,2}$	$\gamma_{4,2}$	$\gamma_{5,2}$	$p(\gamma_2\|Y)$
1	0	0	0	1	1	0.6345
2	0	0	1	0	1	0.01066
3	0	0	1	1	0	0.001587
4	0	0	1	1	1	0.01406
5	0	1	0	0	1	0.1524
6	0	1	0	1	0	0.01973
7	0	1	0	1	1	0.1603
8	0	1	1	1	0	0.0002268
9	1	0	0	0	1	0.002494
10	1	0	0	1	1	0.003855
11	1	1	0	1	1	0.0002268

Figure 4.7: Posterior distribution of the allocation vector, $\boldsymbol{\gamma}_k$, for the first state (left) and the second state (right).

exhibited a simpler relationship between the dependent and the independent variables. During the second state, this dynamic has shifted towards more complex structures, in agreement with the perturbations implied by unstable periods. Accordingly, the weak signals in periods of stability become stronger during those of economic struggle (Figure 4.8).

4.7 Concluding remarks

We have proposed MS-PRR, a novel Bayesian extension of the partial reduced-rank regression model incorporating both time-varying parameters through a Markov-switching mechanism and further flexibility offered by a Gaussian process on complex terms of the regression. Our approach combines the interpretability and parsimony of reduced-rank regression with the adaptability of Gaussian processes for modelling complex dependencies, and the dynamic capabilities of HMRMs for detecting regime changes, particularly in the response clustering. The result is a flexible yet structured framework for modelling multivariate time series with evolving relationships among the responses and the covariates.

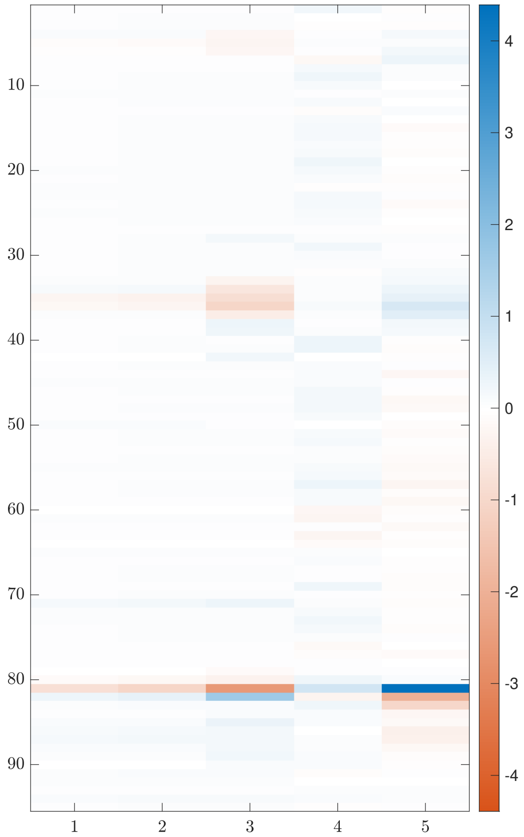


Figure 4.8: Posterior mean of the estimated model. Responses are labelled on the horizontal axis, and time points on the vertical axis.

Through simulation studies, we demonstrated the model's effectiveness in recovering both latent regime transitions and the associated response grouping structures, while providing uncertainty quantification about these estimations. In the empirical application to U.S. macroeconomic data, the proposed method successfully identified periods of changes in the economy, corresponding to major economic crises.

Overall, the MS-PRR framework is an interpretable approach towards intricate structures in reduced-rank regression in terms of potential groups in the responses and time-varying dynamics, offering both methodological innovations and empirical relevance. Further extensions of the model may consider more scalable inference methods for high-frequency data settings, which increases the domain of applications to other financial data or medicine, where the granularity of the data is often daily or even per second.

Chapter 5

Conclusions

This thesis contributes to the development of fully Bayesian methodologies for reduced-rank regression models. The primary focus has been on extending the traditional reduced-rank regression framework to accommodate uncertainty quantification in rank estimation, sparsity, and structural patterns in the responses, including their evolution over time.

In Chapter 2, we introduced Bayesian Rank Estimation and Covariate Selection (BRECS), a novel method that jointly estimates the rank of the coefficient matrix and performs variable selection in a fully Bayesian approach. BRECS addresses existing limitations in the literature, including the need for post-processing steps in rank selection, by incorporating a mixture prior and global-local shrinkage priors for inducing sparsity. We also proposed two uncertainty quantification tools: the Posterior Inclusion Probability (PIP) index and the Relevance Index (RI), which provide summaries for variable selection and their quantification of uncertainty.

Chapter 3 presents the first Bayesian approach to partial reduced-rank regression, where the group structures in the response variables are captured under the proposed Bayesian Partial Reduced-Rank (BPRR) regression model. This approach assumes that an unknown subset of the responses is associated with the covariates through a low-rank matrix of regression coefficients, while the remaining group follows a full-rank relationship. BPRR infers these groupings directly from the data and provides uncertainty quantification about these estimates. Therefore, the proposed method enables more flexible modelling of heterogeneous response structures and yields an interpretable and parsimonious model.

Building on the results obtained by applying BPRR to macroeconomic data of the United States, Chapter 4 integrated the PRR framework into a time-dependent context by introducing the Markov-switching Bayesian Partial Reduced-Rank (MS-PRR) regression model. This time-varying extension captures shifts in the grouping structure of the response variables and the corresponding model parameters across different regimes over time. We also increased the method's flexibility by introducing a nonparametric term in the regression replacing the full-rank linear component of the traditional partial reduced-rank model. By incorporating a hidden Markov chain to detect time-varying groupings and a Gaussian process to model complex relationships in the non-low-rank responses, MS-PRR uncovers changes in the structure of the regression, as demonstrated in both simulation studies and real-world macroeconomic data. This temporal flexibility makes the method especially valuable in applications such as economics and finance, where such changes are common.

The methodologies presented in this thesis form a unified and robust framework for modelling complex relationships while asserting quantification of uncertainty in the framework of reduced-rank regression. They address challenges related to model selection, interpretability, and time-varying dynamics in multivariate regression. Through Bayesian inference, we ensure full posterior uncertainty quantification that facilitates more informed decision-making.

Across the models developed in this thesis, default or weakly informative priors were generally adopted for interpretability and computational tractability, ensuring that they aligned with the underlying model structure and inferential goals. The choice of the Dirichlet-Laplace prior in Chapter 2 over other global-local shrinkage priors is discussed in Appendix A.3.4, where, although some variation in the results is observed, the conclusions remained largely robust. Nonetheless, prior calibration, particularly in sparse settings or complex frameworks as the ones presented in this thesis, remains an important aspect. It is advisable to perform targeted sensitivity analyses for key prior assumptions and to incorporate domain knowledge when tuning the (few) hyperparameters involved. Future work could explore automated approaches to prior calibration to further enhance robustness.

Potential avenues for future research include extending the proposed framework to higher-order structures, such as tensor decompositions encountered in reduced-rank vector autoregression (Luo and Griffin, 2024). On the computational side, developing more scalable inference techniques is crucial. Specifically, we observed that sampling the response binary allocation vector in Chapters 3 and 4 poses a challenge in the computational time of the MCMC algorithm, suggesting the need for more efficient strategies such as the approximate Laplace approximation (Rossell et al., 2021, ALA), or the Point-wise implementation of Adaptive Random Neighbourhood Informed proposal (Liang et al., 2023, PARNI). Despite offering computational benefits, ALA is not a consistent estimator of integrals and its applicability is limited to log-concave settings. The models considered in this thesis may not fully satisfy these conditions, thus a thorough investigation into the theoretical implications of adopting ALA or PARNI in this context, or devising a similar strategy remains a direction of future work. The present work mainly focuses on developing methodologies validated through extensive simulation studies and empirical results, wherein formal theoretical guarantees remain an open area of research. In particular, establishing consistency and other asymptotic properties of the model selection procedures, as well as convergence properties of the proposed MCMC samplers, would deepen their understanding and limitations.

In terms of applications, our methodology can be tailored to a wide range of problems in biostatistics and related fields. Within the framework of generalized linear models, the use of appropriate link functions allows reduced-rank regression to be applied in genomic sequencing (Fitzgerald et al., 2022; Banerjee et al., 2019), an area also studied through factor analysis (Xu et al., 2020). The models can be extended to address matrix imputation problems common in medical research (Zhang and Wang, 2012; Sportisse et al., 2020), as well as in network analysis and imaging (Yuchi et al., 2023). Reduced-rank regression has also been widely used in a nutritional context to derive dietary patterns (Weikert and Schulze, 2016), where understanding the latent response structure is essential. Furthermore, a different direction involves embedding the proposed methods into Vector Error Correction Models (VECMs), where the presence of cointegrating relationships can be addressed within a low-rank setup (Hauzenberger et al., 2025), and integrating regime-switching mechanisms would further enhance its applicability economics.

Bibliography

Alquier, P. (2013). Bayesian methods for low-rank matrix estimation: Short survey and theoretical study. In *Proc. 24th Int. Conf. Algorithmic Learning*, pages 309–23, New York. Springer.

Anderson, R. and Bancroft, T. (1952). *Statistical theory in research.* McGraw-Hill, New York.

Anderson, T. W. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *The Annals of Mathematical Statistics*, 22:327–351.

Babacan, S. D., Luessi, M., Molina, R., and Katsaggelos, A. K. (2011). Low-rank matrix completion by variational sparse Bayesian learning. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2188–2191.

Banerjee, S., Zeng, L., Schunkert, H., and Söding, J. (2019). Bayesian multiple logistic regression for case-control gwas. *PLOS Genetics*, 14(12):1–27.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, 103(4):985–991.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490.

Billio, M., Casarin, R., Iacopini, M., and Kaufmann, S. (2023). Bayesian dynamic tensor regression. *Journal of Business & Economic Statistics*, 41(2):429–439.

Bonnerjee, S., Karmakar, S., and Wu, W. B. (2024). Gaussian approximation for nonstationary time series with optimal rate and explicit construction. *The Annals of Statistics*, 52(5):2293–2317.

Buch, G., Schulz, A., Schmidtmann, I., Strauch, K., and Wild, P. S. (2023). A systematic review and evaluation of statistical methods for group variable selection. *Statistics in Medicine*, 42:331–352.

Bunea, F., She, Y., and Wegkamp, M. H. (2011). Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309.

Bura, E. and Cook, R. (2003). Rank estimation in reduced-rank regression. *Journal of Multivariate Analysis*, 87(1):159–176.

Carriero, A., Clark, T. E., and Marcellino, M. (2019). Large Bayesian vector autoregressions with stochastic volatility and non-conjugate priors. *Journal of Econometrics*, 212(1):137–154.

Chakraborty, A., Bhattacharya, A., and Mallick, B. K. (2019). Bayesian sparse multiple regression for simultaneous rank reduction and variable selection. *Biometrika*, 107:205–221.

Chen, K., Dong, H., and Chan, K.-S. (2013). Reduced rank regression via adaptive nuclear norm penalization. *Biometrika*, 100:901–920.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107(500):1533–1545.

Chicco, D. and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1).

Chua, C. L. and Tsiaplias, S. (2018). A Bayesian approach to modeling time-varying cointegration and cointegrating rank. *Journal of Business & Economic Statistics*, 36(2):267–277.

Cogley, T. and Sargent, T. J. (2005). Drifts and volatilities: Monetary policies and outcomes in the post WWII US. *Review of Economic dynamics*, 8(2):262–302.

Corneli, M., Latouche, P., and Rossi, F. (2018). Multiple change points detection and clustering in dynamic networks. *Statistics and Computing*, 28:989–1007.

Creal, D. D., Gramacy, R. B., and Tsay, R. S. (2014). Market-based credit ratings. *Journal of Business & Economic Statistics*, 32(3):430–444.

Cross, J. L., Hou, C., and Poon, A. (2020). Macroeconomic forecasting with large Bayesian VARs: Global-local priors and the illusion of sparsity. *International Journal of Forecasting*, 36:899–915.

Cubadda, G. and Hecq, A. (2021). Reduced rank regression models in economics and finance. *SSRN Electronic Journal*.

Cunningham, J., Ghahramani, Z., and Rasmussen, C. (2012). Gaussian processes for time-marked time-series data. In *Artificial intelligence and statistics*, pages 255–263. PMLR.

Fitzgerald, T., Jones, A., and Engelhardt, B. (2022). A poisson reduced-rank regression model for association mapping in sequencing data. *BMC Bioinformatics*, 23:529–550.

Foygel, R., Horrell, M., Drton, M., and Lafferty, J. (2012). Nonparametric reduced rank regression. In Bartlett, P., Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer Series in Statistics. Springer New York.

Frühwirth-Schnatter, S., Hosszejni, D., and Lopes, H. F. (2025). Sparse Bayesian factor analysis when the number of factors is unknown. *Bayesian Analysis*, 20(1):213–344.

Geels, V., Pratola, M. T., and Herbei, R. (2023). The taxicab sampler: Mcmc for discrete spaces with application to tree models. *Journal of Statistical Computation and Simulation*, 93(5):753–774.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments. In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting, Dedicated to the memory of Morris H. DeGroot, 1931–1989*. Oxford University Press.

Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *Journal of Econometrics*, 75(1):121–146.

Ghosh, J. and Dunson, D. B. (2009). Default prior distributions and efficient posterior computation in Bayesian factor analysis. *Journal of Computational and Graphical Statistics*, 18(2):306–320.

Goh, G., Dey, D. K., and Chen, K. (2017). Bayesian sparse reduced rank multivariate regression. *Journal of Multivariate Analysis*, 157:14–28.

Gudmundsson, G. (1977). Multivariate analysis of economic variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 26(1):48–59.

Hamilton, J. D. (1990). Analysis of time series subject to changes in regime. *Journal of econometrics*, 45(1-2):39–70.

Hans, C., Dobra, A., and West, M. (2007). Shotgun stochastic search for "large p" regression. *Journal of the American Statistical Association*, 102(478):507–516.

Hansen, P. R. (2002). Generalized reduced rank regression. *Social Science Research Network*.

Hauzenberger, N., Pfarrhofer, M., and Rossini, L. (2025). Sparse time-varying parameter vecms with an application to modeling electricity prices. *International Journal of Forecasting*, 41(1):361–376.

Heidelberger, P. and Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 31(6):1109–1144.

Hilafu, H., Safo, S. E., and Haine, L. (2020). Sparse reduced-rank regression for integrating omics data. *BMC Bioinformatics volume*, 21(283).

Huber, F., Koop, G., and Onorante, L. (2021). Inducing sparsity and shrinkage in time-varying parameter models. *Journal of Business & Economic Statistics*, 39(3):669–683.

Izenman, A. (2008). *Modern multivariate statistical techniques: Regression, classification, and manifold learning*. Springer, New York.

Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264.

Jöreskog, K. G. and Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70(351a):631–639.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Kim, K. and Jung, S. (2024). Integrative sparse reduced-rank regression via orthogonal rotation for analysis of high-dimensional multi-source data. *Statistics and Computing*, 34.

Kleibergen, F. and Paap, R. (2002). Priors, posteriors and Bayes factors for a Bayesian analysis of cointegration. *Journal of Econometrics*, 111(2):223–249.

Li, G., Liu, X., and Chen, K. (2019). Integrative Multi-View Regression: Bridging Group-Sparse and Low-Rank Models. *Biometrics*, 75(2):593–602.

Li, Y., Nan, B., and Zhu, J. (2015). Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure. *Biometrics*, 71:354–363.

Lian, H. and Ma, S. (2013). Reduced-rank regression in sparse multivariate varying-coefficient models with high-dimensional covariates. *arXiv preprint arXiv:1309.6058.*

Liang, X., Livingstone, S., and Griffin, J. (2023). Adaptive mcmc for bayesian variable selection in generalised linear models and survival models. *Entropy*, 25(9).

Lim, Y. and Teh, Y. (2007). Variational Bayesian approach to movie rating prediction. In *Knowledge Discovery and Data Mining.*

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica*, pages 41–67.

Luo, S. and Chen, Z. (2020). Feature selection by canonical correlation search in high-dimensional multiresponse models with complex group structures. *Journal of the American Statistical Association*, 115:1227–1235.

Luo, Y. and Griffin, J. E. (2024). Bayesian inference of vector autoregressions with tensor decompositions. *Journal of Business & Economic Statistics*, pages 1–24.

Mai, T. T. and Alquier, P. (2022). Optimal quasi-Bayesian reduced rank regression with incomplete response. *arXiv preprint arXiv:2206.08619.*

Maruotti, A. and Punzo, A. (2017). Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. *Computational Statistics & Data Analysis*, 113:475–496.

Mukherjee, A. (2013). *Topics on reduced rank methods for multivariate regression.* PhD thesis, The University of Michigan.

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective.* MIT press.

Neal, R. M. (1999). Regression and classification using gaussian process priors. In Bernardo, J., Berger, J., Dawid, A., and Smith, A., editors, *Bayesian Statistics*, volume 6, pages 475–502. Oxford University Press.

Ning, Y.-C. B. and Ning, N. (2024). Spike and slab Bayesian sparse principal component analysis. *Statistics and Computing*, 34(3):118.

Omori, Y., Chib, S., Shephard, N., and Nakajima, J. (2007). Stochastic volatility with leverage: Fast and efficient likelihood inference. *Journal of Econometrics*, 140(2):425–449.

Paci, L. and Finazzi, F. (2018). Dynamic model-based clustering for spatio-temporal data. *Statistics and Computing*, 28:359–374.

Pintado, M., Iacopini, M., Rossini, L., and Shestopaloff, A. (2025). Bayesian partial reduced-rank regression. *Journal of Computational and Graphical Statistics*, pages 1–20.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25:111–163.

Rasmussen, C. and Williams, C. (2005). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press.

Ray, P. and Bhattacharya, A. (2018). Signal adaptive variable selector for the horseshoe prior. *arXiv preprint arXiv:1810.09004*.

Reinsel, G. C. and Velu, R. P. (2006). Partially reduced-rank multivariate regression models. *Statistica Sinica*, 16(3):899–917.

Reinsel, G. C., Velu, R. P., and Chen, K. (2022). *Multivariate Reduced-Rank Regression: Theory, Methods and Applications*, volume 225 of *Lecture Notes in Statistics*. Springer, New York, NY, second edition.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the griddy-Gibbs sampler. *Journal of the American Statistical Association*, 87(419):861–868.

Robert, C. P. and Casella, G. (1999). *Monte Carlo statistical methods*, volume 2. Springer.

Rossell, D., Abril, O., and Bhattacharya, A. (2021). Approximate laplace approximations for scalable model selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):853–879.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society Series B*, 63(2):325–338.

Salakhutdinov, R. and Mnih, A. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 880–887, New York, NY, USA. Association for Computing Machinery.

Salibian-Barrera, M. (2023). Robust nonparametric regression: Review and practical considerations. *Econometrics and Statistics*.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461 – 464.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587 – 2619.

She, Y. and Chen, K. (2017). Robust reduced-rank regression. *Biometrika*, 104(3):633–647.

Šmídl, V. and Quinn, A. (2007). On Bayesian principal component analysis. *Computational Statistics & Data Analysis*, 51(9):4101–4123.

Sobczyk, P., Bogdan, M., and Josse, J. (2017). Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood. *Journal of Computational and Graphical Statistics*, 26(4):826–839.

Sportisse, A., Boyer, C., and Josse, J. (2020). Imputation and low-rank estimation with missing not at random data. *Statistics and Computing*, 30(6):1629–1643.

Strachan, R. W. (2003). Valid Bayesian estimation of the cointegrating error correction model. *Journal of Business & Economic Statistics*, 21(1):185–195.

Tanner, M. A. and and, W. H. W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84(407):710–716.

van Dyk, D. A. and Park, T. (2008). Partially collapsed gibbs samplers. *Journal of the American Statistical Association*, 103(482):790–796.

Velu, R. P. (1991). Reduced rank models with two sets of regressors. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 40(1):159–170.

Viroli, C. (2009). Bayesian inference in non-Gaussian factor analysis. *Statistics and Computing*, 19:451–463.

Weikert, C. and Schulze, M. (2016). Evaluating dietary patterns: the role of reduced rank regression. *Current Opinion in Clinical Nutrition and Metabolic Care*, 19:1.

Wolf, C., Meisenheimer, K., Kleinheinrich, M., Borch, A., Dye, S., Gray, M., Wisotzki, L., Bell, E. F., Rix, H.-W., Cimatti, A., Hasinger, G., and Szokoly, G. (2004). A catalogue of the Chandra Deep Field South with multi-colour classification and photometric redshifts from COMBO-17. *Astronomy & Astrophysics*, 421(3):913–936.

Xu, T., Demmer, R. T., and Li, G. (2020). Zero-inflated poisson factor model with application to microbiome read counts. *Biometrics*, 77(1):91–101.

Yang, D., Goh, G., and Wang, H. (2022). A fully Bayesian approach to sparse reduced-rank multivariate regression. *Statistical Modelling*, 22(3):199–200.

Yuchi, H. S., Mak, S., and Xie, Y. (2023). Bayesian Uncertainty Quantification for Low-Rank Matrix Completion. *Bayesian Analysis*, 18(2):491 – 518.

Zhang, Z. and Wang, L. (2012). A note on the robustness of a full Bayesian method for nonignorable missing data analysis. *Brazilian Journal of Probability and Statistics*, 26(3):244 – 264.

Zhu, H., Khondker, Z., Lu, Z., and Ibrahim, J. G. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*, 109(507):977–990.

Čížek, P. and Sadıkoğlu, S. (2020). Robust nonparametric regression: A review. *WIREs Computational Statistics*, 12(3):e1492.

Appendix A

Additional material for Chapter 2

A.1 Details on posterior full conditionals

This section provides a comprehensive derivation of the posterior full conditional distributions of the parameters in the specified model outlined in Chapter 2. The naïve parametrization, (RRn), and the column-sharing (RRcs) approach differ solely in the update process of the rank-u matrices $\mathbf{A}_u \in \mathbb{R}^{q\times u}$ and $\mathbf{B}_u \in \mathbb{R}^{p\times u}$ when the coefficient matrix has a rank of u, i.e. $\mathbf{C}_u = \mathbf{B}_u\mathbf{A}'_u \in \mathbb{R}^{p\times q}$. The specific details and distinctions between these approaches are elucidated in the respective sections below.

Let $\mathbf{A} = \{\mathbf{A}_s : s = 1, \dots, q\}$ and $\mathbf{B} = \{\mathbf{B}_s : s = 1, \dots, p\}$ be the collections of matrices of each possible rank, $\mathbf{w} = (w_1, \dots, w_q)$, $\boldsymbol{\omega} = (\omega_1, \dots, \omega_q)$, $\boldsymbol{\Psi} = \{\psi_{lh} : l = 1, \dots, p \text{ and } h = 1, \dots, p\}$, $\boldsymbol{\Phi} = \{\phi_{lh} : l = 1, \dots, p \text{ and } h = 1, \dots, p\}$, $\boldsymbol{\tau} = (\tau_1, \dots, \tau_q)$, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)$. The full conditional distributions of the parameters are:

$$\begin{aligned}
p(u \mid \mathbf{A}, \mathbf{B}, \mathbf{w}) &\propto p(u \mid \mathbf{w})\, p(\mathbf{A} \mid u)\, p(\mathbf{B} \mid u), \\
p(\mathbf{w} \mid u) &\propto p(\mathbf{w})\, p(u \mid \mathbf{w}), \\
p(\boldsymbol{\Sigma} \mid \mathbf{Y}, \mathbf{A}, \mathbf{B}, u) &\propto p(\boldsymbol{\Sigma})\, p(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}, u), \\
p(\mathbf{A} \mid \mathbf{Y}, \mathbf{B}, \boldsymbol{\Sigma}, u) &\propto p(\mathbf{A} \mid u)\, p(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}, u), \\
p(\mathbf{B} \mid \mathbf{Y}, \boldsymbol{\Sigma}, \mathbf{A}, \Psi, \boldsymbol{\tau}, \Phi, u) &\propto p(\mathbf{B} \mid \Psi, \boldsymbol{\tau}, \Phi, u)\, p(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}, u), \\
p(\boldsymbol{\tau} \mid \mathbf{B}, \Phi, \boldsymbol{\alpha}, u) &\propto p(\boldsymbol{\tau} \mid \boldsymbol{\alpha})\, p(\mathbf{B} \mid \boldsymbol{\tau}, \Phi, u), \\
p(\Psi \mid \mathbf{B}, \Phi, \boldsymbol{\tau}, u) &\propto p(\Psi)\, p(\mathbf{B} \mid \Psi, \boldsymbol{\tau}, \Phi, u), \\
p(\Phi \mid \mathbf{B}, \boldsymbol{\alpha}, u) &\propto p(\Phi \mid \boldsymbol{\alpha})\, p(\mathbf{B} \mid \Phi, u), \\
p(\boldsymbol{\alpha} \mid \boldsymbol{\tau}, \Phi) &\propto p(\boldsymbol{\alpha})\, p(\boldsymbol{\tau} \mid \boldsymbol{\alpha})\, p(\Phi \mid \boldsymbol{\alpha}).
\end{aligned}$$

The derivation in detail of the full conditional distributions for the Gibbs sample is found hereafter.

A.1.1 Sample u

The posterior probability that $\mathbf{C}$ has been generated from the sth component of the mixture distribution (rank($\mathbf{C}$) $= s$) is

$$\mathbb{P}(u = s \mid C, \mathbf{w}) = \frac{w_s\, p(\mathbf{C}_s)}{\sum_{j=1}^{q} w_j\, p(\mathbf{C}_j)}. \tag{A.1}$$

However, the straightforward implementation of Eq. (A.1) may lead to numerical problems when the denominator is close to zero. Hence it is better to work on the logarithmic scale (Frühwirth-Schnatter, 2006).

Define $L_{\max} = \max_{s\in\{1,\dots,q\}} \log p(\mathbf{C}_s)$, and compute for each $s = 1, \dots, q$:

$$p^*(\mathbf{C}_s) = \exp\{\log p(\mathbf{C}_s) - L_{\max}\}, \tag{A.2}$$

as well as the sum $p^*(\mathbf{C}) = \sum_{j=1}^{q} w_j\, p^*(\mathbf{C}_j)$. Then the probability for each value of the rank is given by

$$\mathbb{P}(u = s \mid C, \mathbf{w}) = \frac{w_s\, p^*(\mathbf{C}_s)}{p^*(\mathbf{C})}. \tag{A.3}$$

To sample from Eq. (A.3), we first compute $p(\mathbf{C}_s)$ needed in Eq. (A.2) as

$$\begin{aligned} p(\mathbf{C}_s) &\propto p(\mathbf{A}_s)\, p(\mathbf{B}_s)\, p(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \mathbf{\Sigma}, u, \mathbf{w}) \\ &\propto \prod_{j=1}^{q} \prod_{h=1}^{s} p(a_{jh}) \cdot \prod_{h=1}^{s} \prod_{l=1}^{p} p(b_{lh}) \cdot p(\mathbf{Y} \mid \mathbf{A}, \mathbf{B}, \mathbf{\Sigma}, u, \mathbf{w}), \end{aligned} \tag{A.4}$$

where a_{jh} is the jhth entry of matrix $\mathbf{A}_s \in \mathbb{R}^{q\times s}$, and b_{lh} corresponds to the lhth element of matrix $\mathbf{B}_s \in \mathbb{R}^{p\times s}$.

We define $\tilde{\mathbf{B}}_s$ as the s-rank matrix whose element in row l and column h is given by $b_{lh}(\psi_{lh}\tau_h^2\phi_{lh}^2)^{-1/2}$, and $\mathbf{\Lambda}_s = \operatorname{diag}(\psi_{11}\tau_1^2\phi_{11}^2, \dots, \psi_{1s}\tau_s^2\phi_{1s}^2, \dots, \psi_{p1}\tau_1^2\phi_{p1}^2, \dots, \psi_{ps}\tau_s^2\phi_{ps}^2)$. Then Eq. (A.4) transforms into

$$\begin{aligned} p(\mathbf{C}_s) \propto (2\pi)^{-s(p+q)/2} |\mathbf{\Lambda}_s|^{-1/2} \exp\Big\{-\frac{1}{2}\operatorname{tr}[\mathbf{A}_s'\mathbf{A}_s + \tilde{\mathbf{B}_s}'\tilde{\mathbf{B}_s}]\Big\} \\ \times \exp\Big\{-\frac{1}{2}\operatorname{tr}[\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\mathbf{B}_s\mathbf{A}_s')'(\mathbf{Y} - \mathbf{X}\mathbf{B}_s\mathbf{A}_s')]\Big\}. \end{aligned}$$

A.1.2 Sample w and $\mathbf{\Sigma}$

It is straightforward to derive the posterior distribution of $\mathbf{w}$ and $\mathbf{\Sigma}$ due to the conjugacy of the priors. The weights of the mixture distribution of $\mathbf{C}$ are updated from $\prod_{s=1}^{q} w_s^{\omega_s + \mathbb{I}(u=s)-1}$, a Dirichlet distribution with parameter $\boldsymbol{\omega}^* = (\omega_1 + \mathbb{I}(u = 1), \dots, \omega_q + \mathbb{I}(u = q))$.

The error covariance matrix is updated by sampling from $\mathcal{IW}(\nu^*, \mathbf{\Upsilon}^*)$, where $\nu^* = \nu + n$ and $\mathbf{\Upsilon}^* = \mathbf{\Upsilon} + (\mathbf{Y} - \mathbf{XC})'(\mathbf{Y} - \mathbf{XC})$ are the degrees of freedom and the scale matrix, respectively.

A.1.3 Sample $\mathbf{A}_u$

Naïve parametrization

To obtain the full conditional distribution of $\mathbf{A}_u$, we use the vectorization of the multivariate linear regression model $\mathbf{Y} = \mathbf{X}\mathbf{B}_u\mathbf{A}_u' + \mathbf{E}$, denoting by $\text{vec}(\mathbf{M})$ the vectorization of a matrix $\mathbf{M}$:

$$\text{vec}(\mathbf{Y}') = \text{vec}(\mathbf{A}_u\mathbf{B}_u'\mathbf{X}') + \text{vec}(\mathbf{E}'). \tag{A.5}$$

For three matrices $\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3$ with appropriate dimensions, the following identity holds:

$$\text{vec}(\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3) = (\mathbf{M}_3' \otimes \mathbf{M}_1)\,\text{vec}(\mathbf{M}_2), \tag{A.6}$$

where $\otimes$ is the Kronecker product. Setting $\mathbf{M}_1 = \mathbf{I}_q$, $\mathbf{M}_2 = \mathbf{A}_u$, and $\mathbf{M}_3 = \mathbf{B}_u'\mathbf{X}'$, from Eq. (A.5) follows

$$\text{vec}(\mathbf{Y}') = (\mathbf{X}\mathbf{B}_u \otimes \mathbf{I}_q)\,\text{vec}(\mathbf{A}_u) + \text{vec}(\mathbf{E}'). \tag{A.7}$$

Let $\mathbf{y} = \text{vec}(\mathbf{Y}') \in \mathbb{R}^{nq\times 1}$, $\tilde{\mathbf{a}}_u = \text{vec}(\mathbf{A}_u) \in \mathbb{R}^{qu\times 1}$, and $\boldsymbol{e} = \text{vec}(\mathbf{E}') \in \mathbb{R}^{nq\times 1}$, where $\boldsymbol{e} \sim N_{nq}(\mathbf{0}, \tilde{\boldsymbol{\Sigma}})$ with $\tilde{\boldsymbol{\Sigma}} = \text{diag}(\boldsymbol{\Sigma}, \ldots, \boldsymbol{\Sigma})$. Eq. (A.7) is equivalent to

$$\mathbf{y} = (\mathbf{X}\mathbf{B}_u \otimes \mathbf{I}_q)\,\tilde{\mathbf{a}}_u + \boldsymbol{e}. \tag{A.8}$$

Since $\text{vec}(\mathbf{Y}) \sim \mathcal{N}_{nq}(\text{vec}(\mathbf{X}\mathbf{B}_u\mathbf{A}_u'), \boldsymbol{\Sigma} \otimes \mathbf{I}_n)$, then $\mathbf{y} = \text{vec}(\mathbf{Y}')$ follows a multivariate normal distribution with mean $\text{vec}(\mathbf{A}_u\mathbf{B}_u'\mathbf{X}') = (\mathbf{X}\mathbf{B}_u \otimes \mathbf{I}_q)\,\tilde{\mathbf{a}}_u$ and covariance matrix $\mathbf{I}_n \otimes \boldsymbol{\Sigma} = \tilde{\boldsymbol{\Sigma}}$. We compute the full conditional of $\mathbf{A}_u$ in terms of $\tilde{\mathbf{a}}_u$. If $\tilde{\mathbf{y}} = \tilde{\boldsymbol{\Sigma}}^{-1/2}\mathbf{y}$, $\tilde{\mathbf{X}} = \tilde{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X}\mathbf{B}_u \otimes \mathbf{I}_q)$, and $\boldsymbol{\Omega}_{\mathbf{A}_u} = (\mathbf{I}_{qu} + \tilde{\mathbf{X}}'\tilde{\mathbf{X}})$, subsequently the vector $\tilde{\mathbf{a}}_u$ is sampled from $\mathcal{N}_{qu}(\boldsymbol{\mu}^*_{\mathbf{A}_u}, \boldsymbol{\Sigma}^*_{\mathbf{A}_u})$, where $\boldsymbol{\mu}^*_{\mathbf{A}_u} = \boldsymbol{\Omega}^{-1}_{\mathbf{A}_u}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$ and $\boldsymbol{\Sigma}^*_{\mathbf{A}_u} = \boldsymbol{\Omega}^{-1}_{\mathbf{A}_u}$. Samples are made resorting to the algorithm developed in Rue (2001).

Column-sharing parametrization

The rank-u matrix $\mathbf{A}_u$ is defined as having exactly u columns, i.e. $\mathbf{A}_u = (\mathbf{a}_1, \ldots, \mathbf{a}_u)$, where $\mathbf{a}_h = (a_{1h}, \ldots, a_{qh})$ for each $h = 1, \ldots, u$. To obtain the full conditional distribution of $\mathbf{A}_u$, we sample each column independently from its respective posterior distribution. Therefore, it is necessary to express the likelihood function in terms of the columns of $\mathbf{A}_u$.

Define $\mathbf{y}$, $\tilde{\mathbf{y}}$ and $\tilde{\boldsymbol{\Sigma}}$ as in the previous section, and let $\mathbf{M}_{\mathbf{B}_u}$ denote the resulting matrix from $X\mathbf{B}_u \otimes \mathbf{I}_q$. We allow the superscript (i, q) to denote the iqth columns of a matrix $\mathbf{M} = (m_1, \ldots, m_k)$ with $k \geq q$ columns. For example, if $\mathbf{M}$ comprises 12 columns, then $\mathbf{M}^{(2,3)} = (m_4, m_5, m_6)$, while $\mathbf{M}^{(4,3)} = (m_{10}, m_{11}, m_{12})$. Therefore, the model in Eq. (A.8) expressed in relation to the columns of $\mathbf{A}_u$ is

$$\mathbf{y} = \sum_{i=1}^{u} \mathbf{M}^{(i,q)}_{\mathbf{B}_u}\mathbf{a}_i + \boldsymbol{e}$$

and the distribution of $\mathbf{y}$ is $\mathcal{N}_{nq}(\mathbf{M}_u, \tilde{\boldsymbol{\Sigma}})$, where $\mathbf{M}_u = \sum_{i=1}^{u} \mathbf{M}^{(i,q)}_{\mathbf{B}_u}\mathbf{a}_i$.

Let $\tilde{\mathbf{X}}_i = \tilde{\boldsymbol{\Sigma}}^{-1/2}\mathbf{M}^{(i,q)}_{\mathbf{B}_u}$, $\boldsymbol{\Omega}_{\mathbf{a}_h} = \mathbf{I}_q + \tilde{\mathbf{X}}_h'\tilde{\mathbf{X}}_h$ and $\Delta_{\mathbf{a}_h} = \left(\tilde{\mathbf{y}}' - \sum_{i\neq h}^{u} \mathbf{a}_i'\tilde{\mathbf{X}}_i'\right)\tilde{\mathbf{X}}_h$. Finally, each column $\mathbf{a}_h$ is drawn from a multivariate normal distribution with mean $\boldsymbol{\mu}^*_{\mathbf{a}_h} = \boldsymbol{\Omega}^{-1}_{\mathbf{a}_h}\Delta'_{\mathbf{a}_h}$, and covariance matrix $\boldsymbol{\Sigma}^*_{\mathbf{a}_h} = \boldsymbol{\Omega}^{-1}_{\mathbf{a}_h}$.

A.1.4 Sample $\mathbf{B}_u$

Naïve parametrization

To obtain the full conditional distribution of $\mathbf{B}_u$, we employ the vectorization of the multivariate linear regression model $\mathbf{Y} = \mathbf{X}\mathbf{B}_u\mathbf{A}'_u + \mathbf{E}$, and the identity in Eq. (A.6) with $\mathbf{M}_1 = \mathbf{A}_u$, $\mathbf{M}_2 = \mathbf{B}'_u$, and $\mathbf{M}_3 = \mathbf{X}'$. Let $\tilde{\mathbf{b}}_u = \text{vec}(\mathbf{B}'_u) \in \mathbb{R}^{pu\times 1}$, thus the model is

$$\mathbf{y} = (\mathbf{X} \otimes \mathbf{A}_u)\,\tilde{\mathbf{b}}_u + \boldsymbol{e}.$$

Note that $\tilde{\mathbf{b}}_u \sim \mathcal{N}_{pu}(\mathbf{0}, \boldsymbol{\Lambda}_u)$, where $\boldsymbol{\Lambda}_u$ is defined as in Eq. (A.4). Through similar reasoning as in the previous step, we obtain that $\tilde{\mathbf{b}}_u$ is multivariate normally distributed with mean $\boldsymbol{\mu}^*_{\mathbf{B}_u} = \boldsymbol{\Omega}^{-1}_{\mathbf{B}_u}\tilde{\mathbf{X}}'\tilde{\mathbf{y}}$ and covariance matrix $\boldsymbol{\Sigma}^*_{\mathbf{B}_u} = \boldsymbol{\Omega}^{-1}_{\mathbf{B}_u}$, with $\boldsymbol{\Omega}_{\mathbf{B}_u} = (\boldsymbol{\Lambda}_u^{-1} + \tilde{\mathbf{X}}'\tilde{\mathbf{X}})$, $\tilde{\mathbf{X}} = \tilde{\boldsymbol{\Sigma}}^{-1/2}(\mathbf{X} \otimes \mathbf{A}_u)$, and $\tilde{\mathbf{y}}$ is defined as in the previous step. The algorithm from Bhattacharya et al. (2016) is adapted to obtain the desired draws.

Column-sharing parametrization

For every $h = 1.\dots,u$, the prior distribution of column $\mathbf{b}_h$ of matrix $\mathbf{B}_u = (\mathbf{b}_1,\dots,\mathbf{b}_u)$ is a multivariate normal with mean the zero vector and the covariance matrix denoted by $\boldsymbol{\Lambda}_{\mathbf{b}_h} = \text{diag}(\psi_{1h}\tau^2_{1h}\phi^2_{1h},\dots,\psi_{ph}\tau^2_{ph}\phi^2_{ph})$. We follow the same reasoning as in the update of $\mathbf{A}_u$ in RRcs to derive the posterior distribution of each column $\mathbf{b}_h$. Consequently, the likelihood function requires to be formulated in relation to the columns of $\mathbf{B}_u$.

Let $\mathbf{y}^* = \text{vec}(\mathbf{Y})$, $\boldsymbol{e}^* = \text{vec}(\mathbf{E})$, $\mathbf{M}_{\mathbf{A}_u} = \mathbf{A}_u \otimes \mathbf{X}$, and $\tilde{\boldsymbol{\Sigma}}^* = \boldsymbol{\Sigma} \otimes \mathbf{I}_n$. Notice that $\mathbf{y}^*$ follows the multivariate normal distribution $\mathcal{N}_{nq}(\mathbf{M}_{\mathbf{A}_u}\,\text{vec}(\mathbf{B}_u),\,\tilde{\boldsymbol{\Sigma}}^*)$. Using the previous superscript notation, we express $\mathbf{y}^*$ in terms of the columns of $\mathbf{B}_u$ as:

$$\mathbf{y}^* = \sum_{i=1}^{u} \mathbf{M}^{(i,p)}_{\mathbf{A}u}\mathbf{b}_i + \boldsymbol{e}^*.$$

Lastly, by denoting $\tilde{\mathbf{y}} = \tilde{\boldsymbol{\Sigma}^*}^{-1/2}\mathbf{y}^*$, $\tilde{\mathbf{X}}^*_i = \tilde{\boldsymbol{\Sigma}^*}^{-1/2}\mathbf{M}^{(i,p)}_{\mathbf{A}_u}$, $\boldsymbol{\Omega}_{\mathbf{b}_h} = \boldsymbol{\Lambda}^{-1}_{\mathbf{b}_h} + \tilde{\mathbf{X}}^{*\prime}_h\tilde{\mathbf{X}}^*_h$ and $\Delta_{\mathbf{b}_h} = \left(\tilde{\mathbf{y}^*}' - \sum_{i\neq h}^{u} \mathbf{b}'_i\tilde{\mathbf{X}}^{*\prime}_i\right)\tilde{\mathbf{X}}^*_h$, we sample $\mathbf{b}_h$ from $\mathcal{N}_p(\boldsymbol{\mu}^*_{\mathbf{b}_h},\,\boldsymbol{\Sigma}^*_{\mathbf{b}_h})$, where $\boldsymbol{\mu}^*_{\mathbf{b}_h} = \boldsymbol{\Omega}^{-1}_{\mathbf{b}_h}\Delta'_{\mathbf{b}_h}$ and $\boldsymbol{\Sigma}^*_{\mathbf{b}_h} = \boldsymbol{\Omega}^{-1}_{\mathbf{b}_h}$.

A.1.5 Update the hyperparameters $\boldsymbol{\tau}$, Ψ, Φ and $\boldsymbol{\alpha}$

The posterior distributions of the global shrinkage parameters are

$$\begin{aligned}\tau_h \mid b_{lh}, \phi_{lh}, \alpha_h &\sim \text{GiG}\left(p(\alpha_h - 1),\, 1,\, 2\sum_{l=1}^{p}\frac{|b_{lh}|}{\phi_{lh}}\right),\\ \psi_{lh} \mid b_{lh}, \phi_{lh}, \tau_h &\sim \text{iG}\left(\frac{\phi_{lh}\tau_h}{|b_{lh}|},\, 1\right)\end{aligned}$$

for $l = 1,\dots,p$ and $h = 1,\dots,q$, where GiG$(\cdot)$ denotes the generalised inverse Gaussian distribution, and iG$(\cdot)$ the inverse Gaussian distribution. The update of ϕ_{lh} is carried out by first obtaining samples for $T_{l1}, \ldots, T_{lq}$ from

$$T_{lh} \mid b_{lh}, \alpha_h \sim \text{GiG}\left((\alpha_h - 1),\, 1,\, 2\,|b_{lh}|\right)$$

afterwards set $\phi_{lh} = \frac{T_{lh}}{\sum_{i=1}^{p} T_{ih}}$. Finally, the log of the full conditional distribution for α_h is given by

$$\log p(\alpha_h \mid -) = -p \log \Gamma(\alpha_h) + \alpha_h \left(p \log \tau_h - p \log 2 + \sum_{l=1}^{p} \phi_{lh} \right) + c,$$

where c is a constant independent of α_h.

A.2 PIP uncertainty index sensitivity analysis

This section examines four different definitions of the PIP uncertainty index, including the one presented in Chapter 2 and three alternative formulations.

The sparse posterior estimate of the coefficient matrix allows exact zero elements. For the jkth coefficient ($j = 1, \ldots, p,\ \ k = 1, \ldots, q$), we obtain a sparse draw at each iteration of the MCMC, and subsequently set the posterior estimate of the jkth entry to 0 if the proportion of nonzero draws is less than 0.5 ($\text{PIP}_{jk} \leq 0.5$), or to the posterior mean of these samples otherwise. For ease of notation, we omit the jk subscript in the remaining lines of this section, allowing PIP to denote any PIP_{jk}. Values close to 0 (or 1) indicate low uncertainty about the exclusion (or inclusion) of the coefficient, while a probability around 0.5 signals insufficient evidence for either conclusion. To provide a straightforward and easily interpreted manner to quantify uncertainty about this decision, we defined the PIP uncertainty index as:

$$\zeta_{(1)} = 1 - 2\big|\text{PIP} - 0.5\big|. \tag{A.9}$$

This choice of ζ is based primarily on two features. First, the image of ζ is the interval $[0, 1]$, where 0 represents low uncertainty and 1 means high uncertainty. Second, the PIP values of 0 and 1 should be mapped to 0 (lowest uncertainty), as opposed to a PIP of 0.5 corresponding to a PIP uncertainty index of 1 (highest uncertainty). Evidently, our restrictions on the definition of ζ permits alternative representations. Therefore, we provide the additional formulations of ζ:

$$\zeta_{(2)} = 1 - 4(\text{PIP} - 0.5)^2,$$
$$\zeta_{(3)} = \frac{\log\left(\big|\text{PIP} - 0.5\big| + \epsilon\right) + \log(0.5)}{\min(\log\left(\big|\text{PIP} - 0.5\big| + \epsilon\right) + \log(0.5))},$$
$$\zeta_{(4)} = \frac{\log\left(\log\left(\big|\text{PIP} - 0.5\big| + \epsilon\right) + \epsilon\right) + \log\left(\log\left(0.5 + \epsilon\right)\right)}{\max(\log\left(\log\left(\big|\text{PIP} - 0.5\big| + \epsilon\right) + \epsilon\right) + \log\left(\log\left(0.5 + \epsilon\right)\right))},$$

where ϵ is an arbitrarily small positive number for numerical stability. The four specifications of ζ are plotted in Figure A.1 and in Figure A.2.

The first definition we presented in Eq. (A.9), a piecewise linear function, poses an advantage over the remaining three through the linearity property. The set of PIP values that are evaluated by $\zeta_{(1)}$ to a low uncertainty ζ ($\zeta \leq 1/3$) is the same size as those PIP values that result in medium ($1/3 < \zeta \leq 2/3$), and high ($\zeta > 2/3$) uncertainty. The remaining three specifications of

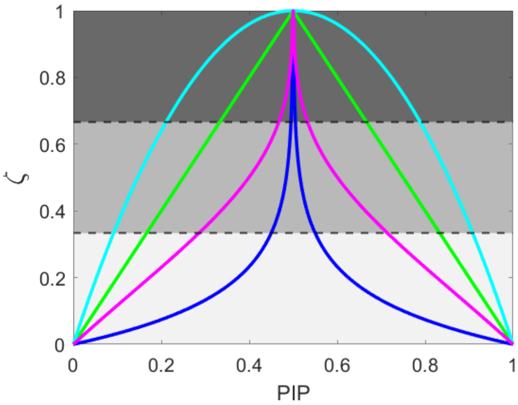


Figure A.1: Plot of the four definitions of PIP uncertainty index, ζ, as a function of the PIP: $\zeta_{(1)}$ (green), $\zeta_{(2)}$ (cyan), $\zeta_{(3)}$ (blue), $\zeta_{(4)}$ (magenta). The shaded regions in a grey-colour scale indicate low ($\zeta \leq 1/3$), medium ($1/3 < \zeta \leq 2/3$), or high ($\zeta > 2/3$) uncertainty.

ζ are nonlinear functions, and therefore the portions in the domain corresponding to each level of uncertainty are of different lengths. Consequently, more than half of the points are allocated to a high uncertainty region in $\zeta_{(2)}$, whereas 24% of PIP values are assigned medium uncertainty, and 18.4% evaluate to a low level. The functions $\zeta_{(3)}$ and $\zeta_{(4)}$ exhibit the opposite behaviour, since the set of PIP values corresponding to low uncertainty is bigger than the set of points mapped to a medium range, which in turn comprises more points than the PIP interval of high uncertainty (see Figure A.2). Hence, the choice of $\zeta_{(1)}$ as the PIP uncertainty index is not biased towards greater portions of the domain corresponding to a particular level of uncertainty, and the relationship between PIP and ζ becomes transparent. Where the particular needs of the researcher demand for a nonlinear behaviour of ζ that could result from an optimistic or pessimistic approach to the quantification of uncertainty, the three aforementioned specifications could be considered as suitable options.

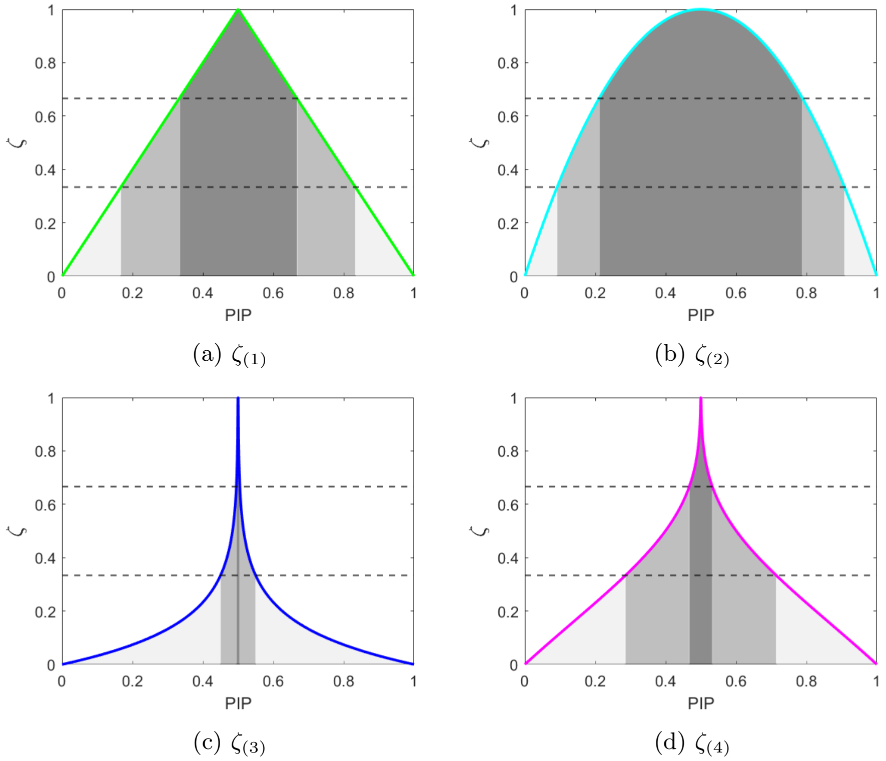
A.3 Simulation experiments

A.3.1 Additional simulation results

In Section 2.4, we conducted a comparative analysis of our methodology under the RRcs parametrization against other state-of-the-art methodologies. Here, we present the findings pertaining to the RRn parametrization across various simulation scenarios (see Table A.1). As elucidated in Section 2.4, RRcs yields superior estimations of the rank and of the coefficient matrix compared to RRn, as evidenced by the Mean Squared Error (MSE) metric. In instances of smaller dimensions, RRn tends to underestimate the rank, whereas RRcs exhibits improved estimations. With increasing dimensions of p and q, both parametrizations yield more conservative rank estimates in sparse scenarios. However, RRcs achieves overall significantly lower errors while promoting a more parsimonious model.

A.3.2 CODA analysis

We perform a CODA analysis for the posterior distribution of the rank (Table A.2), and the estimates of the matrix $\mathbf{C}$ (Table A.3). In both studies, we consider an MCMC chain of $50,000$



(a) $\zeta_{(1)}$ (b) $\zeta_{(2)}$

(c) $\zeta_{(3)}$ (d) $\zeta_{(4)}$

Figure A.2: Function ζ in each of the cases considered: the piece-wise linear function $\zeta_{(1)}$ (panel a), and the nonlinear definitions $\zeta_{(2)}$, $\zeta_{(3)}$ and $\zeta_{(4)}$ (panels b, c and d, respectively). The width of the shaded areas is in accordance with the portion in the domain that evaluates to low (light grey), medium (medium grey) or high (dark grey) uncertainty. The values of ζ that delimit these regions are plotted in horizontal dashed lines.

iterations to perform the Geweke convergence diagnostic test and the Heidelberger and Welch's convergence diagnostic test.

Table A.2 shows that the p-values for the Geweke convergence test and Heidelberg's stationarity test applied to the rank estimation analysis are greater than 0.05, suggesting adequate convergence of the estimated rank chain in all the scenarios considered. These findings hold across all scenarios and thinning levels. By inspecting the ratio between the half-width of the 95% confidence interval for the mean and the mean (last column), a value greater than 0.1 suggests the need for a longer chain. This happens only for the scenario in the third row of Table A.2, which however does pass all the other CODA tests.

Regarding the convergence diagnostics for the coefficient matrix estimates, given the high number of entries, pq, instead of reporting all the individual results, we show some summary measures as follows. Specifically, for each of the above-mentioned tests, we compute in Table A.3 the proportion of the elements of $\mathbf{C}$ passing the specified test. Therefore, the closest the values in each column are to 1, the larger the share of coefficients that pass the corresponding convergence test. The results in Table A.3 report a good performance according to Heidelberg's stationarity test and the half-width test, as shown by the high percentages (above 70%). Moreover, Geweke's diagnostic suggests an average good convergence for the entries of $\mathbf{C}$ in the first and third scenarios but suggests increasing the length of the MCMC.

			Sparse DGP				Non-sparse DGP			
			$\mathbf{\Sigma}_{ind}$		$\mathbf{\Sigma}_{corr}$		$\mathbf{\Sigma}_{ind}$		$\mathbf{\Sigma}_{corr}$	
(q,p)	**X**	**Measure**	**RRcs**	**RRn**	**RRcs**	**RRn**	**RRcs**	**RRn**	**RRcs**	**RRn**
(5,10)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.280	1.020	1.320	1.020	1.280	1.040	1.380
		MSE	0.453	0.297	0.527	3.841	1.163	0.878	1.080	0.776
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.120	1.000	1.140	1.000	1.340	1.020	1.420
		MSE	0.490	0.456	0.484	0.375	1.284	1.006	1.328	0.939
(5,15)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.040	1.000	1.360	1.020	1.420	1.060	1.540
		MSE	0.261	0.241	0.288	2.592	1.204	0.763	1.052	0.634
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.240	1.000	1.280	1.040	1.280	1.020	1.620
		MSE	0.346	0.262	0.367	0.270	1.051	0.905	1.376	0.878
(5,20)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.100	1.000	1.280	1.020	1.460	1.000	1.640
		MSE	0.182	0.168	0.237	0.192	1.057	0.760	1.272	0.736
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.200	1.000	1.080	1.000	1.180	1.020	1.420
		MSE	0.235	0.188	0.297	0.256	1.163	0.983	1.383	0.985
(5,50)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.120	1.000	1.140	1.000	1.220	1.000	1.560
		MSE	0.218	0.187	0.217	0.200	1.112	0.787	1.203	0.666
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.060	1.000	1.060	1.000	1.300	1.000	1.300
		MSE	0.231	0.197	0.252	0.243	1.353	0.998	1.529	1.097
(10,10)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.080	1.000	1.220	1.000	1.260	1.000	1.380
		MSE	0.579	0.523	0.441	0.374	1.243	1.017	1.046	0.785
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.120	1.000	1.140	1.000	1.280	1.000	1.420
		MSE	0.544	0.471	0.588	0.493	1.252	0.926	1.308	0.955
(10,15)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.140	1.000	1.160	1.000	1.380	1.000	1.520
		MSE	0.311	0.269	0.371	0.318	1.259	0.953	1.151	0.835
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.140	1.000	1.100	1.000	1.160	1.000	1.480
		MSE	0.436	0.372	0.349	0.324	1.272	1.158	1.583	1.082
(10,20)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.040	1.000	1.040	1.000	1.380	1.000	1.940
		MSE	0.213	0.203	0.234	0.227	1.210	0.888	1.383	0.747
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.060	1.000	1.060	1.000	1.400	1.000	1.600
		MSE	0.303	0.289	0.309	0.279	1.480	1.071	1.395	0.827
(10,50)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.040	1.000	1.040	1.000	1.660	1.000	1.760
		MSE	0.217	0.208	0.217	0.207	1.419	0.815	1.730	0.909
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.040	1.000	1.040	1.000	1.460	1.000	1.520
		MSE	0.301	0.286	0.256	0.245	1.637	1.143	1.701	0.978
(10,75)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.000	1.000	1.120	1.000	1.700	1.000	1.600
		MSE	0.261	0.260	0.248	0.230	1.597	0.894	1.471	1.002
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.040	1.000	1.100	1.000	1.520	1.000	1.660
		MSE	0.275	0.266	0.298	0.280	2.090	1.320	2.065	1.210
(10,100)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.040	1.000	1.080	1.000	1.420	1.000	1.420
		MSE	0.261	0.253	0.248	0.229	1.949	1.452	1.997	1.473
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.060	1.000	1.060	1.000	1.380	1.000	1.600
		MSE	0.267	0.257	0.327	0.311	2.088	1.597	2.223	1.382
(10,500)	$\mathbf{X}_{ind}$	$\hat{r}$	1.000	1.480	1.000	1.000	1.000	3.200	1.000	3.220
		MSE	0.239	0.190	0.250	0.280	4.379	11.559	4.546	8.214
	$\mathbf{X}_{corr}$	$\hat{r}$	1.000	1.400	1.000	1.140	1.000	2.540	1.000	2.880
		MSE	0.247	0.237	0.274	0.259	4.335	17.480	4.037	22.265

Table A.1: Comparison of the estimated rank ($\hat{r}$) and mean squared error (MSE) obtained by RRcs against RRn for different values of (q, p). In all settings, $n = 100$ and $r_0 = 3$. When $p \in \{10, 15, 20\}$, $p^* = 5$, and in the case where $p \in \{50, 75, 100, 500\}$, $p^* = 0.2p$. We present the average estimates over 50 repetitions for independent errors ($\mathbf{\Sigma}_{ind}$), correlated errors ($\mathbf{\Sigma}_{corr}$), independent regressors ($\mathbf{X}_{ind}$), and correlated regressors ($\mathbf{X}_{corr}$).

Simulation	MCMC iterations	Thinning	Geweke p-value	Heidelberger	
				Stationarity test p-value	Ratio
$(q, p) = (10, 20)$, $r_0 = 3$,$n = 100$, sparse	50,000	1	0.946	0.287	0.048
		10	0.976	0.317	0.054
$(q, p) = (10, 20)$, $r_0 = 3$, $n = 100$, non-sparse	50,000	1	0.132	0.105	0.086
		10	0.092	0.171	0.093
$(q, p) = (10, 15)$, $r_0 = 5$, $n = 100$, non-sparse	50,000	1	0.305	0.282	0.116
		10	0.271	0.252	0.126
$(q, p) = (10, 30)$, $r_0 = 5$, $n = 100$, non-sparse	50,000	1	0.185	0.942	0.077
		10	0.217	0.968	0.080

Table A.2: CODA analysis of u. We studied the various scenarios under the first column. For thinning, we consider the values 1 (no thinning) and 10. Geweke's diagnostic tests the assumption that the first 10% and the last 50% parts of the Markov chain are independent, reporting the results in the fourth column. We consider Heidelberger's stationarity test (5th column) and the half-width test. The latter indicates whether the length of the sample is long enough to estimate the mean with sufficient accuracy (< 0.1).

Simulation	MCMC iterations	Geweke Prop	Heidelberger	
			Stationarity test Prop	Half-width test Prop
$(q, p) = (10, 20)$, $r_0 = 3$, $n = 100$, sparse	50,000	0.980	1.000	0.570
$(q, p) = (10, 20)$, $r_0 = 3$, $n = 100$, non-sparse	50,000	0.315	0.915	0.874
$(q, p) = (10, 15)$, $r_0 = 5$, $n = 100$, non-sparse	50,000	0.667	0.767	0.713
$(q, p) = (10, 30)$, $r_0 = 5$, $n = 100$, non-sparse	50,000	0.280	0.990	0.781

Table A.3: CODA analysis of $\hat{\mathbf{C}}$. We studied the settings presented in the first column, and report the proportion of times that the Markov chain passed Geweke's test (p-value ≥ 0.05) in the third column. The proportion of samples passing Heidelberg's stationarity test and half-width test are stated in the last two columns.

A.3.3 Bura and Cook test

In this subsection, we have implemented the test by Bura and Cook (2003). However, since the method elaborates on the OLS estimate of a multiple linear regression model, it could not be run for the cases $p \geq n$. The test results are reported in Table A.4.

				sparse		non-sparse	
(q,p)	r_0	**X**	**Measure**	$\boldsymbol{\Sigma}_{ind}$	$\boldsymbol{\Sigma}_{corr}$	$\boldsymbol{\Sigma}_{ind}$	$\boldsymbol{\Sigma}_{corr}$
(5,15)	3	$\mathbf{X}_{ind}$	$\hat{r}$	2.950	2.850	3.300	3.000
			$r_{\%}$	0.550	0.750	0.700	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	3.100	3.250	3.250	3.150
			$r_{\%}$	0.600	0.700	0.750	0.850
	5	$\mathbf{X}_{ind}$	$\hat{r}$	4.050	4.050	4.650	4.700
			$r_{\%}$	0.200	0.150	0.650	0.700
		$\mathbf{X}_{corr}$	$\hat{r}$	4.000	4.000	4.800	4.800
			$r_{\%}$	0.200	0.200	0.800	0.800

Table A.4: Comparison of the estimated rank ($\hat{r}$), share of correctly estimated rank ($r_{\%}$), obtained by the test of Bura and Cook (2003) for $(q, p) = (5, 15)$, true rank r_0 and $n = 50$. The sparse DGP considers $p^* = 5$, while the non-sparse DGP has $p^* = p$. We present the average estimates over 20 repetitions for independent errors ($\boldsymbol{\Sigma}_{ind}$), correlated errors ($\boldsymbol{\Sigma}_{corr}$), independent regressors ($\mathbf{X}_{ind}$), and correlated regressors ($\mathbf{X}_{corr}$).

A.3.4 Horseshoe prior

The results presented in Tables 2.2 and 2.3 of Section 2.4.2 for our proposed approach under the column-sharing parameterization use the Dirichlet-Laplace (DL) prior on the columns of matrix $\mathbf{B}$. In this section, we also present the results using the horseshoe (HS) prior in Tables A.5 and A.6.

The MSE shows a good performance in estimating the coefficient matrix $\mathbf{C}$, compared to the other methods; however, the rank is typically overestimated across all the settings. This behaviour signals a potential over-shrinkage of the columns of $\mathbf{B}$, which motivates the need for a larger number of components (i.e., rank) to estimate $\mathbf{C} = \mathbf{B}\mathbf{A}'$. In the non-sparse DGP setting with $(q, p) = (10, 50)$ the performance of the DL and the HS priors is similar, whereas in the other cases, the accuracy of our method using the DL prior is improved.

				$\mathbf{\Sigma}_{ind}$				$\mathbf{\Sigma}_{corr}$			
(q,p)	r_0	**X**	**Measure**	**ANN**	**RRRR**	**RRcs-DL**	**RRcs-HS**	**ANN**	**RRRR**	**RRcs-DL**	**RRcs-HS**
(5,15)	3	$\mathbf{X}_{ind}$	$\hat{r}$	2.320	2.420	1.860	2.660	2.420	2.460	1.720	2.680
			$r_{\%}$	0.400	0.480	0.160	0.000	0.480	0.500	0.180	0.000
			MSE($\mathbf{C}$)	0.030	0.030	0.161	0.153	0.027	0.028	0.189	0.180
			$\varrho(\mathbf{C})$	0.745	0.709	0.984	0.985	0.674	0.670	0.988	0.989
		$\mathbf{X}_{corr}$	$\hat{r}$	2.280	2.240	1.820	2.700	2.020	1.940	1.580	2.740
			$r_{\%}$	0.340	0.300	0.180	0.000	0.180	0.140	0.200	0.000
			MSE($\mathbf{C}$)	0.043	0.047	0.251	0.192	0.046	0.051	0.173	0.139
			$\varrho(\mathbf{C})$	0.779	0.804	0.995	0.994	0.898	0.917	0.986	0.991
	5	$\mathbf{X}_{ind}$	$\hat{r}$	3.160	3.280	2.380	2.700	3.240	3.240	2.100	2.540
			$r_{\%}$	0.000	0.020	0.240	0.400	0.000	0.000	0.120	0.360
			MSE($\mathbf{C}$)	0.036	0.035	0.376	0.342	0.031	0.034	0.394	0.388
			$\varrho(\mathbf{C})$	1.000	0.984	0.380	0.400	1.000	1.000	0.419	0.539
		$\mathbf{X}_{corr}$	$\hat{r}$	2.700	2.940	2.400	2.720	2.900	2.920	2.120	2.660
			$r_{\%}$	0.000	0.000	0.300	0.400	0.000	0.020	0.180	0.400
			MSE($\mathbf{C}$)	0.072	0.063	0.470	0.400	0.063	0.065	0.463	0.434
			$\varrho(\mathbf{C})$	1.000	1.000	0.440	0.581	1.000	0.987	0.442	0.471
(5,50)	3	$\mathbf{X}_{ind}$	$\hat{r}$	0.160	5.000	1.540	2.600	0.240	5.000	1.260	2.440
			$r_{\%}$	0.000	0.000	0.100	0.000	0.020	0.000	0.060	0.040
			MSE($\mathbf{C}$)	1.010	144.533	0.148	0.107	0.667	17.328	0.183	0.155
			$\varrho(\mathbf{C})$	1.000	1.000	1.000	1.000	0.998	1.000	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	0.380	5.000	1.720	2.600	0.300	5.000	1.360	2.560
			$r_{\%}$	0.000	0.000	0.140	0.000	0.000	0.000	0.060	0.020
			MSE($\mathbf{C}$)	2.197	32.051	0.165	0.145	1.181	32.620	0.209	0.178
			$\varrho(\mathbf{C})$	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	5	$\mathbf{X}_{ind}$	$\hat{r}$	0.060	5.000	2.360	2.540	0.020	5.000	2.040	2.580
			$r_{\%}$	0.000	1.000	0.280	0.380	0.000	1.000	0.120	0.380
			MSE($\mathbf{C}$)	1.849	813.710	0.309	0.302	0.917	102.402	0.298	0.277
			$\varrho(\mathbf{C})$	1.000	0.987	0.901	0.896	1.000	0.991	0.912	0.894
		$\mathbf{X}_{corr}$	$\hat{r}$	0.160	5.000	2.560	2.560	0.040	5.000	2.340	2.600
			$r_{\%}$	0.000	1.000	0.380	0.380	0.000	1.000	0.260	0.400
			MSE($\mathbf{C}$)	6.032	65.801	0.371	0.363	0.930	52.111	0.336	0.323
			$\varrho(\mathbf{C})$	1.000	0.992	0.904	0.929	1.000	0.988	0.925	0.906
(10,50)	3	$\mathbf{X}_{ind}$	$\hat{r}$	1.360	10.000	1.000	2.320	1.820	10.000	1.000	1.520
			$r_{\%}$	0.220	0.000	0.000	0.020	0.460	0.000	0.000	0.120
			MSE($\mathbf{C}$)	8.372	236.725	0.222	0.170	16.348	197.125	0.251	0.222
			$\varrho(\mathbf{C})$	0.962	1.000	1.000	1.000	0.943	1.000	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	1.260	10.000	1.040	2.640	1.600	10.000	1.000	1.980
			$r_{\%}$	0.220	0.000	0.000	0.020	0.220	0.000	0.000	0.120
			MSE($\mathbf{C}$)	36.481	13.903	0.300	0.232	3.239	43.166	0.281	0.228
			$\varrho(\mathbf{C})$	0.977	1.000	1.000	1.000	0.971	1.000	0.998	1.000
	5	$\mathbf{X}_{ind}$	$\hat{r}$	0.080	10.000	1.040	2.120	0.020	10.000	1.000	1.380
			$r_{\%}$	0.000	0.000	0.000	0.020	0.000	0.000	0.000	0.040
			MSE($\mathbf{C}$)	1.001	88.157	0.522	0.404	0.917	51.975	0.511	0.465
			$\varrho(\mathbf{C})$	1.000	1.000	0.998	0.999	1.000	1.000	0.999	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	0.480	10.000	1.140	2.960	0.500	10.000	1.020	2.060
			$r_{\%}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
			MSE($\mathbf{C}$)	1.363	64.436	0.559	0.402	2.441	30.059	0.618	0.525
			$\varrho(\mathbf{C})$	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000

Table A.5: Comparison of the estimated rank ($\hat{r}$), share of correctly estimated rank ($r_{\%}$), mean squared error of $\mathbf{C}$ (MSE($\mathbf{C}$)), and $\varrho(\mathbf{C}) = \|\hat{\mathbf{C}}(\hat{\mathbf{C}}'\hat{\mathbf{C}})^{+}\hat{\mathbf{C}}' - \mathbf{C}(\mathbf{C}'\mathbf{C})^{+}\mathbf{C}'\|_2$ obtained by RRcs with Dirichlet-Laplace and horseshoe prior (RRcs-DL and RRcs-HS) against ANN (Chen et al., 2013) and RRRR (She and Chen, 2017) for different values of (q, p) and true rank r_0. In all settings, $n = 50$, and the DGP is sparse with $p^* = 5$ if $p = 15$, while $p^* = 10$ if $p = 50$. We present the average estimates over 50 repetitions for independent errors ($\mathbf{\Sigma}_{ind}$), correlated errors ($\mathbf{\Sigma}_{corr}$), independent regressors ($\mathbf{X}_{ind}$), and correlated regressors ($\mathbf{X}_{corr}$).

				$\boldsymbol{\Sigma}_{ind}$				$\boldsymbol{\Sigma}_{corr}$			
(q,p)	r_0	**X**	**Measure**	**ANN**	**RRRR**	**RRcs-DL**	**RRcs-HS**	**ANN**	**RRRR**	**RRcs-DL**	**RRcs-HS**
(5,15)	3	$\mathbf{X}_{ind}$	$\hat{r}$	2.940	2.880	2.060	2.540	2.940	2.960	1.960	2.360
			$r_{\%}$	0.940	0.880	0.120	0.000	0.940	0.960	0.160	0.000
			MSE($\mathbf{C}$)	0.023	0.026	0.799	0.764	0.021	0.023	0.675	0.797
			$\varrho(\mathbf{C})$	0.260	0.292	1.000	1.000	0.228	0.223	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	2.860	2.880	2.160	2.620	2.940	2.960	1.860	2.600
			$r_{\%}$	0.860	0.840	0.120	0.000	0.940	0.920	0.060	0.000
			MSE($\mathbf{C}$)	0.042	0.045	0.697	0.674	0.040	0.043	0.748	0.777
			$\varrho(\mathbf{C})$	0.340	0.357	1.000	1.000	0.296	0.319	1.000	1.000
	5	$\mathbf{X}_{ind}$	$\hat{r}$	4.000	4.360	2.540	2.600	3.940	4.380	2.600	2.660
			$r_{\%}$	0.000	0.380	0.360	0.380	0.000	0.420	0.360	0.400
			MSE($\mathbf{C}$)	0.058	0.037	1.594	1.545	0.051	0.033	1.624	1.722
			$\varrho(\mathbf{C})$	1.000	0.743	0.880	0.830	1.000	0.730	0.888	0.832
		$\mathbf{X}_{corr}$	$\hat{r}$	3.940	4.240	2.580	2.680	3.880	4.160	2.600	2.640
			$r_{\%}$	0.000	0.320	0.380	0.400	0.000	0.280	0.400	0.400
			MSE($\mathbf{C}$)	0.088	0.070	1.633	1.555	0.080	0.064	1.724	1.686
			$\varrho(\mathbf{C})$	1.000	0.806	0.891	0.855	1.000	0.855	0.898	0.860
(5,50)	3	$\mathbf{X}_{ind}$	$\hat{r}$	0.940	5.000	1.280	2.560	1.620	5.000	1.160	2.400
			$r_{\%}$	0.260	0.000	0.020	0.020	0.360	0.000	0.000	0.000
			MSE($\mathbf{C}$)	4.957	17.005	1.285	0.974	98.964	52.488	1.535	1.202
			$\varrho(\mathbf{C})$	0.953	1.000	1.000	1.000	0.954	1.000	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	0.800	5.000	1.860	2.380	1.620	5.000	1.400	2.600
			$r_{\%}$	0.200	0.000	0.020	0.020	0.420	0.000	0.000	0.000
			MSE($\mathbf{C}$)	15.744	155.578	1.395	1.217	303602.928	34.301	1.653	1.337
			$\varrho(\mathbf{C})$	0.969	1.000	1.000	1.000	0.973	1.000	1.000	1.000
	5	$\mathbf{X}_{ind}$	$\hat{r}$	0.000	5.000	2.600	2.560	0.160	5.000	2.520	2.560
			$r_{\%}$	0.000	1.000	0.400	0.380	0.000	1.000	0.380	0.360
			MSE($\mathbf{C}$)	5.179	228.422	2.586	2.233	5.526	580.312	2.741	2.301
			$\varrho(\mathbf{C})$	1.000	0.972	0.990	0.933	1.000	0.952	0.990	0.926
		$\mathbf{X}_{corr}$	$\hat{r}$	0.180	5.000	2.600	2.600	0.420	5.000	2.600	2.520
			$r_{\%}$	0.000	1.000	0.400	0.400	0.000	1.000	0.400	0.380
			MSE($\mathbf{C}$)	12.083	36.098	3.064	2.639	463.520	128.832	3.293	2.842
			$\varrho(\mathbf{C})$	1.000	0.962	0.989	0.928	1.000	0.978	0.989	0.931
(10,50)	3	$\mathbf{X}_{ind}$	$\hat{r}$	3.000	10.000	1.260	1.260	3.200	10.000	1.220	1.020
			$r_{\%}$	1.000	0.000	0.040	0.000	0.800	0.000	0.080	0.000
			MSE($\mathbf{C}$)	1391.891	494.411	1.570	1.693	25.856	99.000	1.711	1.756
			$\varrho(\mathbf{C})$	0.730	1.000	1.000	1.000	0.782	1.000	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	3.000	10.000	1.240	1.560	3.260	10.000	1.180	1.180
			$r_{\%}$	1.000	0.000	0.020	0.020	0.740	0.000	0.000	0.000
			MSE($\mathbf{C}$)	298.131	76.197	2.062	2.167	25.326	84.624	1.884	1.956
			$\varrho(\mathbf{C})$	0.806	1.000	1.000	1.000	0.839	1.000	1.000	1.000
	5	$\mathbf{X}_{ind}$	$\hat{r}$	2.420	10.000	1.280	1.000	4.060	10.000	1.060	1.000
			$r_{\%}$	0.480	0.000	0.000	0.000	0.500	0.000	0.000	0.000
			MSE($\mathbf{C}$)	25.933	171.279	3.548	3.697	18.649	43.215	4.064	3.898
			$\varrho(\mathbf{C})$	0.922	1.000	1.000	1.000	0.919	1.000	1.000	1.000
		$\mathbf{X}_{corr}$	$\hat{r}$	2.200	10.000	1.200	1.060	2.680	10.000	1.220	1.020
			$r_{\%}$	0.400	0.000	0.000	0.000	0.380	0.000	0.000	0.000
			MSE($\mathbf{C}$)	10.417	47.036	3.913	3.981	545.300	67.874	4.231	4.262
			$\varrho(\mathbf{C})$	0.953	1.000	1.000	1.000	0.948	1.000	1.000	1.000

Table A.6: Comparison of the estimated rank ($\hat{r}$), share of correctly estimated rank ($r_{\%}$), mean squared error of $\mathbf{C}$ (MSE($\mathbf{C}$)), and $\varrho(\mathbf{C}) = \|\hat{\mathbf{C}}(\hat{\mathbf{C}}'\hat{\mathbf{C}})^{+}\hat{\mathbf{C}}' - \mathbf{C}(\mathbf{C}'\mathbf{C})^{+}\mathbf{C}'\|_2$ obtained by RRcs with Dirichlet-Laplace and horseshoe prior (RRcs-DL and RRcs-HS) against ANN (Chen et al., 2013) and RRRR (She and Chen, 2017) for different values of (q, p) and true rank r_0. In all settings, $n = 50$, and the DGP is non-sparse with $p^* = p$. We present the average estimates over 50 repetitions for independent errors ($\boldsymbol{\Sigma}_{ind}$), correlated errors ($\boldsymbol{\Sigma}_{corr}$), independent regressors ($\mathbf{X}_{ind}$), and correlated regressors ($\mathbf{X}_{corr}$).

A.4 Further results on COMBO-17 dataset

The application of the methodology to the COMBO-17 galaxy dataset illustrates extensively the performance of the proposed approach. In this section, we provide further comprehension of the

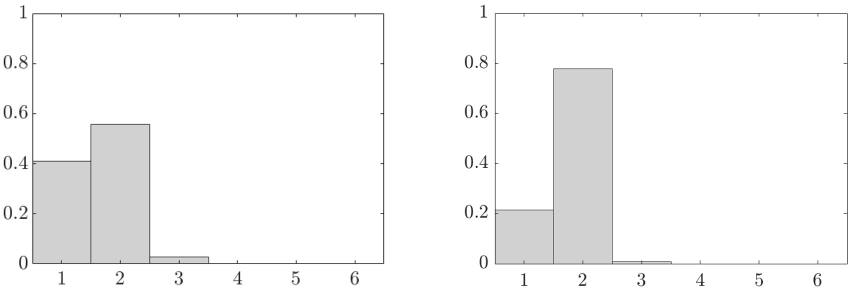
j	**Name**	**Description**
1	UjMAG	$M_{abs,gal}$ in Johnson U
2	BjMAG	$M_{abs,gal}$ in Johnson B
3	VjMAG	$M_{abs,gal}$ in Johnson V
4	usMAG	$M_{abs,gal}$ in SDSS u
5	gsMAG	$M_{abs,gal}$ in SDSS g
6	rsMAG	$M_{abs,gal}$ in SDSS r
7	UbMAG	$M_{abs,gal}$ in Bessell U
8	BbMAG	$M_{abs,gal}$ in Bessell B
9	VbMAG	$M_{abs,gal}$ in Bessell V
10	S280MAG	$M_{abs,gal}$ in 280/40
11	W420FE	photon flux in filter 420 in run E
12	W462FE	photon flux in filter 462 in run E
13	W485FD	photon flux in filter 485 in run D
14	W518FE	photon flux in filter 518 in run E
15	W571FS	photon flux in filter 571 combined
16	W604FE	photon flux in filter 604 in run E
17	W646FD	photon flux in filter 646 in run D
18	W696FE	photon flux in filter 696 in run E
19	W753FE	photon flux in filter 753 in run E
20	W815FS	photon flux in filter 815 combined
21	W856FD	photon flux in filter 856 in run D
22	W914FD	photon flux in filter 914 in run D
23	W914FE	photon flux in filter 914 in run E

Table A.7: Name and description of covariates in the COMBO-17 galaxy dataset. Covariates $1-10$ are the absolute magnitude $M_{abs,gal}$ of the galaxy in the indicated passband, and covariates $11-23$ are the photon flux in different filters and various runs of the experiment.

covariates studied in the model (Table A.7), and show the posterior concentration of the rank when the number of data points increases (Figure A.3), results obtained with the full dataset (Table A.8 and Figure A.4), and plots of the probability mass function and tail distribution of RI for all covariates (Figures A.5 and A.6).

The COMBO-17 object catalogue of the Chandra Deep Field South lists photometry, astrometry and morphological features of more than $60,000$ celestial objects. We restrict the analysis to $3,438$ objects, all classified as "Galaxies", with no missing values. Measurement errors and redundant variables were omitted, resulting in a total of 29 variables divided into 23 covariates and 6 responses, as done in Izenman (2008). Table A.7 provides the description of the 23 explanatory variables, whereas deeper comprehension of how the measurements were acquired, the mentioned passbands, filters and experimental runs is found in Wolf et al. (2004).

We demonstrated in the simulation results in Chapter 2 how the posterior distribution of the rank concentrates around the estimated value as the number of data points is increased. The application of the methodology to the actual dataset on galaxy photometry illustrates high concentration when the observations change from the sub-sample with $n = 500$ observations to the full set (Figure A.3).

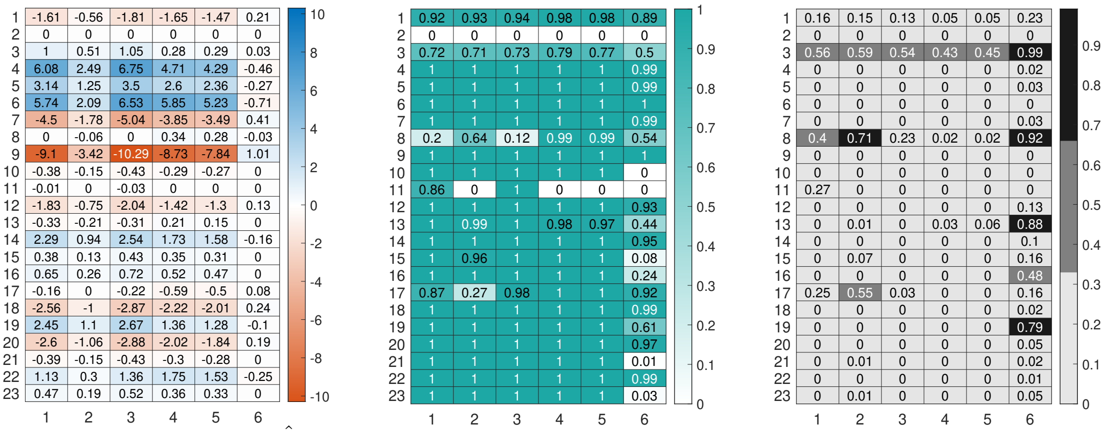


(a) $n = 500$

(b) $n = 3,438$

Figure A.3: Posterior distribution of the estimated rank in the sub-sample of 500 observations of the galaxy dataset (panel a), and the full set (panel b). The plots exhibit posterior concentration of the rank around 2 as the number of data points is increased.

The estimated sparse coefficient matrix $\hat{\mathbf{C}}$ in the sub-sample with 500 data points sets 28% of the coefficients to zero and excludes four covariates from the model. Meanwhile, the full model encompasses 14% of its entries as zeros, and only one complete row of zeros (Figure A.4a). The uncertainty about this decision is low, as reported by the PIP and the PIP uncertainty index in Figures A.4b and A.4c.



(a) Sparse estimate $\hat{\mathbf{C}}$. (b) PIP of $\mathbf{C}_{jk}$. (c) PIP uncertainty index ζ_{jk}.

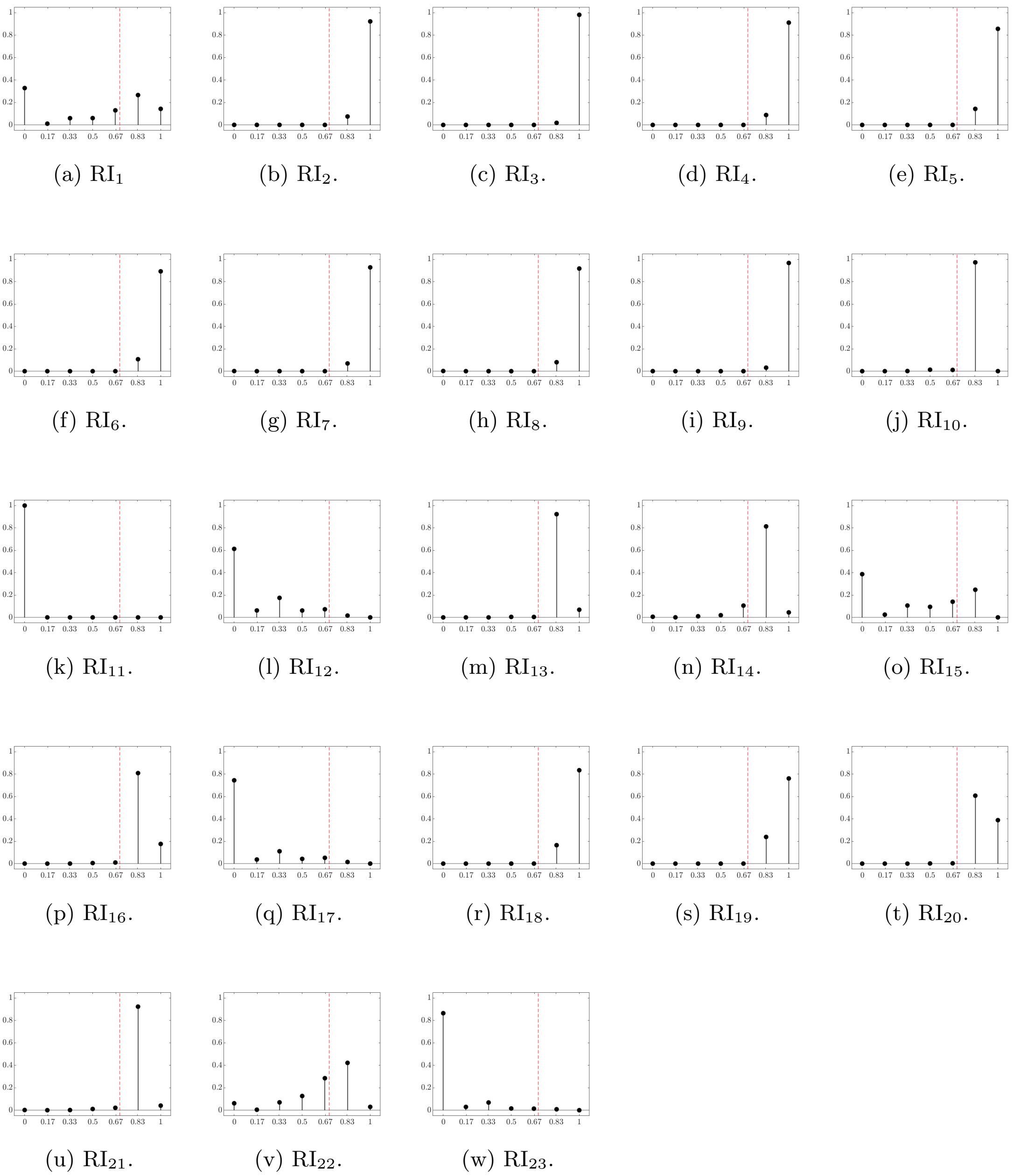
Figure A.4: Sparse estimate $\hat{\mathbf{C}}$ of the coefficient matrix $\mathbf{C}$ of the linear regression model with the full galaxy dataset (panel a), the uncertainty about this estimation through the posterior inclusion probabilities (panel b), and the PIP uncertainty index, in a grey-colour scale according to low ($\zeta_{jk} \leq 1/3$), medium ($1/3 < \zeta_{jk} \leq 2/3$), or high ($\zeta_{jk} > 2/3$) uncertainty (panel c).

A measure that synthesises the relevance of each covariate over all responses is the relevance index. Table A.8 presents the summary statistics of the distribution of RI in the full dataset, as a method to identify the relevant covariates in the model when considering their effect across all responses. The uncertainty about the inclusion of a covariate is given by the standard deviation of the relevance index: the higher the dispersion of RI, the higher the uncertainty about inclusion or exclusion. For instance, covariate x_3 is estimated to have an effect on the 6 responses (Figure A.4a). Nonetheless, $\mathbf{C}_{3k} > 1/3$ for all k (Figure A.4c), interpreted as a medium to high uncertainty for all of its entries, and the corresponding relevance index in Table A.8 exhibits the highest variance (std = 0.372).

x_j	**Mode**	**Mean**	**Std**	**Q25**	**Q50**	**Q75**
1	1	0.937	0.165	1	1	1
2	0	0	0.011	0	0	0
3	1	0.704	0.372	0.5	0.833	1
4	1	0.998	0.016	1	1	1
5	1	0.998	0.019	1	1	1
6	1	1	0.004	1	1	1
7	1	0.998	0.019	1	1	1
8	0.667	0.58	0.177	0.5	0.667	0.667
9	1	1	0.005	1	1	1
10	0.833	0.833	0.004	0.833	0.833	0.833
11	0.333	0.311	0.058	0.333	0.333	0.333
12	1	0.989	0.042	1	1	1
13	0.833	0.897	0.106	0.833	0.833	1
14	1	0.992	0.036	1	1	1
15	0.833	0.841	0.042	0.833	0.833	0.833
16	0.833	0.873	0.071	0.833	0.833	0.833
17	0.833	0.842	0.068	0.833	0.833	0.833
18	1	0.999	0.015	1	1	1
19	1	0.934	0.081	0.833	1	1
20	1	0.995	0.027	1	1	1
21	0.833	0.834	0.022	0.833	0.833	0.833
22	1	0.999	0.013	1	1	1
23	0.833	0.837	0.026	0.833	0.833	0.833

Table A.8: Summary statistics of the distribution of RI in the full dataset ($n = 3,438$) for each covariate: mode, mean, standard deviation (Std), and the 25th, 50th and 75th quartiles (Q25, Q50, Q75). The shaded rows identify covariates with high uncertainty ($\text{Std}(\text{RI}_j) \geq 0.30$).

A visual representation of uncertainty in variable selection is illustrated by RI probability mass function, as depicted in Figure A.5 for the analysed sub-sample. When the distribution is left-skewed or right-skewed, there is low uncertainty about the covariate inclusion or exclusion. For example, x_9 and x_{17} present this behaviour. However, a distribution without concentrated mass indicates high uncertainty, depicted in the plots of x_1, x_{15} and x_{22}. A binary decision about inclusion is given by the tail distribution or survival function of RI (Figure A.6).



(a) RI_1

(b) RI_2.

(c) RI_3.

(d) RI_4.

(e) RI_5.

(f) RI_6.

(g) RI_7.

(h) RI_8.

(i) RI_9.

(j) RI_{10}.

(k) RI_{11}.

(l) RI_{12}.

(m) RI_{13}.

(n) RI_{14}.

(o) RI_{15}.

(p) RI_{16}.

(q) RI_{17}.

(r) RI_{18}.

(s) RI_{19}.

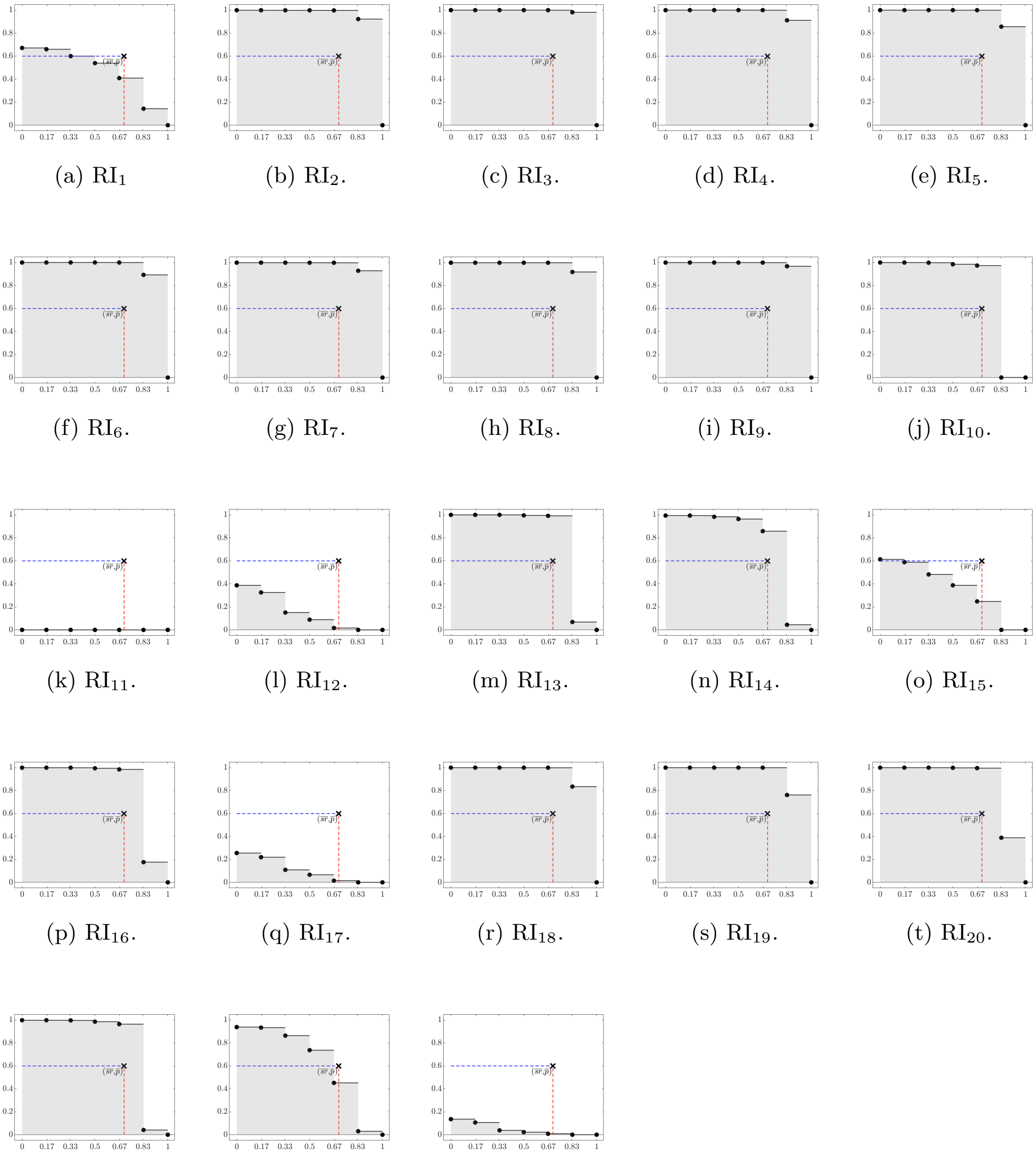
(t) RI_{20}.

(u) RI_{21}.

(v) RI_{22}.

(w) RI_{23}.

Figure A.5: For each covariate in the sub-sample with $n = 500$ data points of the galaxy dataset, the probability mass function of the respective Relevance Index (RI) is reported. Here we used $\overline{sr} = 0.70$ (red vertical line): the covariates with *more* mass (to be specified through $\overline{p}$) located to the right of $\overline{sr}$ are preferred to be included over the counterpart.



(a) RI_1

(b) RI_2.

(c) RI_3.

(d) RI_4.

(e) RI_5.

(f) RI_6.

(g) RI_7.

(h) RI_8.

(i) RI_9.

(j) RI_{10}.

(k) RI_{11}.

(l) RI_{12}.

(m) RI_{13}.

(n) RI_{14}.

(o) RI_{15}.

(p) RI_{16}.

(q) RI_{17}.

(r) RI_{18}.

(s) RI_{19}.

(t) RI_{20}.

(u) RI_{21}.

(v) RI_{22}.

(w) RI_{23}.

Figure A.6: For each covariate in the sub-sample with $n = 500$ data points of the galaxy dataset, the survival function S_{RI} of the respective Relevance Index RI is reported. Here we used $\overline{sr} = 0.70$ (red vertical line) and $\overline{p} = 0.60$ (blue horizontal line). The area below the survival function is shaded: if the point $(\overline{sr}, \overline{p})$ is located outside the shaded area, then the covariate is to be excluded.

A.4.1 Results for $n = 500$ and $n = 1,500$

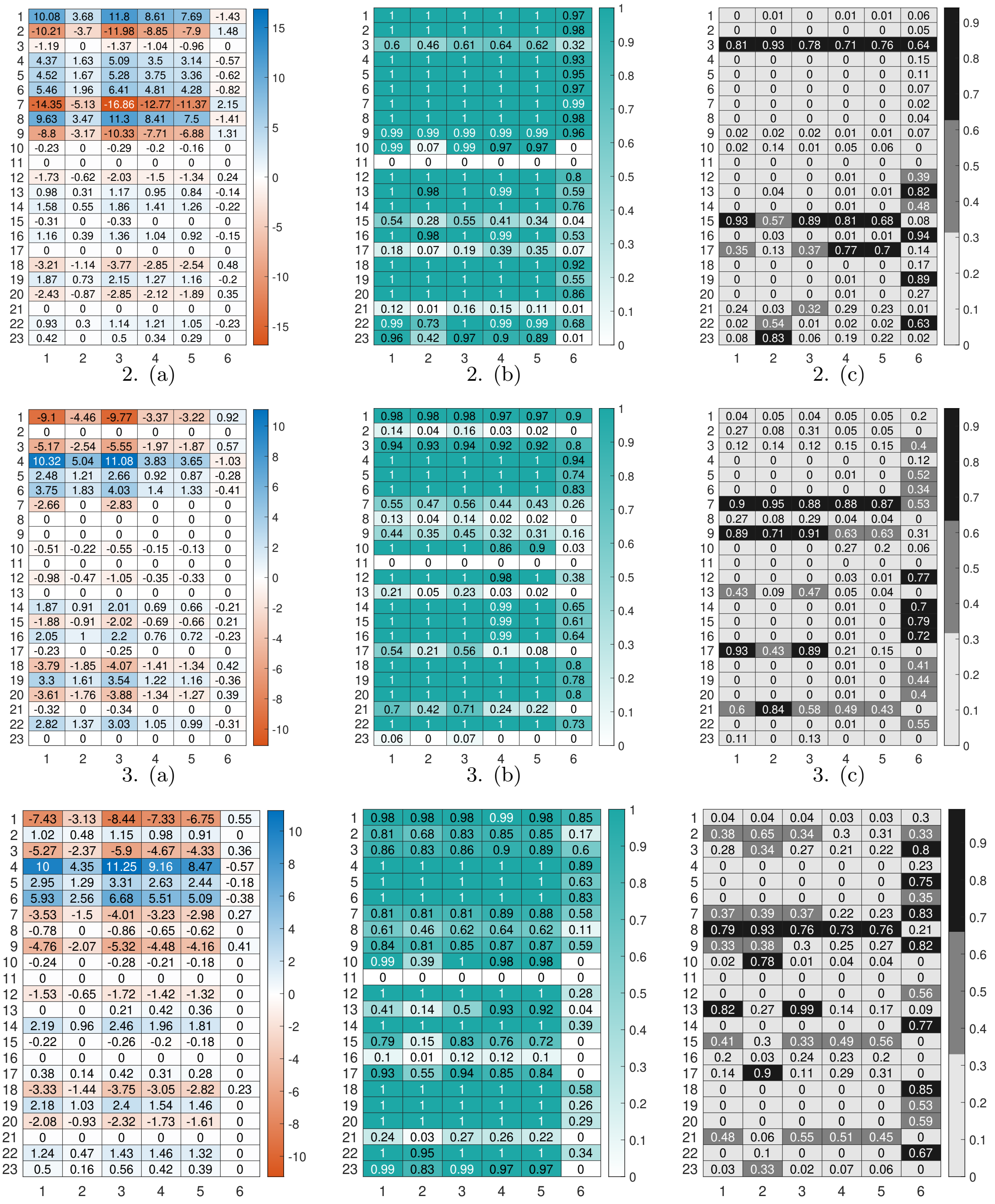
Section 2.5 analysed a sub-sample of the COMBO-17 galaxy dataset consisting of $n = 500$ randomly chosen data points. This has been motivated by the intention of highlighting the ability to quantify the uncertainty of the proposed BRECS method.

As a single random sub-sample might not be representative of the full dataset, in this Section we report the results from the application of BRECS to 5 other randomly drawn sub-samples of size $n = 500$ and 6 sub-samples of size $n = 1,500$. The purpose is twofold: first, we aim to show that the difference between the results in Chapter 2.5 and those with the full dataset (see the previous Section) is due to a small sub-sample size. Second, we argue that the “cross-sample” differences wipe out as the size of the sub-sample increases.

By comparing the plots in Figure A.7, we find an overall similarity in the sparsity pattern of the estimated matrix $\hat{\mathbf{C}}$. However, the entry-wise PIPs and associated uncertainty index ζ show some differences in terms of the precise location of certain zero coefficients (an associated uncertainty). For example, in the 3rd run (subfigure 3), the coefficients of the second covariate have PIPs close to 0 and very low uncertainty, whereas in the 6th run (subfigure 6) all their PIPs except one are greater than 0.50 with medium uncertainty. This is due to the randomness in the selection of the sub-samples, which might result in selecting a handful of data that are affecting the overall results. However, notice that the overall degree of uncertainty shown by the (c) subplots seems similar across all the runs.

Comparing the results for $n = 1,500$ in Figure A.8 to those for $n = 500$, we find evidence of two main changes. First, with $n = 1,500$ we notice a reduction in the number of zero rows compared to the sub-samples with fewer data points, although it remains higher than that of the full dataset. Second, the average level of uncertainty is lower with $n = 1,500$, yet higher than that of the full dataset.

Overall, these results are in line with expectations: the larger the size of the sub-samples, the closer the results get to the full sample. Moreover, as the sub-sample size increases, the uncertainty around parameter classification (zero versus non-zero) decreases.



2. (a)

2. (b)

2. (c)

3. (a)

3. (b)

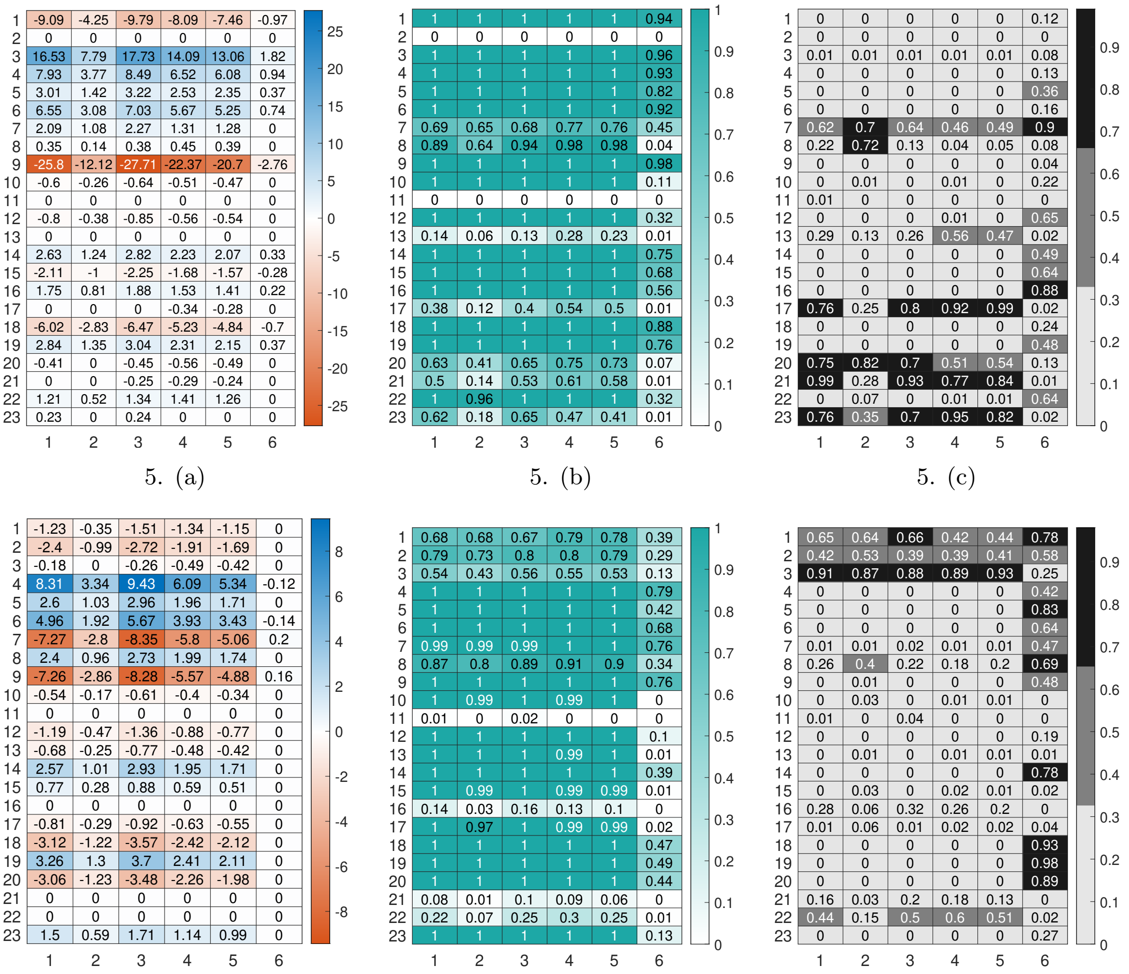
3. (c)

4. (a)

4. (b)

4. (c)

Figure A.7: The methodology was applied six times to the dataset with $n = 500$. The rows correspond to each run (refer to Chapter 2.5 for the results of the first run), where we plot the sparse estimates $\hat{\mathbf{C}}$ of the coefficient matrix $\mathbf{C}$ (panel a), the uncertainty about this estimation through the posterior inclusion probabilities (panel b), and the PIP uncertainty index, in a grey-colour scale according to low ($\zeta_{jk} \leq 1/3$), medium ($1/3 < \zeta_{jk} \leq 2/3$), or high ($\zeta_{jk} > 2/3$) uncertainty (panel c).



5. (a)

	1	2	3	4	5	6
1	-9.09	-4.25	-9.79	-8.09	-7.46	-0.97
2	0	0	0	0	0	0
3	16.53	7.79	17.73	14.09	13.06	1.82
4	7.93	3.77	8.49	6.52	6.08	0.94
5	3.01	1.42	3.22	2.53	2.35	0.37
6	6.55	3.08	7.03	5.67	5.25	0.74
7	2.09	1.08	2.27	1.31	1.28	0
8	0.35	0.14	0.38	0.45	0.39	0
9	-25.8	-12.12	-27.71	-22.37	-20.7	-2.76
10	-0.6	-0.26	-0.64	-0.51	-0.47	0
11	0	0	0	0	0	0
12	-0.8	-0.38	-0.85	-0.56	-0.54	0
13	0	0	0	0	0	0
14	2.63	1.24	2.82	2.23	2.07	0.33
15	-2.11	-1	-2.25	-1.68	-1.57	-0.28
16	1.75	0.81	1.88	1.53	1.41	0.22
17	0	0	0	-0.34	-0.28	0
18	-6.02	-2.83	-6.47	-5.23	-4.84	-0.7
19	2.84	1.35	3.04	2.31	2.15	0.37
20	-0.41	0	-0.45	-0.56	-0.49	0
21	0	0	-0.25	-0.29	-0.24	0
22	1.21	0.52	1.34	1.41	1.26	0
23	0.23	0	0.24	0	0	0

5. (b)

	1	2	3	4	5	6
1	1	1	1	1	1	0.94
2	0	0	0	0	0	0
3	1	1	1	1	1	0.96
4	1	1	1	1	1	0.93
5	1	1	1	1	1	0.82
6	1	1	1	1	1	0.92
7	0.69	0.65	0.68	0.77	0.76	0.45
8	0.89	0.64	0.94	0.98	0.98	0.04
9	1	1	1	1	1	0.98
10	1	1	1	1	1	0.11
11	0	0	0	0	0	0
12	1	1	1	1	1	0.32
13	0.14	0.06	0.13	0.28	0.23	0.01
14	1	1	1	1	1	0.75
15	1	1	1	1	1	0.68
16	1	1	1	1	1	0.56
17	0.38	0.12	0.4	0.54	0.5	0.01
18	1	1	1	1	1	0.88
19	1	1	1	1	1	0.76
20	0.63	0.41	0.65	0.75	0.73	0.07
21	0.5	0.14	0.53	0.61	0.58	0.01
22	1	0.96	1	1	1	0.32
23	0.62	0.18	0.65	0.47	0.41	0.01

5. (c)

	1	2	3	4	5	6
1	0	0	0	0	0	0.12
2	0	0	0	0	0	0
3	0.01	0.01	0.01	0.01	0.01	0.08
4	0	0	0	0	0	0.13
5	0	0	0	0	0	0.36
6	0	0	0	0	0	0.16
7	0.62	0.7	0.64	0.46	0.49	0.9
8	0.22	0.72	0.13	0.04	0.05	0.08
9	0	0	0	0	0	0.04
10	0	0.01	0	0.01	0	0.22
11	0.01	0	0	0	0	0
12	0	0	0	0.01	0	0.65
13	0.29	0.13	0.26	0.56	0.47	0.02
14	0	0	0	0	0	0.49
15	0	0	0	0	0	0.64
16	0	0	0	0	0	0.88
17	0.76	0.25	0.8	0.92	0.99	0.02
18	0	0	0	0	0	0.24
19	0	0	0	0	0	0.48
20	0.75	0.82	0.7	0.51	0.54	0.13
21	0.99	0.28	0.93	0.77	0.84	0.01
22	0	0.07	0	0.01	0.01	0.64
23	0.76	0.35	0.7	0.95	0.82	0.02

6. (a)

	1	2	3	4	5	6
1	-1.23	-0.35	-1.51	-1.34	-1.15	0
2	-2.4	-0.99	-2.72	-1.91	-1.69	0
3	-0.18	0	-0.26	-0.49	-0.42	0
4	8.31	3.34	9.43	6.09	5.34	-0.12
5	2.6	1.03	2.96	1.96	1.71	0
6	4.96	1.92	5.67	3.93	3.43	-0.14
7	-7.27	-2.8	-8.35	-5.8	-5.06	0.2
8	2.4	0.96	2.73	1.99	1.74	0
9	-7.26	-2.86	-8.28	-5.57	-4.88	0.16
10	-0.54	-0.17	-0.61	-0.4	-0.34	0
11	0	0	0	0	0	0
12	-1.19	-0.47	-1.36	-0.88	-0.77	0
13	-0.68	-0.25	-0.77	-0.48	-0.42	0
14	2.57	1.01	2.93	1.95	1.71	0
15	0.77	0.28	0.88	0.59	0.51	0
16	0	0	0	0	0	0
17	-0.81	-0.29	-0.92	-0.63	-0.55	0
18	-3.12	-1.22	-3.57	-2.42	-2.12	0
19	3.26	1.3	3.7	2.41	2.11	0
20	-3.06	-1.23	-3.48	-2.26	-1.98	0
21	0	0	0	0	0	0
22	0	0	0	0	0	0
23	1.5	0.59	1.71	1.14	0.99	0

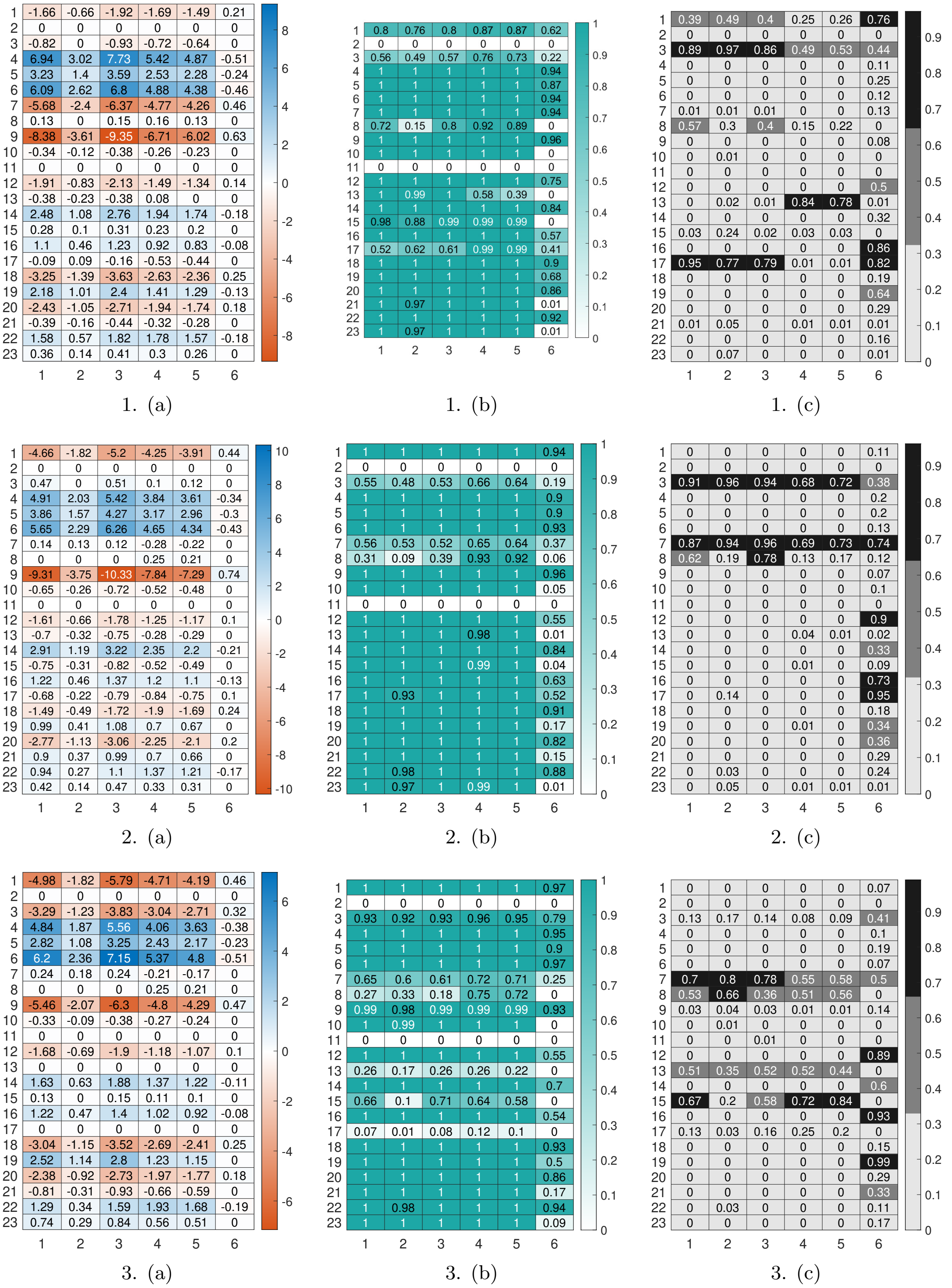
6. (b)

	1	2	3	4	5	6
1	0.68	0.68	0.67	0.79	0.78	0.39
2	0.79	0.73	0.8	0.8	0.79	0.29
3	0.54	0.43	0.56	0.55	0.53	0.13
4	1	1	1	1	1	0.79
5	1	1	1	1	1	0.42
6	1	1	1	1	1	0.68
7	0.99	0.99	0.99	1	1	0.76
8	0.87	0.8	0.89	0.91	0.9	0.34
9	1	1	1	1	1	0.76
10	1	0.99	1	0.99	1	0
11	0.01	0	0.02	0	0	0
12	1	1	1	1	1	0.1
13	1	1	1	0.99	1	0.01
14	1	1	1	1	1	0.39
15	1	0.99	1	0.99	0.99	0.01
16	0.14	0.03	0.16	0.13	0.1	0
17	1	0.97	1	0.99	0.99	0.02
18	1	1	1	1	1	0.47
19	1	1	1	1	1	0.49
20	1	1	1	1	1	0.44
21	0.08	0.01	0.1	0.09	0.06	0
22	0.22	0.07	0.25	0.3	0.25	0.01
23	1	1	1	1	1	0.13

6. (c)

	1	2	3	4	5	6
1	0.65	0.64	0.66	0.42	0.44	0.78
2	0.42	0.53	0.39	0.39	0.41	0.58
3	0.91	0.87	0.88	0.89	0.93	0.25
4	0	0	0	0	0	0.42
5	0	0	0	0	0	0.83
6	0	0	0	0	0	0.64
7	0.01	0.01	0.02	0.01	0.01	0.47
8	0.26	0.4	0.22	0.18	0.2	0.69
9	0	0.01	0	0	0	0.48
10	0	0.03	0	0.01	0.01	0
11	0.01	0	0.04	0	0	0
12	0	0	0	0	0	0.19
13	0	0.01	0	0.01	0.01	0.01
14	0	0	0	0	0	0.78
15	0	0.03	0	0.02	0.01	0.02
16	0.28	0.06	0.32	0.26	0.2	0
17	0.01	0.06	0.01	0.02	0.02	0.04
18	0	0	0	0	0	0.93
19	0	0	0	0	0	0.98
20	0	0	0	0	0	0.89
21	0.16	0.03	0.2	0.18	0.13	0
22	0.44	0.15	0.5	0.6	0.51	0.02
23	0	0	0	0	0	0.27

Figure A.7: The methodology was applied six times to the dataset with $n = 500$. The rows correspond to each run (refer to Chapter 2 for the results of the first run), where we plot the sparse estimates $\hat{\mathbf{C}}$ of the coefficient matrix $\mathbf{C}$ (panel a), the uncertainty about this estimation through the posterior inclusion probabilities (panel b), and the PIP uncertainty index, in a grey-colour scale according to low ($\zeta_{jk} \leq 1/3$), medium ($1/3 < \zeta_{jk} \leq 2/3$), or high ($\zeta_{jk} > 2/3$) uncertainty (panel c) (cont.).



	1	2	3	4	5	6
1	-1.66	-0.66	-1.92	-1.69	-1.49	0.21
2	0	0	0	0	0	0
3	-0.82	0	-0.93	-0.72	-0.64	0
4	6.94	3.02	7.73	5.42	4.87	-0.51
5	3.23	1.4	3.59	2.53	2.28	-0.24
6	6.09	2.62	6.8	4.88	4.38	-0.46
7	-5.68	-2.4	-6.37	-4.77	-4.26	0.46
8	0.13	0	0.15	0.16	0.13	0
9	-8.38	-3.61	-9.35	-6.71	-6.02	0.63
10	-0.34	-0.12	-0.38	-0.26	-0.23	0
11	0	0	0	0	0	0
12	-1.91	-0.83	-2.13	-1.49	-1.34	0.14
13	-0.38	-0.23	-0.38	0.08	0	0
14	2.48	1.08	2.76	1.94	1.74	-0.18
15	0.28	0.1	0.31	0.23	0.2	0
16	1.1	0.46	1.23	0.92	0.83	-0.08
17	-0.09	0.09	-0.16	-0.53	-0.44	0
18	-3.25	-1.39	-3.63	-2.63	-2.36	0.25
19	2.18	1.01	2.4	1.41	1.29	-0.13
20	-2.43	-1.05	-2.71	-1.94	-1.74	0.18
21	-0.39	-0.16	-0.44	-0.32	-0.28	0
22	1.58	0.57	1.82	1.78	1.57	-0.18
23	0.36	0.14	0.41	0.3	0.26	0

1. (a)

	1	2	3	4	5	6
1	0.8	0.76	0.8	0.87	0.87	0.62
2	0	0	0	0	0	0
3	0.56	0.49	0.57	0.76	0.73	0.22
4	1	1	1	1	1	0.94
5	1	1	1	1	1	0.87
6	1	1	1	1	1	0.94
7	1	1	1	1	1	0.94
8	0.72	0.15	0.8	0.92	0.89	0
9	1	1	1	1	1	0.96
10	1	1	1	1	1	0
11	0	0	0	0	0	0
12	1	1	1	1	1	0.75
13	1	0.99	1	0.58	0.39	0
14	1	1	1	1	1	0.84
15	0.98	0.88	0.99	0.99	0.99	0
16	1	1	1	1	1	0.57
17	0.52	0.62	0.61	0.99	0.99	0.41
18	1	1	1	1	1	0.9
19	1	1	1	1	1	0.68
20	1	1	1	1	1	0.86
21	1	0.97	1	1	1	0.01
22	1	1	1	1	1	0.92
23	1	0.97	1	1	1	0.01

1. (b)

	1	2	3	4	5	6
1	0.39	0.49	0.4	0.25	0.26	0.76
2	0	0	0	0	0	0
3	0.89	0.97	0.86	0.49	0.53	0.44
4	0	0	0	0	0	0.11
5	0	0	0	0	0	0.25
6	0	0	0	0	0	0.12
7	0.01	0.01	0.01	0	0	0.13
8	0.57	0.3	0.4	0.15	0.22	0
9	0	0	0	0	0	0.08
10	0	0.01	0	0	0	0
11	0	0	0	0	0	0
12	0	0	0	0	0	0.5
13	0	0.02	0.01	0.84	0.78	0.01
14	0	0	0	0	0	0.32
15	0.03	0.24	0.02	0.03	0.03	0
16	0	0	0	0	0	0.86
17	0.95	0.77	0.79	0.01	0.01	0.82
18	0	0	0	0	0	0.19
19	0	0	0	0	0	0.64
20	0	0	0	0	0	0.29
21	0.01	0.05	0	0.01	0.01	0.01
22	0	0	0	0	0	0.16
23	0	0.07	0	0	0	0.01

1. (c)

	1	2	3	4	5	6
1	-4.66	-1.82	-5.2	-4.25	-3.91	0.44
2	0	0	0	0	0	0
3	0.47	0	0.51	0.1	0.12	0
4	4.91	2.03	5.42	3.84	3.61	-0.34
5	3.86	1.57	4.27	3.17	2.96	-0.3
6	5.65	2.29	6.26	4.65	4.34	-0.43
7	0.14	0.13	0.12	-0.28	-0.22	0
8	0	0	0	0.25	0.21	0
9	-9.31	-3.75	-10.33	-7.84	-7.29	0.74
10	-0.65	-0.26	-0.72	-0.52	-0.48	0
11	0	0	0	0	0	0
12	-1.61	-0.66	-1.78	-1.25	-1.17	0.1
13	-0.7	-0.32	-0.75	-0.28	-0.29	0
14	2.91	1.19	3.22	2.35	2.2	-0.21
15	-0.75	-0.31	-0.82	-0.52	-0.49	0
16	1.22	0.46	1.37	1.2	1.1	-0.13
17	-0.68	-0.22	-0.79	-0.84	-0.75	0.1
18	-1.49	-0.49	-1.72	-1.9	-1.69	0.24
19	0.99	0.41	1.08	0.7	0.67	0
20	-2.77	-1.13	-3.06	-2.25	-2.1	0.2
21	0.9	0.37	0.99	0.7	0.66	0
22	0.94	0.27	1.1	1.37	1.21	-0.17
23	0.42	0.14	0.47	0.33	0.31	0

2. (a)

	1	2	3	4	5	6
1	1	1	1	1	1	0.94
2	0	0	0	0	0	0
3	0.55	0.48	0.53	0.66	0.64	0.19
4	1	1	1	1	1	0.9
5	1	1	1	1	1	0.9
6	1	1	1	1	1	0.93
7	0.56	0.53	0.52	0.65	0.64	0.37
8	0.31	0.09	0.39	0.93	0.92	0.06
9	1	1	1	1	1	0.96
10	1	1	1	1	1	0.05
11	0	0	0	0	0	0
12	1	1	1	1	1	0.55
13	1	1	1	0.98	1	0.01
14	1	1	1	1	1	0.84
15	1	1	1	0.99	1	0.04
16	1	1	1	1	1	0.63
17	1	0.93	1	1	1	0.52
18	1	1	1	1	1	0.91
19	1	1	1	1	1	0.17
20	1	1	1	1	1	0.82
21	1	1	1	1	1	0.15
22	1	0.98	1	1	1	0.88
23	1	0.97	1	0.99	1	0.01

2. (b)

	1	2	3	4	5	6
1	0	0	0	0	0	0.11
2	0	0	0	0	0	0
3	0.91	0.96	0.94	0.68	0.72	0.38
4	0	0	0	0	0	0.2
5	0	0	0	0	0	0.2
6	0	0	0	0	0	0.13
7	0.87	0.94	0.96	0.69	0.73	0.74
8	0.62	0.19	0.78	0.13	0.17	0.12
9	0	0	0	0	0	0.07
10	0	0	0	0	0	0.1
11	0	0	0	0	0	0
12	0	0	0	0	0	0.9
13	0	0	0	0.04	0.01	0.02
14	0	0	0	0	0	0.33
15	0	0	0	0.01	0	0.09
16	0	0	0	0	0	0.73
17	0	0.14	0	0	0	0.95
18	0	0	0	0	0	0.18
19	0	0	0	0.01	0	0.34
20	0	0	0	0	0	0.36
21	0	0	0	0	0	0.29
22	0	0.03	0	0	0	0.24
23	0	0.05	0	0.01	0.01	0.01

2. (c)

	1	2	3	4	5	6
1	-4.98	-1.82	-5.79	-4.71	-4.19	0.46
2	0	0	0	0	0	0
3	-3.29	-1.23	-3.83	-3.04	-2.71	0.32
4	4.84	1.87	5.56	4.06	3.63	-0.38
5	2.82	1.08	3.25	2.43	2.17	-0.23
6	6.2	2.36	7.15	5.37	4.8	-0.51
7	0.24	0.18	0.24	-0.21	-0.17	0
8	0	0	0	0.25	0.21	0
9	-5.46	-2.07	-6.3	-4.8	-4.29	0.47
10	-0.33	-0.09	-0.38	-0.27	-0.24	0
11	0	0	0	0	0	0
12	-1.68	-0.69	-1.9	-1.18	-1.07	0.1
13	0	0	0	0	0	0
14	1.63	0.63	1.88	1.37	1.22	-0.11
15	0.13	0	0.15	0.11	0.1	0
16	1.22	0.47	1.4	1.02	0.92	-0.08
17	0	0	0	0	0	0
18	-3.04	-1.15	-3.52	-2.69	-2.41	0.25
19	2.52	1.14	2.8	1.23	1.15	0
20	-2.38	-0.92	-2.73	-1.97	-1.77	0.18
21	-0.81	-0.31	-0.93	-0.66	-0.59	0
22	1.29	0.34	1.59	1.93	1.68	-0.19
23	0.74	0.29	0.84	0.56	0.51	0

3. (a)

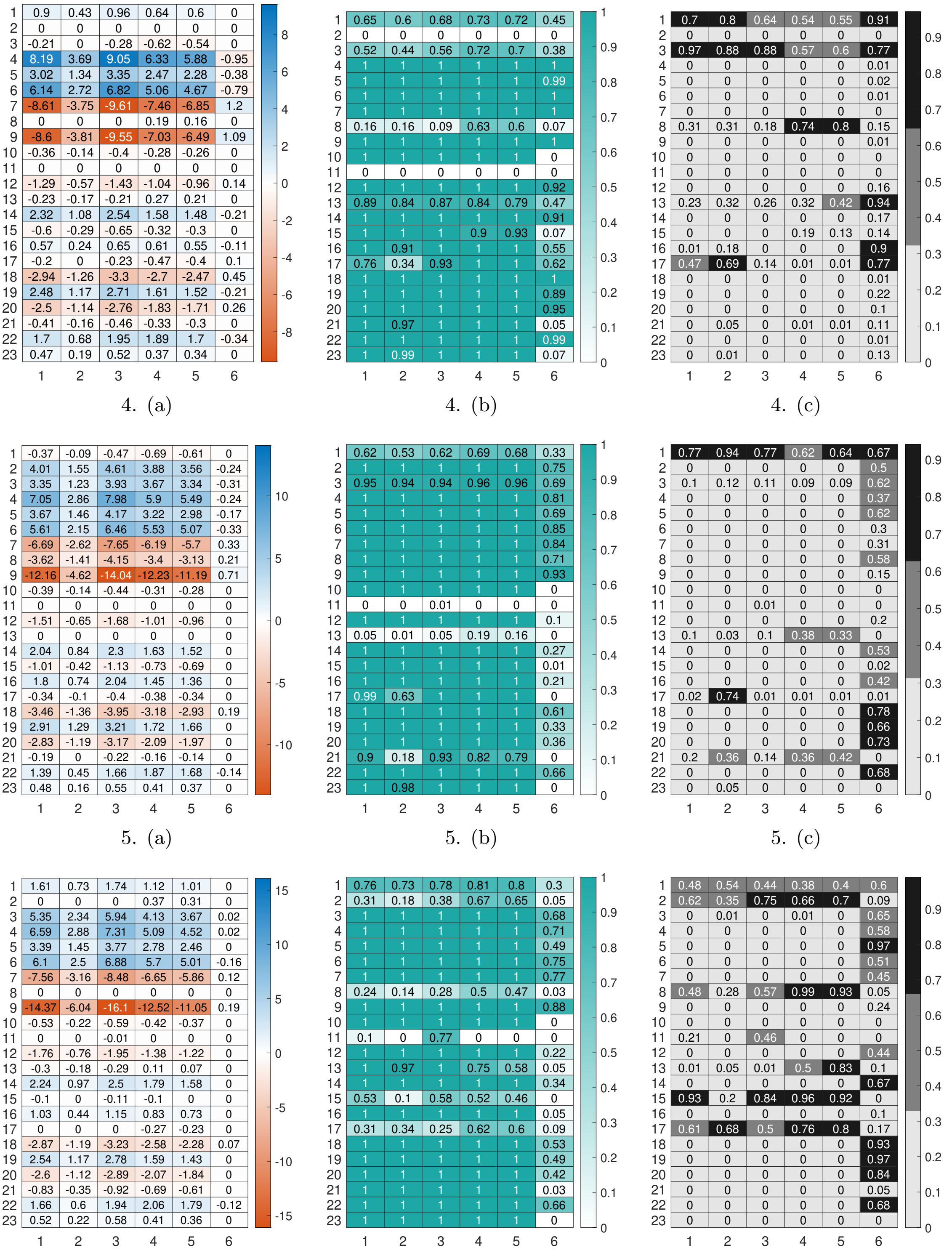
	1	2	3	4	5	6
1	1	1	1	1	1	0.97
2	0	0	0	0	0	0
3	0.93	0.92	0.93	0.96	0.95	0.79
4	1	1	1	1	1	0.95
5	1	1	1	1	1	0.9
6	1	1	1	1	1	0.97
7	0.65	0.6	0.61	0.72	0.71	0.25
8	0.27	0.33	0.18	0.75	0.72	0
9	0.99	0.98	0.99	0.99	0.99	0.93
10	1	0.99	1	1	1	0
11	0	0	0	0	0	0
12	1	1	1	1	1	0.55
13	0.26	0.17	0.26	0.26	0.22	0
14	1	1	1	1	1	0.7
15	0.66	0.1	0.71	0.64	0.58	0
16	1	1	1	1	1	0.54
17	0.07	0.01	0.08	0.12	0.1	0
18	1	1	1	1	1	0.93
19	1	1	1	1	1	0.5
20	1	1	1	1	1	0.86
21	1	1	1	1	1	0.17
22	1	0.98	1	1	1	0.94
23	1	1	1	1	1	0.09

3. (b)

	1	2	3	4	5	6
1	0	0	0	0	0	0.07
2	0	0	0	0	0	0
3	0.13	0.17	0.14	0.08	0.09	0.41
4	0	0	0	0	0	0.1
5	0	0	0	0	0	0.19
6	0	0	0	0	0	0.07
7	0.7	0.8	0.78	0.55	0.58	0.5
8	0.53	0.66	0.36	0.51	0.56	0
9	0.03	0.04	0.03	0.01	0.01	0.14
10	0	0.01	0	0	0	0
11	0	0	0.01	0	0	0
12	0	0	0	0	0	0.89
13	0.51	0.35	0.52	0.52	0.44	0
14	0	0	0	0	0	0.6
15	0.67	0.2	0.58	0.72	0.84	0
16	0	0	0	0	0	0.93
17	0.13	0.03	0.16	0.25	0.2	0
18	0	0	0	0	0	0.15
19	0	0	0	0	0	0.99
20	0	0	0	0	0	0.29
21	0	0	0	0	0	0.33
22	0	0.03	0	0	0	0.11
23	0	0	0	0	0	0.17

3. (c)

Figure A.8: The methodology was applied six times to the dataset with $n = 1500$. The rows correspond to each run (refer to Chapter 2 for the results of the first run), where we plot the sparse estimates $\hat{\mathbf{C}}$ of the coefficient matrix $\mathbf{C}$ (panel a), the uncertainty about this estimation through the posterior inclusion probabilities (panel b), and the PIP uncertainty index, in a grey-colour scale according to low ($\zeta_{jk} \leq 1/3$), medium ($1/3 < \zeta_{jk} \leq 2/3$), or high ($\zeta_{jk} > 2/3$) uncertainty (panel c).



4. (a) 4. (b) 4. (c)

5. (a) 5. (b) 5. (c)

6. (a) 6. (b) 6. (c)

Figure A.8: The methodology was applied six times to the dataset with $n = 1500$. The rows correspond to each run (refer to Chapter 2 for the results of the first run), where we plot the sparse estimates $\hat{\mathbf{C}}$ of the coefficient matrix $\mathbf{C}$ (panel a), the uncertainty about this estimation through the posterior inclusion probabilities (panel b), and the PIP uncertainty index, in a grey-colour scale according to low ($\zeta_{jk} \leq 1/3$), medium ($1/3 < \zeta_{jk} \leq 2/3$), or high ($\zeta_{jk} > 2/3$) uncertainty (panel c) (cont.).

A.4.2 Results for alternative variable selection methods

Concerning variable selection, we propose the RI and suggest a rule-of-thumb based on two tuning parameters for which we provide a concrete interpretation. As an alternative, we have implemented the group lasso method used by Chakraborty et al. (2019) to obtain an indicator for each variable and MCMC iteration. This information was then used to define the quantity $\mathrm{PIP}^*_j \in [0,1]$, which reports the posterior probability of including covariate j. We have also introduced another measure, PIP^{**}_j, which is based on the entry-wise probabilities PIP_{ij}.

Chakraborty et al. (2019) use a version of the SAVS to solve an optimisation problem inducing row sparsity via a group lasso penalty. Specifically, the jth (sparsified) row of the coefficient matrix is obtained as

$$\hat{\mathbf{C}}_R^{(j)} = (\mathbf{X}_j'\mathbf{X}_j)^{-1}\Big(1 - \frac{\mu_j}{2\|\mathbf{X}_j'\mathbf{R}_j\|}\Big)_+ \mathbf{X}_j'\mathbf{R}_j,$$

with $\mathbf{R}_j = \mathbf{X}\mathbf{C} - \sum_{i\neq j}\mathbf{X}_i\hat{\mathbf{C}}_R^{(i)}$ and the weight $\mu_j = \|\mathbf{C}_j\|^{-2}$. By computing the row-sparsified matrix $\hat{\mathbf{C}}_R$ at each iteration, we can define the posterior inclusion probability of each row $j = 1,\dots,p$ as

$$\mathrm{PIP}^*_j = 1 - \frac{1}{M}\sum_{m=1}^{M}\mathbb{I}(\hat{\mathbf{C}}_R^{(j)} = \mathbf{0}),$$

then define the associated uncertainty index, ζ_j, in a way analogous to the scalar ζ_{jk}.

An alternative way to define the PIP row-wise in the context of our proposed method is as follows

$$\begin{aligned}\mathrm{PIP}^{**}_j &= 1 - \frac{1}{M}\sum_{m=1}^{M}\mathbb{I}\big(\bar{C}_{j1}^{(m)} = 0 \wedge \ldots \wedge \bar{C}_{jq}^{(m)} = 0\big)\\ &= \frac{1}{M}\sum_{m=1}^{M}\mathbb{I}\big(\bar{C}_{j1}^{(m)} \neq 0 \vee \ldots \vee \bar{C}_{jq}^{(m)} \neq 0\big).\end{aligned}$$

This would produce potentially a conservative estimate of the number of zero rows as it suffices to have one non-zero coefficient in order to include a covariate. The problem might be exacerbated when q is big and just one coefficient is non-zero, as that would imply the inclusion of the covariate even though $q-1$ response variables are unaffected by it.

Table A.9 presents the application of the row-wise methods for uncertainty quantification about variable inclusion discussed above. The galaxy dataset of Section 2.5.2 comprises 23 covariates, and the relevance of each one is assessed by PIP^* and PIP^{**}. We report as well the associated uncertainty indices ζ^*_j and ζ^{**}_j, computed in the same fashion as ζ_{ij} for PIP_{ij}.

In Section 2.5.2, we apply our *rule of thumb* based on the RI to the galaxy dataset. Under the choice $\overline{sr} = 0.60$ (share of response variables on which the jth covariate is required to have a significant impact) and $\overline{p} = 0.40$ (minimum probability with which this occurs), we exclude covariates 1, 11, 12, 15, 17, 22 and 23 from the model. A decision about variable selection based on PIP^*_j and PIP^{**}_j can be achieved by choosing a threshold for these values. For instance, a minimum posterior probability of inclusion of 0.5 is a natural choice. In this case PIP^*_j agrees on the exclusion of covariates 17 and 23, while PIP^{**}_j points towards excluding covariates 11, 12, 17, and 23. Covariates 1 and 15 exhibit a high uncertainty under our RI approach (see Table 2.4), which is equally observed in Table A.9 above. The standard deviation of RI_{22} is at a medium level, and the conclusion of excluding it from the model would not have been reached for slightly

j	$\mathbf{PIP}^*_\mathbf{j}$	ζ^*_j	$\mathbf{PIP}^{**}_\mathbf{j}$	ζ^{**}_j
1	0.7361	0.5278	0.6714	0.6571
2	0.9998	0.0004	0.9995	0.0011
3	1.0000	0.0000	1.0000	0.0000
4	1.0000	0.0000	1.0000	0.0000
5	1.0000	0.0000	1.0000	0.0000
6	1.0000	0.0000	1.0000	0.0000
7	0.9987	0.0025	0.9987	0.0025
8	0.9982	0.0036	0.9982	0.0036
9	0.9995	0.0011	0.9995	0.0011
10	0.9998	0.0004	0.9998	0.0004
11	0.9415	0.1171	0.0004	0.0007
12	0.6655	0.6689	0.3875	0.7749
13	1.0000	0.0000	1.0000	0.0000
14	0.9986	0.0029	0.9941	0.0118
15	0.7330	0.5339	0.6134	0.7731
16	0.9998	0.0004	0.9996	0.0007
17	0.4790	0.9579	0.2555	0.5110
18	1.0000	0.0000	1.0000	0.0000
19	1.0000	0.0000	1.0000	0.0000
20	1.0000	0.0000	1.0000	0.0000
21	0.9996	0.0007	0.9982	0.0036
22	0.9563	0.0874	0.9382	0.1235
23	0.3271	0.6543	0.1355	0.2711

Table A.9: Quantification of uncertainty about the inclusion of covariate $j = 1, \dots, 23$ from the galaxy dataset (Section 5.2). PIP* and PIP** values close to 1 (0) indicate strong evidence that the covariate is relevant (irrelevant). ζ^* and ζ^{**} values close to 1 (0) indicate high (low) uncertainty about including or not a covariate.

different values of $\overline{sr}$ and/or $\overline{p}$.

There is an apparent contrast between PIP$^*_{11}$ and PIP$^{**}_{11}$. The group lasso penalty of Chakraborty et al. (2019) is a conservative approach compared to our implementation of the SAVS. $\hat{\mathbf{C}}^{11}_R$ did produce zero rows in a few iterations of the MCMC, but for the most part, the estimated values of this covariate were close to zero[1]. The penalisation imposed in $\bar{\mathbf{C}}_{11}$ continuously produced zero estimates, as the otherwise unpenalised values were not considered significant due to their small magnitude.

Finally, we remark that the 0.5 threshold (as any other) may lead to different conclusions according to the criterion chosen. For instance, covariate 12 should be included by PIP*, but excluded according to PIP**. This evidence is motivated by the high uncertainty in the posterior distributions of the coefficients associated to this covariate. However, we believe this uncertainty is hardly captured by the scalar-valued PIP* and PIP**, whereas the proposed RI is a better suited tool to explain and investigate this situation, also emphasising that in such non-extreme situations, the final decision to include/exclude a covariate is potentially highly subjective.

A.4.3 Comparison to other methods

We study a forecasting exercise where we predict a subset of n^* observations of the sub-sample with $n = 500$ examined in Section 5.2. Then, we compute the mean squared error (MSE) of the fitted values, $\hat{\mathbf{y}}_i$, as $\sum_{i=1}^{n^*}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2/q$, and the mean absolute error (MAE) as $\sum_{i=1}^{n^*}|\mathbf{y}_i - \hat{\mathbf{y}}_i|/q$. We perform this procedure by applying our method and the frequentist approaches of Chen et al. (2013) (ANN) and She and Chen (2017) (RRRR). We outperform the latter in most cases in terms of the MSE, with the remaining metrics exhibiting comparable values, albeit with slight differences. Table A.10 provides a summary of the forecast results using an expanding window and a rolling window.

[1] The norm of the average $\hat{\mathbf{C}}^{11}_R$ across iterations is 0.049. For comparison, the next smallest norm value is $\|\hat{\mathbf{C}}^{23}_R\| = 0.134$, and the maximum is $\|\hat{\mathbf{C}}^{3}_R\| = 53.428$.

Forecasting window	**n***	**MSE** **ANN**	 **RRRR**	 **RRcs**	**MAE** **ANN**	 **RRRR**	 **RRcs**
Expanding	400	0.921	1.311	1.085	0.616	0.634	0.682
Expanding	250	0.660	1.925	0.835	0.553	0.585	0.647
Expanding	100	0.482	0.416	0.534	0.534	0.467	0.560
Rolling	250	0.576	1.506	0.726	0.550	0.606	0.632

Table A.10: Mean squared error (MSE) and mean absolute error (MAE) of the true responses versus the fitted values through a forecast with expanding or rolling window, where n^* represents the number of out-of-sample points. The table reports the average errors across the predicted observations for each of the three methods, ANN, RRRR, and RRcs.

Appendix B

Additional material for Chapter 3

B.1 Details on posterior full conditionals

B.1.1 Sample $\mathbf{C}_2$

The full-rank coefficient matrix is sampled in vectorised form, denoted $\boldsymbol{\delta} = \mathrm{vec}(\mathbf{C}_2)$. For ease of notation, let $\tilde{\mathbf{y}}_1 = \mathbf{y} - \mathbf{U}_1\mathbf{c}_{1*}$ in the derivations below, with $(\mathbf{y}, \mathbf{U}_1, \tilde{\boldsymbol{\Sigma}})$ as defined in Eq. (3.5), and $\mathbf{c}_{1*} = \mathrm{vec}(\mathbf{C}_{1*})$. The matrix $\mathbf{C}_{1*} \in \mathbb{R}^{p\times q_\gamma}$ is obtained by the procedure in Section 3.3, where we take the first q_γ columns of matrix $\mathbf{C} \in \mathbb{R}^{p\times q}$ to obtain a $\mathbf{C}_1$ of consistent dimensions. The posterior of $\boldsymbol{\delta}$ is

$$\begin{aligned} p(\boldsymbol{\delta} \mid \mathbf{A}_*, \mathbf{B}_*, \mathbf{y}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}) &\propto p(\boldsymbol{\delta} \mid \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}, \boldsymbol{\gamma})\, p(\mathbf{y} \mid \boldsymbol{\delta}, \mathbf{A}_*, \mathbf{B}_*, \boldsymbol{\Sigma}) \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\delta}'\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\delta}\right\}\exp\left\{-\frac{1}{2}(\tilde{\mathbf{y}}_1 - \mathbf{U}_2\boldsymbol{\delta})'\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{y}}_1 - \mathbf{U}_2\boldsymbol{\delta})\right\}. \end{aligned}$$

Let $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}} = (\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}^{-1} + \mathbf{U}_2'\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{U}_2)^{-1}$, then the above expression becomes

$$\begin{aligned} p(\boldsymbol{\delta} \mid \mathbf{A}_*, \mathbf{B}_*, \mathbf{y}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}) &\propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\delta}'\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\delta} - 2\boldsymbol{\delta}'\mathbf{U}_2'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{y}}_1 + \tilde{\mathbf{y}}_1'\tilde{\boldsymbol{\Sigma}}^{-1}\mathbf{U}_2\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}\mathbf{U}_2'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{y}}_1\right]\right\} \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{\delta} - \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}\mathbf{U}_2'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{y}}_1)'\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}^{-1}(\boldsymbol{\delta} - \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}\mathbf{U}_2'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{y}}_1)\right\}. \end{aligned}$$

Hence, $\boldsymbol{\delta} \mid \mathbf{A}, \mathbf{B}, \mathbf{y}, \boldsymbol{\Sigma}, \boldsymbol{\gamma} \sim \mathcal{N}_{p(q-q_\gamma)}(\overline{\boldsymbol{\mu}}_{\boldsymbol{\delta}}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}})$, where $\overline{\boldsymbol{\mu}}_{\boldsymbol{\delta}} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\delta}}\mathbf{U}_2'\tilde{\boldsymbol{\Sigma}}^{-1}\tilde{\mathbf{y}}_1$.

B.1.2 Sample A

The update of $\mathbf{A} = [\mathbf{I}_r, \mathbf{F}']'$ is done by sampling $\boldsymbol{\alpha}_{\mathbf{F}} = \mathrm{vec}(\mathbf{F}')$, conditional on $(\mathbf{Y}, \boldsymbol{\gamma}, r, \boldsymbol{\Sigma}, \mathbf{B}_*, \mathbf{C}_2)$. Let $\tilde{\mathbf{y}}_2 = \mathbf{y} - \mathbf{U}_2\boldsymbol{\delta}$ in the derivations below, with $(\mathbf{y}, \mathbf{U}_2, \boldsymbol{\delta}, \tilde{\boldsymbol{\Sigma}})$ as defined in Eq. (3.5), $\boldsymbol{\alpha} = \mathrm{vec}(\mathbf{A}')$, and $\mathbf{M}_{\boldsymbol{\alpha}} = \mathbf{U}_1(\mathbf{I}_{q_\gamma} \otimes \mathbf{B}_*)$. Then, the mean of $\tilde{\mathbf{y}}_2$ is $\mathbf{M}_{\boldsymbol{\alpha}}\boldsymbol{\alpha}$, since $\mathbf{c}_1 = \mathrm{vec}(\mathbf{C}_1) = \mathrm{vec}(\mathbf{B}\mathbf{A}') = (\mathbf{I}_{q_\gamma} \otimes \mathbf{B})\,\mathrm{vec}(\mathbf{A}')$. Therefore

$$\begin{aligned} p(\boldsymbol{\alpha}_{\mathbf{F}} \mid \mathbf{y}, \boldsymbol{\gamma}, r, \boldsymbol{\Sigma}, \mathbf{B}_*, \mathbf{C}_2) &\propto p(\boldsymbol{\alpha}_{\mathbf{F}} \mid \boldsymbol{\gamma}, r)\, p(\mathbf{y} \mid \mathbf{A}, \mathbf{B}_*, \mathbf{C}_2, \boldsymbol{\Sigma}) \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}_{\mathbf{F}}'\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\alpha}_{\mathbf{F}}\right\}\exp\left\{-\frac{1}{2}(\tilde{\mathbf{y}}_2 - \mathbf{M}_{\boldsymbol{\alpha}}\boldsymbol{\alpha})'\tilde{\boldsymbol{\Sigma}}^{-1}(\tilde{\mathbf{y}}_2 - \mathbf{M}_{\boldsymbol{\alpha}}\boldsymbol{\alpha})\right\} \\ &\propto \exp\left\{-\frac{1}{2}\boldsymbol{\alpha}_{\mathbf{F}}'\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{-1}\boldsymbol{\alpha}_{\mathbf{F}}\right\}\exp\left\{-\frac{1}{2}\left[-2\boldsymbol{\alpha}'\mathbf{m} + \boldsymbol{\alpha}'\mathbf{H}\boldsymbol{\alpha}\right]\right\}, \end{aligned}$$

where $\mathbf{m} = \mathbf{M}_{\boldsymbol{\alpha}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{y}}_2$, and $\mathbf{H} = \mathbf{M}_{\boldsymbol{\alpha}}' \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{M}_{\boldsymbol{\alpha}}$. Denote by J the index set $\left\{r^2+1, r^2+2, \dots, q_\gamma r\right\}$ indicating the part of $\boldsymbol{\alpha}_{\mathbf{F}}$ within $\boldsymbol{\alpha}$. Note that J has $(q_\gamma - r)r$ elements, while its complement, $\bar{J} = \left\{1, 2, \dots, r^2\right\}$, has r^2 elements. We define $\mathbf{v} = \mathrm{vec}(\mathbf{I}_r)$, and obtain an expression proportional to the posterior distribution of $\boldsymbol{\alpha}_{\mathbf{F}}$ as follows:

$$\begin{aligned} p(\boldsymbol{\alpha}_{\mathbf{F}} \mid \mathbf{y}, \boldsymbol{\gamma}, r, \boldsymbol{\Sigma}, \mathbf{B}_*, \mathbf{C}_2) &\propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\alpha}_{\mathbf{F}}' \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{\alpha}_{\mathbf{F}} + \boldsymbol{\alpha}_{\mathbf{F}}' \mathbf{H}_{[J,J]} \boldsymbol{\alpha}_{\mathbf{F}} - 2\boldsymbol{\alpha}_{\mathbf{F}}' \left(\mathbf{m}_J - \mathbf{H}_{[J,\bar{J}]} \mathbf{v} \right) \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} (\boldsymbol{\alpha}_{\mathbf{F}} - \overline{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}) \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{-1} (\boldsymbol{\alpha}_{\mathbf{F}} - \overline{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}) \right\}, \end{aligned}$$

where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}} = \left(\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}}^{-1} + \mathbf{H}_{[J,J]} \right)^{-1}$, and $\overline{\boldsymbol{\mu}}_{\boldsymbol{\alpha}} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}} \left(\mathbf{m}_J - \mathbf{H}_{[J,\bar{J}]} \mathbf{v} \right)$. Then, $\boldsymbol{\alpha}_{\mathbf{F}} \mid \mathbf{y}, \boldsymbol{\gamma}, r, \boldsymbol{\Sigma}, \mathbf{B}, \mathbf{C}_2 \sim \mathcal{N}_{(q_\gamma - r)r}(\overline{\boldsymbol{\mu}}_{\boldsymbol{\alpha}}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\alpha}})$.

B.1.3 Sample B

The update of $\mathbf{B}$ is done by sampling $\boldsymbol{\beta} = \mathrm{vec}(\mathbf{B})$, conditional on $(\mathbf{Y}, \boldsymbol{\gamma}, r, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{C}_2, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}})$. Denoting $\mathbf{M}_{\boldsymbol{\beta}} = \mathbf{U}_1(\mathbf{A} \otimes \mathbf{I}_p)$, and exploiting the properties of vectorisation and the Kronecker product, an alternative representation of the mean of $\tilde{\mathbf{y}}_2$ is $\mathbf{M}_{\boldsymbol{\beta}} \boldsymbol{\beta}$. Therefore

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\gamma}, r, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{C}_2, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}) &\propto p(\boldsymbol{\beta} \mid \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}, \boldsymbol{\gamma}, r)\, p(\mathbf{y} \mid \mathbf{A}, \mathbf{B}, \boldsymbol{\Sigma}) \\ &\propto \exp\left\{ -\frac{1}{2} \boldsymbol{\beta}' \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta} \right\} \exp\left\{ -\frac{1}{2} (\tilde{\mathbf{y}}_2 - \mathbf{M}_{\boldsymbol{\beta}} \boldsymbol{\beta})' \tilde{\boldsymbol{\Sigma}}^{-1} (\tilde{\mathbf{y}}_2 - \mathbf{M}_{\boldsymbol{\beta}} \boldsymbol{\beta}) \right\}. \end{aligned}$$

Let $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = (\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{-1} + \mathbf{M}_{\boldsymbol{\beta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{M}_{\boldsymbol{\beta}})^{-1}$, then the above expression becomes

$$\begin{aligned} p(\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\gamma}, r, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{C}_2, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}) &\propto \exp\left\{ -\frac{1}{2} \left[\boldsymbol{\beta}' \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} \boldsymbol{\beta} - 2\boldsymbol{\beta}' \mathbf{M}_{\boldsymbol{\beta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{y}}_2 + \tilde{\mathbf{y}}_2' \tilde{\boldsymbol{\Sigma}}^{-1} \mathbf{M}_{\boldsymbol{\beta}} \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} \mathbf{M}_{\boldsymbol{\beta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{y}}_2 \right] \right\} \\ &\propto \exp\left\{ -\frac{1}{2} (\boldsymbol{\beta} - \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} \mathbf{M}_{\boldsymbol{\beta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{y}}_2)' \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} \mathbf{M}_{\boldsymbol{\beta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{y}}_2) \right\}. \end{aligned}$$

Then, $\boldsymbol{\beta} \mid \mathbf{y}, \boldsymbol{\gamma}, r, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{C}_2, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} \sim \mathcal{N}_{pr}(\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}})$, where $\overline{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = \overline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} \mathbf{M}_{\boldsymbol{\beta}}' \tilde{\boldsymbol{\Sigma}}^{-1} \tilde{\mathbf{y}}_2$.

B.1.4 Sample $\boldsymbol{\Sigma}$

The error covariance matrix, conditional on $(\mathbf{C} = \left[\mathbf{C}_1, \mathbf{C}_2\right], \mathbf{Y})$, is sampled from $\mathcal{IW}_q(\overline{\nu}, \overline{\boldsymbol{\Psi}})$, where $\overline{\nu} = \underline{\nu} + n$ and $\overline{\boldsymbol{\Psi}} = \underline{\boldsymbol{\Psi}} + (\mathbf{Y} - \mathbf{XC})'(\mathbf{Y} - \mathbf{XC})$. In fact,

$$\begin{aligned} &p(\boldsymbol{\Sigma} \mid \mathbf{C}, \mathbf{Y}) \propto p(\boldsymbol{\Sigma} \mid \underline{\nu}, \underline{\boldsymbol{\Psi}})\, p(\mathbf{Y} \mid \mathbf{C}, \boldsymbol{\Sigma}) \\ &\quad \propto \frac{|\underline{\boldsymbol{\Psi}}|^{\frac{\underline{\nu}}{2}}}{2^{\frac{\underline{\nu} q}{2}} \Gamma_q(\frac{\underline{\nu}}{2}) |\boldsymbol{\Sigma}|^{\frac{\underline{\nu}+q+1}{2}}} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left[\underline{\boldsymbol{\Psi}} \boldsymbol{\Sigma}^{-1} \right] \right\} |\boldsymbol{\Sigma}|^{-n/2} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \mathbf{XC})'(\mathbf{Y} - \mathbf{XC}) \right] \right\} \\ &\quad \propto |\boldsymbol{\Sigma}|^{-(\underline{\nu}+n+q+1)/2} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left[\underline{\boldsymbol{\Psi}} \boldsymbol{\Sigma}^{-1} + (\mathbf{Y} - \mathbf{XC})'(\mathbf{Y} - \mathbf{XC}) \boldsymbol{\Sigma}^{-1} \right] \right\} \\ &\quad \propto |\boldsymbol{\Sigma}|^{-(\underline{\nu}+n+q+1)/2} \exp\left\{ -\frac{1}{2} \operatorname{tr}\left[(\underline{\boldsymbol{\Psi}} + (\mathbf{Y} - \mathbf{XC})'(\mathbf{Y} - \mathbf{XC})) \boldsymbol{\Sigma}^{-1} \right] \right\}, \end{aligned}$$

which is proportional to the kernel of another inverse Wishart distribution, with the previously mentioned parameters, due to the conjugate nature of the prior.

B.1.5 Sample ρ

To obtain a draw of ρ, the probability of success of the allocation variable $\boldsymbol{\gamma}$, we sample ρ from its posterior distribution given by

$$\begin{aligned} p(\rho \mid \gamma) &\propto p(\rho)\, p(\gamma \mid \rho_i) \\ &\propto \frac{1}{\mathrm{B}(\underline{a}_\rho, \underline{b}_\rho)} \rho^{\underline{a}_\rho - 1} (1-\rho)^{\underline{b}_\rho - 1} \prod_{i=1}^{n} \rho^{\gamma_i} (1-\rho)^{1-\gamma_i} \\ &\propto \rho^{\underline{a}_\rho + q_\gamma - 1} (1-\rho)^{\underline{b}_\rho - q_\gamma}, \end{aligned}$$

where $\mathrm{B}(\cdot,\cdot)$ represents the Beta function. Therefore, the posterior distribution of ρ is Beta with parameters $\overline{a}_\rho = \underline{a}_\rho + q_\gamma$, and $\overline{b}_\rho = \underline{b}_\rho + q - q_\gamma$.

B.2 Derivation of maximum likelihood estimators

Section 3.3.2 describes the update of $(\boldsymbol{\gamma}, r)$, which involves the computation of the maximum likelihood estimator of matrix $\mathbf{C}_1 = \mathbf{B}\mathbf{A}'$ in the Laplace approximation in Eq. (3.8). Noticing that the marginalised likelihood in Eq. (3.6) not only incorporates heteroscedastic errors but also imposes structure restrictions on the model, the standard MLE techniques described in Reinsel et al. (2022) are not suitable.

Hansen (2002) provides the parameters MLEs in a generalized reduced rank regression framework (GRRR), accommodating our necessities in this estimation procedure.

Departing from Eq. (3.6), the GRRR problem considers the regression

$$\mathbf{y}_i = \mathbf{V}_1' \mathbf{A} \mathbf{B}' \mathbf{x}_i + \tilde{\mathbf{e}}_i,$$

where $\tilde{\mathbf{e}}_i$ is the ith column of $\tilde{\mathbf{E}} \in \mathbb{R}^{q \times n}$, $\mathrm{vec}(\tilde{\mathbf{E}}) \sim \mathcal{N}_{nq}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{y}})$ and

$$\begin{aligned} \boldsymbol{\alpha}_{\mathbf{V}_1} &= \mathrm{vec}(\mathbf{V}_1' \mathbf{A}) = \mathbf{G}\psi + \mathbf{g}, \\ \boldsymbol{\beta} &= \mathrm{vec}(\mathbf{B}) = \mathbf{H}\varphi + \mathbf{h}, \end{aligned}$$

for known binary matrices $\mathbf{G}$ and $\mathbf{H}$ and known real vectors $\mathbf{g}$ and $\mathbf{h}$. In our setting, the vector $\boldsymbol{\alpha}_{\mathbf{V}_1}$ is subject to an identification restriction that corresponds to setting some of its entries to either 0 or 1. The binary matrix $\mathbf{G}$ and the vector $\mathbf{g}$ serve to impose these values in the appropriate positions within $\boldsymbol{\alpha}_{\mathbf{V}_1}$, while $\psi = \mathbf{f} = \mathrm{vec}(\mathbf{F})$. Conversely, since we assume no constraints on the matrix $\mathbf{B}$, then $\mathbf{H} = \mathbf{I}_{pr}$ is the identity matrix and $\mathbf{h} = \mathbf{0}_{pr}$ is a null vector. Therefore, the total number of restrictions is $(qr - q_\gamma r + r^2) = r(r + q - q_\gamma)$, and the number of free parameters is the length of $\mathbf{f}$, that is $q_\gamma r - r^2$.

In more detail, $\mathbf{g}$ is a qr-dimensional vector with ones in the index set $\{(k-1)(q+1) + 1, \text{ for each } k = 1, 2, \ldots, r\}$, and zeros in the remaining entries. The $qr \times r(q_\gamma - r)$ matrix $\mathbf{G}$ is a block matrix defined as

$$\begin{aligned} \mathbf{G} &= [\mathbf{G}_1, \mathbf{G}_2, \ldots, \mathbf{G}_r], \\ \mathbf{G}_k &= [\mathbf{0}_{(q_\gamma - r) \times (q(k-1)+r)}, \mathbf{I}_{q_\gamma - r}, \mathbf{0}_{(q_\gamma - r) \times (q(r-k+1) - q_\gamma)}]', \qquad k = 1, 2, \ldots, r. \end{aligned}$$

Example 3. *Consider our PRR model with $q = 4$, $q_\gamma = 3$, $r = 2$. Then, we have*

$$\begin{aligned}
\boldsymbol{\alpha}_{\mathbf{V}_1} &= \underbrace{\mathbf{G}}_{qr\times(q_\gamma r - r^2)} \times \underbrace{\mathbf{f}}_{(q_\gamma r - r^2)\times 1} + \underbrace{\mathbf{g}}_{qr\times 1} \\
&= \underbrace{\mathbf{G}}_{8\times 2} \times \underbrace{\mathbf{f}}_{2\times 1} + \underbrace{\mathbf{g}}_{8\times 1}
\end{aligned}$$

$$\begin{pmatrix} 1.0 \\ 0 \\ f_1 \\ 0 \\ 0 \\ 1.0 \\ f_2 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \times \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} + \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}.$$

Recalling that $\mathbf{y} = \text{vec}(\mathbf{Y})$, $\mathbf{U}_1 = \mathbf{V}_1' \otimes \mathbf{X}$, and $\mathbf{c}_1 = \text{vec}(\mathbf{C}_1) = \text{vec}(\mathbf{BA}')$, the GRRR minimisation problem is

$$\min_{\boldsymbol{\alpha}_{\mathbf{V}_1},\boldsymbol{\beta}} \left\|(\mathbf{y} - \mathbf{U}_1\mathbf{c}_1)'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}(\mathbf{y} - \mathbf{U}_1\mathbf{c}_1)\right\| \text{ subject to } \text{rank}(\mathbf{C}_1) = r,\ \boldsymbol{\alpha}_{\mathbf{V}_1} = \mathbf{GF} + \mathbf{g}.$$

Then, we define the matrices

$$\begin{aligned}
\mathbf{M_B} &= (\mathbf{XB} \otimes \mathbf{I}_q)'\tilde{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}(\mathbf{XB} \otimes \mathbf{I}_q), \\
\mathbf{n_B} &= (\mathbf{XB} \otimes \mathbf{I}_q)'\tilde{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\,\text{vec}(\mathbf{Y}'), \\
\mathbf{M_A} &= \mathbf{K}_{p,r}'(\mathbf{X} \otimes \mathbf{V}_1'\mathbf{A})'\tilde{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}(\mathbf{X} \otimes \mathbf{V}_1'\mathbf{A})\mathbf{K}_{p,r}, \\
\mathbf{n_A} &= \mathbf{K}_{p,r}'(\mathbf{X} \otimes \mathbf{V}_1'\mathbf{A})'\tilde{\boldsymbol{\Sigma}}_{\mathbf{y}}^{-1}\,\text{vec}(\mathbf{Y}'),
\end{aligned}$$

where $\tilde{\boldsymbol{\Sigma}}_{\mathbf{y}} = \mathbf{K}_{n,q}\boldsymbol{\Sigma}_{\mathbf{y}}\mathbf{K}_{n,q}'$, and $\mathbf{K}_{m,n}$ is the $mn \times mn$ commutation matrix, which transforms the vectorisation of a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ into the vectorisation of its transpose, such that $\mathbf{K}_{m,n}\,\text{vec}(\mathbf{M}) = \text{vec}(\mathbf{M}')$.

Finally, the maximum likelihood estimators following the GRRR method are

$$\begin{aligned}
\hat{\boldsymbol{\alpha}}_{\mathbf{V}_1} &= \mathbf{G}(\mathbf{G}'\mathbf{M_B}\mathbf{G})^{-1}\mathbf{G}'(\mathbf{n_B} - \mathbf{M_B}\mathbf{g}) + \mathbf{g}, \\
\hat{\boldsymbol{\beta}} &= \mathbf{M}_{\mathbf{A}}^{-1}\mathbf{n_A}.
\end{aligned}$$

An estimate $(\hat{\boldsymbol{\alpha}}_{\mathbf{V}_1}, \hat{\boldsymbol{\beta}})$ is obtained by employing an iterative procedure: starting from a given initial value, say $\boldsymbol{\beta}^{(0)}$, a new value $\boldsymbol{\alpha}_{\mathbf{V}_1}^{(1)}$ is computed given $\boldsymbol{\beta}^{(0)}$, then $\boldsymbol{\beta}^{(1)}$ is obtained given $\boldsymbol{\alpha}_{\mathbf{V}_1}^{(1)}$. The process is iterated until convergence.

B.3 Metropolized Shotgun Stochastic Search algorithm

The Shotgun Stochastic Search (SSS) algorithm introduced by Hans et al. (2007) is a computationally efficient method for model search in regression with a large number of predictive variables. The algorithm explores high-probability regions of the model space by examining many neighbours of each model selected, without requiring an exhaustive enumeration of all model configurations through a restricted neighbourhood. In our context, we adapt the strategy to perform selection over response variables rather than covariates.

SSS iterates over three steps: (1) define a neighbourhood of proposal models based on the current model, (2) evaluate each proposal in parallel, and (3) select a new model among the proposals. Notably, the two main aspects of SSS are the neighbourhood choice and the model move strategy.

The neighbourhood needs to be large enough to allow moves throughout the model space, which is accomplished by considering each possible variable in one of the proposals at each iteration, resulting in an evaluation of every candidate variable in the context of different regression models. Hence, the neighbourhood is defined to be a one-variable change to the current model.

Let $\boldsymbol{\gamma}$ denote the q-dimensional indicator vector with $\gamma_j = 1$ if variable j is included and $\gamma_j = 0$ otherwise. For a current model of dimension q_γ, the neighbourhood is defined as $\text{nbd}(\boldsymbol{\gamma}) = \{\boldsymbol{\gamma}^+, \boldsymbol{\gamma}^\circ, \boldsymbol{\gamma}^-\}$. The set $\boldsymbol{\gamma}^+$ contains the models of dimension $q_\gamma + 1$, called the "addition" moves, adding any one of the $q - q_\gamma$ variables. The "replacement" moves are denoted by $\boldsymbol{\gamma}^\circ$, where the neighbours replace any one current variable with any one of the $q - q_\gamma$ remaining variables, resulting in no change in the dimension of the model. $\boldsymbol{\gamma}^-$ is the "deletion" moves set, which contains the models of dimension $q_\gamma - 1$ obtained by deleting any current variable. In our framework, we allow moves only to models of different dimensions to the current model, similar to Yang et al. (2022). Consequently, our neighbourhood does not include the set $\boldsymbol{\gamma}^\circ$, which reduces the computational cost while effectively moving to areas that comprise the true model, as shown in the simulation study of Section 3.4.

The SSS algorithm can be easily adapted to become a Metropolis-Hastings algorithm within an MCMC scheme, enabling the identification of regions of high posterior probability. Let us consider a discrete distribution, $P(x)$, known up to a normalising constant, $P(x) \propto Q(x)$. We can use a Metropolis-Hastings algorithm to sample from this distribution restricted to a neighbourhood $\text{nbd}(\cdot)$. Consider the proposal distribution to generate a sample x^* at iteration $m+1$:

$$T(x^*, x^{(m)}) = \frac{P(x^*)\,\mathbb{I}(x^* \in \text{nbd}(x^{(m)}))}{\sum_{x^\dagger \in \text{nbd}(x^{(m)})} P(x^\dagger)} = \frac{Q(x^*)\,\mathbb{I}(x^* \in \text{nbd}(x^{(m)}))}{\sum_{x^\dagger \in \text{nbd}(x^{(m)})} Q(x^\dagger)},$$

where $\mathbb{I}(\cdot)$ is the indicator function. The acceptance probability at iteration $m+1$ is

$$\alpha = \min\left\{1, \frac{\sum_{x^\dagger \in \text{nbd}(x^{(m)})} Q(x^\dagger)}{\sum_{x^\dagger \in \text{nbd}(x^*)} Q(x^\dagger)}\right\}.$$

In our context, x is the allocation vector $\boldsymbol{\gamma}$. $P(x)$ corresponds to the posterior distribution of $\boldsymbol{\gamma}$, $p(\boldsymbol{\gamma} \mid \mathbf{Y}, \boldsymbol{\Sigma}, \rho)$, while $Q(x)$ is the approximate unnormalised posterior $\tilde{f}_{\boldsymbol{\gamma}}(\mathbf{Y} \mid \boldsymbol{\Sigma}, \boldsymbol{\gamma}) p(\boldsymbol{\gamma} \mid \rho)$, where $\tilde{f}_{\boldsymbol{\gamma}}$ is defined as in Eq.(3.11).

B.4 Additional empirical results

B.4.1 Supporting simulation plots

The following figures complement Figure 3.3, presenting the trace and the posterior distribution of $\boldsymbol{\gamma}$ under different simulation settings. Each figure depicts the results from a randomly selected replicate among the 100 independent replicates conducted for the specified simulation scenario.

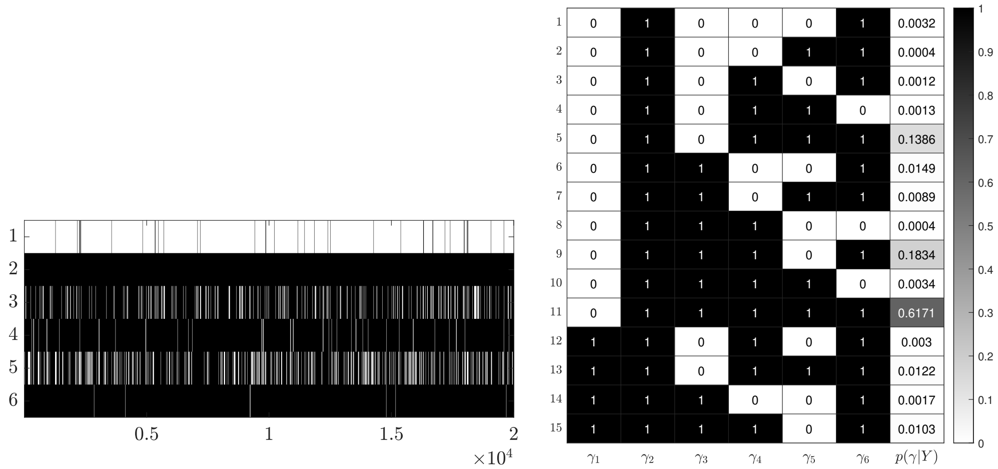


Figure B.1: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 5$, $q = 6$, $q_\gamma = 5$, $r = 2$ and $n = 50$, with true allocation vector $\boldsymbol{\gamma}_0 = (0, 1, 1, 1, 1, 1)$.

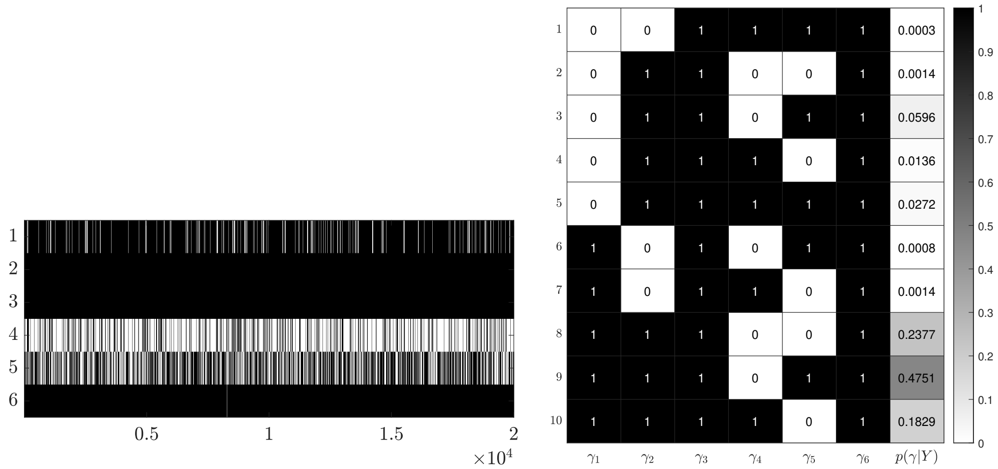
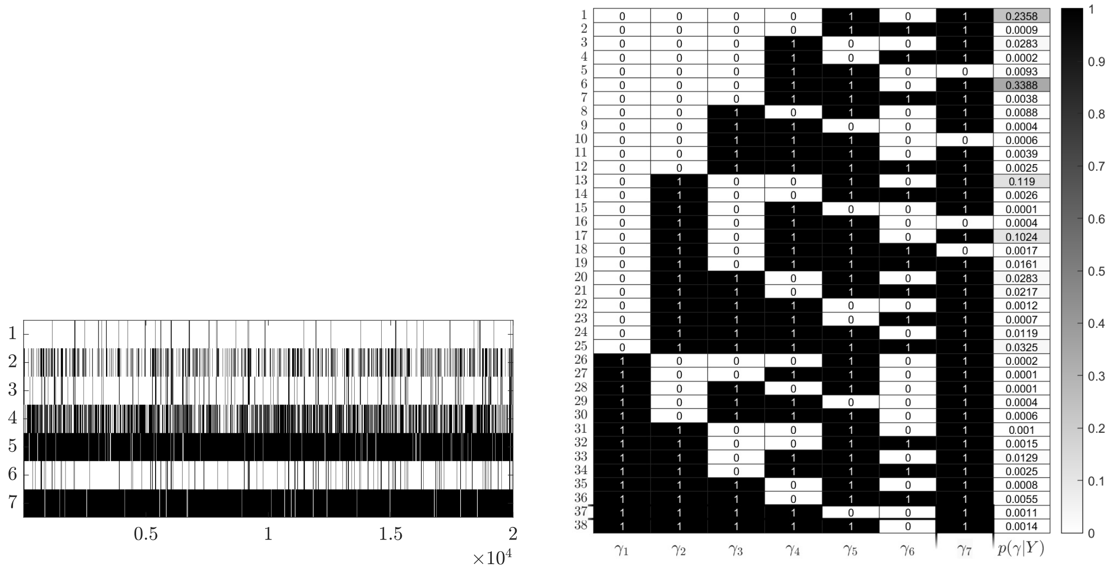
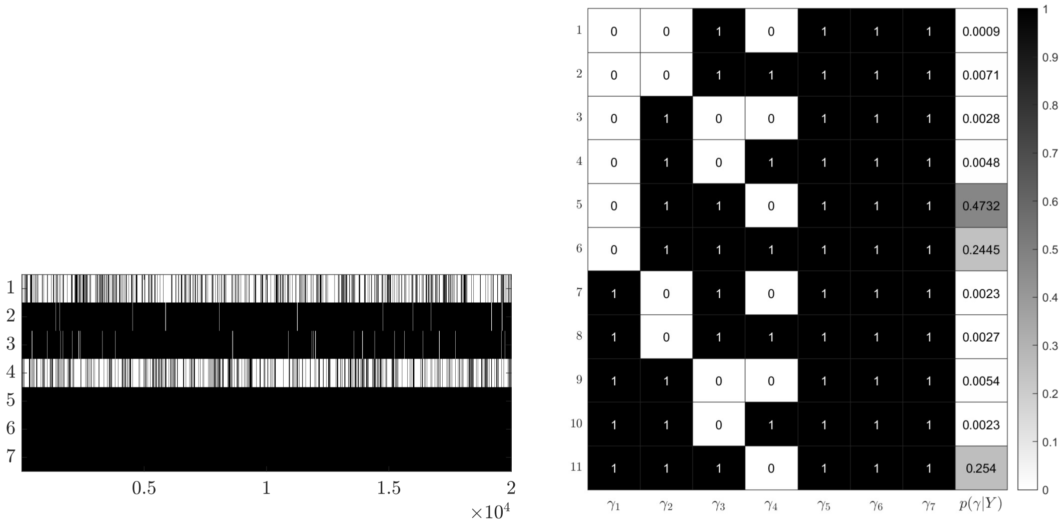


Figure B.2: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 5$, $q = 6$, $q_\gamma = 5$, $r = 2$ and $n = 100$, with true allocation vector $\boldsymbol{\gamma}_0 = (1, 1, 1, 0, 1, 1)$.



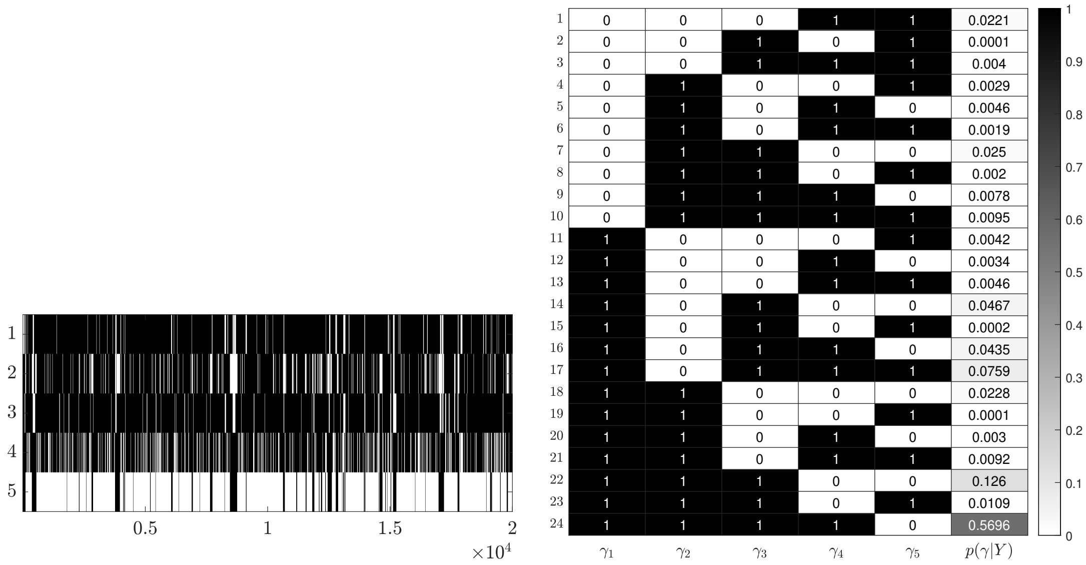
	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	$p(\gamma\|Y)$
1	0	0	0	0	1	0	1	0.2358
2	0	0	0	0	1	1	1	0.0009
3	0	0	0	1	0	0	1	0.0283
4	0	0	0	1	0	1	1	0.0002
5	0	0	0	1	1	0	0	0.0093
6	0	0	0	1	1	0	1	0.3388
7	0	0	0	1	1	1	1	0.0038
8	0	0	1	0	1	0	1	0.0088
9	0	0	1	1	0	0	1	0.0004
10	0	0	1	1	1	0	0	0.0006
11	0	0	1	1	1	0	1	0.0039
12	0	0	1	1	1	1	1	0.0025
13	0	1	0	0	1	0	1	0.119
14	0	1	0	0	1	1	1	0.0026
15	0	1	0	1	0	0	1	0.0001
16	0	1	0	1	1	0	0	0.0004
17	0	1	0	1	1	0	1	0.1024
18	0	1	0	1	1	1	0	0.0017
19	0	1	0	1	1	1	1	0.0161
20	0	1	1	0	1	0	1	0.0283
21	0	1	1	0	1	1	1	0.0217
22	0	1	1	1	0	0	1	0.0012
23	0	1	1	1	0	1	1	0.0007
24	0	1	1	1	1	0	1	0.0119
25	0	1	1	1	1	1	1	0.0325
26	1	0	0	0	1	0	1	0.0002
27	1	0	0	1	1	0	1	0.0001
28	1	0	1	0	1	0	1	0.0001
29	1	0	1	1	0	0	1	0.0004
30	1	0	1	1	1	0	1	0.0006
31	1	1	0	0	1	0	1	0.001
32	1	1	0	0	1	1	1	0.0015
33	1	1	0	1	1	0	1	0.0129
34	1	1	0	1	1	1	1	0.0025
35	1	1	1	0	1	0	1	0.0008
36	1	1	1	0	1	1	1	0.0055
37	1	1	1	1	0	0	1	0.0011
38	1	1	1	1	1	0	1	0.0014

Figure B.3: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 5$, $q = 7$, $q_\gamma = 3$, $r = 1$ and $n = 50$, with true allocation vector $\boldsymbol{\gamma}_0 = (0, 0, 0, 1, 1, 0, 1)$.



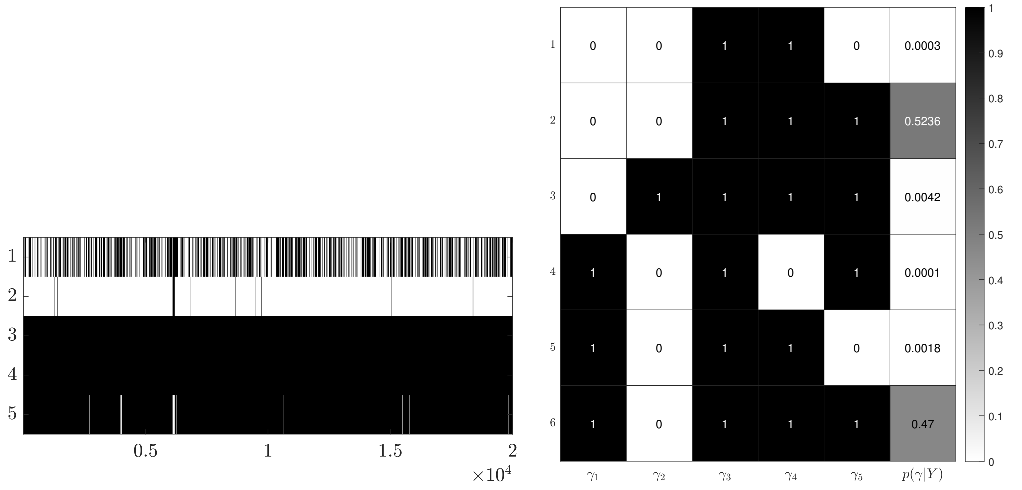
	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	$p(\gamma\|Y)$
1	0	0	1	0	1	1	1	0.0009
2	0	0	1	1	1	1	1	0.0071
3	0	1	0	0	1	1	1	0.0028
4	0	1	0	1	1	1	1	0.0048
5	0	1	1	0	1	1	1	0.4732
6	0	1	1	1	1	1	1	0.2445
7	1	0	1	0	1	1	1	0.0023
8	1	0	1	1	1	1	1	0.0027
9	1	1	0	0	1	1	1	0.0054
10	1	1	0	1	1	1	1	0.0023
11	1	1	1	0	1	1	1	0.254

Figure B.4: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 5$, $q = 7$, $q_\gamma = 5$, $r = 2$ and $n = 50$, with true allocation vector $\boldsymbol{\gamma}_0 = (0, 1, 1, 0, 1, 1, 1)$.



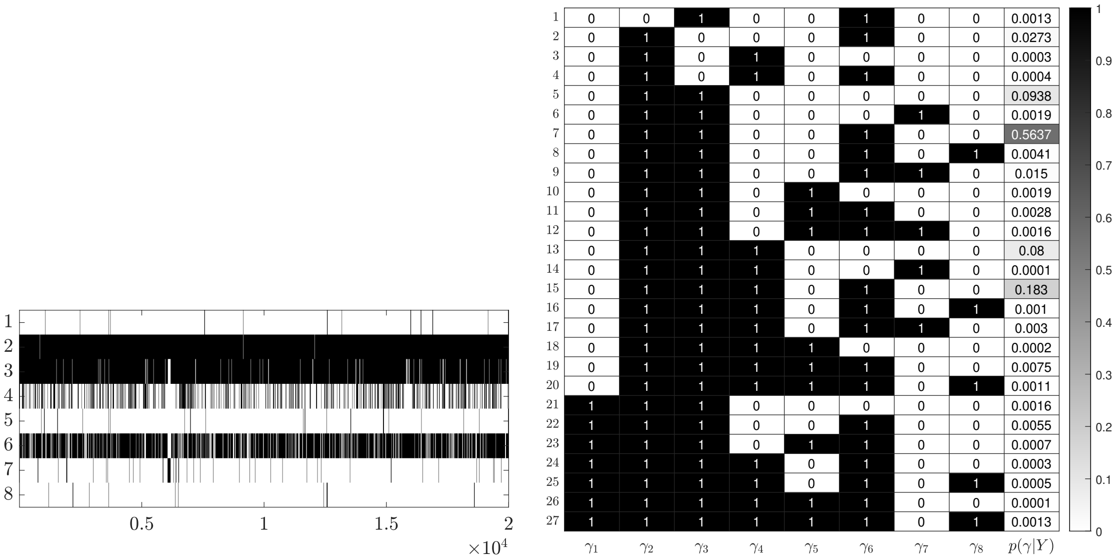
	γ_1	γ_2	γ_3	γ_4	γ_5	$p(\gamma\|Y)$
1	0	0	0	1	1	0.0221
2	0	0	1	0	1	0.0001
3	0	0	1	1	1	0.004
4	0	1	0	0	1	0.0029
5	0	1	0	1	0	0.0046
6	0	1	0	1	1	0.0019
7	0	1	1	0	0	0.025
8	0	1	1	0	1	0.002
9	0	1	1	1	0	0.0078
10	0	1	1	1	1	0.0095
11	1	0	0	0	1	0.0042
12	1	0	0	1	0	0.0034
13	1	0	0	1	1	0.0046
14	1	0	1	0	0	0.0467
15	1	0	1	0	1	0.0002
16	1	0	1	1	0	0.0435
17	1	0	1	1	1	0.0759
18	1	1	0	0	0	0.0228
19	1	1	0	0	1	0.0001
20	1	1	0	1	0	0.003
21	1	1	0	1	1	0.0092
22	1	1	1	0	0	0.126
23	1	1	1	0	1	0.0109
24	1	1	1	1	0	0.5696

Figure B.5: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 10$, $q = 5$, $q_\gamma = 3$, $r = 1$ and $n = 20$, with true allocation vector $\boldsymbol{\gamma}_0 = (1, 1, 1, 0, 0)$.



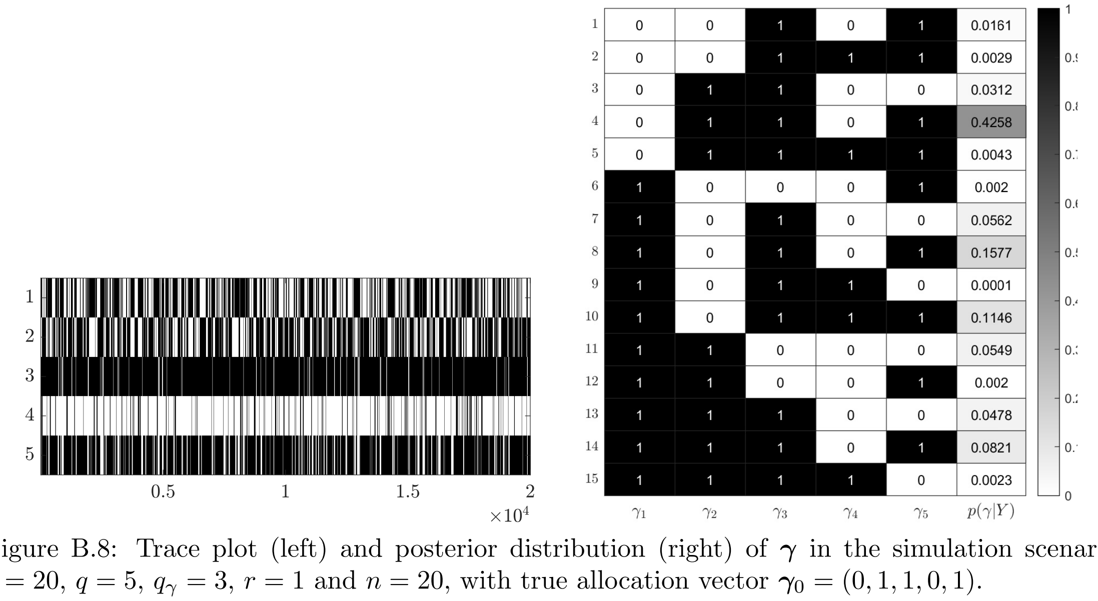
	γ_1	γ_2	γ_3	γ_4	γ_5	$p(\gamma\|Y)$
1	0	0	1	1	0	0.0003
2	0	0	1	1	1	0.5236
3	0	1	1	1	1	0.0042
4	1	0	1	0	1	0.0001
5	1	0	1	1	0	0.0018
6	1	0	1	1	1	0.47

Figure B.6: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 10$, $q = 5$, $q_\gamma = 3$, $r = 1$ and $n = 100$, with true allocation vector $\boldsymbol{\gamma}_0 = (0, 0, 1, 1, 1)$.



	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8	$p(\gamma\|Y)$
1	0	0	1	0	0	1	0	0	0.0013
2	0	1	0	0	0	1	0	0	0.0273
3	0	1	0	1	0	0	0	0	0.0003
4	0	1	0	1	0	1	0	0	0.0004
5	0	1	1	0	0	0	0	0	0.0938
6	0	1	1	0	0	0	1	0	0.0019
7	0	1	1	0	0	1	0	0	0.5637
8	0	1	1	0	0	1	0	1	0.0041
9	0	1	1	0	0	1	1	0	0.015
10	0	1	1	0	1	0	0	0	0.0019
11	0	1	1	0	1	1	0	0	0.0028
12	0	1	1	0	1	1	1	0	0.0016
13	0	1	1	1	0	0	0	0	0.08
14	0	1	1	1	0	0	1	0	0.0001
15	0	1	1	1	0	1	0	0	0.183
16	0	1	1	1	0	1	0	1	0.001
17	0	1	1	1	0	1	1	0	0.003
18	0	1	1	1	1	0	0	0	0.0002
19	0	1	1	1	1	1	0	0	0.0075
20	0	1	1	1	1	1	0	1	0.0011
21	1	1	1	0	0	0	0	0	0.0016
22	1	1	1	0	0	1	0	0	0.0055
23	1	1	1	0	1	1	0	0	0.0007
24	1	1	1	1	0	1	0	0	0.0003
25	1	1	1	1	0	1	0	1	0.0005
26	1	1	1	1	1	1	0	0	0.0001
27	1	1	1	1	1	1	0	1	0.0013

Figure B.7: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 10$, $q = 8$, $q_\gamma = 3$, $r = 1$ and $n = 20$, with true allocation vector $\boldsymbol{\gamma}_0 = (0, 1, 1, 0, 0, 1, 0, 0)$.



	γ_1	γ_2	γ_3	γ_4	γ_5	$p(\gamma\|Y)$
1	0	0	1	0	1	0.0161
2	0	0	1	1	1	0.0029
3	0	1	1	0	0	0.0312
4	0	1	1	0	1	0.4258
5	0	1	1	1	1	0.0043
6	1	0	0	0	1	0.002
7	1	0	1	0	0	0.0562
8	1	0	1	0	1	0.1577
9	1	0	1	1	0	0.0001
10	1	0	1	1	1	0.1146
11	1	1	0	0	0	0.0549
12	1	1	0	0	1	0.002
13	1	1	1	0	0	0.0478
14	1	1	1	0	1	0.0821
15	1	1	1	1	0	0.0023

Figure B.8: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 20$, $q = 5$, $q_\gamma = 3$, $r = 1$ and $n = 20$, with true allocation vector $\boldsymbol{\gamma}_0 = (0, 1, 1, 0, 1)$.

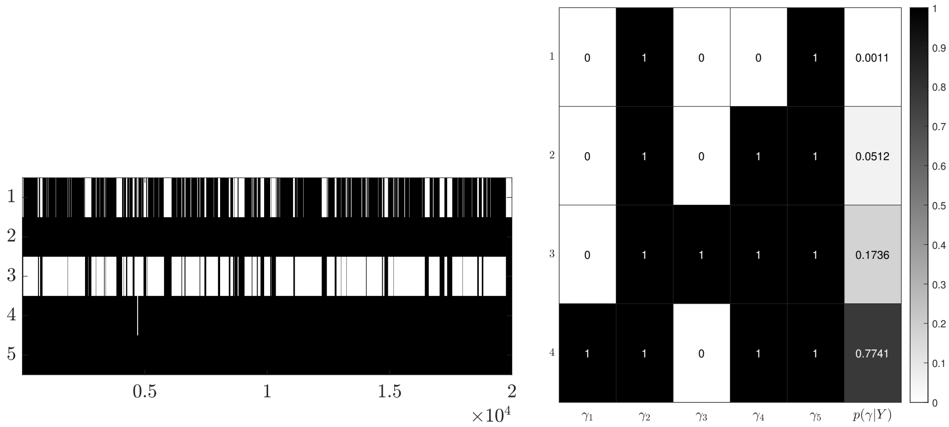


Figure B.9: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 20$, $q = 5$, $q_\gamma = 3$, $r = 1$ and $n = 50$, with true allocation vector $\boldsymbol{\gamma}_0 = (0, 1, 0, 1, 1)$.

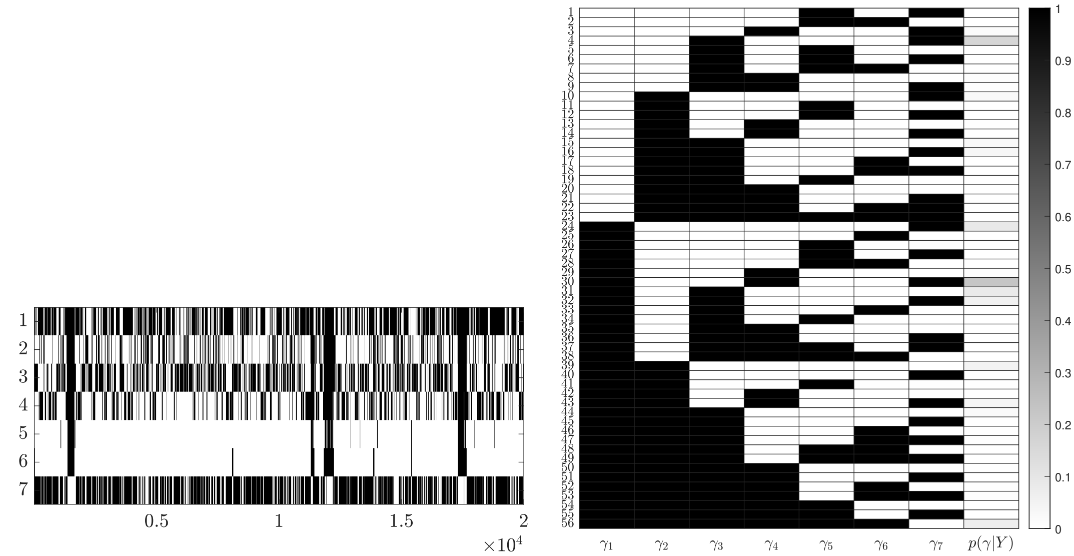
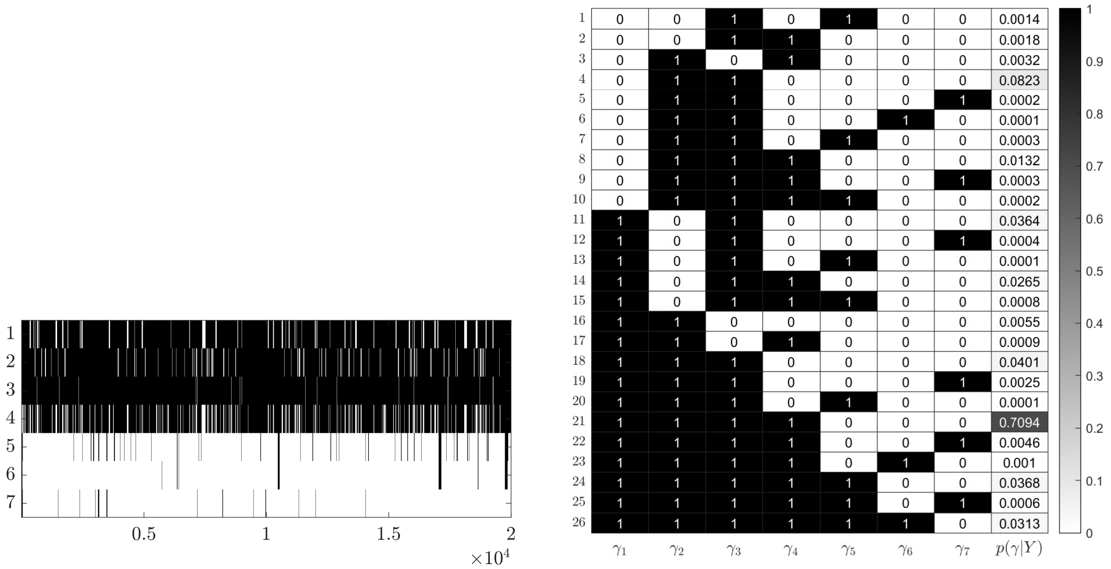
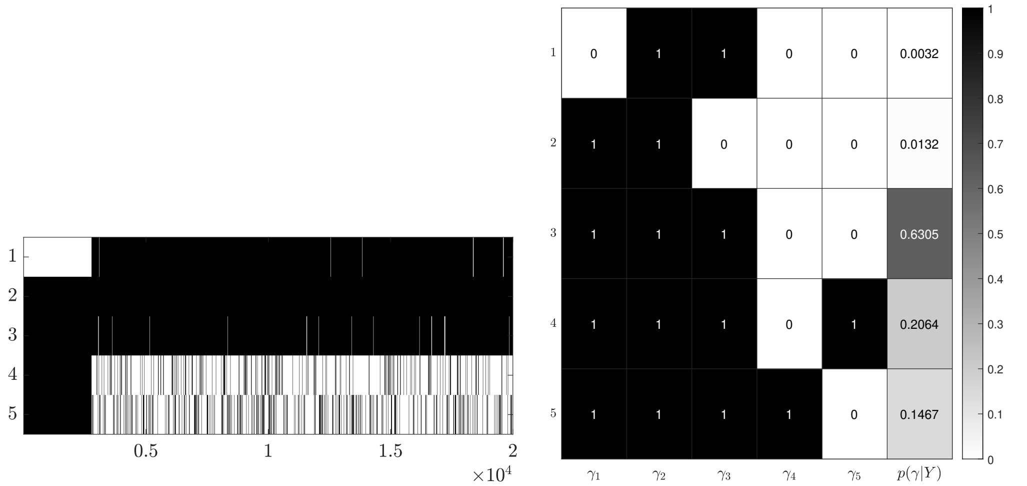


Figure B.10: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 20$, $q = 7$, $q_\gamma = 3$, $r = 1$ and $n = 20$, with true allocation vector $\boldsymbol{\gamma}_0 = (1, 0, 0, 1, 0, 0, 1)$.



	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	$p(\gamma\|Y)$
1	0	0	1	0	1	0	0	0.0014
2	0	0	1	1	0	0	0	0.0018
3	0	1	0	1	0	0	0	0.0032
4	0	1	1	0	0	0	0	0.0823
5	0	1	1	0	0	0	1	0.0002
6	0	1	1	0	0	1	0	0.0001
7	0	1	1	0	1	0	0	0.0003
8	0	1	1	1	0	0	0	0.0132
9	0	1	1	1	0	0	1	0.0003
10	0	1	1	1	1	0	0	0.0002
11	1	0	1	0	0	0	0	0.0364
12	1	0	1	0	0	0	1	0.0004
13	1	0	1	0	1	0	0	0.0001
14	1	0	1	1	0	0	0	0.0265
15	1	0	1	1	1	0	0	0.0008
16	1	1	0	0	0	0	0	0.0055
17	1	1	0	1	0	0	0	0.0009
18	1	1	1	0	0	0	0	0.0401
19	1	1	1	0	0	0	1	0.0025
20	1	1	1	0	1	0	0	0.0001
21	1	1	1	1	0	0	0	0.7094
22	1	1	1	1	0	0	1	0.0046
23	1	1	1	1	0	1	0	0.001
24	1	1	1	1	1	0	0	0.0368
25	1	1	1	1	1	0	1	0.0006
26	1	1	1	1	1	1	0	0.0313

Figure B.11: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 20$, $q = 7$, $q_\gamma = 5$, $r = 2$ and $n = 20$, with true allocation vector $\boldsymbol{\gamma}_0 = (1, 1, 1, 1, 0, 0, 1)$.



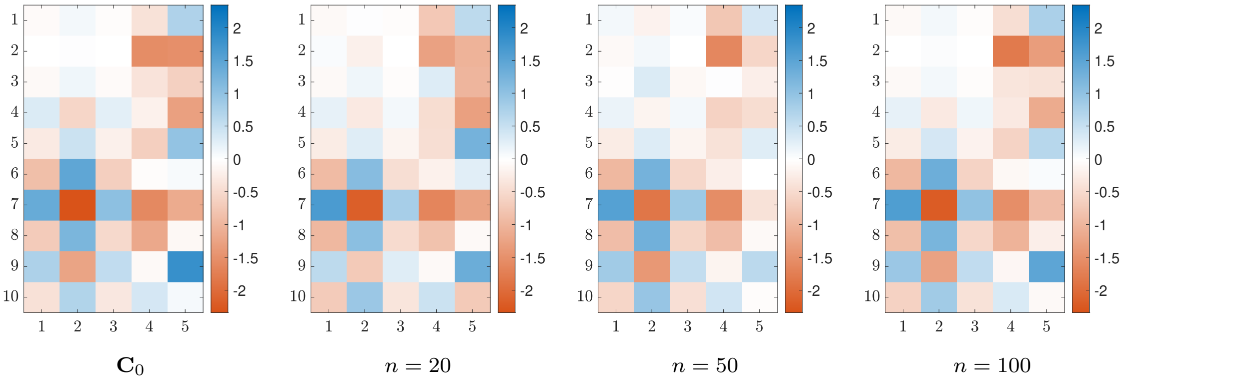
	γ_1	γ_2	γ_3	γ_4	γ_5	$p(\gamma\|Y)$
1	0	1	1	0	0	0.0032
2	1	1	0	0	0	0.0132
3	1	1	1	0	0	0.6305
4	1	1	1	0	1	0.2064
5	1	1	1	1	0	0.1467

Figure B.12: Trace plot (left) and posterior distribution (right) of $\boldsymbol{\gamma}$ in the simulation scenario $p = 50$, $q = 5$, $q_\gamma = 3$, $r = 1$ and $n = 50$, with true allocation vector $\boldsymbol{\gamma}_0 = (1, 1, 1, 0, 0)$.

B.4.2 Posterior concentration

Considering the choice of the prior distributions, we have adopted uniform priors over the set of possible outcomes for the allocation vector $\boldsymbol{\gamma}$ and the rank r. Then, we followed the standard practice in linear regression models and relied on the multivariate Gaussian and inverse Wishart for the vectorised coefficient matrices and the covariance matrix, respectively, choosing the hyperparameters to define vague priors with large variances. Therefore, we consider the proposed prior structure to have limited influence on the results, and to be conservative from a concentration perspective.

To empirically validate this claim, Figure B.13 reports the posterior mean of the coefficient matrix $\mathbf{C}$ for different values of the sample size, and Figure B.14 shows the posterior density of randomly chosen entries of $\mathbf{C}$. As n increases, the distribution concentrates around the mean, showcasing the posterior contraction property of the algorithm. This property is present as well in the parameter $\boldsymbol{\gamma}$, where an increase in n leads to a concentrated posterior around the true value $\boldsymbol{\gamma}_0$ (Figure B.15).



$\mathbf{C}_0$ $\quad n = 20$ $\quad n = 50$ $\quad n = 100$

Figure B.13: True coefficient matrix (first left) and estimated $\mathbf{C}$ in the simulation scenario where $p = 10$, $q = 5$, $q_\gamma = 3$ and $r = 1$, for different sample sizes $n = 20, 50, 100$ (left to right).

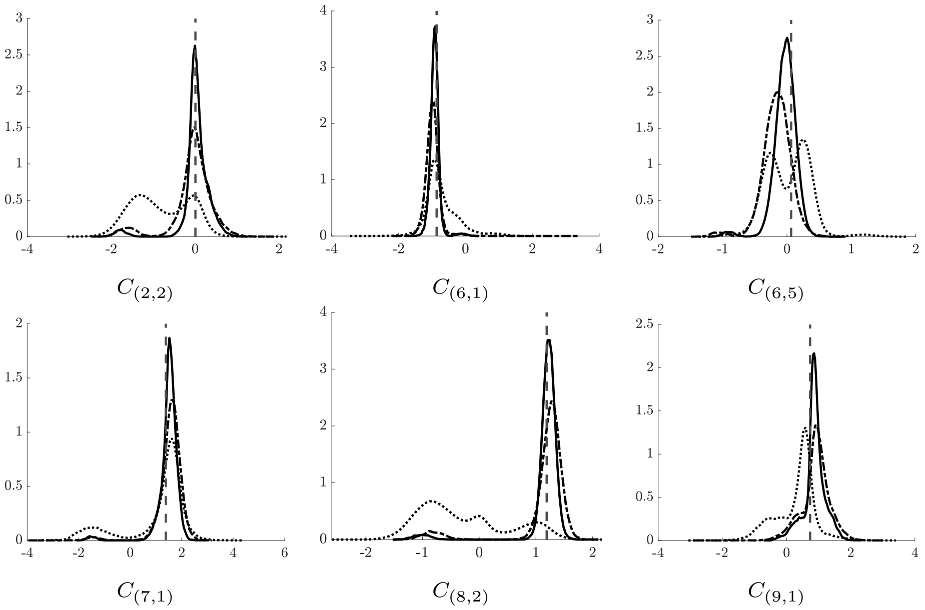
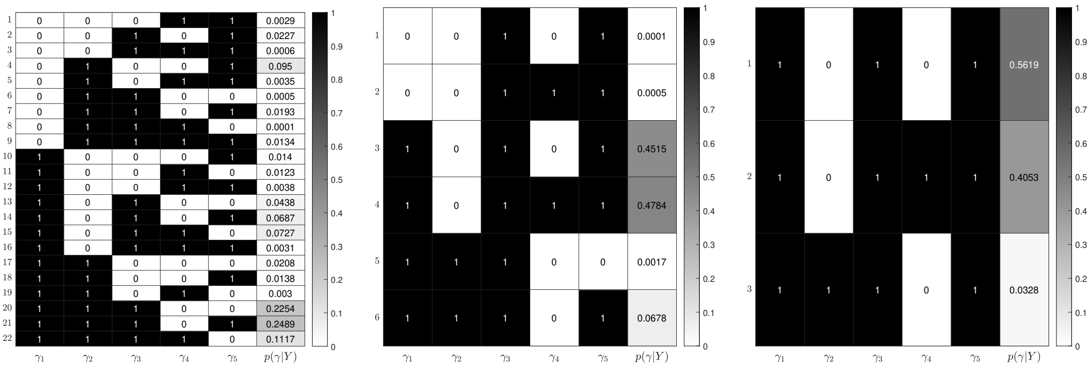


Figure B.14: Posterior density of randomly selected entries $C_{(i,j)}$ obtained via kernel density estimation, in the simulation scenario where $p = 10$, $q = 5$, $q_\gamma = 3$ and $r = 1$, for different sample sizes $n = 20$ (dotted), 50 (dashed-dotted) and 100 (solid). The true value is represented by a vertical dashed line.



	γ_1	γ_2	γ_3	γ_4	γ_5	$p(\gamma\|Y)$
1	0	0	0	1	1	0.0029
2	0	0	1	0	1	0.0227
3	0	0	1	1	1	0.0006
4	0	1	0	0	1	0.095
5	0	1	0	1	1	0.0035
6	0	1	1	0	0	0.0005
7	0	1	1	0	1	0.0193
8	0	1	1	1	0	0.0001
9	0	1	1	1	1	0.0134
10	1	0	0	0	1	0.014
11	1	0	0	1	0	0.0123
12	1	0	0	1	1	0.0038
13	1	0	1	0	0	0.0438
14	1	0	1	0	1	0.0687
15	1	0	1	1	0	0.0727
16	1	0	1	1	1	0.0031
17	1	1	0	0	0	0.0208
18	1	1	0	0	1	0.0138
19	1	1	0	1	0	0.003
20	1	1	1	0	0	0.2254
21	1	1	1	0	1	0.2489
22	1	1	1	1	0	0.1117

$n = 20$

	γ_1	γ_2	γ_3	γ_4	γ_5	$p(\gamma\|Y)$
1	0	0	1	0	1	0.0001
2	0	0	1	1	1	0.0005
3	1	0	1	0	1	0.4515
4	1	0	1	1	1	0.4784
5	1	1	1	0	0	0.0017
6	1	1	1	0	1	0.0678

$n = 50$

	γ_1	γ_2	γ_3	γ_4	γ_5	$p(\gamma\|Y)$
1	1	0	1	0	1	0.5619
2	1	0	1	1	1	0.4053
3	1	1	1	0	1	0.0328

$n = 100$

Figure B.15: Posterior distribution of $\boldsymbol{\gamma}$ (where the true value is $\boldsymbol{\gamma}_0 = (1, 0, 1, 0, 1)$) in the simulation scenario where $p = 10$, $q = 5$, $q_\gamma = 3$ and $r = 1$, for different sample sizes $n = 20, 50, 100$ (left to right).

B.5 Macroeconomic application

B.5.1 US quarterly macroeconomic data

The variables analyzed in section 3.5 have been sourced from two different websites: the Federal Reserve Economic Data (FRED) provided by the Federal Reserve Bank of St. Louis (https://fred.stlouisfed.org/) and the Organisation for Economic Co-operation and Development (OECD) (https://data.oecd.org/). The ten variables considered are listed below, along with their respective sources and names:
FRED

- y_1: Industrial Production: Total Index (INDPRO).
- y_2: Personal consumption expenditures: Nondurable goods: Food and beverages purchased for off-premises consumption (DFXARC1Q027SBEA).
- y_3: Unemployment Rate (UNRATE).
- x_1: Civilian Labor Force Level (CLF16OV).
- x_2: Employed full time: Median usual weekly real earnings: Wage and salary workers: 16 years and over (LES1252881600Q).

OECD: Quarterly GDP and components - expenditure approach - volume and price indices

- y_4: Imports of goods and services (P7): Index (IX), Chain linked volume (LR).
- y_5: Exports of goods and services (P6): Index (IX), Chain linked volume (LR).
- x_3: Imports of goods and services (P7): Index (IX), Deflator (DR).
- x_4: Exports of goods and services (P6): Index (IX), Deflator (DR).
- x_5: Final consumption expenditure (P3): Index (IX), Deflator (DR).

B.5.2 Stochastic volatility

The application of BPRR to the quaterly macroeconomic data of the United States incorporates a time-dependence structure in the covariance matrix through a stochastic volatility process. Figures B.16-B.17 show the mean path of the log stochastic volatility, $\mathbf{h}_j = (h_{j1}, \ldots, h_{jn})'$, for each response variable $j = 1, \ldots, q$.

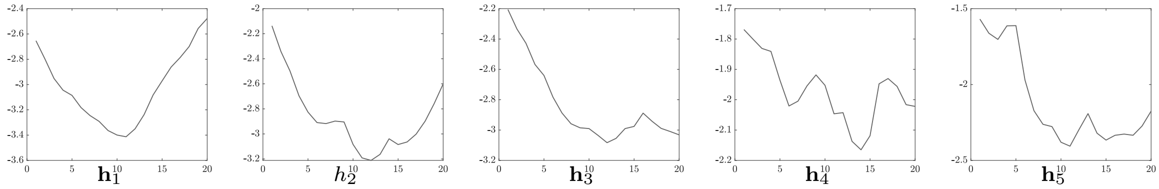


Figure B.16: Mean log volatility $\mathbf{h}_j$ across iterations for each response variable $j = 1, \ldots, 5$ in the period 2014Q1 - 2018Q4

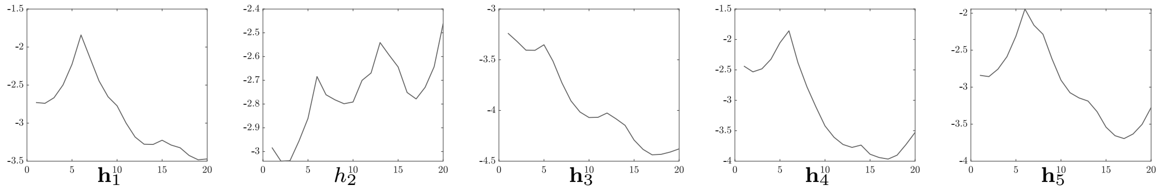


Figure B.17: Mean log volatility $\mathbf{h}_j$ across iterations for each response variable $j = 1, \ldots, 5$ in the period 2019Q1 - 2023Q4

Appendix C

Additional material for Chapter 4

C.1 Sampling the latent states

Let $\mathbf{Y}^{(t)}$ denote the matrix of observations from the first time point to time t, and $\Theta = (\boldsymbol{\gamma}, \mathbf{A}, \mathbf{B}, \mathbf{f}, \boldsymbol{\Sigma}, \mathbf{r}, \boldsymbol{\gamma}, \boldsymbol{\Xi})$. First, the filtered probabilities $Pr(s_t = k \mid \mathbf{Y}^{(t)}, \Theta)$ are computed for each $l = 1, \ldots, K$ and $t = 1, \ldots, T$ as

$$Pr(s_t = l \mid \mathbf{Y}^{(t)}, \Theta) = \frac{p(\mathbf{y}_t \mid s_t = l, \mathbf{Y}^{(t-1)}, \Theta)\, Pr(s_t = l \mid \mathbf{Y}^{(t-1)}, \Theta)}{p(\mathbf{y}_t \mid \mathbf{Y}^{(t-1)}, \Theta)},$$

where $p(\mathbf{y}_t \mid \mathbf{Y}^{(t-1)}, \Theta) = \sum_{k=1}^{K} p(\mathbf{y}_t \mid s_t = k, \mathbf{Y}^{(t-1)}, \Theta)\, Pr(s_t = k \mid \mathbf{Y}^{(t-1)}, \Theta)$.

The one-step ahead prediction of s_t is obtained from

$$Pr(s_t = l \mid \mathbf{Y}^{(t-1)}, \Theta) = \sum_{k=1}^{K} \xi^{*}_{kl}(t-1)\, Pr(s_{t-1} = l \mid \mathbf{Y}^{(t-1)}, \Theta),$$

for each $l = 1, \ldots, K$, where $\xi^{*}_{kl}(t-1) = Pr(s_t = l \mid s_{t-1} = k, \mathbf{Y}^{(t-1)}, \Theta)$. At $t = 1$, the probability distribution of s_1 conditional on the parameters and $\mathbf{Y}^{0}$ is

$$Pr(s_1 = l \mid \mathbf{Y}^{0}, \Theta) = \sum_{k=1}^{K} \xi^{*}_{kl}(0) Pr(s_0 = k \mid \boldsymbol{\xi}),$$

where $\xi^{*}_{kl}(0) = Pr(s_1 = l \mid s_0 = k, \Theta)$.

Thereafter, s_T is sampled from the filtered state probability distribution $Pr(s_T = k \mid \mathbf{Y}^{(T)}, \Theta)$. Finally, the state of the hidden Markov chain at time $t = T-1, T-2, \ldots, 0$ is sampled from the conditional distribution

$$Pr(s_t = k \mid s_{t+1}, \mathbf{Y}^{(t)}, \Theta) = \frac{\xi^{*}_{kl}(t)\, Pr(s_t = k \mid \mathbf{Y}^{(t)}, \Theta)}{\sum_{k=1}^{K} \xi^{*}_{kl}(t)\, Pr(s_t = k \mid \mathbf{Y}^{(t)}, \Theta)},$$

where $\xi^{*}_{kl}(t) = Pr(s_{t+1} = l \mid s_t = k, \mathbf{Y}^{(t)}, \Theta)$.