Journal of Machine Learning Research 25 (2024) 1-52 Submitted 6/22; Revised 11/23; Published 4/24

Flexible Bayesian Product Mixture Models for Vector Autoregressions

Suprateek Kundu SKUNDU2@MDANDERSON.ORG
Department of Biostatistics, The University of Texas MD Anderson Cancer Center
University of Texas
Houston, TX 77030, USA

Joshua Lukemire JOSHUA.LUKEMIRE@EMORY.EDU
Department of Biostatistics and Bioinformatics
Emory University
Atlanta, GA 30322, USA

Editor: Qiang Liu

Abstract

Bayesian non-parametric methods based on Dirichlet process mixtures have seen tremendous success in various domains and are appealing in being able to borrow information by clustering samples that share identical parameters. However, such methods can face hurdles in heterogeneous settings where objects are expected to cluster only along a subset of axes or where clusters of samples share only a subset of identical parameters. We overcome such limitations by developing a novel class of product of Dirichlet process location-scale mixtures that enables independent clustering at multiple scales, which results in varying levels of information sharing across samples. First, we develop the approach for independent multivariate data. Subsequently we generalize it to multivariate time-series data under the framework of multi-subject Vector Autoregressive (VAR) models that is our primary focus, which go beyond parametric single-subject VAR models. We establish posterior consistency and develop efficient posterior computation for implementation. Extensive numerical studies involving VAR models show distinct advantages over competing methods in terms of estimation, clustering, and feature selection accuracy. Our resting state fMRI analysis from the Human Connectome Project reveals biologically interpretable connectivity differences between distinct intelligence groups, while another air pollution application illustrates the superior forecasting accuracy compared to alternate methods.

Keywords: Dirichlet process mixtures, spatio-temporal data, functional magnetic resonance imaging, human connectome project, vector auto-regressive models

1. Introduction

Multivariate time-series data routinely arise in diverse application areas such as finance (Cramer and Miller, 1978), econometrics (Engle and Watson, 1981), air pollution forecasting (Nath et al., 2021) and medical imaging (Kundu and Risk, 2021), among other domains. In order to tackle such data, a rich body of work on modeling autocorrelations and temporal cross-correlations between variables with multivariate outcomes has been developed, of which vector autoregressive (VAR) models are widely used (Lütkepohl, 2005). Our focus in this paper is on Bayesian VAR modeling, which was initially heavily motivated by

econometric research (Doan et al., 1984) and has since seen a rich development (Korobilis, 2013). More recently, Bayesian VAR models have been adopted with increasing prominence in biomedical research including patient-level predictive modeling (Lu et al., 2018) and functional Magnetic Resonance Imaging (fMRI) applications (Gorrostieta et al., 2013; Chiang et al., 2017) in neuroimaging studies. However, existing Bayesian VAR literature has primarily focused on methodological and computational developments, with limited theoretical investigations. Recently, Ghosh et al. (2018) addressed this gap by establishing posterior consistency for the autocovariance matrix in parametric Bayesian VAR models based on single subject data.

The vast majority of the Bayesian VAR literature involves Gaussian assumptions and parametric prior specifications that may not be sufficiently flexible in characterizing the underlying probability distributions with non-regular features. For example, it is known that the nature of shocks in econometric analysis may not always be Gaussian (Weise, 1999). Similarly, flexible VAR modeling is necessary for analyzing heterogeneous multi-subject data in neuroimaging studies, where parametric VAR models may prove inadequate (see our Human Connectome Project (HCP) application in Section 6). Non-Gaussianity is also observed in air pollution data captured via sensors (Kim et al., 2013), where it is often of interest to perform forecasting using VAR models (Hajmohammadi and Heydecker, 2021). Such parametric VAR models may result in inaccurate performance when parametric assumptions are violated or even mis-specified. To bypass parametric constraints in VAR models, some recent articles relaxed Gaussianity assumptions (Jeliazkov, 2013). Recently, Bayesian nonparametric VAR models were proposed by Kalli and Griffin (2018) involving single subject data, where the mixing weights of the transition density depend on the previous lags. On the other hand, Billio et al. (2019) proposed Dirichlet process mixture of normal-Gamma priors on the VAR autocovariance elements. Unlike for the parametric case, the non-parametric methods are more robust to mis-specification and can potentially cater to a large class of models. However, the above approaches were applied to small or moderate dimensional data with limited or no emphasis on pooling information across samples and with negligible or no theoretical investigations.

Existing literature has largely ignored the problem of developing provably flexible non-parametric Bayesian VAR methodology to model heterogeneous multi-subject time-series data, to our knowledge. Such approaches are desirable over single-subject VAR analyses in terms of being able to pool information across samples in a flexible manner that can accommodate arbitrary probability distributions. They also facilitate robust and reproducible parameter estimates and provide a natural foundation for conduct inferences to test for differences across samples via credible intervals, which may not be straightforward under single-subject analysis. Although there is some literature on parametric VAR modeling of multi-subject data, these existing approaches typically require *a priori* knowledge of class labels (Gorrostieta et al., 2013; Chiang et al., 2017; Kook et al., 2021). Hence, they have a limited ability to accommodate heterogeneity within each class and may result in poor performance when the class labels are mis-specified due to no clear distinction between groups. Moreover, they clearly suffer from the aforementioned pitfalls of parametric methods.

Motivated by the above discussions, we propose a broad class of novel Bayesian non-parametric models that specify Dirichlet process (DP) mixture priors independently on mutually exclusive subsets of model parameters. Our specification results in a product of

Dirichlet process mixture (PDPM) priors. A key feature of the proposed approach is the ability to allow differential clustering at multiple scales, which enables clusters of samples that share only a subset of common model parameters resulting in greater flexibility. We develop several variants of the proposed approach that encourage differential degrees of heterogeneity via different modes of multiscale clustering by altering the manner in which the parameter space is partitioned. First, we develop the PDPM approach in the generic setting of multivariate density estimation for kernel mixtures of the form $\int K(x;\Theta)dP(\Theta)$, and establish posterior consistency properties. We also provide a toy example that illustrates the distinct numerical advantages of the product mixture models compared to traditional DP mixtures in terms of clustering accuracy. Subsequently, we generalize the proposed PDPM approach to multivariate time-series data under the framework of a VAR model, which is our primary focus in this article. In such settings, the multiscale clustering approach becomes even more relevant given the large number of parameters in the autocovariance matrix whose dimension grows quadratically with the vector dimension. Starting from a VAR model that allows for limited differences in clustering across multiple scales and greater model parsimony, we eventually develop a variant that is able to independently cluster row-specific parameters, which provides greater flexibility in practical applications. By specifying appropriate base measures in the DP prior, it is possible to enable appropriate shrinkage for the autocovariance elements that facilitates feature selection. Additional dimension reduction is also possible via a low rank representation for the residual covariance.

By designing non-parametric Bayesian VAR models based on heterogeneous multi-subject data, we are able to relax the parametric assumptions and provide a more flexible characterisation of heterogeneity via unsupervised clustering. The proposed methods are particularly desirable in terms of being able to bypass any restrictive assumptions such as the presence of replicated samples, which is routinely assumed in Bayesian non-parametric literature (Tokdar, 2006; Durante et al., 2017), but may potentially lead to inadequate characterization of heterogeneity. In particular, replicated samples are structured to share fully identical sets of model parameters within a given cluster, which may not be realistic in applications where heterogeneous samples are often effectively clustered only along a subset of directions with the remaining axes being uninformative/redundant for clustering (Agrawal et al., 2005).

Another appealing feature of the proposed approach is the associated posterior consistency properties for density estimation, as the number of samples (n) grows to ∞. We note that such theoretical results for VAR models involving multivariate time-series data represent non-trivial extensions of the rich theoretical properties established in the Bayesian non-parametric literature for independent outcomes (Tokdar, 2006; Canale and De Blasi, 2017). We resolve the significant challenges arising from the non-parametric Bayesian theoretical analysis by establishing Kullback-Leibler properties for VAR models, and constructing carefully designed sieves that are shown to satisfy certain entropy bounds and tail prior probability conditions under the product of DP priors. Moreover, we show that the theoretical results hold for commonly used base measures that enable straightforward posterior computation, and subsequently outline the computational complexity.

We develop an efficient and scalable Markov chain Monte Carlo (MCMC) implementation for the proposed class of models in the VAR framework. In addition, we illustrate the sharp numerical advantages and efficient mixing under the proposed non-parametric

Bayesian VAR approach in terms of parameter estimation and recovering the true clusters, compared to competing state-of-the-art methods. Further, the inferential capability of the proposed approach is evident from accurate feature selection of the autocovariance elements. Our analysis of resting state fMRI data from a subset of individuals in the HCP study infers several effective connectivity differences between the high and low fluid intelligence groups that are supported by existing evidence in literature. Moreover, the analysis under the proposed approach produces biologically reproducible estimates that are consistent across repeated neuroimaging scans from the same samples. In contrast, a single subject VAR analysis is able to identify only negligible effective connectivity differences across groups, which seems biologically implausible. Using a second data application example involving air pollution data from the Environment Protection Agency (EPA), we illustrate the considerable advantages in forecasting accuracy under the proposed approach compared to a parametric VAR model even when the dimension of the outcome is small.

The rest of the article is structured as follows. Section 2 develops the product of DP mixtures for independently distributed multivariate data and establishes posterior consistency properties. Section 3 extends the methodology to multivariate time-series data under a VAR framework, along with illustrating theoretical properties. Section 4 describes the posterior computation scheme. Section 5 reports results from extensive simulation studies involving VAR models. Sections 6 and 7 describe our analysis of the neuroimaging data from the HCP as well as air pollution data from the EPA. Section 8 contains additional discussions. Appendices are provided that contain other relevant details.

2. Product of DP Mixtures for Multivariate Data

2.1 A Primer on DP mixture approaches

Consider i.i.d. random vectors $\mathbf{x}_i, i, = 1, \ldots, n$, each of dimension $D \times 1$, and denote the collection of vectors as $X_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Non-parametric Bayesian literature has often focused on modeling these vectors under a DP location mixture or location-scale framework. Such approaches (Escobar and West, 1995) often specify $\mathbf{x}_i \sim N(\boldsymbol{\mu}_i, \Sigma_i), (\boldsymbol{\mu}_i, \Sigma_i) \sim P, P \sim DP(MP_0), i = 1, \ldots, n$, where $\Sigma_i \in S_{D\times D}$ denotes the covariance for subject i, $S_{D\times D}$ denotes the space of all $D \times D$ symmetric positive definite matrices, P_0 denotes the base measure of the DP, and M is the precision parameter. We note that alternative choices other than the Gaussian kernel may also be used but are not considered here for simplicity. The resulting DP location-scale mixture induces the unknown probability density $f_P(\mathbf{x}) = \int \phi_\Sigma(\mathbf{x} - \boldsymbol{\mu}) dP(\boldsymbol{\mu}, \Sigma)$, where $\phi_\Sigma(\cdot - \boldsymbol{\mu})$ denotes the density of a D-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ. Given that $P \sim DP(MP_0)$, the proposed method results in probability distributions on the class of densities $\mathcal{F} = \{f_P\}$, which can also be seen from the result $f_P(\mathbf{x}) = \sum_{h=1}^{\infty} \pi_h \phi_{\Sigma_h}(\mathbf{x} - \boldsymbol{\mu}_h)$, where $(\boldsymbol{\mu}_h, \Sigma_h) \sim P_0$ and $\pi_h = \nu_h \prod_{l=1}^{h-1}(1 - \nu_l), \nu_h \sim Be(1, M)$, using Sethuraman's (1994) stick-breaking representation.

The above commonly used DP mixture specification results in clusters of replicated samples that share identical sets of parameters $(\boldsymbol{\mu}, \Sigma)$, which allows for pooling of information across samples resulting in robust learning. While such a clustering mechanism is routinely used and often backed by posterior consistency guarantees, it may not be well-equipped to succeed for more heterogeneous settings where the clustering is dictated by a small number

of axes or subspaces, with the other axes being irrelevant to clustering. A more reasonable approach is to allow differential clustering at multiple scales that does not constrain samples to share fully identical parameter sets, but instead allows subsets of parameters to cluster independently resulting in partially overlapping clusters. Such a multi-scale clustering approach results in a more accurate characterization of heterogeneity that is expected to improve finite sample performance. The above arguments form the basis of the proposed product mixture priors in this article.

2.2 Proposed Methodology and Properties

We propose a class of novel product mixture priors that is equipped to perform differential clustering at multiple scales. Consider equally sized mutually exclusive and exhaustive subsets of the full parameter set denoted as $\boldsymbol{\mu} = \cup_{m_1=1}^{\mathcal{M}_\mu} \boldsymbol{\mu}_{m_1}$, $\tilde{\boldsymbol{\sigma}} = \cup_{m_2=1}^{\mathcal{M}_\sigma} \boldsymbol{\sigma}_{m_2}$, where $\tilde{\boldsymbol{\sigma}}$ denotes the vectorized upper triangular matrix of Σ, and $\{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{\mathcal{M}_\mu}\}$ represent subsets of equal cardinality, and similarly for $\{\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{\mathcal{M}_\sigma}\}$. Consider specifying the following product of DP priors on the parameters:

$$\begin{aligned} &\boldsymbol{\mu}_{m_1} \overset{indep}{\sim} P_\mu,\ m_1 = 1, \ldots, \mathcal{M}_\mu,\ \boldsymbol{\sigma}_{m_2} \overset{indep}{\sim} P_\sigma,\ m_2 = 1, \ldots, \mathcal{M}_\sigma,\ \Sigma \in S_{D\times D}, \\ &P_\mu \sim DP(\alpha_1 P_1^*),\ P_\sigma \sim DP(\alpha_2 P_2^*), \end{aligned} \tag{1}$$

where each component is assigned independent priors $\boldsymbol{\mu}_{m_1} \overset{indep}{\sim} P_\mu, \boldsymbol{\sigma}_{m_2} \overset{indep}{\sim} P_\sigma$, that follow Dirichlet process with base measures P_1^* and P_2^* respectively, with corresponding precision parameters α_1, α_2. The specification (1) results in a product of DP priors on the original parameters $(\boldsymbol{\mu}, \Sigma)$ that is denoted by Π^*, where the exact prior depends on the way in which the partitions are defined. Hence, one can obtain a class of product mixture priors by tweaking the partition structure to reflect the most appropriate setting for the data at hand. The product priors in (1) induce priors Π on $\mathcal{F}$ via the relationship:

$$\begin{aligned} f_P(\mathbf{x}) &= \int\int \phi_\Sigma\big(\mathbf{x} - \boldsymbol{\mu}\big) d\Pi^*(\boldsymbol{\mu}, \Sigma) \\ &= \sum_{h_{11},\ldots,h_{1\mathcal{M}_\mu}=1}^{\infty} \sum_{h_\sigma=1}^{\infty} \pi_{1,h_{11}} \ldots \pi_{\mathcal{M}_\mu,h_{1\mathcal{M}_\mu}} \pi_{\sigma,h_\sigma} \phi_{\Sigma_{h_\sigma}}\big(\mathbf{x} - (\boldsymbol{\mu}_{1,h_{11}}^T, \ldots, \boldsymbol{\mu}_{\mathcal{M}_\mu,h_{1\mathcal{M}_\mu}}^T)^T\big), \end{aligned} \tag{2}$$

where the second equality is obtained by Sethuraman's (1994) stick breaking representation with $\pi_{h_{1m_1}} = \nu_{h_{1m_1}} \prod_{l_1<h_{1m_1}}(1-\nu_{l_1})$, $\nu_{l_1} \sim Be(1, \alpha_1)$, $\pi_{\sigma,h_\sigma} = \nu_{\sigma,h_\sigma} \prod_{l_2<h_\sigma}(1-\nu_{\sigma,l_2})$, $\nu_{\sigma,h_\sigma} \sim Be(1, \alpha_2)$, and further $\Sigma_{h_\sigma} \sim P_2^*$, $\boldsymbol{\mu}_{m_1} \sim P_1^*$, and $\mathcal{M}_\sigma$ is assumed to be one in the above expression. The choice of $\mathcal{M}_\sigma = 1$ is guided by practical considerations in VAR models that is our primary focus (next section) where the residual covariance matrix often has a sparse or even diagonal structure after regressing out the lag effects of previous time points. However our treatment can be generalized to $\mathcal{M}_\sigma > 1$ in a straightforward manner. We note that the above form in (2) follows the generic kernel mixture representation $\int K(x;\Theta) dP(\Theta)$ that is commonly considered in non-parametric Bayesian density estimation literature (Wu and Ghosal, 2008). We denote the resulting class of priors on $\mathcal{F}$ arising from (2) as the product of Dirichlet process mixtures (PDPM).

The most straightforward case of the prior in (1) is given as $\Pi^*(\boldsymbol{\mu}, \Sigma) = P_\mu(\boldsymbol{\mu}) \times P_\sigma(\Sigma)$, which specifies independent priors on the mean and covariance parameters without further

partitioning these parameters (i.e. $\mathcal{M}_\mu = 1, \mathcal{M}_\sigma = 1$). The PDPM operates by clustering the mean and covariance parameters independently, which suggests separate modes of pooling information across samples for the mean and covariance. Such a multiscale clustering approach results in greater flexibility and a more accurate characterization of heterogeneity compared to replicated samples with fully identical parameter sets via allowing for dedicated clusters of samples that share common mean signatures (but not necessarily for the covariance), along with independently constructed subgroups of samples that share common patterns in the covariance (but not necessarily for the mean). As a more flexible generalization, one can consider the *generalized PDPM (gPDPM)* model that specifies independent DP priors for each element of the mean vector, i.e. $\Pi^*(\boldsymbol{\mu}, \Sigma) = \prod_{m=1}^{D} P_\mu(\mu_m) \times P_\sigma(\tilde{\boldsymbol{\sigma}})$. The gPDPM approach allows separate clustering for each element of $\boldsymbol{\mu}$ across samples, which enables differential clustering along various axis and provides a more granular approach for pooling information, albeit at the cost of a larger number of model parameters . The multiscale clustering approach is expected to excel in settings where the clustering is dictated by a subset of axes in the mean with the other axes being redundant towards clustering. The above discussions highlight the advantages of the multiscale clustering aspect under the proposed product mixture modeling methodology, and provides the central motivation for this article. A schematic representation of the above ideas is presented in Figure 1.

Toy Example: We illustrate the advantages of the multiscale clustering approach using a toy example. Multivariate data $Y_i \sim N_D(\boldsymbol{\mu}_i, \Sigma_i)$, for $i = 1, \ldots, 250$, was generated such that the mean across samples were identical except the first d elements, where $d \approx \frac{D}{3}$. For the first d elements of $\boldsymbol{\mu}$, there were 5 clusters, each with a corresponding $d-$vector of $\boldsymbol{\mu}$ values. Similarly, 5 clusters were generated for $\Sigma(D \times D)$ that were constructed independent of the mean. We used the standard DPM and the proposed PDPM to fit these data, and evaluate the clustering performance across varying dimensions. The posterior computation steps for both approaches are just simplified versions of the posterior computation for the PDPM-VAR model that will be introduced in the sequel (omitted here for conciseness). For each method, we evaluated the adjusted Rand index for clustering the mean vector that is averaged over all MCMC iterations (Rand, 1971). Figure 1c illustrates the clustering accuracy. Unsurprisingly, the PDPM offers significant improvement over the DPM for such a heterogeneous clustering setup across varying dimensions D, which clearly illustrates the considerable advantages of the proposed approach. The standard DPM approach allocates identical mean and covariance parameters for all samples within each cluster, which can not tackle the differential clustering allocations between the mean and covariance and hence results in spurious clusters that adversely affect the overall clustering accuracy.

Theoretical Properties: From a theoretical perspective, it is possible to show that the proposed product of DP approach leads to posterior consistency for Kullback-Leibler neighborhoods, under reasonable assumptions on the true density f_0 that are routinely assumed in multivariate density estimation literature (Wu and Ghosal, 2008). This is not surprising given that the density follows a generic kernel mixture representation $\int K(x; \Theta) dP(\Theta)$ commonly encountered in the literature. Some additional notations are provided below. Denote the Euclidean norm for a vector as $||\cdot||$, and denote the spectral norm of a matrix as $||\cdot||_2$. Further, denote the eigenvalues of a $D \times D$ positive definite matrix Σ in decreasing order as $\lambda_1(\Sigma) \geq \ldots \geq \lambda_D(\Sigma)$. Let $a \lesssim b$ imply that a is less than b upto a constant, and let $\lfloor \cdot \rfloor$

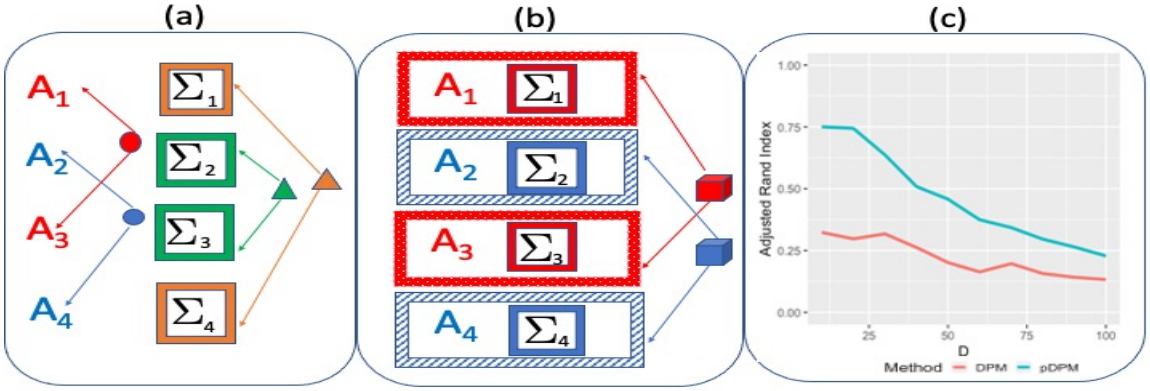

Figure 1: A schematic representation of the product of DP mixture prior. Panel (a) illustrates the product prior that separately clusters the mean (represented by A) into red and blue clusters and the covariance (Σ) into green and saffron clusters. Panel (b) represents the traditional DP mixture prior that forms clusters containing samples having identical values for both the mean and covariance parameters. Panel (c) illustrates the results from the toy example under the traditional DPM and the proposed PDPM, in terms of the change in clustering accuracy with varying dimension when the clustering is dictated by a subset of axes.

denote the floor operator. Denote the Kullback-Leibler (KL) divergence between densities $f, g \in \mathcal{F}$ as $KL(f,g) = \int \log(f/g)f$. Denote the set of natural numbers as $\mathbb{N}$.

We now establish our result on positive prior support for Kullback-Leibler neighborhoods under the product of DP prior below, based on the following assumptions.

(C1) $0 < f_0(\mathbf{x}) < M$ for some constant M and all $\mathbf{x} \in \Re^D$;
(C2) $\int f_0(\mathbf{x}) \log(f_0(\mathbf{x}))d\mathbf{x} < \infty$;
(C3) $\int f_0(\mathbf{x}) \log(f_0(\mathbf{x})/\phi_\delta(\mathbf{x}))d\mathbf{x} < \infty$ where $\phi_\delta(\mathbf{x}) = \inf_{||\mathbf{t}-\mathbf{x}||<\delta} f_0(\mathbf{t})$ for some $\delta > 0$;
(C4) for some $\eta > 0$, $\int ||\mathbf{x}||^{2(1+\eta)} f_0(\mathbf{x})d\mathbf{x} < \infty$.

Assumptions (C1)-(C4) are similar to routinely used conditions in non-parametric Bayesian literature for establishing posterior consistency properties. For example, these conditions were proposed in Wu and Ghosal (2008) for establishing Kullback-Leibler convergence properties for location-scale mixtures. Since then, they have been used extensively in related literature such as Wu and Ghosal (2010) for multivariate location mixtures, in Canale and De Blasi (2017) for showing strong consistency properties for multivariate location-scale mixtures, as well as for conditional density estimation (Pati et al., 2013), and Dirichlet mixtures of exponential power densities (Scricciolo, 2011), among others. Condition (C1) simply implies that the true density f_0 is bounded which is a reasonable and mild assumption. Conditions (C2) and (C3) are subtle, but are also mild, as noted in Pati et al. (2013). For example, condition (C2) should be satisfied by appropriate location-scale mixtures of normals. Condition (C4) imposes a minor tail restriction that should be satisfied by the t-distribution with suitable degrees of freedom, among others.

Lemma 1: *Let $f_0 \in \mathcal{F}$ and assume conditions (C1)-(C4) hold. Then for the prior defined in (1), we have $\Pi\big(f \in \mathcal{F} : \int \log\big(f_0/f\big) f_0 \leq \eta^*\big) \geq 0$, for any $\eta^* > 0$.*

Remark 1: Lemma 1 provides weak consistency guarantees by establishing positive prior support for arbitrarily small Kullback-Leibler neighborhoods of f_0, as per Schwartz (1965).

An outline of the proof is provided in the Appendix. Although weak consistency is useful, a more appealing feature is strong consistency, ensuring that the posterior distribution concentrates in arbitrarily small L_1 neighborhoods of the true density. The next result states that under certain tail conditions on the base measure of the DP priors, it is possible to derive strong posterior consistency corresponding to the PDPM priors.

Theorem 1: *Suppose f_0 satisfies the conditions of Lemma 1. Then the posterior is strongly consistent at f_0 under the PDPM and gPDPM priors Π in (2) with base measures that satisfy the conditions: (i) $P_2^*(\lambda_1(\Sigma_{h_\sigma}^{-1}) > x^*) \lesssim \exp(-c_1(x^*)^{c_2}), P_2^*(\lambda_D(\Sigma_{h_\sigma}^{-1}) < 1/x^*) \lesssim (x^*)^{-c_3}, P_2^*\big(\frac{\lambda_1(\Sigma^{-1})}{\lambda_D(\Sigma_{h_\sigma}^{-1})} > x^*\big) \lesssim (x^*)^{-\kappa}$, for some constants c_1, c_2, c_3, κ, and all clusters h_σ; (ii) $P_1^*(||\boldsymbol{\mu}_{m_1,h_{1m_1}}|| > x^*) \lesssim (x^*)^{-2(r+1)}$ for all clusters h_{1m_1}, where $m_1 = 1, \ldots, \mathcal{M}_\mu$.*

Theorem 1 provides explicit conditions on the PDPM prior that will ensure strong consistency, corresponding to any true density f_0 lying in the weak neighborhood of the prior Π on the set of densities $\mathcal{F}$. The proof is provided in the Appendix. The tail conditions on the base measures in Theorem 1 are very reasonable and hold for commonly used distributions (such as Gaussian and Laplace) on the mean, as well as the inverse-Wishart distribution on the covariance (see Lemmas 2-3 in the sequel). It should be noted that the procedure for proving the strong consistency result in Theorem 1 corresponding to non-compact space of densities $\mathcal{F}$ relies on carefully designed sieves $\mathcal{F}_n$ that are compact subsets of $\mathcal{F}$ but that grow with n to eventually cover all of $\mathcal{F}$ as $n \to \infty$. These sieves must satisfy certain sufficient conditions for the strong consistency result to hold. These sufficient conditions are motivated from ideas in Theorem 5 of Ghosal and Van Der Vaart (2007), and were derived by Shen et al. (2013) for location mixtures and in Theorem 1 of Canale and De Blasi (2017) for location-scale mixtures. For clarity, we restate the result in Canale and De Blasi (2017) as Theorem 2 in our paper that will be leveraged to establish the posterior consistency under our set-ups in Sections 2 and 3 (Theorems 1 and 5 respectively).

Denote the entropy of a space of densities $\mathcal{G} \subset \mathcal{F}$ as $N(\epsilon, \mathcal{G}, d)$, which is defined (in terms of the metric d) as the minimum integer N for which there exists densities $f_1, \ldots, f_N \in \mathcal{F}$ satisfying $\mathcal{G} \subset \cup_{j=1}^N \{f : d(f, f_j) < \epsilon\}$. The distance metric used to study convergence in the space $\mathcal{F}$ is evaluated in terms of the Hellinger distance (defined $d(f,g) = \big[\int(\sqrt{f} - \sqrt{g})^2\big]^{1/2}$), as well as the L_1 metric (defined as $||f - g||_1 = \int |f - g|$). Further, denote F_0^n as the n-products measure, where F_0 is the probability measure corresponding to f_0.

Theorem 2 (Canale and De Blasi): *Consider sieves $\mathcal{F}_n \subset \mathcal{F}$ with $\mathcal{F}_n \uparrow \mathcal{F}$ as $n \to \infty$, where $\mathcal{F}_n = \cup_j \mathcal{F}_{n,j}$, with (2A) $\Pi(\mathcal{F}_n^c) \lesssim e^{-bn}$; and (2B) $\sum_j \sqrt{N(2\epsilon, \mathcal{F}_{n,j}, d)}\sqrt{\Pi(\mathcal{F}_{n,j})} e^{-(4-c)n\epsilon^2} \to 0$, for $b, c, \epsilon > 0$. Then $\Pi(f : d(f_0, f) > 8\epsilon \mid X_n) \to 0$ in F_0^n-probability for any f_0 in the weak support of Π defined in (1).*

In Theorem 2, condition (*2A*) suggests that the sieve should grow with the sample size such that only small neighborhoods with exponentially small prior probabilities are ex-

cluded. On the other hand, condition (*2B*) reflects the summability condition that involves smaller subsets $\mathcal{F}_{n,j}$ that cover the sieve $\mathcal{F}_n$ under the union operation. It places constraints on the growth rate of the metric entropy in a manner that the weighted sum of the square root of metric entropy of $\mathcal{F}_{n,j}$ (weighted by the corresponding square root of prior probabilities) go towards zero with increasing n. We construct such sieves in the proof of Theorem 1 in the Appendix, and illustrate that the conditions (2A) and (2B) are satisfied, which results in strong consistency.

Model (1) lays the foundation for the novel PDPM priors, that potentially has a wide array of applications, and can likely be generalized to most frameworks that involve clustering under Dirichlet process mixtures. We are now well positioned to turn our focus on the primary goal in this article, which is to develop a provably flexible non-parametric Bayesian methodology for multivariate time-series data modeled under a VAR framework, which is one of the first such set of results in literature, to our knowledge.

3. Extension to Vector Autoregressive Models

3.1 Proposed Model

Consider the data matrix $X_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, where $\mathbf{x}_{it}$ represents the $(D \times 1)$ temporally dependent multivariate measurement for the i-th subject at the t-th time point $(i = 1, \ldots, n, t = 1, \ldots, T)$. Note that our model can easily accommodate subject-specific scan lengths (T_i); however we will assume $T_i = T$ from hereon in, to ease the exposition. Throughout, we will also assume a fixed dimension (D), and a pre-specified number of time scans (T), which is consistent with the routinely used fixed dimensional assumptions in the literature on non-parametric modeling of location-scale mixtures. Consider the VAR model:

$$\mathbf{x}_{it} = \sum_{k=1}^{\min\{t-1,K\}} A_{ik}\mathbf{x}_{i,t-1} + \boldsymbol{\epsilon}_{it},\ \boldsymbol{\epsilon}_{it} \sim N(\mathbf{0}, \Sigma_i),\ i = 1, \ldots, n,\ t = 2, \ldots, T, \tag{3}$$

where A_{ik} denotes the $D \times D$ matrix of autocovariance parameters for subject i at lag k $(k = 1, \ldots, K)$, $\Sigma_i \in S_{D\times D}$ denotes the time-invariant residual covariance for subject i, and the lag order (K) is pre-specified as per standard practice in the VAR model literature (Ghosh et al., 2018). Model (3) implies that the mean of $\mathbf{x}_t$ depends on $\mathbf{x}_{t-1}, \ldots, \mathbf{x}_1$ when $t \leq K$ and on $\mathbf{x}_{t-1}, \ldots, \mathbf{x}_{t-K}$ for $t > K$, with $\mathbf{x}_{i1} \sim N(0, \Sigma_i)$ as per convention. As is common in practice, the intercept term is fixed to be zero and not included in (3).

In order to understand the properties of (3), it is imperative to note that the likelihood for the i-th sample can be written as a product of conditional densities as

$$L(X_i \mid \Theta_i, \Sigma_i) = \prod_{t=2}^{T} \phi_{\Sigma_i}\left(\mathbf{x}_{it} - \sum_{k=1}^{\min\{t-1,K\}} A_{ik}\mathbf{x}_{i,t-k}\right) \times \phi_{\Sigma_i}\big(\mathbf{x}_{i1}\big),\ i = 1, \ldots, n, \tag{4}$$

where Θ_i denotes the collection of autocovariance matrices for sample i across lags. For example, the likelihood for the ith sample under a VAR(2) model may be written as $\prod_{t=3}^{T} \phi_{\Sigma_i}\big(\mathbf{x}_{it} - \sum_{k=1}^{2} A_{ik}\mathbf{x}_{i,t-k}\big) \times \phi_{\Sigma_i}\big(\mathbf{x}_{i2} - A_{i1}\mathbf{x}_{i,1}\big) \times \phi_{\Sigma_i}\big(\mathbf{x}_{i1}\big)$. The above likelihood in (4) will be used throughout in our treatment of VAR models. We note that (4) is a different

way of representing the likelihood compared to the linear regression framework that is often used in single subject VAR models (Ghosh et al., 2018).

Our goal involves multi-subject VAR analysis by proposing suitable priors on (Θ_i, Σ_i) in (3) to leverage common patterns of information across samples in an unsupervised and flexible manner. A natural framework for pooling information across subjects is via clustering, which also inherently results in model parsimony that is particularly important in our settings where the number of parameters grow with n. Such a clustering approach should enable straightforward posterior computation and result in theoretical guarantees. To this end, we extend the PDPM methodology to the case of multivariate time-series data that imposes independent DP mixture priors separately on Θ and Σ to induce multiscale clustering. Depending on the manner of the DP prior specification on the autocovariance elements, one can obtain different variants of the proposed method that allow for varying degrees of model parsimony and varying levels of information sharing within samples, via different patterns of autocovariance clusters. Such a multi-scale clustering approach is particularly relevant in the context of VAR models where the dimension of the autocovariance matrix increases quadratically with the outcome dimension D, making it imperative to avoid the assumption of replicated samples that is embedded in typical mixture modeling approaches. The resulting PDPM approach leads to a more fitting characterization of heterogeneity and greater accuracy, as illustrated via extensive numerical studies involving VAR models in the sequel. In addition, appropriate base measures in the DP can be chosen to encourage shrinkage in the autocovariance elements that facilitate feature selection, as well as to induce low rank decomposition for the residual covariance resulting in additional model parsimony.

Product of DP mixtures for VAR models: In the following specifications, we will omit subscript i where appropriate, for notational convenience and as per convention (Wu and Ghosal, 2008; Canale and De Blasi, 2017). We propose the following PDPM prior

$$\Theta = \{vec(A_1), \ldots, vec(A_K)\} \sim P_\Theta,\ P_\Theta \sim DP(\alpha_1 P_1^*), \Sigma \sim P_\mathcal{S},\ P_\mathcal{S} \sim DP(\alpha_2 P_2^*), \tag{5}$$

where α_1, α_2, represent precision parameters in the Dirichlet process, the base measure P_1^* belongs to the space of probability measures $\mathcal{P}_1$ on $\mathcal{D}_1 = \underbrace{\Re^{D^2\times 1} \times \ldots \times \Re^{D^2\times 1}}_{\text{K}}$, and the base measure P_2^* belongs to the space of probability measures $\mathcal{P}_2$ on $\mathcal{D}_2 = S_{D\times D}$. Model (5) specifies unknown distributions P_Θ and $P_\mathcal{S}$ on model parameters, that are modeled under independent DP priors. The resulting product of DP priors in (5) is defined on the space of densities $\mathcal{P}$ with domain $\mathcal{D}_1 \times \mathcal{D}_2$ and may be expressed as $\Pi^*(\Theta, \Sigma) = P_\mathcal{S}(\Sigma) \times P_\Theta(\Theta)$. This prior specification translates to a *product of DP mixture of VAR (PDPM-VAR)* models that induces a prior Π on the space of probability densities $\mathcal{F}$ for the data matrix X as follows:

$$\begin{aligned} f_P(X) &= \int\int \prod_{t=1}^{T} \phi_\Sigma\Big(\mathbf{x}_t - \sum_{k=1}^{\min\{t-1,K\}} A_k \mathbf{x}_{t-k}\Big) dP_\Theta(\Theta) dP_\mathcal{S}(\Sigma) \\ &= \sum_{h_1=1}^{\infty}\sum_{h_\sigma=1}^{\infty} \pi_{h_1}\pi_{\sigma,h_\sigma} \prod_{t=1}^{T} \phi_{\Sigma_{h_\sigma}}\Big(\mathbf{x}_t - \sum_{k=1}^{\min\{t-1,K\}} A_{k,h_1} \mathbf{x}_{t-k}\Big), \end{aligned} \tag{6}$$

where $\pi_{h_1} = \nu_{h_1}\prod_{l_1<h_1}(1-\nu_{l_1})$, $\nu_{h_1} \sim Be(1,\alpha_1)$, $\pi_{\sigma,h_\sigma} = \nu_{\sigma,h_\sigma}\prod_{l_2<h_\sigma}(1-\nu_{\sigma,l_2})$, $\nu_{\sigma,h_\sigma} \sim Be(1,\alpha_2)$, and further $\Sigma_{h_\sigma} \sim P_2^*$, $(vec(A_{1,h_1}), \ldots, vec(A_{K,h_1})) \sim P_1^*$. We consider a broad

class of base measures to study theoretical properties (Section 3.2), but for implementation we focus on specific choices for (P_1^*, P_2^*) that facilitate posterior computations (Section 4).

While (5) provides a greater degree of flexibility in terms of accommodating heterogeneity compared to existing DP mixture approaches, there is further scope for generalizing this approach to accommodate additional heterogeneity in lag-specific and row-specific relationships. Such generalizations become particularly important when clusters of samples tend to share common autocovariance elements for some but not all lags or have identical elements for only a subset of rows/nodes in the autocovariance matrices in practical applications. For example, the latter scenario arises when the effective clustering for the autocovariance elements is confined to a subset of rows in the matrix A, with the remaining rows being irrelevant with respect to clustering. Such aspects are routinely encountered in heterogeneous and high-dimensional clustering problems (Agrawal et al., 2005), such as our VAR settings of interest where the number of autocovariance parameters increase quadratically with the outcome dimension (D). We now generalize the PDPM-VAR method below to account for such heterogeneous settings.

Generalization across autocovariance rows: It is possible to generalize the PDPM-VAR model in (5) in a manner that relaxes the restriction to have fully identical autocovariance matrices for all samples within a given autocovariance cluster. In particular, consider an approach that specifies independent priors on the VAR model parameters corresponding to each row of the autocovariance matrices, which results in row-specific clustering patterns. In particular, denote $A_{k,d'\bullet}$ as the d'-th row of A_k and consider the following specification

$$vec\{A'_{1,d'\bullet}, \ldots, A'_{K,d'\bullet}\} \overset{indep}{\sim} P_{\Theta_{d'}},\ P_{\Theta_{d'}} \sim DP(\alpha^*_{d'} P^{**}_{1d'}),\ \Sigma \sim P_{\mathcal{S}},\ P_{\mathcal{S}} \sim DP(\alpha_2 P_2^*), d' = 1, \ldots, D, \quad (7)$$

where A' denotes the transpose of A, and the row-specific priors $P_{\Theta_{d'}}(vec\{A'_{1,d'\bullet}, \ldots, A'_{K,d'\bullet}\})$ are specified independently for each row and jointly across lags. The product of DP prior in (7) is expressed as $\Pi^*(\Theta, \Sigma) = P_{\mathcal{S}}(\Sigma) \times \prod_{d'=1}^{D} P_{\Theta_{d'}}(vec\{A'_{1,d'\bullet}, \ldots, A'_{K,d'\bullet}\})$, and results in the row-generalized PDPM-VAR (rgPDPM-VAR) model that induces priors on $\mathcal{F}$ via

$$f_P(X) = \sum_{h_{1,1}=1}^{\infty} \ldots \sum_{h_{1,D}=1}^{\infty} \sum_{h_\sigma=1}^{\infty} (\pi_{\sigma,h_\sigma} \prod_{d'=1}^{D} \pi^*_{d',h_{1d'}}) \prod_{t=1}^{T} \phi_{\Sigma_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{\min\{t-1,K\}} A_{k,h_{11},\ldots,h_{1D}} \mathbf{x}_{t-k}\big), \quad (8)$$

where $A_{k,h_{11},\ldots,h_{1D}}$ denotes the autocovariance matrix at lag k that assigns the $h_{1d'}$-th mixture component to the d'-th row with prior probability $\pi^*_{d',h_{1d'}} = \nu^*_{d',h_{1d'}} \prod_{l_{1d'}<h_{1d'}} (1 - \nu^*_{l_{1d'}})$, where $\nu^*_{d',h_{1d'}} \sim Be(1, \alpha^*_{d'})$, $vec\{A'_{1,d'\bullet,h_{1,d'}}, \ldots, A'_{K,d'\bullet,h_{1,d'}}\} \overset{indep}{\sim} P^{**}_{1d'}$ and $A_{k,d'\bullet,h_{1,d'}}$ denotes the d'-th row for the matrix A_k that takes values from the $h_{1,d'}$-th mixture component. Further, $\Sigma_{h_\sigma} \sim P_2^*$ with prior probability $\pi_{\sigma,h_\sigma} = \nu_{\sigma,h_\sigma} \prod_{l_2<h_\sigma} (1 - \nu_{\sigma,l_2})$ and $\nu_{\sigma,h_\sigma} \sim Be(1, \alpha_2)$.

In the scenario when multiple rows have identical clustering configurations, the rgPDPM-VAR model is able to identify clusters of samples that share identical autocovariance elements corresponding to a subset of nodes only, but exhibit variations corresponding to the remaining autocovariance rows. We note that for our motivating neuroimaging applications, this scenario translates to identical effective connectivity corresponding to a subset of brain regions within a autocovariance cluster, while the remaining brain regions are allowed to

exhibit varying connectivity profiles within this cluster. By allowing row-specific clustering patterns in the autocovariance matrix, the rgPDPM-VAR approach results in a more complete characterization of heterogeneity compared to the PDPM-VAR. Additional generalizations are also possible; for example, one may extend specification (7) to impose row- and lag-specific priors. However, such extensions may result in a rapid rise in parameters that presents potential computational issues, and hence are not considered further.

Generalization across lags: For the second extension, we specify independent DP priors for the autocovariance matrices at each lag, which results in lag-specific clustering as follows:

$$vec(A_k) \overset{indep}{\sim} P_{\Theta_k},\ P_{\Theta_k} \sim DP(\alpha_{1k} P^*_{1k}),\ \Sigma \sim P_{\mathcal{S}},\ P_{\mathcal{S}} \sim DP(\alpha_2 P^*_2),\ k = 1, \ldots, K, \quad (9)$$

where P_{Θ_k} denotes the unknown density for $vec(A_k)$ that is modeled under a DP prior with base measure P^*_{1k} and precision parameter $\alpha_{1k} (k = 1, \ldots, K)$, and the prior on the residual covariance parameters is defined similarly to (5), but with the understanding that α_2 and P^*_2 in the DP priors in (9) and (5) are allowed to be distinct. The resulting product of DP priors in (9) may be expressed as $\Pi^*(\Theta, \Sigma) = P_{\mathcal{S}}(\Sigma) \times \prod_{k=1}^{K} P_{\Theta_k}(A_k)$. As under the PDPM-VAR, specification (9) induces a prior on the space of densities $\mathcal{F}$ via

$$f_P(X) = \sum_{h_{11}, \ldots, h_{1K}=1}^{\infty} \sum_{h_\sigma=1}^{\infty} \pi_{\sigma,h_\sigma} \Big(\prod_{k=1}^{K} \pi_{k,h_{1k}} \Big) \prod_{t=1}^{T} \phi_{\Sigma_{h_\sigma}} \Big(\mathbf{x}_t - \sum_{k=1}^{\min\{t-1,K\}} A_{k,h_{1k}} \mathbf{x}_{t-k} \Big), \quad (10)$$

where $\pi_{k,h_{1k}} = \nu_{k,h_{1k}} \prod_{l_{k,1k} < h_{k,1k}} (1 - \nu_{k,l_{1k}})\ (k = 1, \ldots, K)$, $\pi_{\sigma,h_\sigma} = \nu_{\sigma,h_\sigma} \prod_{l_2 < h_\sigma} (1 - \nu_{\sigma,l_2})$ and $\nu_{k,h_{1k}} \sim Be(1, \alpha_{1k}), \nu_{\sigma,h_\sigma} \sim Be(1, \alpha_2)$, and further $vec(A_{k,h_{1k}}) \sim P^*_{1k}, \Sigma_{h_\sigma} \sim P^*_2$ for $k = 1, \ldots, K$, using the stick-breaking construction in Sethuraman (1994). We denote the model under (3) and (9) as the lag-generalized product of DP mixture of VAR (lgPDPM-VAR) model and note that this model reduces to the PDPM-VAR for lag 1 models. This approach is expected to be less flexible compared to the rgPDPM-VAR method in general, but may exhibit some advantages when the clustering patterns are distinct across lags.

3.2 Theoretical Properties

Notations and Definitions: In this section we will establish posterior consistency properties of the proposed product of DP mixture of VAR models. We will assume that the $D \times T$ data matrices $X_1, \ldots, X_n$, are i.i.d. under some true density $f_0 \in \mathcal{F}$. We note that the theoretical derivations corresponding to VAR models involving multivariate time-series data are more involved than the independent multivariate outcome settings in Section 2 that has been the focus of existing non-parametric Bayesian density estimation literature. Moreover, our theoretical results assume fixed T (finite time set-up) with growing number of samples, which is in contrast to theoretical settings in parametric VAR analysis for single subjects that rely on growing T (Ghosh et al., 2018).

Throughout the article, we will assume the following reasonable regularity conditions on $f_0(\mathbf{x}_t \mid X_{1:(t-1)})$ that reflect counterparts of the assumptions (C1)-(C4) corresponding to multivariate density estimation. Here, $f_0(\mathbf{x}_t \mid X_{1:(t-1)})$ denotes the true conditional density of $\mathbf{x}_t$ that depends on previous time scans up to a certain known lag (K).

(A0) The form of the true density satisfies $f_0(X) = \big\{\prod_{t=1}^{T} f_0(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1)\big\} = \big\{\prod_{t=1}^{T} f_0(\mathbf{x}_t \mid X_{1:(t-1)})\big\}$, for all $X \in \Re^{D\times T}$.
(A1) $0 < f_0(X) < M$ for some constant M and for all $X \in \Re^{D\times T}$.
(A2) $|\int f_0\big(\mathbf{x}_t \mid X_{1:(t-1)}\big) \log\big(f_0(\mathbf{x}_t \mid X_{1:(t-1)})\big) d\mathbf{x}_t| < \infty$, point-wise for $X_{1:(t-1)}$ for all t.
(A3) For all t and some $\delta > 0$, $\int f_0\big(\mathbf{x}_t \mid X_{1:(t-1)}\big) \log\big(\frac{f_0(\mathbf{x}_t|X_{1:(t-1)})}{\phi^*_\delta(\mathbf{x}_t|X_{1:(t-1)})}\big) d\mathbf{x}_t < \infty$, where $\phi^*_\delta(\mathbf{x}_t \mid X_{1:(t-1)}) = \inf_{||\mathbf{r}-\mathbf{x}_t||<\delta} f_0(\mathbf{r} \mid X_{1:(t-1)})$, point-wise for $X_{1:(t-1)}$.
(A4) For all t and some $\eta > 0$, $\int ||\mathbf{x}_t||^{2(1+\eta)} f_0(\mathbf{x}_t \mid X_{1:(t-1)}) d\mathbf{x}_t < \infty$, point-wise for $X_{1:(t-1)}$.

Condition *(A0)* expresses the true density as a product of conditional densities, subject to a known K, where the true conditional density only depends on $\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots, \mathbf{x}_{t-K}$, when $t > K$ and depends on $\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots, \mathbf{x}_1$ for $t \leq K$. Condition *(A1)* assumes that the true density is bounded. Further, assumptions (A1)-(A4) are reminiscent of conditions used for conditional density estimation in Pati et al. (2013) who focused on dependent stick-breaking processes. In the special case when the true density corresponds to a VAR structure, *(A0)-(A4)* would imply (among other things) that the true VAR parameters are well-behaved and satisfy stability conditions so that the true density does not blow up to ∞ or attenuate to zero.

The following Theorem formally states the result on positive prior support under the above assumptions. The proof is provided in the Appendix and uses key results in Wu and Ghosal (2008) for multivariate density estimation under DP mixtures.

Theorem 3: *Suppose assumptions* $(A0) - (A4)$ *are satisfied. Then the product of DP mixture priors* Π *specified in (5), (7), and (9) satisfies the Kullback-Leibler property, i.e.* $\Pi\Big(f \in \mathcal{F} : \int \log\big(f_0/f\big) f_0 \leq \eta^*\Big) \geq 0$, *for any* $\eta^* > 0$.

The next goal is to establish strong consistency for the proposed approach. We will again leverage the sufficiency conditions in Theorem 2 that rely on careful sieve constructions. In practice, it may not be straightforward to construct such sieves for the matrix-variate density estimation case, since the metric entropy depends on a number of terms including the sample size n, dimension D, as well as T (see Theorem 4). A major contribution of our work is to construct appropriate sieves using the stick-breaking representation and inspired by the ideas implemented in Shen et al. (2013), which satisfy the conditions in Theorem 2.

Sieve Constructions: The sieves are constructed so as to allow the norm of the elements in the autocovariance matrices, as well as the condition number of the residual covariance matrices, to increase with sample size at an appropriate rate that satisfies the conditions in Theorem 2. We note that the condition number of a matrix frequently appears in the random matrix literature (Edelman, 1988) and is defined as the ratio of the largest to the smallest eigen values, i.e. $\lambda_1(\Sigma)/\lambda_D(\Sigma) = \lambda_1(\Sigma^{-1})/\lambda_D(\Sigma^{-1})$. For our purposes, we construct the following sieves corresponding to the PDPM-VAR model in (3) and (5) as:

$$\mathcal{F}_n = \Bigg\{ f_p : P = \sum_{h_1 \geq 1} \sum_{h_\sigma \geq 1} \pi_{h_1} \pi_{\sigma,h_\sigma} \delta_{\Theta_{h_1}, \Sigma_{h_\sigma}} : \sum_{h_1 > H_n} \pi_{h_1} < \epsilon_1, \sum_{h_\sigma > H_n} \pi_{\sigma,h_\sigma} < \epsilon_2, \text{and for} \\ h_\sigma \leq H_n, \ \underline{\sigma}_n^2 \leq \lambda_{D,\Sigma_{h_\sigma}} \leq \lambda_{1,h_\sigma} \leq \underline{\sigma}_n^2 (1 + \epsilon/\sqrt{D})^{M_n}, \ 1 < \frac{\lambda_{1,h_\sigma}}{\lambda_{D,h_\sigma}} \leq n^{H_n} \Bigg\},$$

$$\mathcal{F}_{n,\mathbf{jl}} = \Bigg\{ f_p \in \mathcal{F}_n : \text{for } h_1, h_\sigma \leq H_n, \underline{a_{h_1,j}} \leq ||vec(A_{k,h_1})|| \leq \bar{a}_{h_1,j} \ \forall k, \ \underline{u_{h_\sigma,l}} \leq \frac{\lambda_{1,h_\sigma}}{\lambda_{D,h_\sigma}} \leq u_{h_\sigma,l} \Bigg\}, \tag{11}$$

where δ_θ denotes the probability measure degenerate at θ, $\lambda_{d',\Sigma_{h_\sigma}}$ is a shorthand for $\lambda_{d'}(\Sigma_{h_\sigma})$, i.e. the eigen values corresponding to Σ_{h_σ}, j, l are integers that are $\leq H_n$ for a given n, the sequences $\{H_n\}, \{M_n\}, \{\underline{\sigma}_n\} \{\underline{a_{h_1,j}}\}, \{\bar{a}_{h_1,j}\}, \{\underline{u_{h_\sigma,j}}\}, \{u_{h_\sigma,j}\}$ grow to ∞ with n and are chosen appropriately such that $\mathcal{F}_n \subset \cup_{\mathbf{j},\mathbf{l}} \mathcal{F}_{n,\mathbf{jl}}$, and further, $\mathcal{F}_n \uparrow \mathcal{F}$ as $n \to \infty$. Moreover, the sieves corresponding to rgPDPM-VAR in (3) and (7) are constructed as:

$$\mathcal{F}_n = \Bigg\{ f_p : P = \sum_{h_{1,1}=1}^{\infty} .. \sum_{h_{1,D}=1}^{\infty} \sum_{h_\sigma=1}^{\infty} (\pi_{\sigma,h_\sigma} \prod_{d'=1}^{D} \pi^*_{d',h_{1d'}}) \delta_{\Theta_{h_{1d'}}, \Sigma_{h_\sigma}} : \sum_{h_{1d'} > H_n} \pi^*_{d',h_{1d'}} < \epsilon_1, \ \forall d' \leq D,$$

$$\sum_{h_\sigma > H_n} \pi_{\sigma,h_\sigma} < \epsilon_2, \text{ and for } h_\sigma \leq H_n, \ \underline{\sigma}_n^2 \leq \lambda_{D,h_\sigma} \leq \lambda_{1,h_\sigma} \leq \underline{\sigma}_n^2 (1 + \epsilon/\sqrt{D})^{M_n}, \ 1 < \frac{\lambda_{1,h_\sigma}}{\lambda_{D,h_\sigma}} \leq n^{H_n} \Bigg\},$$

$$\mathcal{F}_{n,\mathbf{jl}} = \Bigg\{ f_p \in \mathcal{F}_n : \underline{a}_{h_{1d'},j} \leq ||vec(A_{k,d'\bullet,h_{1,d'}})|| \leq \bar{a}_{h_{1d'},j} \text{ for } h_{11}, \ldots, h_{1D} \leq H_n,$$

$$\text{and } d' = 1, \ldots, D, \text{ and } \underline{u}_{h_\sigma,l} \leq \frac{\lambda_{1,h_\sigma}}{\lambda_{D,h_\sigma}} \leq u_{h_\sigma,l}, \text{ for } h_\sigma \leq H_n \Bigg\}, \tag{12}$$

and the sieves for the lgPDPM-VAR model in (3) and (9) are constructed similarly as:

$$\mathcal{F}_n = \Bigg\{ f_p : P = \sum_{h_{11}=1}^{\infty} \ldots \sum_{h_{1K}=1}^{\infty} \sum_{h_\sigma=1}^{\infty} \pi_{\sigma,h_\sigma} \big(\prod_{k=1}^{K} \pi_{k,h_{1k}} \big) \delta_{\Theta_{h_{1k}}, \Sigma_{h_\sigma}} : \sum_{h_{1,1k} > H_n} \pi_{k,h_{1k}} < \epsilon_1, \forall k = 1, \ldots, K,$$

$$\sum_{h_\sigma > H_n} \pi_{\sigma,h_\sigma} < \epsilon_2, \text{ and for } h_\sigma \leq H_n, \ \underline{\sigma}_n^2 \leq \lambda_D(\Sigma_{h_\sigma}) \leq \lambda_1(\Sigma_{h_\sigma}) \leq \underline{\sigma}_n^2 (1 + \epsilon/\sqrt{D})^{M_n}, \ 1 < \frac{\lambda_{1,h_\sigma}}{\lambda_{D,h_\sigma}} \leq n^{H_n} \Bigg\},$$

$$\mathcal{F}_{n,\mathbf{jl}} = \Bigg\{ f_p \in \mathcal{F}_n : \underline{a}_{h_{1k},j} \leq ||vec(A_{k,h_{1k}})|| \leq \bar{a}_{h_{1k},j} \text{ for all } h_{1k} \leq H_n, \underline{u}_{h_\sigma,l} \leq \frac{\lambda_{1,h_\sigma}}{\lambda_{D,h_\sigma}} \leq u_{h_\sigma,l}, \ h_\sigma \leq H_n \Bigg\}, \tag{13}$$

where $\mathcal{F}_n \subset \cup_{\mathbf{j},\mathbf{l}} \mathcal{F}_{n,\mathbf{jl}}$ and $\mathcal{F}_n \uparrow \mathcal{F}$ as $n \to \infty$, and it is understood that the sequences $\{H_n\}, \{M_n\}, \{\underline{\sigma}_n\}, \{\underline{a_{h_1,j}}\}, \{\bar{a}_{h_1,j}\}, \{\underline{u_{h_\sigma,l}}\}, \{u_{h_\sigma,l}\}$ are chosen appropriately and can be specific to sieves corresponding to PDPM-VAR, lgPDPM-VAR or rgPDPM-VAR. The following results establish entropy bounds that are vital to establishing strong consistency.

Theorem 4: *The entropy bound for sieves (11) satisfies* $N(\epsilon, \mathcal{F}_{n,\mathbf{jl}}, ||\cdot||_1) \lesssim \left(\frac{M^D}{\epsilon^{-C_1}} \right)^{H_n} \times \prod_{h_\sigma \leq H_n} \left\{ \frac{2D u_{h_\sigma,l}}{\epsilon^2} \right\}^{D(D-1)/2} \times \prod_{h_1 \leq H_n} \left\{ \left(\frac{C^*_{h_{1,j},h_{\sigma,l}} \bar{a}_{h_1,j}}{\underline{\sigma}_n \epsilon} + 1 \right)^{D^2} - \left(\frac{C^*_{h_{1,j},h_{\sigma,l}} \underline{a}_{h_1,j}}{\underline{\sigma}_n \epsilon} - 1 \right)^{D^2} \right\}^K$, *where constants* $C_1 > 0$ *and* $C^*_{h_{1,j},h_{\sigma,l}} > 0$ *depend on* (D, T, K).

Corollary 1: *The entropy for sieves in (12) and (13) corresponding to rgPDPM-VAR and lgPDPM-VAR respectively, satisfy* $N(\epsilon, \mathcal{F}_{n,\mathbf{jl}}, ||\cdot||_1) \lesssim \mathcal{K}^* \left(\frac{M^D}{\epsilon^{-C_1}}\right)^{H_n} \times \prod_{h_\sigma \leq H_n} \left\{\frac{2Du_{h_\sigma,l}}{\epsilon^2}\right\}^{D(D-1)/2}$, *where* $\mathcal{K}^*$ *is understood to vary depending on the specific variant of the PDPM model used.*

Having established the entropy bounds, the next step is to propose sensible base measures that satisfy the tail conditions and summability constraints in Theorem 2. These base measures include some commonly used choices as discussed in the sequel.
(B1) The base measures corresponding to $P_{\mathcal{S}}$ in (5), (7) and (9) satisfy $P_2^*(\lambda_1(\Sigma_{h_\sigma}^{-1}) > x^*) \lesssim \exp(-c_1(x^*)^{c_2}), P_2^*(\lambda_D(\Sigma_{h_\sigma}^{-1}) < 1/x^*) \lesssim (x^*)^{-c_3}, P_2^*\big(\frac{\lambda_1(\Sigma_{h_\sigma}^{-1})}{\lambda_D(\Sigma_{h_\sigma}^{-1})} > x^*\big) \lesssim (x^*)^{-\kappa}$, for some positive constants c_1, c_2, c_3, κ, and corresponding to the cluster h_σ.
(B2) The base measure corresponding to P_Θ specifies independence across lags, and satisfies the following tail conditions: (i) under PDPM-VAR, $P_1^*(||vec(A_{k,h_1})|| > x^*) \lesssim (x^*)^{-2(r+1)}$ for cluster h_1; (ii) under lgPDPM-VAR, $P_{1k}^*(||vec(A_{k,h_{1k}})|| > x^*) \lesssim (x^*)^{-2(r+1)}$ for cluster h_{1k}; and (iii) under rgPDPM-VAR, $P_{1d'}^*(||vec\{A'_{1,d'\bullet,h_{1,d'}}, \ldots, A'_{K,d'\bullet,h_{1,d'}}\}|| > x^*) \lesssim (x^*)^{-2(r^*+1)}$ corresponding to cluster $h_{1d'}$, for some constants $r, r^* > 0$, and $d' = 1, \ldots, D$.

The above conditions on the base measures are very reasonable and hold for commonly used distributions on autocovariance matrices (such as Gaussian and Laplace), as well as inverse-Wishart distribution corresponding to P_2^*. These tail conditions are also satisfied by certain low rank decompositions for the covariance, such as a factor model form ($\Sigma = \Lambda\Lambda^T + \Omega$ where the $D \times B$ matrix Λ contains $B << D$ factor loadings), which is particularly suitable for scaling up the approach to higher dimensions. Such low rank representations are routinely used for dimension reduction in the factor model literature (Ghosh and Dunson, 2009). Denote $\mathcal{A}_{d',h_{1,d'}} = vec\{A'_{1,d'\bullet,h_{1,d'}}, \ldots, A'_{K,d'\bullet,h_{1,d'}}\}$ and let DE denote a double exponential prior. The following Lemmas formalize the above discussions on the base measures.

Lemma 2: *Condition (B2) holds when* $P_1^*(vec(A_{k,h_1}))$ *is specified as* $N_{D^2}(vec(A_k); \boldsymbol{\mu}, \Lambda)$ *with* $\Lambda \sim IW(\Lambda_0, \nu_\lambda)$ *corresponding to PDPM-VAR, and for a similar choice of* $P_{1k}^*(vec(A_{k,h_{1k}}))$ *under lgPDPM-VAR. It is also satisfied when* $P_{1d'}^{**}(\mathcal{A}_{d',h_{1,d'}}) = \prod_{k=1}^K N_D(A_{k,d'\bullet,h_{1,d'}}; \boldsymbol{\mu}, \Lambda_{d'}), \Lambda_{d'} \sim IW(\Lambda_{0d'}, \nu_{\lambda,d'})$, *under rgPDPM-VAR. Further, (B2) also holds if the above base measures are changed to a product of independent DE(*λ*) priors with suitably large* λ.

Lemma 3: *Condition (B1) holds when for cluster* h_σ, $P_2^*(\Sigma_{h_\sigma}) = IW(\Sigma_{h_\sigma}; \Sigma_0, \nu_\sigma)$, *as well as under the low rank representation* $\Sigma_{h_\sigma} = \Gamma_{h_\sigma}\Gamma_{h_\sigma}^T + \Omega_{h_\sigma}$ *where* Γ_{h_σ} *is* $D \times B$ *and* $\Omega_{h_\sigma} = diag(\sigma_{1,h_\sigma}^2, \ldots, \sigma_{D,h_\sigma}^2)$, *and* $P_2^*(\Sigma_{h_\sigma}) = \big\{\prod_{d'=1}^D \prod_{m'=1}^B N(\gamma_{d'm',h_\sigma}; 0, 1)\big\}\big\{\prod_{j=1}^D Ga(\sigma_{j,h_\sigma}^{-2}; a_\sigma, b_\sigma)\big\}$.

The proof of Lemma 2 is provided in the Appendix, while that of Lemma 3 follows directly from Corollaries 1 and 2 in Canale and De Blasi (2017). We note that $B << D$ in Lemma 3 ensures a reduced rank structure on the residual covariance matrix.

One can now use the entropy bounds derived in Theorem 4 and Corollary 1 along with tail conditions in *(B1)-(B2)* to establish our strong consistency under a broad class of base measures, by applying Theorem 2. Our strong consistency result is stated below.

Theorem 5: *Suppose Theorem 3 holds, and (B1)-(B2) are satisfied. Then for suitably large constants* r, r^*, κ, *the posterior distributions corresponding to the PDPM-VAR, lgPDPM-*

VAR and rgPDPM-VAR are strongly consistent at f_0 under suitable choice of sequences $\{H_n\}, \{M_n\}, \{\underline{\sigma}_n\}\{\underline{a_{h_1,j}}\}, \{\bar{a}_{h_1,j}\}, \{\underline{u_{h_\sigma,j}}\}, \{u_{h_\sigma,j}\}$ in the sieves (11), (12), and (13).

Remark 3: In mathematical terms, strong posterior consistency can be written as $\Pi(\{f : d(f, f_0) > \epsilon_f\} \mid X^{(1)}, \ldots, X^{(n)}) \to 0$ as $n \to \infty$ in F_0^n probability for any $\epsilon_f > 0$.

Remark 4: While Theorem 5 is stated in terms of general class of base measures that satisfy *(B1)-(B2)*, we rely on commonly used base measures outlined in Lemmas 2-3 for implementing the proposed approach. We elaborate these choices in the next section.

4. Posterior Computation

We outline the posterior computation steps to fit all proposed VAR models that is the main focus of this work. Our approach alternates between sampling parameters related to the autocovariance matrices and the residual covariance matrix. For all models, we update the autocovariance parameters row-wise for one outcome at a time. For the PDPM-VAR, rgPDPM-VAR, and lgPDPM-VAR we use a hierarchical representation of Laplace base measures (Park and Casella, 2008). Under these base measures, these autocovariance elements follow independent $DE(\lambda)$ distributions (Park and Casella, 2008). Explicit details are provided in Appendix C.

In order to scale up the implementation of the proposed method to high dimensional applications, we use a reduced rank factor model representation for the residual covariance matrix in our implementation, which provides a desired balance between computational scalability and theoretical flexibility. In particular, such a low rank structure on the residual covariance does not adversely impact the accuracy of parameter estimates compared to an unstructured covariance matrix, in our experience involving extensive numerical experiments with true unstructured residual covariances. Moreover, the results for estimation of the autocovariance terms are not particularly sensitive to the choice of rank. Further, it is considerably more flexible and results in greater accuracy compared to a diagonal residual covariance that is routinely used in VAR literature (Kook et al., 2021) but may be restrictive in practical applications. In particular, we specify $\Sigma_i = \mathbf{\Gamma}_i\mathbf{\Gamma}_i' + \mathbf{\Psi}_i$, where $\mathbf{\Gamma}_i$ is a $D \times B$ factor loadings matrix with $B(<< D)$ factors, and $\mathbf{\Psi}_i$ is diag$\{\sigma_{i,1}^2, \ldots, \sigma_{i,D}^2\}$.To facilitate posterior computation, we use the following parameter expanded version of the model,

$$\mathbf{x}_{i,t} = \sum_{k=1}^{\min\{t-1,K\}} A_{ik}\mathbf{x}_{i,t-k} + \mathbf{\Gamma}_i^*\boldsymbol{\eta}_{i,t}^* + \boldsymbol{\epsilon}_{i,t}^*,\ \boldsymbol{\eta}_{i,t}^* \sim N(0, \mathbf{\Xi}_i),\ \boldsymbol{\epsilon}_{i,t}^* \sim N(0, \mathbf{\Psi}_i), \tag{14}$$

where $\mathbf{\Xi}_i = \text{diag}\{\ \xi_{i,1}, \ldots,\ \xi_{i,B}\}$. Under the low rank representation, we impose DP mixture priors on $(\mathbf{\Gamma}_i^*,\ \mathbf{\Xi}_i,\ \mathbf{\Psi}_i)$ leading to a mixture prior on Σ_i. This corresponds to the prior $\Sigma_i \sim \sum_{h_\sigma=1}^{\infty} \pi_{\sigma,h_\sigma}\delta_{(\mathbf{\Gamma}_{h_\sigma}^*, \mathbf{\Xi}_{h_\sigma}, \mathbf{\Psi}_{h_\sigma})}$, where $(\mathbf{\Gamma}_{h_\sigma}^*, \mathbf{\Xi}_{h_\sigma}, \mathbf{\Psi}_{h_\sigma}) \sim P_2^* \equiv P_{\mathbf{\Gamma}^*} \times P_{\mathbf{\Xi}} \times P_{\mathbf{\Psi}}$. Here $P_{\mathbf{\Gamma}^*}$ is a product of independent standard normal distributions, $P_{\mathbf{\Xi}}$ is a product of independent $Gamma(1/2, 1/2)$ distributions yielding a half-Cauchy prior on the diagonal elements of $\mathbf{\Gamma}$ and a Cauchy prior on the lower-off-diagonal elements as in Ghosh and Dunson (2009), and the inverse of the diagonal elements of $\mathbf{\Psi}$ have independent $Gamma(\alpha_\sigma, \beta_\sigma)$ priors.

Computational Cost of the PDPM-VAR Approaches: The proposed approaches are quite efficient as long as the number of nodes is not overly large. The computation is driven by

the need to sample the rows of A_i and the terms in the low rank representation for Σ_i as per equation (14). We briefly discuss the computational costs below. In practice, the computational costs are quite manageable, as demonstrated by the HCP analysis in the sequel.

First, for the PDPM-VAR and rgPDPM-VAR, the rows of A_i are updated one at a time, separately for each cluster. This corresponds to performing D draws, each from a DK-variate multivariate normal distribution, thus updating A_i requires D $\mathcal{O}(D^3K^3)$ steps for the PDPM-VAR and rgPDPM-VAR. We note that this can also be parallelized over the outcomes as a possible direction for future work. The draw for the lgPDPM-VAR is more complicated. Although we still sample one row of A_i at a time, we now must sample across all clusters, since different subjects can belong to different collections of clusters for a given row due to the lag-specific clustering structure. This means that we must draw from a $H_{all}D$ dimensional multivariate normal distribution, where H_{all} is the total number of clusters across all lags. Therefore, this update requires D $\mathcal{O}(H_{all}^3D^3)$ computational steps. We note that if one does not use the low rank decomposition in (14) but instead imposes a diagonal residual covariance structure, then the computational complexity remains similar.

In the scenario that an unstructured residual covariance structure is imposed but without a low rank decomposition, the computational complexity rapidly increases to $\mathcal{O}(HD^6K^3)$ due to the need to sample from a KD^2-variate normal distribution for each cluster. This may result in prohibitive computational costs for higher dimensions. Therefore, sampling from the low rank representation in model (14) is desirable, especially given that sampling the latent factors and their loading is computationally straightforward. In particular, each subject has a $B \times T_i$ dimensional matrix of latent factors to sample, where B is generally small and the time points are independent. Thus this simplifies to sampling a set of T_i B-dimensional vectors, all of which have the same posterior covariance. Since T_i is large relative to B, this means that the draw for the η_i terms is driven by the cost to generate standard normal draws (each requiring $\mathcal{O}(1)$ operations), and subsequently multiply them with the lower triangular matrix from the Cholesky factorization of the posterior covariance of size $B \times B$. Given that this must only be calculated once per cluster, the cost per subject is $\mathcal{O}(DBT)$. Thus the overall cost across all samples is $\mathcal{O}(HB^3) + \mathcal{O}(nDBT)$.

5. Simulation Studies

We compared the performance under the proposed approaches to a state-of-the-art single-subject VAR approach, as well as an ad-hoc clustering extension of the single subject VAR model that is able to borrow information across samples. We generated data for $n = 100$, 200, $D = 100$, $T_i = 250$, and different levels of sparsity within the autocovariance matrices were considered (75% and 90%). For each data generation setting we generate 25 simulation replicates, and for all settings the true VAR model involved $K = 2$ lags. We consider four settings for generating the subject-level autocovariance matrices that differ with respect to the clustering structure. Settings 1-3 represent the PDPM-VAR, lgPDPM-VAR and rgPDPM-VAR scenarios respectively, while Setting 4 represents a more heterogeneous setting that is obtained via introducing additional random noise to the autocovariance elements generated under Setting 3. For Setting 1, we use 3 autocovariance clusters, for Setting 2 we use 3 clusters for lag 1 and 2 clusters for lag 2, and for Settings 3-4 we vary

the true number of clusters randomly (between $2-5$) across rows of the autocovariance matrix, and the elements of these matrices are generated randomly in order to ensure a stable time series. In Setting 3, subjects within a cluster share the exact same elements for the corresponding rows of the autocovariance matrix, whereas in Setting 4 the subject-level rows of the autocovariance matrix within a cluster are random deviations from a shared mean row. Each cluster's residual covariance matrix was generated from an Inverse Wishart distribution with D degrees of freedom and diagonal scale matrix with elements equal to $D/2$. Subject level time courses were obtained by starting with random values for the multivariate observation at the first time point, and subsequently generating future observations from the assumed true VAR model. For a subject, an additional 5 time scans were generated after the initial T_i observations to evaluate forecasting accuracy.

5.1 Approaches and Performance Metrics

We compare the proposed approaches to the single subject Bayesian VAR (SS-VAR) model developed in Ghosh et al. (2018), which separately models the time courses for each subject. We also consider a two-stage clustering extension of this method, where we first estimated subject specific autocovariance and residual covariances under the single VAR approach by Ghosh et al. (2018) and then applied the k–means clustering separately to the vectorized autocovariance and residual covariance matrix estimates. We choose the number of clusters to maximize the silhouette score (Rousseeuw, 1987) and we then allocate each sample to one of the k clusters that is based on both the autocovariance terms and the residual covariance estimates from the initial SS-VAR fit. We subsequently concatenate the time courses across all subjects within the same cluster in order to borrow information within cluster, and finally re-fit the SS-VAR model to this concatenated data separately for each cluster. Since the true clustering structure was assumed to be unknown when fitting the model, it was not possible to compare the performance with existing multi-subject VAR modeling methods that assume known groupings (Chiang et al., 2017; Kook et al., 2021).

We evaluate performance in terms of (1) autocovariance estimation accuracy, (2) clustering accuracy, (3) feature selection for identifying structural zeros in the autocovariance matrices, and (4) forecasting accuracy. Following Ghosh et al. (2018), we measure estimation accuracy using the relative L2 error of the estimates to the true estimates. Clustering accuracy is measured using the adjusted Rand index (Rand, 1971), which measures agreement between the assigned and true cluster labels, adjusted for chance agreement. Feature selection performance is evaluated via area under the receiver operating characteristic (RoC) curve and precision recall curve (PRC). To calculate both curves we considered a sequence of significance thresholds, and for each threshold, we examined the corresponding credible interval to infer the significance. The corresponding sequence of sensitivity versus 1-specificity values were plotted over varying thresholds in order to obtain the ROC curve, while the PRC was obtained by plotting the positive predictive value ($1-FDR$) against sensitivity (FDR denotes the false discovery rate). Finally, forecasting accuracy is measured via the relative L2 error of the predicted time courses for time scans $T_i+1,\ldots,T_i+5$. The MCMC chains converged for all methods as assessed using Dickey-Fuller tests of stationarity, although the results are not displayed due to space constraints.

5.2 Simulation Results

Simulation results are presented in Figures 2–3. Due to space constraints, we provide the simulation results for the most challenging case ($D = 100$, $T = 250$) at the 75% sparsity level here, but the results under the 90% sparsity case were quite similar. Several general patterns are clear from the results. First, the clustering performance for the autocovariance depends heavily on the true clustering structure (Figure 2, Panel A), with the PDPM-VAR, lgPDPM-VAR, and rgPDPM-VAR generally outperforming the other approaches when the data is generated from Settings 1-3 respectively. However, the rgPDPM-VAR often has close to optimal clustering performance when the PDPM-VAR is the true model and it also performs the best for the heterogeneous Setting 4, which reflects the generalizability of this variant. Critically, when there are differences in clustering across the different outcomes corresponding to more heterogeneous scenarios (Settings 3-4), only the rgPDPM-VAR is able to achieve a good clustering score. Finally, across all settings, the SS-VAR with clustering has the worst performance, demonstrating that the ad hoc two-stage analysis procedure is not able to accurately pool information across subjects.

The areas under the ROC and PR curves (Figure 2, Panels C and D) illustrate a consistently superior feature selection performance under the three proposed variants compared to the single subject VAR model with and without clustering. As expected, the PDPM-VAR, lgPDPM-VAR and rgPDPM-VAR approaches have higher area under the ROC and PR curves when the data is generated from Settings 1-3 respectively. In addition, the rgPDPM-VAR often has comparable area under the curve with PDPM-VAR for $n = 200$ under Setting 1 and the best performance under the more heterogeneous Setting 4. These results imply the ability of rgPDPM-VAR to accurately identify the sparsity structure of the autocovariance with a low risk of false discoveries for data with unknown clustering.

When estimating the residual covariance matrices (Figure 3, Panel A), all three proposed approaches are able to heavily outperform the SS-VAR model. The performance under the three proposed approaches is generally comparable, with the rgPDPM-VAR outperforming the others in the more heterogeneous Settings 3 and 4. In addition, the SS-VAR approach with initial clustering has a higher relative error compared to the rgPDPM-VAR for the vast majority of cases, although it occassionally has a slightly improved performance in Setting 3. We conjecture that this is due to the assumed full rank structure for the residual covariance that is modeled via an inverse-Wishart distribution under the SS-VAR, which aligns with the true data generation scenario, in contrast to the assumed low-rank structure on the PDPM-VAR. Unfortunately, the SS-VAR approach with clustering has extremely poor performance in terms of autocovariance estimation (Figure 3 B), while the PDPM-VAR, lgPDPM-VAR, and rgPDPM-VAR approaches typically have the lowest errors when the data is generated from Settings 1-3 respectively. The rgPDPM-VAR method also has the best autocovariance estimation performance under the more heterogeneous Setting 4.

Figure 2 Panel B displays the forecasting error for each of the autocovariance clustering setups, averaged over the sparsity level and the number of time points per subject. With the exception of Setting 2 where lgPDPM-VAR performs best, the rgPDPM-VAR approach has the best or close to optimal forecasting performance for other settings. Moreover, in the more heterogeneous Settings 3-4, the SS-VAR method with initial clustering has better forecasting accuracy compared to the PDPM-VAR and lgPDPM-VAR approaches, although

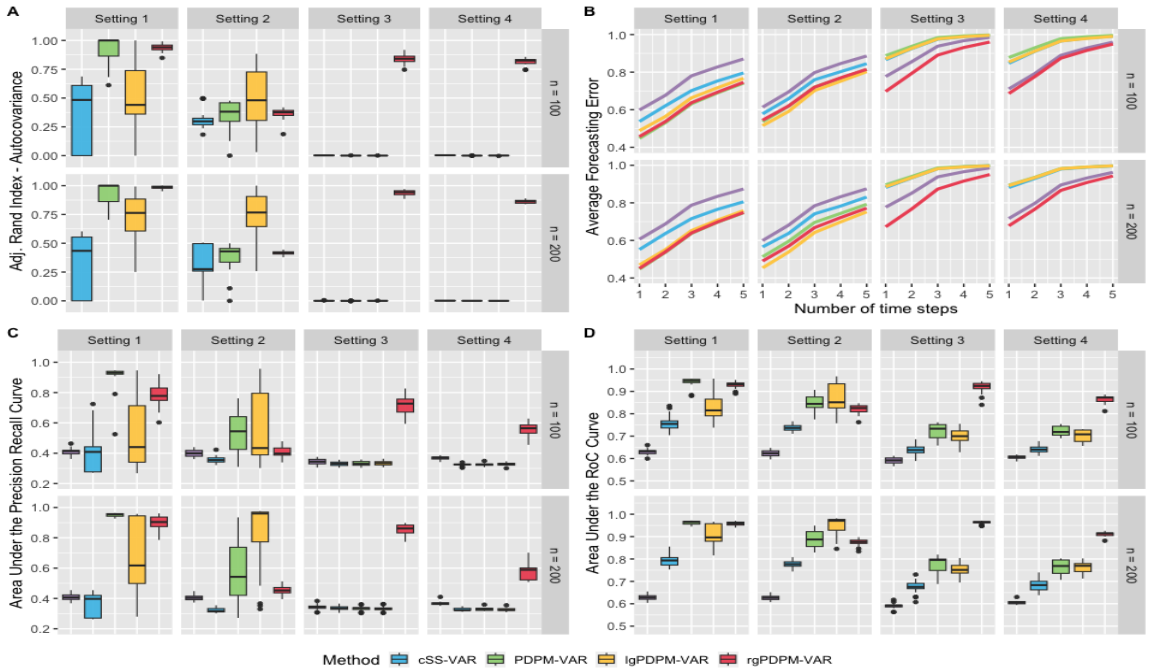

Figure 2: Simulation results for $D = 100$, $T = 250$ case with sparsity level 0.75. Panel A displays the adjusted Rand index for clustering the autocovariance. Panel B displays the forecasting error. Panels C and D display the area under the PR and RoC curves for identifying autocovariance non-zero elements.

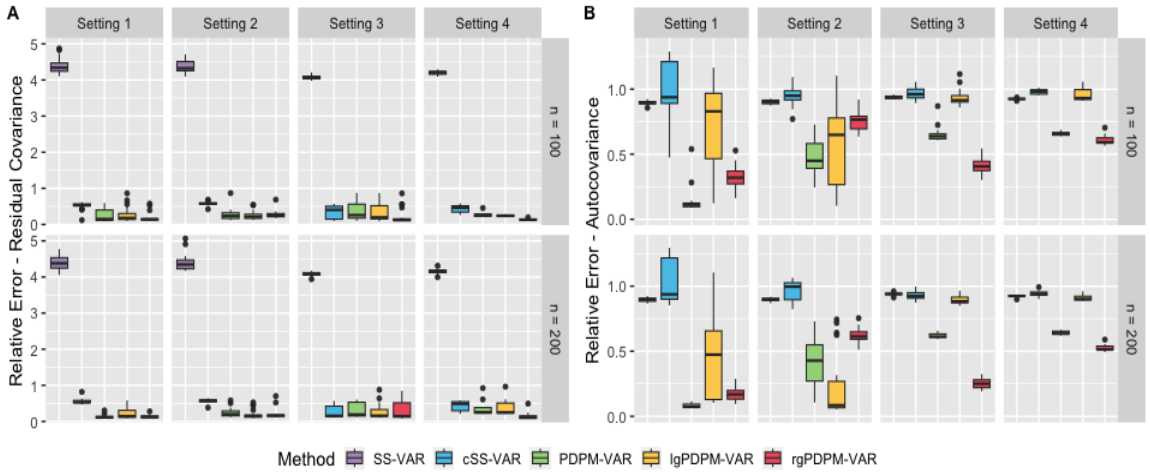

Figure 3: Relative L1 error for estimating the residual covariance (Panel A) and the subject-specific autocovariance matrices (Panel B) for $D = 100$, $T = 250$ case with sparsity level 0.75.

it can not outperform the rgPDPM-VAR method. The relative forecasting performance levels off for greater than three time steps for all approaches, as expected.
MCMC Diagnostics: The MCMC was implemented via a Gibbs sampler and exhibited good mixing as measured by the effective sample size (Appendix Section 12). Moreover, we varied the dimension of the low rank representation of the covariance and found that the results were not particularly sensitive to this choice, however these results are excluded due to space constraints.
Synopsis of findings: Overall, the rgPDPM-VAR provides a desirable balance between model parsimony and accurate estimation and inference for various degrees of heterogeneity across samples. The advantages under the rgPDPM-VAR are most pronounced under the heterogeneous Settings 3 and 4, and it often has close to optimal performance in Setting 1 for larger n. This illustrates the advantages of pooling information across subjects, while accommodating varying levels of heterogeneity at the level of the rows of the autocovariance matrix. While the SS-VAR approach with ad-hoc clustering is also able to pool information, it is highly sensitive to the clustering accuracy in the first step, and it can not capture clustering uncertainty, resulting in inferior performance.

6. Analysis of Human Connectome Project Data

6.1 Analysis Description

We use the rgPDPM-VAR approach to investigate effective connectivity differences between individuals with high and low fluid intelligence (FI) using a subset of resting-state fMRI data from the Human Connectome Project. Preprocessing details for these data can be found in (Smith et al., 2013). We adopt the 360-region Glasser atlas for parcellation as in Akiki and Abdallah (2019), where each node has a corresponding time course with $T = 1200$. We centered and scaled the subject level time courses for each node before analysis, and verified that each node's time course was stationary using Dickey-Fuller tests. We grouped the brain nodes into one of 6 well known functional brain networks (Akiki and Abdallah, 2019). These networks corresponded to the central executive (67 nodes), default mode (96 nodes), dorsal salience (23 nodes), somatomotor (55 nodes), ventral salience (49 nodes), and the visual network (70 nodes). We fit a separate VAR model with lag-1 on each of these networks, which corresponds to six separate VAR analyses. We selected the lag 1 model following previous literature on VAR models applied to fMRI data (Kook et al., 2021), and based on the lower temporal resolution of the fMRI data. We restrict our analysis to a subset of samples with the highest 10% and lowest 10% fluid intelligence scores, with $n = 306$ samples. We note that the grouping information was only used for post-model fitting comparisons in effective connectivity across groups. We used 1500 burn-in and 3500 MCMC iterations.

To the best of our knowledge, our approach for analyzing fluid intelligence-related effective connectivity differences using heterogeneous multi-subject data is one of the first such attempts. Most existing approaches involve a single-subject VAR analysis, and subsequently these estimates are combined to estimate between-subject variations and examine group differences (Deshpande et al., 2009). There are a handful of approaches for estimating effective connectivity by pooling information across multiple subjects, however they assume known groups (Chiang et al., 2017) with limited heterogeneity within groups, and

have similar limitations as outlined in the Introduction. Our analysis using the rgPDPM-VAR model is able to compute effective connectivity for multiple samples without any given group labels and can account for heterogeneity in an unsupervised manner. We compare the performance with a SS-VAR approach that analyses each sample separately, and subsequently performs permutation tests to assess significant differences (10,000 permutations). For both methods, false discovery rate control was applied to obtained significant elements.

In addition to investigating effective connectivity differences, we are interested in the clustering reliability and biological reproducibility of our findings. We report clustering reliability over two distinct MCMC runs, that are designed to evaluate the reliability of the clusters discovered by rgPDPM-VAR. As discussed in the introduction, for heterogeneous multivariate measurements, one can expect a subset of nodes/rows to drive the clustering whereas for other nodes the clustering patterns likely hold little information. We calculate the ARI for the node-level clustering across the two MCMC runs to investigate this aspect of clustering reliability. To assess biological reproducibility, we conduct our VAR analysis for two scans collected from each individual using different phase-encodings (LR1 and RL1), with the expectation that the parameter estimates should be similar corresponding to the two scans. We examine the correlation of the estimated autocovariance elements across the two runs (LR1 and RL1) under both the SS-VAR and the rgPDPM-VAR, with high correlation providing evidence that the findings are reproducible.

6.2 Results

Figure 4 displays heatmaps of the significant autocovariance differences between the low and high FI groups under the rgPDPM-VAR, after appropriate FDR control. Several patterns are clear from Figure 4. First, the rgPDPM-VAR is able to identify a large number of significant differences between the two groups after FDR control. Second, the rgPDPM-VAR finds a large number of strong differences along the diagonal. These correspond to AR(1) coefficients, and it seems sensible that if there are differences between groups at Lag 1 that they would be strongly related to each nodes' own time course. Thirdly, the strongest differences were observed corresponding to the nodes in the Dorsal Salience network, as illustrated in Table 1. These nodes were identified by looking at columns of the autocovariance matrix with a large proportion of significant elements, which accounts for the varying sizes for the 6 networks. These findings are consistent with previous evidence, which have suggested the dorsal salience and attention networks to be highly related to fluid intelligence (Santarnecchi et al., 2017). We note that in contrast, only one significantly different effective connectivity difference between the high and low fluid intelligence groups was reported under the SS-VAR approach. Such results are clearly biologically implausible. Our overall findings point to the advantages of performing a multi-subject analysis accounting for heterogeneity, over a single subject analysis.

To examine biological reproducibility, the right–hand side of Figure 5 displays histograms of the correlations of the rows of the autocovariance matrices across the two analyses corresponding to the LR1 and RL1 fMRI scans, under the SS-VAR and rgPDPM-VAR. The estimates under the rgPDPM-VAR exhibit a very high degree of correlation, almost entirely > 0.8. On the other hand, the correlation for the majority of the elements is less than 0.5 under SS-VAR, with only 10 elements registering a correlation greater than

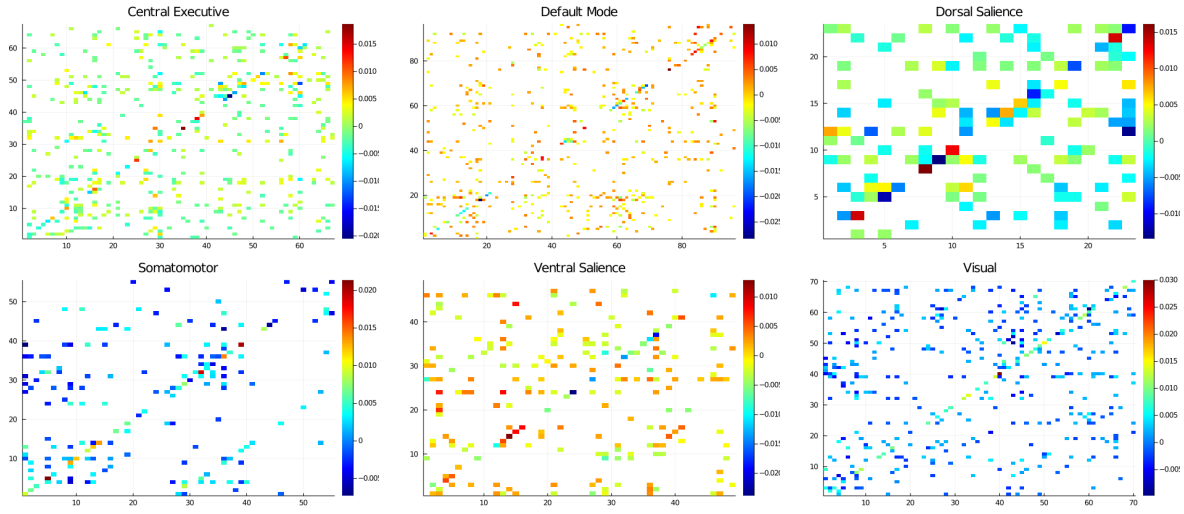

Figure 4: Elements of the autocovariance matrices exhibiting significant differences between the low and high FI groups. The color of the element represents the strength of the mean difference between groups (high FI − low FI), with white elements corresponding to non-significant elements.

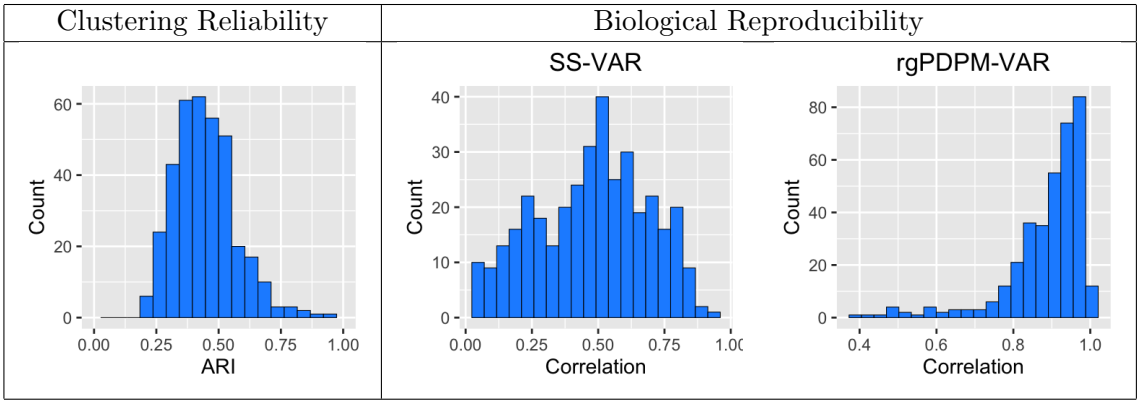

Figure 5: Adjusted Rand index across two runs of the analysis of the HCP LR1 data for assessing clustering reliability (left). Correlation between the rows of A_i across two different HCP data sets (right). The first run of the analysis was on the LR1 phase-encoding data and the second on RL1.

Node	Network	Prop. FI Diff.	Node	Network	Prop. FI Diff.
L_PF	Dorsal Salience	0.43	R_PFt	Dorsal Salience	0.26
R_7Am	Dorsal Salience	0.35	L_PEF	Dorsal Salience	0.26
L_IFSa	Dorsal Salience	0.35	L_TE2p	Dorsal Salience	0.26
L_PHT	Dorsal Salience	0.35	R_V3A	Visual	0.23
R_PHT	Dorsal Salience	0.30	R_PSL	Ventral Salience	0.22
R_PF	Dorsal Salience	0.30	R_7PL	Dorsal Salience	0.22
L_6a	Dorsal Salience	0.30	R_6r	Dorsal Salience	0.22
L_PFt	Dorsal Salience	0.30	L_7Am	Dorsal Salience	0.22
R_PGs	Central Executive	0.28	L_6r	Dorsal Salience	0.22
R_IFSa	Dorsal Salience	0.26	L_V3A	Visual	0.21

Table 1: Table of the 20 nodes with large proportion of significant effects on other nodes within their network. Note that the proportion is used instead of the raw count to account for the different network sizes.

0.8, which implied considerably lower reproducibility overall compared to the multi-subject analysis. Moreover a non-negligible number of nodes had weak reproducibility with correlations less than 0.25 under SS-VAR. In addition, Figure 5 (left) displays a histogram of the ARI for clustering each node across two separate runs of MCMC on the LR1 data, which illustrates clustering reliability. In general, most nodes exhibited 9–10 clusters. As hypothesized in the Introduction, we see a pattern in which a subset of nodes exhibited very high clustering reliability across runs (> 0.7), which supports our hypothesis that only some nodes contribute meaningfully towards clustering of samples. On the other hand, most of the nodes exhibited relatively moderate clustering reliability (ARI ≈ 0.5), which indicates much higher clustering than chance, but not fully consistent clustering across all subjects corresponding to these nodes. We note that given the strong biological reproducibility results, the moderate or low clustering reliability for a subset of nodes in our analysis should be attributed to the fact that these nodes are irrelevant to clustering. This provides further justification for using the rgPDPM-VAR, which is designed to accommodate exactly this kind of clustering structure. Finally, the computation for the analysis is very efficient. On average, a single MCMC iteration required 1.8 (centeral executive), 3.8 (default mode), 0.3 (dorsal salience), 1.26 (somatomotor), 1.01 (ventral salience), and 1.9 (visual) seconds on an 8 core 2021 M1 Macbook Air.

7. Forecasting of Air Quality Data

While the fMRI study described above focuses on connectivity between brain regions, fMRI studies are not generally concerned with forecasting accuracy. To demonstrate the forecasting accuracy of our method, we next apply the proposed methods to an open source air quality data set from the EPA (`https://www.epa.gov/outdoor-air-quality-data`). The data consist of daily measurements from air quality monitors spread across the United States. For our purposes, we consider the air quality index time series for nitrogen dioxide (NO_2), ozone (O_3), and carbon monoxide (CO). We used data from sensors having 100 days of consecutive data in 2000 from May 20th to August 27th. A simple kernel regression density plot for the data for each pollutant produced non-Gaussian curves, which motivates the use of non-parametric Bayesian analysis over parametric forecasting models. While our

method does not require the data from each sensor to be overlapping, this restriction helps ensure that the forecasting results are due only to the method and not so some other data-dependent difference. Additionally, the relatively short time span helps reduce concerns about non-stationarity. For each time course, the time series were differenced (5 steps), de-meaned, and outliers were replaced using the `tsoutliers` function in R (López-de Lacalle, 2024). The time courses were then checked for stationarity using a Dickey-Fuller test, and sensors for which the time courses were not stationary were removed from the data. After this procedure, we had 41 sensors with complete data.

We fit the rgPDPM-VAR model to this data using 1500 burn-in samples and 3500 MCMC samples. The concentration parameter was set to 5 to encourage more clusters to spawn. The resulting forecasting accuracy was measured in terms of the relative error and was 0.549 at step 1, 0.877 at step 2, and 0.929 at step 3. Thus the model shows considerable forecasting performance for 1 step forecasting, and naturally this performance degrades as the number of forecasting steps increases. As a comparison, we also used the SS-VAR approach of Ghosh et al. (2018) to analyse each sensor separately. The SS-VAR model illustrates decent performance, but is considerably worse that the rgPDPM-VAR, with a relative forecasting error of 0.598 at step 1, 0.897 at step 2, and 0.959 at step 3. This suggests that even for a small number of nodes, pooling information across sensors/subjects under the rgPDPM-VAR can still provide forecasting benefits. We note that other choices could have been made for the number of differencing steps. However, the overall relative performance between the two methods stays similar, with the proposed rgPDPM-VAR consistently performing better over different settings, as reported below.

Differencing Level	1 Step Forecasting		2 Step Forecasting		3 Step Forecasting	
	rgPDPM-VAR	SS-VAR	rgPDPM-VAR	SS-VAR	rgPDPM-VAR	SS-VAR
1	1.202 (0.043)	1.204 (0.120)	1.022 (0.089)	1.054 (0.146)	1.002 (0.037)	1.009 (0.053)
2	0.718 (0.064)	0.731 (0.118)	0.995 (0.131)	1.034 (0.140)	1.022 (0.021)	1.022 (0.041)
3	0.644 (0.143)	0.696 (0.254)	0.913 (0.376)	0.934 (0.411)	1.057 (0.129)	1.067 (0.178)
4	0.616 (0.133)	0.703 (0.226)	0.889 (0.212)	0.932 (0.256)	0.962 (0.227)	1.000 (0.498)
5	0.549 (0.118)	0.598 (0.167)	0.877 (0.247)	0.897 (0.237)	0.929 (0.127)	0.959 (0.150)

Table 2: Forecasting results under the rgPDPM-VAR and SS-VAR models for the air quality data with different levels of differencing. The values in the cells represent the relative L2 error between the model-based predictions and the true values, and the standard deviation is given in parenthesis.

8. Discussion

In this work, we developed a non-parametric Bayesian framework for i.i.d. multivariate data as well as multivariate time-series data, which provides a fundamentally novel way to borrow information across samples, via a class of novel product of DP mixture priors. The proposed approach employs multi-scale clustering to flexibly borrow information across heterogeneous samples that bypasses restrictive parametric assumptions and the requirement of replicated samples. The method is implemented via an efficient MCMC sampling scheme and computational complexity calculations are presented. The distinct numerical advantages over existing methods are illustrated via extensive numerical examples. The

proposed class of methods are shown to have desirable posterior consistency properties that are derived based on novel sieve constructions and careful entropy calculations. Unlike single-subject parametric VAR modeling (Ghosh et al., 2018) that relies on an increasing T to establish posterior consistency, the proposed Bayesian non-parametric analysis focuses on posterior consistency corresponding to density estimation as $n \to \infty$. While it would be interesting to explore posterior consistency under our set-up as both $T \to \infty, n \to \infty$, there are potential theoretical challenges to be encountered. For example, increasing the number of timepoints T directly has an impact on the expression of the true density f_0, as well as on the form of $f_P(X)$. The latter has a direct impact on the sieve entropy (Theorem 4) that is intricately tied to the sufficient conditions for posterior consistency in Theorem 2. We plan to explore such aspects in future work.

While this work introduced several variants of the PDPM-VAR model, there are numerous potential extensions that lie within our class of models. In particular, future directions might investigate possible generalizations intended to induce sparsity in the parameter estimates. For example, a spike and slab prior could be used to model the autocovariance elements, with the slab component modeled using a DP mixtures. Additionally, the models could be generalized to accommodate even higher levels of heterogeneity, such as clustering individual autocovariance elements separately. However, such extensions may involve a massive computational burden. The proposed approaches, particularly the rgPDPM-VAR, seem to strike a desirable balance between computational complexity, clustering flexibility and model parsimony, with theoretical guarantees and appealing practical performance. Finally, we note that the proposed product of DP priors provides a viable improvement over traditional DP mixture models and it should have wide applicability to other types of settings that go beyond the VAR framework, which is of immediate interest in this article. We expect to pursue these directions in future research. For example, the VAR model structure could be relaxed to a dynamic linear model framework that can cater to non-stationary time-series, using a set of lower dimensional latent time courses to model the observed data (Sevestre and Trognon, 1996). Such an approach would potentially enrich the kinds of time courses that could be described by incorporating time-varying relationships.

Acknowledgments

The authors gratefully acknowledge support from NIH awards R01AG071174 and R01MH120299.

Appendix

Appendix A. Proofs of Results

Proof of Lemma 1: An outline for the proof of Lemma 1 is provided, which follows similar steps as the proof of Lemma 1 in Canale and De Blasi (2017). One may write $KL(f_0, f_P) = KL(f_0, f_{P_\epsilon}) + KL(f_{P_\epsilon}, f_P)$ and then show that each of the terms in the right hand side can be made exceedingly small with positive probability, for some compactly supported P_ϵ. The compact support for P_ϵ is taken as $[-\mu^*, \mu^*]^D \times \{\Sigma \in \mathcal{S} : \underline{\sigma}^2 < \lambda_l(\Sigma) < \bar{\sigma}^2, 1 \leq l \leq p\}$,

for some constants $\mu^* > 0$ and $0 < \underline{\sigma} < \bar{\sigma}$, and eigenvalues denoted as $\lambda_l, l = 1, \ldots, p$, the second term can also be shown to be infinitesimally small with positive probability.

Proof of Theorem 1: The proof relies on two parts, i.e calculating the entropy bounds and calculating the prior probability of the constructed sieves, and then using them to show the summability condition in Theorem 2 holds. We will illustrate the proof for the case of $\mathcal{M}_\mu = 1$, and extensions to higher values of $\mathcal{M}_\mu$ are straightforward.

First, we will construct sieves of the following form -

$$\begin{aligned}\mathcal{F}_n &= \Big\{ f_p : P = \sum_{h_1 \geq 1} \sum_{h_\sigma \geq 1} \pi_{h_1} \pi_{\sigma,h_\sigma} \delta_{\boldsymbol{\mu}_{h_1}, \Sigma_{h_\sigma}} : \sum_{h_1 > H_n} \pi_{h_1} < \epsilon, \sum_{h_\sigma > H_n} \pi_{\sigma,h_\sigma} < \epsilon, \text{ and for} \\ &\quad h_\sigma \leq H_n,\ \underline{\sigma}_n^2 \leq \lambda_D, \lambda_1 \leq \underline{\sigma}_n^2 (1 + \epsilon/\sqrt{D})^{M_n},\ 1 < \frac{\lambda_1(\Sigma_{h_\sigma})}{\lambda_D(\Sigma_{h_\sigma})} \leq n^{H_n} \Big\}, \qquad (15) \\ \mathcal{F}_{n,\mathbf{j},\mathbf{l}} &= \Big\{ f_p \in \mathcal{F}_n :\ \text{for } h_1, h_\sigma \leq H_n, n^{H_n^2}(j_{h_1} - 1) = \underline{a}_{h_1,j} \leq ||\boldsymbol{\mu}_{h_1}|| \leq \bar{a}_{h_1,j} = n^{H_n^2} j_{h_1}, \\ &\quad k \in \{1, \ldots, K\},\ n^{l_{h_\sigma} - 1} < \frac{\lambda_1(\Sigma_{h_\sigma})}{\lambda_D(\Sigma_{h_\sigma})} \leq n^{l_{h_\sigma}} \Big\}, \qquad (16)\end{aligned}$$

where $M_n = \underline{\sigma}^{-2c_2} = n$ and $H_n = \lfloor Cn\epsilon^2 / \log(n) \rfloor$ for some positive constant C, and clearly $\mathcal{F}_n \subset \cup_{\mathbf{j},\mathbf{l}} \mathcal{F}_{n,\mathbf{j},\mathbf{l}}$. Using similar techniques to those used in Lemma 6 in the Appendix, it is possible to show the tail condition (2A) holds in Theorem 2, i.e. $\Pi(\mathcal{F}_n^c) \leq e^{-b^* n}$. Further, using similar steps as in the proof of Lemma 4 in the Appendix, the distance between the two densities f_{P_1} and f_{P_2} can be expressed as $||f_{P_1} - f_{P_2}||_1 \leq \epsilon^2 + \sum_{h_1, h_\sigma < H_n} |\pi_{h_1}^{(1)} \pi_{\sigma,h_\sigma}^{(1)} - \pi_{h_1}^{(2)} \pi_{\sigma,h_\sigma}^{(2)}| + \epsilon + \sum_{h_1, h_\sigma \leq H_n} \pi_{h_1}^{(1)} \pi_{\sigma,h_\sigma}^{(1)} \left\| \phi_{\Sigma_{h_\sigma}^{(1)}}\big(\mathbf{x} - \boldsymbol{\mu}_{h_1}^{(1)}\big) - \phi_{\Sigma_{h_\sigma}^{(2)}}\big(\mathbf{x}_t - \boldsymbol{\mu}_{h_1}^{(2)}\big) \right\|_1$, where $||\cdot||_1$ denotes the L_1 norm. Using similar steps as in the proof of Lemma 2 in Canale and De Blasi (2017), it is possible to show that $\left\| \phi_{\Sigma_{h_\sigma}^{(1)}}\big(\mathbf{x} - \boldsymbol{\mu}_{h_1}^{(1)}\big) - \phi_{\Sigma_{h_\sigma}^{(2)}}\big(\mathbf{x}_t - \boldsymbol{\mu}_{h_1}^{(2)}\big) \right\|_1 \leq \sqrt{\frac{2}{\pi}} \frac{||\boldsymbol{\mu}_{h_1}^{(1)} - \boldsymbol{\mu}_{h_1}^{(2)}||}{\sqrt{\lambda_D(\Sigma_{h_\sigma}^{(2)})}}$ $+ \left\{ \sum_{k=1}^{D} \frac{\lambda_k(\Sigma_{h_\sigma}^{(1)})}{\lambda_k(\Sigma_{h_\sigma}^{(2)})} - \log \frac{\lambda_k(\Sigma_{h_\sigma}^{(1)})}{\lambda_k(\Sigma_{h_\sigma}^{(2)})} - 1 \right\}^{1/2} + \left\{ 2D ||O_{h_\sigma}^{(1)} - O_{h_\sigma}^{(2)}||_2 \frac{\lambda_1(\Sigma_{h_\sigma}^{(1)})}{\lambda_D(\Sigma_{h_\sigma}^{(2)})} \right\}$, where $O(j')$ represents the matrix of orthogonal vectors in the spectral decomposition of $\Sigma^{(j')}, j' = 1, 2$. Finally, we need to establish an upper bound for the term $\sum_{h_1, h_\sigma < H_n} |\pi_{h_1}^{(1)} \pi_{\sigma,h_\sigma}^{(1)} - \pi_{h_1}^{(2)} \pi_{\sigma,h_\sigma}^{(2)}|$, which is given by Lemma 5 as $\sum_{h_1, h_\sigma < H_n} \left| \tilde{\pi}_{h_1}^{(1)} \tilde{\pi}_{\sigma,h_\sigma}^{(1)} - \pi_{h_1}^{(2)} \pi_{\sigma,h_\sigma}^{(2)} \right| + \left| 1 - (1 - \epsilon)^2 \right|$, where $\tilde{\pi} = \frac{\pi_h}{(1 - \sum_{h > H} \pi_h)}$.

For a given $f_P \in \mathcal{F}_{n,\mathbf{jl}}$ with $P = \sum_{h_1 \geq 1} \sum_{h_\sigma \geq 1} \pi_{h_1}^{(1)} \pi_{\sigma,h_\sigma}^{(1)} \delta_{\left(\Theta_{h_1}^{(1)}, \Sigma_{h_\sigma}^{(1)}\right)}$ and $\Sigma_h = (O_h \Lambda_h O_h^T)^{-1}$ where $\Lambda_h = diag(\lambda_{h,1}, \ldots, \lambda_{h,D})$, we will construct another density, $f_{\hat{P}}$, where $\hat{P} = \sum_{h_1 \geq 1} \sum_{h_\sigma \geq 1} \pi_{h_1}^{(1)} \pi_{\sigma,h_\sigma}^{(1)} \delta_{\left(\hat{\Theta}_{h_1}^{(1)}, \hat{\Sigma}_{h_\sigma}^{(1)}\right)}$ within the ϵ-net and then compute the cardinality of the ϵ-net set to derive an upper bound for the entropy of sieve $\mathcal{F}_{n,\mathbf{jl}}$. To construct such a density, we will choose:

1. $\hat{\boldsymbol{\mu}}_{h_1} \in \hat{\mathcal{R}}_{h_1}, h_1 = 1, \ldots, H$, where $\hat{\mathcal{R}}_{h_1}$ is a ϵ^*-net of $\mathcal{R}_{h_1} := \{\boldsymbol{\mu} \in \Re^D : \underline{\mu}_{h_1,j} \leq ||\boldsymbol{\mu}|| \leq \bar{\mu}_{h_1,j}\}$, such that $||\boldsymbol{\mu}_{h_1} - \hat{\boldsymbol{\mu}}_{h_1}|| \leq \underline{\sigma}_n \epsilon$, $k = 1, \ldots, K$, where $\bar{\mu}, \underline{\mu}$, and $\underline{\sigma}_n$ correspond to the sieve boundaries in (16).

2. $\{\hat{\pi}_{h_1}\hat{\pi}_{h_\sigma}, h_1, h_\sigma \leq H_n\} \in \hat{\Delta}$, where $\hat{\Delta}$ is a ϵ-net of a H_n^2 dimensional probability simplex such that $\sum_{h_1,h_\sigma \leq H_n} |\tilde{\pi}_{h_1}\tilde{\pi}_{h_\sigma} - \hat{\pi}_{h_1}\hat{\pi}_{h_\sigma}| \leq \epsilon$, and $\tilde{\pi}_h = \frac{\pi_h}{\sum_{h \leq H_n} \pi_h}, h \leq H_n$.

3. $\hat{O}_h \in \hat{\mathcal{O}}_h$, where $\hat{\mathcal{O}}_h$ is a δ_h-net of the set $\mathcal{O}_h$ defined as the set of $D \times D$ orthogonal matrices with respect to the spectral norm $||\cdot||_2$ with $\delta_h = \epsilon^2/(2Du_{h,l})$ such that $||O_h - \hat{O}_h||_2 \leq T\delta_h$.

4. $(m_{h,1}, ..., m_{h_D}) \in \{1, ..., M\}^D, h = 1, ..., H$, such that $\hat{\lambda}_{h,l} = \{\underline{\sigma}^2(1 + \epsilon\sqrt{D})^{m_{h,l}-1}\}^{-1}$ will satisfy $1 \leq \hat{\lambda}_{h,l}/\lambda_{h,l} < (1 + \epsilon/\sqrt{D})$.

Under this construction, it can be shown that $||f_P - f_{\hat{P}}||_1 < C^*\epsilon$ for some constant C^*, by employing some additional algebra and similar arguments as in the proof of Lemma 2 in Canale and De Blasi (2017) and Theorem 4 in our paper. Further, the cardinality of the ϵ-net can be computed by noting that $\#(\hat{\Delta}) \lesssim \epsilon^{-H_n^2}$ for $j = 1, 2, \#(\hat{\mathcal{O}}_h) \lesssim \delta_h^{-D(D-1)/2}, \#(\hat{\mathcal{R}}_{k,h}) \lesssim [(\frac{\bar{a}_h}{\epsilon^*} + 1)^D - (\frac{a_h}{\epsilon^*} - 1)^D]$. Using these quantities, one can write the upper bound for the exponential of the entropy bound as,

$$
\begin{aligned}
&(M)^{DH_n}\epsilon^{-H_n^2} \times \prod_{h_1 \leq H_n} \{(\frac{\bar{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} + 1)^D - (\frac{\underline{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} - 1)^D\} \prod_{h_\sigma \leq H_n} \{\frac{2Du_{h_\sigma,l}}{\epsilon^2}\}^{D(D-1)/2} \\
&\approx \exp\left\{DH_n \log(M) + H_n^2 \log(\frac{1}{\epsilon}) + \frac{D(D-1)}{2}\log(n^{l_{h_\sigma}}) + \frac{D(D-1)}{2}\log(\frac{1}{\epsilon})\right\},
\end{aligned}
$$

when n and j_{h_1} is large, and using the definitions of $\underline{a}$ and $\bar{a}$ defined in (16), and following similar steps as in the proof of Theorem 2 in Canale and De Blasi (2017) to show that $\left[\left(\frac{\bar{a}_{h_1,j}}{\underline{\sigma}_n\epsilon/2} + 1\right)^D - \left(\frac{\underline{a}_{h_1,j}}{\underline{\sigma}_n\epsilon/2} - 1\right)^D\right] \lesssim \left[\frac{n^{(H_n + \frac{1}{2c_2})D} j_{h1}^{D-1}}{(\epsilon)^D}\right]$. Further, using similar steps as in (33) in the proof of Theorem 5, we have $\Pi(\mathcal{F}_{n,\mathbf{jl}}) \leq \prod_{h_1 \leq H_n} P_1^*\big(||\boldsymbol{\mu}_{h_1}|| > n^{H_n^2}(j_{h_1} - 1)\big) \prod_{h_\sigma \leq H_n} P_2^*\big(\lambda_1(\Sigma)/\lambda_D(\Sigma) > n^{(l_{h_\sigma}-1)}\big) \lesssim \prod_{h_1 \leq H_n} \left\{\left(n^{H_n^2}(j_{h_1} - 1)\right)^{-1_{(j_{h_1} \geq 2)}2(r+1)}\right\}$
$\times \prod_{h_\sigma \leq H_n} (n^{(l_{h_\sigma}-1)})^{-1_{(l_{h_\sigma} \geq 1)}\kappa} \approx \left\{n^{-2H_n^3(r+1)} \prod_{h_1 \leq H_n} \left(j_{h_1} - 1\right)^{-1_{(j_{h_1} \geq 2)}2(r+1)}\right\}$
$\times \left\{\prod_{h_\sigma \leq H_n} (n^{\kappa(l_{h_\sigma}-1)})^{-1_{(l_{h_\sigma} \geq 1)}}\right\}$, for large n.

Finally, using similar steps as in Lemma 7 in the Appendix it is possible to show that the summability condition in Theorem 2 holds. This proves the strong consistency result for the product mixture of DP priors for multivariate density estimation.

Proof of Theorem 3: We will use the conditions in Lemmas 2-4 and Theorem 2 in Wu and Ghosal (2008) to prove our results. Note that $f_0(X) = \prod_{t=1}^T f_0(\mathbf{x}_t \mid X_{1:(t-1)})$, where $f_0(\mathbf{x}_t \mid X_{1:(t-1)}) = f_0(\mathbf{x}_1)$ for $t = 1$ by convention. As a shorthand notation, we will denote $\frac{f_0(\mathbf{x})}{f_P(\mathbf{x})} = (f_0/f_P)(\mathbf{x})$ in the following proof. For any $P \in \mathcal{P}$, note that the KL divergence can

be expressed as $KL(f_0(X), f_P(X)) = \int f_0(X) \log\big((f_0/f_P)(X)\big) d\mathbf{x}_1 \ldots d\mathbf{x}_T$

$$
\begin{aligned}
&= \int \prod_{t=1}^{T} f_0\big(\mathbf{x}_t \mid X_{1:(t-1)}\big) \log\big(\prod_{t=1}^{T} (f_0/f_P)(\mathbf{x}_t \mid X_{1:(t-1)})\big) d\mathbf{x}_1 \ldots d\mathbf{x}_T \\
&= \int \prod_{t=1}^{T} f_0\big(\mathbf{x}_t \mid X_{1:(t-1)}\big) \times \Big\{ \sum_{t=1}^{T} \log\big((f_0/f_P)(\mathbf{x}_t \mid X_{1:(t-1)})\big) \Big\} d\mathbf{x}_1 \ldots d\mathbf{x}_T \\
&= \sum_{t=1}^{T} \Big[\prod_{t^*>t} \underbrace{\int f_0\big(\mathbf{x}_{t^*} \mid X_{1:(t^*-1)}\big) d\mathbf{x}_{t^*}}_{=1} \times \int \Big\{ \int f_0\big(\mathbf{x}_t \mid X_{1:(t-1)}\big) \log\big((f_0/f_P)(\mathbf{x}_t \mid X_{1:(t-1)})\big) d\mathbf{x}_t \\
&\quad \times \prod_{t'<t} f_0\big(\mathbf{x}_{t'} \mid X_{1:(t'-1)}\big) d\mathbf{x}_{t'} \Big\} \Big] \\
&= \sum_{t=1}^{T} \Big\{ \int KL\big(f_0\big(\mathbf{x}_t \mid X_{1:(t-1)}\big), f_P\big(\mathbf{x}_t \mid X_{1:(t-1)}\big)\big) \prod_{t'=1}^{t-1} f_0(\mathbf{x}_{t'} \mid X_{1:(t'-1)}) d\mathbf{x}_{t'-1} \ldots d\mathbf{x}_1 \Big\}, \quad (17)
\end{aligned}
$$

which is a sum of integrals involving Kullback-Leibler divergences of conditional densities. In the following derivations, we will use the shorthand notation $KL_{(f_0,f_P)}(\mathbf{x}_t \mid X_{1:(t-1)})$ to denote $KL\big(f_0\big(\mathbf{x}_t \mid X_{1:(t-1)}\big), f_P\big(\mathbf{x}_t \mid X_{1:(t-1)}\big)\big)$ where convenient. We will prove that $KL_{(f_0,f_P)}(\mathbf{x}_t \mid X_{1:(t-1)})$ is infinitesmall ($< \epsilon$) with positive probability pointwise for $X_{1:(t-1)}$ for $t = 1, \ldots, T$, which will imply that the above sum on the right hand side of the equality also becomes infinitesmall for fixed T, and hence we will prove our Theorem.

As a first step, we will define P_ϵ on a compact set $\{\mathbf{\Theta} : -a \leq A_k(j,j') \leq a,\ k = 1, \ldots, K,$ and $||\sum_{k=1}^{K} A_k \mathbf{x}_{t-k}|| < m\} \times \{\Sigma \in \mathcal{S} : h_m = m^{-\eta} \leq \lambda_D(\Sigma) < \ldots < \lambda_1(\Sigma) \leq \bar{M},\ t = 1, \ldots, T\} = \mathcal{D}_1^* \times \mathcal{D}_2^*$, such that $P_\epsilon(\mathcal{D}_1^* \times \mathcal{D}_2^*) = 1$, for some $m, \eta > 0$. We will construct P_ϵ such that it ensures that the upper bounds for the terms in the right hand side of the above equality are arbitrarily small with positive prior probability:

$$
KL_{(f_0,f_P)}\big(\mathbf{x}_t \mid X_{1:(t-1)}\big) = KL_{(f_0,f_{P_\epsilon})}\big(\mathbf{x}_t \mid X_{1:(t-1)}\big) + KL_{(f_{P_\epsilon},f_P)}\big(\mathbf{x}_t \mid X_{1:(t-1)}\big). \quad (18)
$$

Write $\mathbf{r} = \sum_k A_k \mathbf{x}_{t-k}$, and $t_m^{-1} = \int\limits_{||\mathbf{r}||<m} f_0\big(\mathbf{r} \mid X_{1:(t-1)}\big) d(\mathbf{r})$ and define $t_m^{-1} f_{P_\epsilon}\big(\mathbf{x}_t \mid X_{1:(t-1)}\big)$

$$
\begin{aligned}
&= \int\limits_{||\mathbf{r}||<m} \phi_\Sigma\big(\mathbf{x}_t - \mathbf{r} \mid X_{1:(t-1)}\big) f_0\big(\mathbf{r}_{t-k} \mid X_{1:(t-1)}\big) d(\mathbf{r}) \geq \int\limits_{||\mathbf{r}||<m} \phi_{h_m I_D}\big(\mathbf{x}_t - \mathbf{r} \mid X_{1:(t-1)}\big) f_0\big(\mathbf{r} \mid X_{1:(t-1)}\big) d(\mathbf{r}) \\
&\times \big(\frac{\lambda_D(\Sigma)}{\lambda_1(\Sigma)}\big)^{(D-1)/2} = t_m \int\limits_{||\mathbf{x}_t - \boldsymbol{\theta} h_m||<m} \phi_{h_m I_D}\big(\boldsymbol{\theta} \mid X_{1:(t-1)}\big) f_0\big(\mathbf{x}_t - \boldsymbol{\theta} h_m \mid X_{1:(t-1)}\big) d\boldsymbol{\theta} \times \big(\frac{\lambda_D(\Sigma)}{\lambda_1(\Sigma)}\big)^{(D-1)/2}
\end{aligned}
$$

where $\boldsymbol{\theta} = (\mathbf{x}_t - \sum_{k=1}^{K} A_k \mathbf{x}_{t-k})/h_m$, and the inequality in the second last step is derived using the fact that $\big(\frac{\lambda_D(\Sigma)}{\lambda_1(\Sigma)}\big)^{(D-1)/2} \phi_{\lambda_D(\Sigma) I_D}\big(\mathbf{x}\big) \leq \phi_\Sigma\big(\mathbf{x}\big) \leq \big(\frac{\lambda_1(\Sigma)}{\lambda_D(\Sigma)}\big)^{(D-1)/2} \phi_{\lambda_1(\Sigma) I_D}\big(\mathbf{x}\big)$. Using the above, we have $KL_{(f_0,f_{P_\epsilon})} \leq \big(\frac{\lambda_1(\Sigma)}{\lambda_D(\Sigma)}\big)^{(D-1)/2} KL_{(f_0,f_{P_{\epsilon,\Sigma=h_m I_D}})}$, where $f_{P_{\epsilon,\Sigma=h_m I_D}}$ has the same form as f_{P_ϵ} but with Σ set to $h_m I_D$. Hence, the next step is to show that $KL(f_0, f_{P_{\epsilon,\Sigma=h_m I_D}})(\mathbf{x}_t \mid X_{1:(t-1)})$ can be made arbitrarily small with positive prior probability, point-wise for all $X_{1:(t-1)}$, which will help establish that the first term on the right hand side of (18) is negligible with positive prior probability.

Since when $m \to \infty$ and $h_m \to 0$, $\phi_{h_m I_D}\big(\mathbf{x}_t - \sum_{k=1}^K A_k \mathbf{x}_{t-k} \mid X_{1:(t-1)}\big)$ takes non-zero values only when $\sum_{k=1}^K A_k \mathbf{x}_{t-k} \to \mathbf{x}_t$, we obtain $\phi_{h_m I_D}\big(\mathbf{x}_t - \sum_{k=1}^K A_k \mathbf{x}_{t-k} \mid X_{1:(t-1)}\big) f_0\big(\sum_{k=1}^K A_k \mathbf{x}_{t-k} \mid X_{1:(t-1)}\big) \to f_0(\mathbf{x}_t \mid X_{1:(t-1)})$ as $m \to \infty$ and $h_m \to 0$, which also implies $f_{P_{\epsilon,\Sigma=h_m I_D}}(\mathbf{x}_t \mid X_{1:(t-1)}) \to f_0(\mathbf{x}_t \mid X_{1:(t-1)})$, point-wise for all $X_{1:(t-1)}$. We will combine the above convergence with the fact the $\log\big(\frac{f_0(\mathbf{x}_t|X_{1:(t-1)})}{f_{P_{\epsilon,\Sigma=h_m I_D}}(\mathbf{x}_t|X_{1:(t-1)})}\big)$ is bounded and integrable (as shown in the sequel) and subsequently apply the DCT to achieve our result. As a next step, note $KL_{(f_0, f_{P_{\epsilon,\Sigma=h_m I_D}})}(\mathbf{x}_t \mid X_{1:(t-1)}) = \int\limits_{||\mathbf{x}_t \le m||} f_0(\mathbf{x}_t \mid X_{1:(t-1)}) \log\big(\frac{f_0(\mathbf{x}_t|X_{1:(t-1)})}{f_{P_{\epsilon,\Sigma=h_m I_D}}(\mathbf{x}_t|X_{1:(t-1)})}\big) d\mathbf{x}_t +$ $\int\limits_{||\mathbf{x}_t > m||} f_0(\mathbf{x}_t \mid X_{1:(t-1)}) \log\big(\frac{f_0(\mathbf{x}_t|X_{1:(t-1)})}{f_{P_{\epsilon,\Sigma=h_m I_D}}(\mathbf{x}_t|X_{1:(t-1)})}\big) d\mathbf{x}_t$. Using similar arguments as in the Proof of Theorem 2 in Wu and Ghosal (2008), for $||\mathbf{x}_t|| > m$, we have $f_{P_{\epsilon,\Sigma=h_m I_D}}(\mathbf{x}_t \mid X_{1:(t-1)})$

$$
\begin{aligned}
&\ge t_m \int\limits_{||\sum_k A_k \mathbf{x}_{t-k}|| < m} \phi_{h_m I_D}\big(\mathbf{x}_t + m\mathbf{x}_t/||\mathbf{x}_t|| \mid X_{1:(t-1)}\big) f_0\big(\mathbf{r} \mid X_{1:(t-1)}\big) d\Big(\sum_{k=1}^K A_k \mathbf{x}_{t-k}\Big) \\
&= \phi_{h_m I_D}\big(\mathbf{x}_t + m\mathbf{x}_t/||\mathbf{x}_t||\big) \times \underbrace{\Big\{ t_m \int\limits_{||\sum_k A_k \mathbf{x}_{t-k}|| < m} f_0\big(\mathbf{r} \mid X_{1:(t-1)}\big) d\Big(\sum_{k=1}^K A_k \mathbf{x}_{t-k}\Big) \Big\}}_{=1} \\
&= \phi_{h_m I_D}\big(\mathbf{x}_t + m\mathbf{x}_t/||\mathbf{x}_t||\big) = m^\eta \phi_{I_D}\big(m^\eta \mathbf{x}_t + m^{1+\eta}\mathbf{x}_t/||\mathbf{x}_t||\big) \ge ||\mathbf{x}_t||^\eta \phi_{I_D}(2||\mathbf{x}_t||\mathbf{x}_t).
\end{aligned}
$$

Similarly for $||\mathbf{x}_t|| \le m$, it is possible to show that given a constant $\delta > 0$, $f_{P_{\epsilon,\Sigma=h_m I_D}}(\mathbf{x}_t \mid X_{1:(t-1)}) \ge c \inf\limits_{||\mathbf{r}-\mathbf{x}_t|| < \delta} f_0(\mathbf{r} \mid X_{1:(t-1)})$, using similar steps as in the proof of Theorem 2 in Wu and Ghosal (2008). Hence for a given constant $R < m$, we have

$$
\log\left(\frac{f_0(\mathbf{x}_t \mid X_{1:(t-1)})}{f_{P_{\epsilon,\Sigma=h_m I_D}}(\mathbf{x}_t \mid X_{1:(t-1)})}\right) = \xi(\mathbf{x}_t; X_{1:(t-1)}) \le \begin{cases} r_1^* = \log\left(\frac{f_0(\mathbf{x}_t|X_{1:(t-1)})}{c \inf\limits_{||\mathbf{r}-\mathbf{x}_t|| < \delta} f_0(\mathbf{r}|X_{1:(t-1)})}\right), & ||\mathbf{x}_t|| < R, \\ \max\left\{ \log\left(\frac{f_0(\mathbf{x}_t|X_{1:(t-1)})}{||\mathbf{x}_t||^\eta \phi_{I_D}(2||\mathbf{x}_t||\mathbf{x}_t)}\right), r_1^* \right\} & ||\mathbf{x}_t|| \ge R. \end{cases} \tag{19}
$$

Further note that $f_{P_{\epsilon,\Sigma=h_m I_D}}(\mathbf{x}_t \mid X_{1:(t-1)}) \le Mt_1$ which implies $\log\big(\frac{f_0(\mathbf{x}_t|X_{1:(t-1)})}{f_{P_{\epsilon,\Sigma=h_m I_D}}(\mathbf{x}_t|X_{1:(t-1)})}\big) \ge \log\big(\frac{f_0(\mathbf{x}_t|X_{1:(t-1)})}{M\phi_1^*}\big)$, where $\phi_1^* = \int\limits_{||\mathbf{x}_t|| < 1} f_0(\mathbf{x}_t \mid X_{1:(t-1)})$ the lower bound is < 0. Combining this fact with the upper bound in (19), it is possible to write

$$
KL\left(\frac{f_0(\mathbf{x}_t \mid X_{1:(t-1)})}{f_{P_{\epsilon,\Sigma=h_m I_D}}(\mathbf{x}_t \mid X_{1:(t-1)})}\right) \le \int f_0(\mathbf{x}_t \mid X_{1:(t-1)}) \max\big\{\xi(\mathbf{x}_t; X_{1:(t-1)}), \big|\log\big(\frac{f_0}{M\phi_1^*}\big)\big|\big\}.
$$

Using assumptions *(A2)-(A4)* and similar steps as in the proof of Theorem 2 in Wu and Ghosal (2008), it is possible to show that the RHS is bounded, point-wise for all $X_{1:(t-1)}$. Hence using DCT, the term $KL_{(f_0, f_{P_{\epsilon,\Sigma=h_m I_D}})}\big(\mathbf{x}_t \mid X_{1:(t-1)}\big)$ can be made arbitrarily small

with positive prior probability. Therefore, we see that for any $\epsilon > 0$, there exists m_ϵ such that $KL_{(f_0, f_{P_\epsilon})}\big(\mathbf{x}_t \mid X_{1:(t-1)}\big) \leq \big(\frac{\lambda_1(\Sigma)}{\lambda_D(\Sigma)}\big)^{(D-1)/2} KL_{(f_0, f_{P_{\epsilon,\Sigma=h_{m_\epsilon} I_D}})}\big(\mathbf{x}_t \mid X_{1:(t-1)}\big) \leq \epsilon/2$. Hence the first term in (18) is bounded by $\epsilon/2$ with positive prior probability.

To show that the second term in (18) is negligible, we will demonstrate that the conditions in Lemma 3 of Wu and Ghosal (2008) are satisfied. Using similar arguments as in Wu and Ghosal (2008) as well as the proof of Lemma 1 in Canale and De Blasi (2017), it is possible to show that the weak support of Π^* contains any compactly supported $\mathcal{P}$. Since P_ϵ is compactly supported by definition, it belongs to the weak support of Π^*. Next, condition (A7) of Lemma 3 in Wu and Ghosal (2008) requires $\log(f_{P_\epsilon})$ and $\log \inf_{A_1,\ldots,A_K,\Sigma} \phi_\Sigma\big(\mathbf{x}_t - \sum_{k=1}^K A_k \mathbf{x}_{t-k}\big)$ to be f_0- integrable. Note that for $||\mathbf{x}_t|| < m$, $\log \inf_{A_1,\ldots,A_K,\Sigma} \phi_\Sigma\big(\mathbf{x}_t - \sum_{k=1}^K A_k \mathbf{x}_{t-k}\big)$ is bounded, and for $||\mathbf{x}_t|| > m$, $\inf_{A_1,\ldots,A_K,\Sigma} \phi_\Sigma(\mathbf{x}_t - \sum_{k=1}^K A_k \mathbf{x}_{t-k}) \leq \bar{M}^{-D} \phi(\exp\{-\frac{4||\mathbf{x}_t||^2}{2m^{-\eta}}\})$, which is f_0 integrable. A similar upper bound can be applied for bounding $|\log(f_{P_\epsilon})|$ for $||\mathbf{x}_t|| > m$ that implies that $|\log(f_{P_\epsilon})|$ is f_0 integrable, which satisfies condition (A7) in Lemma 3 of Wu and Ghosal (2008). Note that the above statements hold point-wise for all $X_{1:(t-1)}$.

Condition (A8) of Lemma 3 in Wu and Ghosal (2008) corresponding to the second term in (18) is clearly satisfied since the multivariate normal kernel $\phi_\Sigma\big(\mathbf{x}_t - \sum_{k=1}^K A_k \mathbf{x}_{t-k}\big)$ is bounded away from zero for $\mathbf{x}_t$ in a compact set of $\Re^d$ and $(A_1, \ldots, A_K, \Sigma) \in \mathcal{D}_1^* \times \mathcal{D}_2^*$. To show condition (A9) in Lemma 3 in Wu and Ghosal (2008), we will need to show that the kernel $\phi_\Sigma(\mathbf{x}_t - \sum_{k=1}^K A_k \mathbf{x}_{t-k})$ is equicontinuous as a family of functions of $\{A_1, \ldots, A_K, \Sigma\}$ for $\mathbf{x}_t$ lying on a compact subset of $\Re^D$ and conditional on given values of $X_{1:(t-1)}$. Note that for two distinct sets of parameters $(\boldsymbol{\Theta}, \Sigma)$ and $(\boldsymbol{\Theta}', \Sigma')$,

$$\Big|\phi_\Sigma\big(\mathbf{x}_t - \sum_{k=1}^K A_k \mathbf{x}_{t-k}\big) - \phi_{\Sigma_t'}\big(\mathbf{x}_t - \sum_{k=1}^K A_k' \mathbf{x}_{t-k}\big)\Big| \leq \Big|\phi_{\Sigma'}\big(\mathbf{x}_t - \sum_{k=1}^{K-1} A_k \mathbf{x}_{t-k} - A_K' \mathbf{x}_{t-K}\big) -$$
$$\phi_\Sigma\big(\mathbf{x}_t - \sum_{k=1}^K A_k \mathbf{x}_{t-k}\big)\Big| + \Big|\phi_{\Sigma'}\big(\mathbf{x}_t - \sum_{k=1}^{K-1} A_k \mathbf{x}_{t-k} - A_K' \mathbf{x}_{t-K}\big) - \phi_{\Sigma'}\big(\mathbf{x}_t - \sum_{k=1}^K A_k' \mathbf{x}_{t-k}\big)\Big|. \quad (20)$$

The first term in (20) can be shown to be arbitrarily small when (A_K', Σ') lies within a small neighborhood of (A_K, Σ) (or equivalently $A_K \mathbf{x}_{t-K}$ is in a neighborhood of $A_K' \mathbf{x}_{t-K}$ pointwise for $\mathbf{x}_{t-K}$, and Σ lies in a neighborhood of Σ') for $\mathbf{x}_t \in C \subset \Re^D$, where C is a compact subset of $\Re^D$, using arguments similar to the last part of the proof of Theorem 2 in Wu and Ghosal (2008). The second term in (20) can be decomposed using similar and repeated iterative steps, and hence shown to be arbitrarily small. Hence the equicontinuity condition holds, and all conditions of Lemma 3 in Wu and Ghosal (2008) are satisfied, pointwise for all $X_{1:(t-1)}$. This implies that the second term in (18) is less than or equal to $\epsilon/2$. Combining this with the upper bound on the first term in (18), it is possible to show that $KL(f_0, f_{P_\epsilon})(\mathbf{x}_t \mid X_{1:(t-1)}) < \epsilon$, point-wise almost everywhere for $X_{1:(t-1)}$. Combining the above results and using the expression in (17), we have $KL(f_0(X), f(X)) \leq \epsilon$. Finally, we note that this bound holds with positive prior probability under the PDPM-VAR, lgPDPM-VAR and rgPDPM-VAR models, thus yielding the desired result in Theorem 3.

Proof of Theorem 4: The proof of this result is based on modifications of the arguments in the proof of Proposition 2 in Shen et al. (2013) and Lemma 2 in Canale and De Blasi (2017). We will show that for every $f_P \in \mathcal{F}_{n,\mathbf{jl}}$, it is possible to find another density $f_{\hat{P}}$ belonging to $\hat{\mathcal{G}}$ (the ϵ-net over $\mathcal{F}_{n,\mathbf{jl}}$) such that $||f_P - f_{\hat{P}}||_1 \leq \epsilon$. Since $N(\epsilon, \mathcal{F}_{n,\mathbf{jl}}, ||\cdot||_1)$ is the minimum cardinality of the ϵ-net over $\mathcal{F}_{n,\mathbf{jl}}$, we will be able to obtain a desired upper bound on the entropy if the number of balls required to cover the ϵ-net $\hat{\mathcal{G}}$ is bounded. Let us consider $P_1 = \sum_{h_1 \geq 1} \sum_{h_\sigma \geq 1} \pi^{(1)}_{h_1} \pi^{(1)}_{\sigma,h_\sigma} \delta_{\left(\Theta^{(1)}_{h_1}, \Sigma^{(1)}_{h_\sigma}\right)}$ and $P_2 = \sum_{h_1 \geq 1} \sum_{h_\sigma \geq 1} \pi^{(2)}_{h_1} \pi^{(2)}_{\sigma,h_\sigma} \delta_{\left(\Theta^{(2)}_{h_1}, \Sigma^{(2)}_{h_\sigma}\right)}$. Using Lemma 4 (elaborated in the sequel), the distance between the corresponding densities can be expressed as

$$
\begin{aligned}
||f_{P_1} - f_{P_2}||_1 \leq\ & 2\epsilon^2 + \sum_{h_1, h_\sigma < H_n} |\pi^{(1)}_{h_1} \pi^{(1)}_{\sigma,h_\sigma} - \pi^{(2)}_{h_1} \pi^{(2)}_{\sigma,h_\sigma}| + 4\epsilon \\
& + \sum_{h_1, h_\sigma \leq H_n} \pi^{(1)}_{h_1} \pi^{(1)}_{\sigma,h_\sigma} \left\| \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{t-k}\big) - \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1} \mathbf{x}_{t-k}\big) \right\|_1 \qquad (21)
\end{aligned}
$$

Let us first investigate the upper bound for the last term on the right hand side of (21). Note that

$$
\begin{aligned}
& \left\| \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{t-k}\big) - \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1} \mathbf{x}_{t-k}\big) \right\|_1 \\
\leq\ & \left\| \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{t-k}\big) - \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1} \mathbf{x}_{t-k}\big) \right\|_1 \\
+\ & \left\| \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{t-k}\big) - \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{t-k}\big) \right\|_1, \qquad (22)
\end{aligned}
$$

using the triangle inequality. The first term on the right hand side of (22) can be bounded above using the following steps. In particular, the first term is $\leq$

$$
\begin{aligned}
& \left\| \left\{ \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_T - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{T-k}\big) - \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_T - \sum_{k=1}^{K} A^{(2)}_{k,h_1} \mathbf{x}_{T-k}\big) \right\} \prod_{t=1}^{T-1} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{t-k}\big) \right\|_1 \\
& + \left\| \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_T - \sum_{k=1}^{K} A^{(2)}_{k,h_1} \mathbf{x}_{T-k}\big) \left\{ \prod_{t=1}^{T-1} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{t-k}\big) - \prod_{t=1}^{T-1} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1} \mathbf{x}_{t-k}\big) \right\} \right\|_1. \qquad (23)
\end{aligned}
$$

The first term in (23) may be written as $\int \left\{ \int \left| \left\{ \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_T - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{T-k}\big) - \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_T - \sum_{k=1}^{K} A^{(2)}_{k,h_1} \mathbf{x}_{T-k}\big) \right\} \right| d\mathbf{x}_T \right\} \prod_{t=1}^{T-1} \phi_{\Sigma^{(2)}_{h_\sigma}}\left(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{t-k}\right) d\mathbf{x}_{T-1} \ldots d\mathbf{x}_1$, which is $\leq \frac{2}{\pi\sqrt{\lambda_D(\Sigma^{(2)}_{h_\sigma})}} \times \int \ldots \int \left\{ \left\| \sum_{k=1}^{K} \left(A^{(1)}_{k,h_1} - A^{(2)}_{k,h_1} \right) \mathbf{x}_{T-k} \right\| \times \prod_{t=1}^{T-1} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1} \mathbf{x}_{t-k}\big) \right\} d\mathbf{x}_{T-1} \ldots d\mathbf{x}_1$, where we have used the well-known result $||\phi_\Sigma\big(\mathbf{x}_T - \boldsymbol{\mu}_1\big) - \phi_\Sigma\big(\mathbf{x}_T - \boldsymbol{\mu}_2\big)||_1 \leq \frac{2}{\pi\sqrt{\lambda_D(\Sigma)}} ||\boldsymbol{\mu}_1 -$

$\boldsymbol{\mu}_2||$. One can write $\left\| \sum_{k=1}^{K} \left(A^{(1)}_{k,h_1} - A^{(2)}_{k,h_1} \right) \mathbf{x}_{T-k} \right\| = \sqrt{\sum_{k=1}^{K} \sum_{l=1}^{D} \left\{ \sum_{l'=1}^{D} \left(A^{(1)}_{k,h_1}(l,l') - A^{(2)}_{k,h_1}(l,l') \right) x_{T-k,l'} \right\}^2}$
$\leq \sqrt{\sum_{k=1}^{K} \sum_{l=1}^{D} \left(\sum_{l'=1}^{D} \left(A^{(1)}_{k,h_1}(l,l') - A^{(2)}_{k,h_1}(l,l') \right)^2 \right) ||\mathbf{x}_{T-k}||^2} = \sqrt{\sum_{k=1}^{K} ||vec\big(A^{(1)}_{k,h_1} - A^{(2)}_{k,h_1}\big)||^2 \times ||\mathbf{x}_{T-k}||^2}$
$\leq \sum_{k=1}^{K} ||vec(A^{(1)}_{k,h_1} - A^{(2)}_{k,h_1})|| \times ||\mathbf{x}_{T-k}||$, where the second to last inequality is obtained using Cauchy-Schwarz inequality, and the last inequality uses the fact $\sum_{k=1}^{K} (a^*)_k^2 \leq (\sum_{k=1}^{K} |a^*|_k)^2$. Hence, the first term in (23) has an upper bound $2\big(\pi \sqrt{\lambda_D(\Sigma^{(2)}_{h_\sigma})}\big)^{-1} \times \mathcal{K}^*$, where $\mathcal{K}^* = \sum_{k=1}^{K} ||vec(A^{(1)}_{k,h_1} - A^{(2)}_{k,h_1})|| \int \left\{ ||\mathbf{x}_{T-k}|| \prod_{t=1}^{T-1} \phi_{\Sigma^{(2)}_{h_\sigma}} \big(\mathbf{x}_t - \sum_{k'=1}^{K} A^{(1)}_{k',h_1} \mathbf{x}_{t-k'} \big) \right\} d\mathbf{x}_{T-1} \dots d\mathbf{x}_1$. Using Cauchy-Schwarz, $\int \left\{ ||\mathbf{x}_{T-k}|| \prod_{t=1}^{T-1} \phi_{\Sigma^{(2)}_{h_\sigma}} \big(\mathbf{x}_t - \sum_{k'=1}^{K} A^{(1)}_{k',h_1} \mathbf{x}_{t-k'} \big) \right\} d\mathbf{x}_{T-1} \dots d\mathbf{x}_1 \leq \sqrt{\tilde{\zeta}} \times \sqrt{\zeta_{1:(T-k)}}$, where $\tilde{\zeta} = \int \prod_{t=T-k+1}^{T-1} \left\{ \phi_{\Sigma^{(2)}_{h_\sigma}} \big(\mathbf{x}_t - \sum_{k'=1}^{K} A^{(1)}_{k',h_1} \mathbf{x}_{t-k'} \big) \right\}^2 d\mathbf{x}_{T-1} \dots d\mathbf{x}_{T-k+1}$ and $\zeta_{1:(T-k)} = \int ||\mathbf{x}_{T-k}||^2 \left\{ \prod_{t=1}^{T-k} \phi_{\Sigma^{(2)}_{h_\sigma}} \big(\mathbf{x}_t - \sum_{k'=1}^{K} A^{(1)}_{k',h_1} \mathbf{x}_{t-k'} \big) \right\}^2 d\mathbf{x}_{T-k} \dots d\mathbf{x}_1$.

Noting that $\phi_\Sigma(\mathbf{x}) \leq \big(\frac{\lambda_1(\Sigma)}{\lambda_D(\Sigma)}\big)^{(D-1)/2} \phi_{\lambda_1(\Sigma) I_D}(\mathbf{x})$ and the fact that $\left(\frac{\lambda_1(\Sigma^{(2)}_{h_\sigma})}{\lambda_D(\Sigma^{(2)}_{h_\sigma})} \right) \leq u_{h_\sigma,l}$ for all densities belonging to $\mathcal{F}_{n,\mathbf{jl}}$, one can write $\tilde{\zeta} \leq \int \left\{ \prod_{t=T-k+1}^{T-1} (u_{h_\sigma,l})^{(D-1)/2} \phi_{\lambda_D(\Sigma^{(2)}_{h_\sigma}) I_D} \big(\mathbf{x}_t - \sum_{k'=1}^{K} A^{(1)}_{k',h_1} \mathbf{x}_{t-k'} \big) \right\}^2 d\mathbf{x}_{T-1} \dots d\mathbf{x}_{T-k+1} = \tilde{\zeta^*} \times \frac{(u_{h_\sigma,l})^{(D-1)k}}{\lambda_D^{Dk/2}(\Sigma^{(2)}_{h_\sigma})} \int \left\{ \prod_{t=T-k+1}^{T-1} \phi_{\frac{1}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma}) I_D} \big(\mathbf{x}_t - \sum_{k'=1}^{K} A^{(1)}_{k',h_1} \mathbf{x}_{t-k'} \big) d\mathbf{x}_{T-1} \dots d\mathbf{x}_{T-k+1} \right\}$, which is equal to $\tilde{\zeta^*} \times \frac{(u_{h_\sigma,l})^{(D-1)k}}{\lambda_D^{Dk/2}(\Sigma^{(2)}_{h_\sigma})}$ since the integral is one, where $\tilde{\zeta^*}$ is some constant.

Next, we need to derive an upper bound for $\zeta_{1:(T-k)}$. Note that $\zeta_{1:(T-k)} \leq \zeta^* \times \frac{(u_{h_\sigma,l})^{(D-1)(T-k)}}{\lambda_D^{D(T-k)/2}(\Sigma^{(2)}_{h_\sigma})} \int \left\{ ||\mathbf{x}_{T-k}||^2 \prod_{t=1}^{T-k} \phi_{\frac{1}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma}) I_D} \big(\mathbf{x}_t - \sum_{k'=1}^{K} A^{(1)}_{k',h_1} \mathbf{x}_{t-k'} \big) d\mathbf{x}_{T-k} \right\} d\mathbf{x}_{T-k-1} \dots d\mathbf{x}_1 = \zeta^* \times \frac{(u_{h_\sigma,l})^{(D-1)(T-k)}}{\lambda_D^{D(T-k)/2}(\Sigma^{(2)}_{h_\sigma})} I_{1:(T-k)}$, where ζ^* is some constant and $I_{1:(T-k)}$ denotes the integral term. Therefore, $\mathcal{K}^* \leq \sum_{k=1}^{K} ||vec(A^{(1)}_{k,h_1} - A^{(2)}_{k,h_1})|| \times \sqrt{\tilde{\zeta^*} \times \frac{(u_{h_\sigma,l})^{(D-1)k}}{\lambda_D^{Dk/2}(\Sigma^{(2)}_{h_\sigma})} \times \zeta^* \times \frac{(u_{h_\sigma,l})^{(D-1)(T-k)}}{\lambda_D^{D(T-k)/2}(\Sigma^{(2)}_{h_\sigma})} I_{1:(T-k)}} = \tilde{\zeta^{**}} \sum_{k=1}^{K} ||vec(A^{(1)}_{k,h_1} - A^{(2)}_{k,h_1})|| \times \frac{(u_{h_\sigma,l})^{(D-1)T/2}}{\lambda_D^{DT/2}(\Sigma^{(2)}_{h_\sigma})} \times \sqrt{I_{1:(T-k)}}$. Recalling that $I_{1:(T-k)} = \int \left\{ ||\mathbf{x}_{T-1}||^2 \prod_{t=1}^{T-1} \phi_{\lambda_D(\Sigma^{(2)}_{h_\sigma}) I_D} \big(\mathbf{x}_t - \sum_{k'=1}^{K} A^{(1)}_{k',h_1} \mathbf{x}_{t-k'} \big) d\mathbf{x}_{T-1} \right\} d\mathbf{x}_{T-2} \dots d\mathbf{x}_1$, we note that $I_{1:(T-k)}$ can be bounded using the relationship $E\big(\frac{||\mathbf{x}_{T-k}||^2}{\frac{1}{2}\lambda_D\left(\Sigma^{(2)}_{h_\sigma}\right)}\big) = D + \sum_{l'=1}^{D} \left(\sum_{k'=1}^{K} A^{(1)}_{k',h_1}(l',\cdot) \mathbf{x}_{T-k-k'} \right)^2 \leq$

$D + K\bar{a}^2_{h_1,j}\big(\sum_{k'=1}^{K} D||\mathbf{x}_{T-k-k'}||^2\big)$. In other words, $I_{1:(T-k)} \leq$

$$\frac{D}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma}) + K\bar{a}^2_{h_1,j}\int \big(\sum_{k'=1}^{K} D||\mathbf{x}_{T-k-k'}||^2\big)\prod_{t=1}^{T-k-1}\phi_{\frac{1}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma})I_D}\big(\mathbf{x}_t - \sum_{k'=1}^{K} A^{(1)}_{k',h_1}\mathbf{x}_{t-k'}\big)d\mathbf{x}_{T-k-1}\dots d\mathbf{x}_1,$$
$$= \frac{D}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma}) + KD\bar{a}^2_{h_1,j}I^*_{1:(T-k-1)} = \frac{D}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma}) + KD\bar{a}^2_{h_1,j}\big\{\tilde{I}^*_{1:(T-k-1),1} + \ldots + \tilde{I}^*_{1:(T-k-1),K}\big\},$$

using Cauchy-Schwarz inequality and the sieve constructions, and where $I^*_{1:(T-k-1)}$ denotes the integral in the first line of the above upper bound, which is decomposable into K different integrals denoted as $\{\tilde{I}^*_{1:(T-k-1),1}, \ldots, \tilde{I}^*_{1:(T-k-1),K}\}$. For $k' = 1$, the term $\tilde{I}^*_{1:(T-k-1),1}$ in the above integral can be written as

$$\begin{aligned} = & \int ||\mathbf{x}_{T-k-1}||^2\phi_{\frac{1}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma})I_D}\big(\mathbf{x}_{T-k-1} - \sum_{k'=1}^{K} A^{(1)}_{k',h_1}\mathbf{x}_{t-k-1-k'}\big)d\mathbf{x}_{T-k-1} \\ & \times \int \Big\{\prod_{t=1}^{T-k-2}\phi_{\frac{1}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma})I_D}\big(\mathbf{x}_t - \sum_{k'=1}^{K} A^{(1)}_{k',h_1}\mathbf{x}_{t-k'}\big)d\mathbf{x}_{T-k-2}\dots d\mathbf{x}_1\Big\} \\ & \leq \frac{D}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma}) + KD\bar{a}^2_{h_1,j}I^*_{1:(T-k-2)} = \frac{D}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma}) + KD\bar{a}^2_{h_1,j} \times \{\tilde{I}^*_{1:(T-k-2),1} + \ldots + \tilde{I}^*_{1:(T-k-2),K}\}, \end{aligned}$$

using similar notations as previously used. Similarly, when $k' = 2$, the term $\tilde{I}^*_{1:(T-k-1),2}$ in the upper bound for $I_{1:(T-k)}$ may be written as $\tilde{I}^*_{1:(T-k-1),2} \leq \frac{D}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma}) + KD\bar{a}^2_{h_1,j}I^*_{1:(T-k-3)} = \frac{D}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma}) + KD\bar{a}^2_{h_1,j} \times \{\tilde{I}^*_{1:(T-k-3),1} + \ldots + \tilde{I}^*_{1:(T-k-3),K}\}$, using similar notation as above.

The number of terms in the above expression can be derived via a decision tree. For lag K, the total number of terms will be less than or equal to K^{T-1}. In the special case when $K = 1$, the upper bound for I_{T-1} is given as

$$\begin{aligned} I_{1:(T-1)} \quad & \leq D\frac{1}{2}\lambda_D(\Sigma^{(2)}_{h_\sigma})\Big\{1 + KD\bar{a}^2_{h_1,j} + (KD\bar{a}^2_{h_1,j})^2 + \ldots + (KD\bar{a}^2_{h_1,j})^{T-2}\Big\} + (KD\bar{a}^2_{h_1,j})^{T-1} \\ & \leq D\lambda_D(\Sigma^{(2)}_{h_\sigma})(KD\bar{a}^2_{h_1,j})^{T-2} + (KD\bar{a}^2_{h_1,j})^{T-1} \leq c^*(KD\bar{a}^2_{h_1,j})^{T-1}, \end{aligned} \tag{24}$$

where c^* is some constant. Given that there are a total of K^{T-1} terms, each being bounded as in (24), the total upper bound for lag K is given as

$$I_{1:(T-1)} \leq D\lambda_D(\Sigma^{(2)}_{h_\sigma})(KD\bar{a}^2_{h_1,j})^{T-2} + (KD\bar{a}^2_{h_1,j})^{T-1} \leq c^*(KD\bar{a}^2_{h_1,j})^{T-1} \times K^{T-1} = c^*(K^2D\bar{a}^2_{h_1,j})^{T-1}.$$

Hence using previous calculations, $\mathcal{K}^* \leq \frac{c^{**}}{\underline{\sigma}_n^{(DT+1)}} \times \Big\{\sum_{k=1}^{K}||vec(A^{(1)}_{k,h_1} - A^{(2)}_{k,h_1})||(u_{h_\sigma,l})^{T(D-1)/2} \times \big(DK^2\bar{a}^2_{h_1,j}\big)^{T-1}\Big\}$. The second term in (23) is equal to $\Big\|\phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_T - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{T-k}\big)\Big\{\prod_{t=1}^{T-1}\phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1}\mathbf{x}_{t-k}\big) - \prod_{t=1}^{T-1}\phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big)\Big\}\Big\|_1 = \Big\|\prod_{t=1}^{T-1}\phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1}\mathbf{x}_{t-k}\big) - \prod_{t=1}^{T-1}\phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big)\Big\|_1$, since $\Big\|\phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_T - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{T-k}\big)\Big\|_1 = 1$. Hence the

first term in the upper bound in (22) can be written as

$$\begin{aligned}\mathcal{K}^* &\lesssim c^{**}\frac{\sum_{k=1}^{K}||vec(A_{k,h_1}^{(1)}-A_{k,h_1}^{(2)})||(u_{h_\sigma,l})^{T(D-1)/2}}{\underline{\sigma}_n^{DT+1}}\times\left(DK^2\bar{a}_{h_1,j}^2\right)^{T-1}\\ &+\Big\|\prod_{t=1}^{T-1}\phi_{\Sigma_{h_\sigma}^{(2)}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)-\prod_{t=1}^{T-1}\phi_{\Sigma_{h_\sigma}^{(2)}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(2)}\mathbf{x}_{t-k}\big)\Big\|_1\\ &\lesssim c^{**}T\sum_{k=1}^{K}||vec(A_{k,h_1}^{(1)}-A_{k,h_1}^{(2)})||(u_{h_\sigma,l})^{T(D-1)/2}\times\left\{\frac{\underline{\sigma}_n^{TD+1}}{\left(DK^2\bar{a}_{h_1,j}^2\right)^{T-1}}\right\}^{-1}\end{aligned}\tag{25}$$

using similar steps to obtain an upper bound for $\Big\|\prod_{t=1}^{T-1}\phi_{\Sigma_{h_\sigma}^{(2)}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)-\prod_{t=1}^{T-1}\phi_{\Sigma_{h_\sigma}^{(2)}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(2)}\mathbf{x}_{t-k}\big)\Big\|_1$. For the second term in (22), note that

$$\begin{aligned}&\Big\|\prod_{t=1}^{T}\phi_{\Sigma_{h_\sigma}^{(1)}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)-\prod_{t=1}^{T}\phi_{\Sigma_{h_\sigma}^{(2)}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)\Big\|_1\leq\\ &\Big\|\prod_{t=1}^{T}\phi_{\tilde{\Sigma}_{h_\sigma}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)-\prod_{t=1}^{T}\phi_{\Sigma_{h_\sigma}^{(2)}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)\Big\|_1+\\ &\Big\|\prod_{t=1}^{T}\phi_{\tilde{\Sigma}_{h_\sigma}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)-\prod_{t=1}^{T}\phi_{\Sigma_{h_\sigma}^{(1)}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)\Big\|_1,\end{aligned}\tag{26}$$

where $\tilde{\Sigma}_{h_\sigma}=\left(O_{t,h_\sigma}^{(2)}\Lambda_{t,h_\sigma}^{(1)}(O_{th_\sigma}^{(2)})'\right)^{-1}$, and $\Sigma_{h_\sigma}^{(j)}=\left(O_{t,h_\sigma}^{(j)}\Lambda_{t,h_\sigma}^{(j)}(O_{t,h_\sigma}^{(j)})'\right)^{-1}$, $j=1,2$. Using Csiszar's inequality the first term in (26) $\leq$

$$\begin{aligned}&\sqrt{2\int\ldots\int\log\left(\prod_{t=1}^{T}\frac{\phi_{\tilde{\Sigma}_{h_\sigma}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)}{\phi_{\Sigma_{h_\sigma}^{(2)}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)}\right)\left\{\prod_{t=1}^{T}\phi_{\tilde{\Sigma}_{h_\sigma}}\big(\mathbf{x}_t-\sum_{k=1}^{K}A_{k,h_1}^{(1)}\mathbf{x}_{t-k}\big)\right\}d\mathbf{x}_T\ldots d\mathbf{x}_1}\\ &=\sqrt{2\times\frac{1}{2}\sum_{t=1}^{T}\left\{\log\det\big(\tilde{\Sigma}_{h_\sigma}^{-1}\Sigma_{h_\sigma}^{(2)}\big)+tr\big((\Sigma_{h_\sigma}^{(2)})^{-1}\tilde{\Sigma}_{h_\sigma}\big)-D\right\}}\\ &=\sqrt{\sum_{t=1}^{T}\sum_{d'=1}^{D}\left\{\frac{\lambda_{d'}(\Sigma_{h_\sigma}^{(1)})}{\lambda_{d'}(\Sigma_{h_\sigma}^{(2)})}-\log\big(\frac{\lambda_{d'}(\Sigma_{h_\sigma}^{(1)})}{\lambda_{d'}(\Sigma_{h_\sigma}^{(2)})}\big)-1\right\}}.\end{aligned}\tag{27}$$

Similarly, the second term in R.H.S. of (26) $\leq\sqrt{\sum_{t=1}^{T}\left\{\log\det\big(\tilde{\Sigma}_{h_\sigma}^{-1}\Sigma_{h_\sigma}^{(1)}\big)+tr\big((\Sigma_{h_\sigma}^{(1)})^{-1}\tilde{\Sigma}_{h_\sigma}\big)-D\right\}}$. Using similar steps as in the proof of Lemma 2 in Canale and De Blasi (2017), it is possible to show that the second term in the R.H.S. of (26) (and hence the last term in the R.H.S. of

(22)), has an upper bound given by $\sqrt{2T||O^{(1)}_{h_\sigma} - O^{(2)}_{h_\sigma}||_2 \frac{\lambda_1(\Sigma^{(1)}_{h_\sigma})}{\lambda_D(\Sigma^{(1)}_{h_\sigma})}}$. Finally, we need to establish an upper bound for the term $\sum_{h_1,h_\sigma<H_n} |\pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} - \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}|$ in (21), which is given by Lemma 5 as $\sum_{h_1,h_\sigma<H_n} \left|\tilde{\pi}^{(1)}_{h_1}\tilde{\pi}^{(1)}_{\sigma,h_\sigma} - \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}\right| + \left|1-(1-\epsilon)^2\right|$, where $\tilde{\pi} = \frac{\pi_h}{(1-\sum_{h>H}\pi_h)}$.

For a given $f_P \in \mathcal{F}_{n,\mathbf{jl}}$ with $P = \sum_{h_1\geq 1}\sum_{h_\sigma\geq 1} \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma}\delta_{\left(\Theta^{(1)}_{h_1},\Sigma^{(1)}_{h_\sigma}\right)}$ and $\Sigma_h = (O_h\Lambda_h O_h^T)^{-1}$ where $\Lambda_h = diag(\lambda_{h,1},\ldots,\lambda_{h,D})$, we will construct another density $f_{\hat{P}}$ with $\hat{P} = \sum_{h_1\geq 1}\sum_{h_\sigma\geq 1} \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma}\delta_{\left(\hat{\Theta}^{(1)}_{h_1},\hat{\Sigma}^{(1)}_{h_\sigma}\right)}$ within the ϵ-net and then compute the cardinality of the ϵ-net set to derive an upper bound for the entropy of sieve $\mathcal{F}_{n,\mathbf{jl}}$. To construct such a density, we will choose:

1. $\hat{A}_{k,h_1} \in \hat{\mathcal{R}}_{h_1}, h_1 = 1,\ldots,H$, where $\hat{\mathcal{R}}_{h_1}$ is a ϵ^*-net of $\mathcal{R}_{h_1} := \{A \in \Re^{D\times D} : \underline{a}_{h_1,j} \leq ||vec(A)|| \leq \bar{a}_{h_1,j}\}$, such that $||vec(A)_{k,h_1} - vec(\hat{A})_{k,h_1}|| \leq \epsilon^*$, $k = 1,\ldots,K$, where $\epsilon^* = \pi\epsilon\left\{\frac{\underline{\sigma}_n^{TD+1}}{2T(u_{h_\sigma,l})^{T(D-1)/2}\times\left(DK^2\bar{a}^2_{h_1,j}\right)^{T-1}}\right\}$ using (25) and the fact that $\underline{\sigma}_n < 1$ for large n.

2. $\{\hat{\pi}_{h_1}\hat{\pi}_{h_\sigma}, h_1, h_\sigma \leq H_n\} \in \hat{\Delta}$, where $\hat{\Delta}$ is a ϵ-net of a H_n^2 dimensional probability simplex such that $\sum_{h_1,h_\sigma\leq H_n} |\tilde{\pi}_{h_1}\tilde{\pi}_{h_\sigma} - \hat{\pi}_{h_1}\hat{\pi}_{h_\sigma}| \leq \epsilon$, and $\tilde{\pi}_h = \frac{\pi_h}{\sum_{h\leq H_n}\pi_h}, h \leq H_n$.

3. $\hat{O}_h \in \hat{\mathcal{O}}_h$, where $\hat{\mathcal{O}}_h$ is a δ_h-net of the set $\mathcal{O}_h$ defined as the set of $D\times D$ orthogonal matrices with respect to the spectral norm $||\cdot||_2$ with $\delta_h = \epsilon^2/(2Du_{h,l})$ such that $||O_h - \hat{O}_h||_2 \leq T\delta_h$.

4. $(m_{h,1},...,m_{h_D}) \in \{1,...,M\}^D, h = 1,...,H$, such that $\hat{\lambda}_{h,l} = \{\underline{\sigma}^2(1+\epsilon\sqrt{D})^{m_{h,l}-1}\}^{-1}$ will satisfy $1 \leq \hat{\lambda}_{h,l}/\lambda_{h,l} < (1+\epsilon/\sqrt{D})$.

Using this construction, the term in (27) is shown to be bounded by $\sqrt{T\sum_{d'=1}^{D}\left\{\left(\frac{\hat{\lambda}_{h,D-d'+1}}{\lambda_{h,D-d'+1}} - 1\right)^2\right\}}$. Moreover under this construction, it can be shown that $||f_P - f_{\hat{P}}||_1 < C^*\epsilon$ for some constant C^*, by employing some additional algebra and the above arguments. Further, the cardinality of the ϵ-net can be computed by noting that $\#(\hat{\Delta}) \lesssim \epsilon^{-H_n^2}$ for $j = 1,2, \#(\hat{\mathcal{O}}_h) \lesssim \delta_h^{-D(D-1)/2}, \#(\hat{\mathcal{R}}_{k,h}) \lesssim [(\frac{\bar{a}_h}{\epsilon^*}+1)^{D^2} - (\frac{\underline{a}_h}{\epsilon^*}-1)^{D^2}]$. Using these quantities, one can write the upper bound for the exponential of the entropy bound as $\lesssim (M)^{DH_n}\epsilon^{-H_n^2}\times\mathcal{K}^*$, where

$$\mathcal{K}^* = \prod_{h_1\leq H_n}\left\{(\frac{\bar{a}_{h_1,j}}{\epsilon^*}+1)^{D^2} - (\frac{\underline{a}_{h_1,j}}{\epsilon^*}-1)^{D^2}\right\}^K \prod_{h_\sigma\leq H_n}\left\{\frac{2Du_{h_\sigma,l}}{\epsilon^2}\right\}^{D(D-1)/2}$$

$$\lesssim \prod_{h_\sigma\leq H_n}\left\{\frac{2Du_{h_\sigma,l}}{\epsilon^2}\right\}^{D(D-1)/2} \prod_{h_1\leq H_n}\left\{\left(\frac{C^*_{h_{1,j},h_{\sigma,l}}\bar{a}_{h_1,j}}{\underline{\sigma}_n\epsilon}+1\right)^{D^2} - \left(\frac{C^*_{h_{1,j},h_{\sigma,l}}\underline{a}_{h_1,j}}{\underline{\sigma}_n\epsilon}-1\right)^{D^2}\right\}^K$$

and $C^*_{h_{1,j},h_{\sigma,l}} = \frac{2}{\pi}\left\{T(u_{h_\sigma,l})^{(T)(D-1)/2}\right\} \times \frac{\left(DK^2\bar{a}^2_{h_1,j}\right)^{T-1}}{\underline{\sigma}_n^{TD}}$.

Proof of Corollary 1: For computing the sieve entropy bound corresponding to lgPDPM-VAR, note that using similar calculations as in (21), one can show that $|f_{P_1} - f_{P_2}||_1 =$

$$\sum_{h_{11}} \cdots \sum_{h_{1K}} \sum_{h_\sigma < H_n} \Big(\prod_{k=1}^{K} \pi^{(1)}_{k,h_{1k}} \Big) \pi^{(1)}_{\sigma,h_\sigma} \Big\| \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}} \big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_{1k}} \mathbf{x}_{t-k} \big) - \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}} \big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_{1k}} \mathbf{x}_{t-k} \big) \Big\|_1$$

$$+ \sum_{h_{11}} \cdots \sum_{h_{1K}} \sum_{h_\sigma < H_n} \Big| \Big(\prod_{k=1}^{K} \pi^{(1)}_{k,h_{1k}} \Big) \pi^{(1)}_{\sigma,h_\sigma} - \Big(\prod_{k=1}^{K} \pi^{(2)}_{k,h_{1k}} \Big) \pi^{(2)}_{\sigma,h_\sigma} \Big| + K\epsilon_1^K + L\epsilon_1, \tag{28}$$

for some constant L. Using similar calculations as under the PDPM-VAR case, $N(\epsilon_1, \mathcal{F}_{n,\mathbf{jl}}, ||\cdot ||_1) \lesssim \Big(\frac{M^D}{\epsilon_1^K} \Big)^{H_n} \prod_{k=1}^{K} \prod_{h_{1k} \le H_n} \big\{ (\frac{\bar{a}_{h_{1k},j}}{\epsilon_1^*} + 1)^{D^2} - (\frac{\underline{a}_{h_{1k},j}}{\epsilon_1^*} - 1)^{D^2} \big\} \prod_{h_\sigma \le H_n} \big\{ \frac{2Du_{h_\sigma,l}}{\epsilon_1^2} \big\}^{D(D-1)/2} \lesssim \Big(\frac{M^D}{\epsilon_1^{C_1^*}} \Big)^{H_n} \times \prod_{h_\sigma \le H_n} \big\{ \frac{2Du_{h_\sigma,l}}{\epsilon_1^2} \big\}^{D(D-1)/2} \prod_{k=1}^{K} \prod_{h_{1k} \le H_n} \Big\{ \big(\frac{C^{**}_{h_{1,j},h_{\sigma,l}} \bar{a}_{h_{1k},j}}{\underline{\sigma}_n \epsilon_1} + 1 \big)^{D^2} - \big(\frac{C^{**}_{h_{1,j},h_{\sigma,l}} \underline{a}_{h_{1k},j}}{\underline{\sigma}_n \epsilon_1} - 1 \big)^{D^2} \Big\}$, where $C^{**}_{h_{1,j},h_{\sigma,l}} = \frac{2}{\pi} \Big\{ \frac{T(u_{h_\sigma,l})^{(T-1)(D-1)/2}}{\underline{\sigma}_n^{TD}} \Big\} \times \big(DK^2 \max \{ \bar{a}^2_{h_{11},j}, \ldots, \bar{a}^2_{h_{1K},j} \} \big)^{T-1}$.

Similarly, for computing the entropy bound corresponding to sieves in the rgPDPM-VAR model, note that using similar calculations as before, one can show that the densities satisfy $||f_{P_1} - f_{P_2}||_1 = \sum_{h_{11} \le H_n} \cdots \sum_{h_{1D} \le H_n} \sum_{h_\sigma < H_n} \Big| \big(\prod_{d'=1}^{D} \pi^{*(1)}_{d',h_{1d'}} \big) \pi^{(1)}_{\sigma,h_\sigma} - \big(\prod_{d'=1}^{D} \pi^{*(2)}_{d',h_{1d'}} \big) \pi^{(2)}_{\sigma,h_\sigma} \Big| + \sum_{h_{11} \le H_n} \cdots \sum_{h_{1D} \le H_n} \sum_{h_\sigma \le H_n} \big(\prod_{d'=1}^{D} \pi^{*(1)}_{d',h_{1d'}} \big) \pi^{(1)}_{\sigma,h_\sigma} \times \Big\| \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}} \big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_{11},\ldots,h_{1d}} \mathbf{x}_{t-k} \big) - \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}} \big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_{11},\ldots,h_{1d}} \mathbf{x}_{t-k} \big) \Big\|_1 + K\epsilon_2^K + L_2^* \epsilon_2$, for some constant L_2^*. Similar calculations as before yield the bound for exponential of the entropy $N(\epsilon_2, \mathcal{F}_{n,\mathbf{jl}}, ||\cdot||_1))$ as:

$$\lesssim \Big(\frac{M^D}{\epsilon_2^D} \Big)^{H_n} \prod_{h_\sigma \le H_n} \big\{ \frac{2Du_{h_\sigma,l}}{\epsilon_2^2} \big\}^{D(D-1)/2} \prod_{d'=1}^{D} \prod_{h_{1d'} \le H_n} \big\{ (\frac{\bar{a}_{h_{1d'},j}}{\epsilon_2^*} + 1)^D - (\frac{\underline{a}_{h_{1d'},j}}{\epsilon_2^*} - 1)^D \big\}^K \lesssim \Big(\frac{M^D}{\epsilon_2^{C_1^*}} \Big)^{H_n}$$

$$\times \prod_{h_\sigma \le H_n} \big\{ \frac{2Du_{h_\sigma,l}}{\epsilon_2^2} \big\}^{D(D-1)/2} \prod_{d'=1}^{D} \prod_{h_{1d'} \le H_n} \left\{ \left(\frac{\tilde{C}^*_{h_{1,j},h_{\sigma,l}} \bar{a}_{h_{1d'},j}}{\underline{\sigma}_n \epsilon_2} + 1 \right)^D - \left(\frac{\tilde{C}^*_{h_{1,j},h_{\sigma,l}} \underline{a}_{h_{1d'},j}}{\underline{\sigma}_n \epsilon_2} - 1 \right)^D \right\}^K \tag{29}$$

where $\tilde{C}^*_{h_{1,j},h_{\sigma,l}} = \frac{2}{\pi} \Big\{ \frac{T(u_{h_\sigma,l})^{(T-1)(D-1)/2}}{\underline{\sigma}_n^{TD}} \Big\} \times \big(DK^2 \max \{ \bar{a}^2_{h_{11},j}, \ldots, \bar{a}^2_{h_{1d},j} \} \big)^{T-1}$.

Proof of Theorem 5: The proof follows using Theorem 2 and the entropy bounds established in Theorem 3. For the case of PDPM-VAR, consider the sieves

$$\mathcal{F}_{n,\mathbf{j},\mathbf{l}} = \Big\{ f_p \in \mathcal{F}_n : \text{for } h_1, h_\sigma \le H_n, n^{H_n^2}(j_{h_1} - 1) \le ||vec(A_{k,h_1})|| \le n^{H_n^2} j_{h_1}, \forall k,$$

$$n^{l_{h_\sigma} - 1} < \frac{\lambda_1(\Sigma_{h_\sigma})}{\lambda_D(\Sigma_{h_\sigma})} \le n^{l_{h_\sigma}} \Big\},\ \mathcal{F}_n = \Big\{ f_p : P = \sum_{h_1 \ge 1} \sum_{h_\sigma \ge 1} \pi_{h_1} \pi_{\sigma,h_\sigma} \delta_{\Theta_{h_1},\Sigma_{h_\sigma}} : \sum_{h_1 > H_n} \pi_{h_1} < \epsilon,$$

$$\sum_{h_\sigma > H_n} \pi_{\sigma,h_\sigma} < \epsilon, \text{for } h_\sigma \le H_n,\ \underline{\sigma}_n^2 \le \lambda_D, \lambda_1 \le \underline{\sigma}_n^2 (1 + \epsilon/\sqrt{D})^{M_n},\ 1 < \frac{\lambda_1}{\lambda_D} \le n^{H_n} \Big\}, \tag{30}$$

where $M_n = \underline{\sigma}^{-2c_2} = n$ and $H_n = \lfloor Cn\epsilon^2/\log(n) \rfloor$ for some positive constant C, and clearly $\mathcal{F}_n \subset \cup_{\mathbf{j},\mathbf{l}} \mathcal{F}_{n,\mathbf{j},\mathbf{l}}$. Comparing to the notations used in the manuscript, we note that $\underline{a}_{h_1,j} = n^{H_n}(j_{h_1} - 1), \bar{a}_{h_1,j} = n^{H_n} j_{h_1}$, and $u_{h_\sigma,l} = n^{l_{h_\sigma}}$, for integers $(j_1, \ldots, j_{H_n}) \in \mathbb{N}$ and $(l_1, \ldots, l_{H_n}) \in \{1, \ldots, H_n\}$.

Using Lemma 6 in the Appendix, it is clear that condition (2A) holds for the PDPM-VAR. Next, we will derive the entropy bounds and the complement of the prior probability for the sieves and illustrate that the summability condition *(2B)* in Theorem 2 holds.

Now, using Theorem 4, the upper bound on the entropy term is $\lesssim (M^D \epsilon^{-C_1})^{H_n} \times \prod_{h_\sigma=1}^{H_n} \left\{\frac{2Du_{h_\sigma,l}}{\epsilon^2}\right\}^{D(D-1)/2} \times \mathcal{K}^*$, $\mathcal{K}^* = \prod_{h_1=1}^{H_n} \left\{ \left(\frac{C^*_{h_1,j,h_\sigma,l}\bar{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} + 1\right)^{D^2} - \left(\frac{C^*_{h_1,j,h_\sigma,l}\underline{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} - 1\right)^{D^2} \right\}^K$,

$$
\begin{aligned}
\text{i.e. } \mathcal{K}^* &\approx \prod_{h_1 \le H_n} \left(C^*_{h_1,j,h_\sigma,l}\right)^{KD^2} \left\{ \left(\frac{\bar{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} + o_n(1)\right)^{D^2} - \left(\frac{\underline{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} - o_n(1)\right)^{D^2} \right\}^K \\
&\lesssim \prod_{h_1 \le H_n} \left(C^*_{h_1,j,h_\sigma,l}\right)^{KD^2} \left\{ \left(\frac{\bar{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} + 1\right)^{D^2} - \left(\frac{\underline{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} - 1\right)^{D^2} \right\}^K \lesssim \left\{(u_{h_\sigma,l})^{(T-1)(D-1)/2}\right\}^{KD^2H_n} \\
&\times \prod_{h_1 \le H_n} \left[\frac{2}{\pi}\left(T(u_{h_\sigma,l})^{(T)(D-1)/2}\right) \times \frac{\left(DK^2\bar{a}^2_{h_1,j}\right)^{T-1}}{\underline{\sigma}_n^{TD}}\right]^{KD^2} \left\{ \left(\frac{\bar{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} + 1\right)^{D^2} - \left(\frac{\underline{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} - 1\right)^{D^2} \right\}^K, \quad (31)
\end{aligned}
$$

where $o_n(1)$ is a vanishing term with increasing sample size. Using similar steps as in the proof of Theorem 2 in Canale and De Blasi (2017), it is possible to show that $\left[\left(\frac{\bar{a}_{h_1,j}}{\underline{\sigma}_n\epsilon/2} + 1\right)^{D^2} - \left(\frac{\underline{a}_{h_1,j}}{\underline{\sigma}_n\epsilon/2} - 1\right)^{D^2}\right]^K \lesssim \left[\frac{n^{(H_n + \frac{1}{2c_2})D^2} j_{h1}^{D^2-1}}{(\epsilon)^{D^2}}\right]^K$, when n and j_{h_1} are large. Hence, when n is large enough,

$$
\begin{aligned}
&\left[\frac{2T}{\pi}\frac{\left(DK^2\bar{a}^2_{h_1,j}\right)^{T-1}}{\underline{\sigma}_n^{TD}}\right]^{KD^2} \left\{ \left(\frac{\bar{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} + 1\right)^{D^2} - \left(\frac{\underline{a}_{h_1,j}}{\underline{\sigma}_n\epsilon} - 1\right)^{D^2} \right\}^K \\
&\le \left\{\frac{n^{(H_n^2 + \frac{1}{2c_2})D^2} j_{h1}^{D^2-1}}{(\epsilon)^{D^2}}\right\}^K \times \left[\frac{2T}{\pi}\left(\frac{DK^2\bar{a}^2_{h_1,j}}{\underline{\sigma}_n}\right)^{T-1} \times \left(\underline{\sigma}_n\right)^{-TD}\right]^{KD^2} \\
&\approx \mathcal{C}\left\{\frac{n^{(H_n^2 + \frac{1}{2c_2})D^2} j_{h1}^{D^2-1}}{(\epsilon)^{D^2}}\right\}^K \times \left(n^{2H_n^2 + 1/(2c_2)} j_{h_1}^2\right)^{KD^2(T-1)} \times \left(n^{TD/c_2}\right)^{KD^2} \le \mathcal{C} \times \left(\frac{1}{\epsilon}\right)^{KD^2} \times \\
&\exp\left\{H_n^2 KD^2(2T-3)\log(n) + \frac{1}{c_2}\left(KD^2(T-1) + TKD^3\right)\log(n)\right\} \times \left(j_{h1}\right)^{2KD^2(T-2) + K(D^2-1)},
\end{aligned}
$$

where $\mathcal{C}$ is a constant independent of n. Further,

$$
\begin{aligned}
&\left\{\frac{2Du_{h_\sigma,l}}{\epsilon^2}\right\}^{D(D-1)/2} \left\{(u_{h_\sigma,l})^{(T)(D-1)/2}\right\}^{KD^2H_n} = \mathcal{C}_1 (u_{h_\sigma,l})^{D(D-1)/2 + KD^2H_n(T)(D-1)/2} \\
&\approx \mathcal{C}_1 \exp\left\{\left(\frac{D(D-1)}{2} + KD^2H_n(T)\frac{D-1}{2}\right)\log(n^{l_{h_\sigma}})\right\} \times \left(\frac{1}{\epsilon}\right)^{D(D-1)},
\end{aligned}
$$

for large n, where the constant $\mathcal{C}_1$ that does not depend on n.

Hence we have

$$N(\epsilon, \mathcal{F}_{n,\mathbf{jl}}, ||\cdot||_1) \lesssim \mathcal{C}_1^{H_n} \exp\left\{ DH_n \log(M) + H_n \log\frac{C_1}{\epsilon} \right\}$$
$$\times \quad \exp\left\{ H_n^3(2KD^2(T-2)+KD^2)\log(n) + \frac{H_n}{c_2}\big(KD^2(T-1)+TKD^3\big)\log(n) \right\}$$
$$\times \quad \left\{ \prod_{h_\sigma \le H_n} (n^{l_{h_\sigma}})^{\frac{D(D-1)}{2}+KD^2H_n(T)\frac{D-1}{2}} \right\} \left\{ \prod_{h_1 \le H_n} (j_{h1})^{2KD^2(T-2)+K(D^2-1)} \right\} \times \left(\frac{1}{\epsilon}\right)^{H_n(KD^2+D(D-1))} \quad (32)$$

Further, under the specification $P_1^*(A_1,\ldots,A_K) = \prod_{k=1}^K P_1^*(A_k)$, we have

$$\Pi(\mathcal{F}_{n,\mathbf{jl}}) \le \prod_{h_1 \le H_n} P_1^*\big(||vec(A_k)|| > n^{H_n^2}(j_{h_1}-1),\ \forall k\big) \prod_{h_\sigma \le H_n} P_2^*\big(\lambda_1(\Sigma)/\lambda_D(\Sigma) > n^{(l_{h_\sigma}-1)}\big),$$
$$\lesssim \quad \prod_{h_1 \le H_n} \left\{ \big(n^{H_n^2}(j_{h_1}-1)\big)^{-1_{(j_{h_1}\ge 2)}2(r+1)} \right\}^K \times \prod_{h_\sigma \le H_n} (n^{(l_{h_\sigma}-1)})^{-1_{(l_{h_\sigma}\ge 1)}\kappa}$$
$$\approx \quad \left\{ n^{-2H_n^3(r+1)K} \prod_{h_1 \le H_n} \big(j_{h_1}-1\big)^{-1_{(j_{h_1}\ge 2)}2K(r+1)} \right\} \times \left\{ \prod_{h_\sigma \le H_n} (n^{\kappa(l_{h_\sigma}-1)})^{-1_{(l_{h_\sigma}\ge 1)}} \right\}, \text{ for large } n. \quad (33)$$

Under Lemma 7 in the Appendix, we can show that for the PDPM-VAR, $\sum_{j_{h_1}\in N}\sum_{1\le l_{h_\sigma}\le H_n} \sqrt{N(\epsilon,\mathcal{F}_{n,j_{h_1}l_{h_\sigma}})\Pi(\mathcal{F}_{n,j_{h_1}l_{h_\sigma}})}e^{-(4-c)n\epsilon^2} \to 0$ as $n\to\infty$ for a suitable choice of constants. Hence the condition $(2B)$ in Theorem 2 is satisfied and the strong posterior consistency result is proved corresponding to the PDPM-VAR model.

To prove the summability result for the lgPDPM-VAR approach, consider the sieves defined in the manuscript as

$$\mathcal{F}_n = \Bigg\{ f_p : P = \sum_{h_{11}=1}^{\infty} \cdots \sum_{h_{1K}=1}^{\infty} \sum_{h_\sigma=1}^{\infty} \pi_{\sigma,h_\sigma} \Big(\prod_{k=1}^K \pi_{k,h_{1k}} \Big) \delta_{\Theta_{h_{1k}},\Sigma_{h_\sigma}} : \sum_{h_{1,1k}>H_n} \pi_{h_{1,1k}} < \epsilon_1, 1\le k\le K,$$
$$\sum_{h_\sigma > H_n} \pi_{\sigma,h_\sigma} < \epsilon_2, \underline{\sigma}_n^2 \le \lambda_D(\Sigma_{h_\sigma}), \lambda_1(\Sigma_{h_\sigma}) \le \underline{\sigma}_n^2(1+\epsilon/\sqrt{D})^{M_n},\ 1 < \frac{\lambda_1(\Sigma_{h_\sigma})}{\lambda_D(\Sigma_{h_\sigma})} \le n^{H_n},\ h_\sigma \le H_n \Bigg\},$$
$$\mathcal{F}_{n,\mathbf{j},\mathbf{l}} = \Bigg\{ f_p \in \mathcal{F}_n :\ \text{for } h_{11},\ldots,h_{1K} \le H_n, n^{H_n^2}(j_{h_{1k}}-1) \le ||vec(A_{kh_{1k}})|| \le n^{H_n^2} j_{h_{1k}}, \quad (34)$$
$$\text{and for } h_\sigma \le H_n, n^{l_{h_\sigma}-1} \le \frac{\lambda_1(\Sigma_{h_\sigma})}{\lambda_D(\Sigma_{h_\sigma})} \le n^{l_{h_\sigma}} \Bigg\}$$

Now note that the prior probability on the sieve $\mathcal{F}_n$ satisfies

$$\Pi(\mathcal{F}_n^c) \le Pr\Big(\sum_{h_\sigma > H_n} \pi_{\sigma,h_\sigma} > \epsilon_1 \Big) + \sum_{k=1}^K Pr\Big(\sum_{h_{1k}>H_n} \pi_{k,h_{1k}} > \epsilon_1 \Big) + H_n P_2^*\Big(\lambda_D(\Sigma) \le \underline{\sigma}_n^2 \Big) +$$
$$H_n P_2^*\Big(\lambda_1(\Sigma) > \underline{\sigma}_n^2(1+\epsilon_1/\sqrt{D})^{M_n} \Big) + H_n P_2^*\left(\frac{\lambda_1(\Sigma_{h_\sigma})}{\lambda_D(\Sigma_{h_\sigma})} > n^{H_n} \right) \lesssim e^{-b^* n},$$

using similar techniques used to derive (37) in Lemma 6 of the Appendix. Hence, condition $(2A)$ in Theorem 2 holds. Further, one can rewrite (33) as

$$\Pi(\mathcal{F}_{n,\mathbf{jl}}) \lesssim \left\{ n^{-2H_n^3(r+1)K} \prod_{k=1}^{K} \prod_{h_{1k} \leq H_n} \left(j_{h_{1k}} - 1\right)^{-1_{(j_{h_{1k}} \geq 2)} 2(r+1)} \right\} \times \left\{ \prod_{h_\sigma \leq H_n} (n^{\kappa(l_{h_\sigma}-1)})^{-1_{(l_{h_\sigma} \geq 1)}} \right\},$$

for large n. Moreover, using similar steps as in the proof for Lemma 7 corresponding to the PDPM-VAR model, it is possible to show that the entropy bound in Corollary 1 satisfies $\sqrt{N(\epsilon_1, \mathcal{F}_{n,\mathbf{j}_{h_1}\mathbf{l}_{h_\sigma}}, ||\cdot||_1)\Pi(\mathcal{F}_{n,\mathbf{j}_{h_1}\mathbf{l}_{h_\sigma}})}$

$$\begin{aligned}
&\lesssim \sqrt{\mathcal{C}_1^{H_n}} \exp\left\{ H_n \log(\frac{C_1^*}{\epsilon_1}) + \frac{H_n}{2}\big(KD^2 + D(D-1)\big)\log(\frac{1}{\epsilon_1}) \right\} \\
\times \quad & \exp\left\{ \frac{Cn\epsilon^2}{2}\big(D + \frac{1}{2c_2}KD^2(T-1+TD)\big) \right\} \times n^{KH_n^3\left(D^2(T-2)+\frac{1}{2}D^2-r\right)-KH_n^3} \\
\times \quad & \left\{ \prod_{k=1}^{K} \prod_{h_{1k} \leq H_n} (j_{h_{1k}})^{D^2(T-2)+\frac{1}{2}(D^2-1)} \right\} \left\{ \prod_{k=1}^{K} \prod_{h_{1k} \leq H_n} \left(j_{h_{1k}} - 1\right)^{-1_{(j_{h_{1k}} \geq 2)}(r+1)} \right\} \\
\times \quad & \prod_{h_\sigma \leq H_n} \left\{ n^{l_{h_\sigma}} \right\}^{\frac{D(D-1)}{4} + KD^2H_n(T-1)\frac{D-1}{4}} \times \left\{ \prod_{h_\sigma \leq H_n} (n^{\kappa(l_{h_\sigma}-1)/2})^{-1_{(l_{h_\sigma} \geq 1)}} \right\}.
\end{aligned}$$

Using the above expressions and similar arguments as in (38), it is straightforward to show that condition $(2B)$ in Theorem 2 holds. Hence the result is proved.

The proof for the rgPDPM-VAR proceeds in a similar fashion by noting that the prior probability for the complement of the sieves defined in the manuscript can be written as

$$\begin{aligned}
&\Pi(\mathcal{F}_n^c) \leq H_n P_2^*\left(\lambda_1(\Sigma) > \underline{\sigma}_n^2(1+\epsilon/\sqrt{D})^{M_n}\right) + H_n P_2^*\left(\frac{\lambda_1(\Sigma_{h_\sigma})}{\lambda_D(\Sigma_{h_\sigma})} > n^{H_n}\right) + H_n P_2^*\left(\lambda_D(\Sigma) \leq \underline{\sigma}_n^2\right) \\
&+Pr\left(\sum_{h_\sigma > H_n} \pi_{\sigma,h_\sigma} > \epsilon_2\right) + Pr\left(\sum_{h_{11} > H_n} \pi^*_{1,h_{11}} > \epsilon_2\right) + \cdots + Pr\left(\sum_{h_{1D} > H_n} \pi^*_{D,h_{1D}} > \epsilon_2\right) \lesssim e^{-b^* n},
\end{aligned}$$

using similar techniques used to derive (37) in Lemma 6 in the Appendix. Hence, condition $(2A)$ in Theorem 2 holds. Also define $\mathcal{F}_{n,\mathbf{j},\mathbf{l}}$ such that $\mathcal{F}_n \subset \cup_{\mathbf{j},\mathbf{l}} \mathcal{F}_{n,\mathbf{j},\mathbf{l}}$ and

$$\begin{aligned}
\mathcal{F}_{n,\mathbf{j},\mathbf{l}} = \Bigg\{ & f_p \in \mathcal{F}_n : \text{for } h_{11}, \ldots, h_{1D}, n^{H_n^2}(j_{h_{1d'}} - 1) \leq ||vec(A_{kh_{1k}})|| \leq n^{H_n^2} j_{h_{1d'}}, \\
& d' = 1, \ldots, D, \text{ and for } h_\sigma \leq H_n, \quad n^{l_{h_\sigma}-1} \leq \frac{\lambda_1(\Sigma_{h_\sigma})}{\lambda_D(\Sigma_{h_\sigma})} \leq n^{l_{h_\sigma}} \Bigg\} \quad . \qquad (35)
\end{aligned}$$

Given the entropy bound in (29) and using similar calculations as in Lemma 7 for the PDPM-VAR case, it is possible to show that $(2B)$ in Theorem 2 holds. The strong consistency result follows under rgPDPM-VAR once conditions $(2A)$-$(2B)$ in Theorem 2 are satisfied.

Proof of Lemma 2: For the case with independent double exponential priors involving shrinkage parameter λ, note that $P\left(a^2(k,l) \leq \frac{(x^*)^2}{D^2}, \text{for all } 1 \leq k, l \leq D\right) \leq P(||vec(A_k)|| \leq$

x^*). Further for large positive $x^* > D$, and denoting $\pi(s \mid \lambda) = \frac{1}{2}\sqrt{\lambda}\exp(-\sqrt{\lambda}\frac{s}{2})$

$$P_1^*\left(a^2(k,l) > \frac{(x^*)^2}{D^2}\right) = P\left(|a(k,l)| > \frac{(x^*)}{D}\right) = 2\int_{x^*/D}^{\infty}\int \frac{1}{\sqrt{2\pi s}}\exp(-\frac{1}{2s}a^2)\pi(s \mid \lambda)\ ds\ da$$
$$\leq 2\int \frac{1}{\sqrt{2\pi}s(x^*/D)}\exp(-\frac{1}{2s}(x^*/D)^2)\pi(s \mid \lambda)ds \leq 2\int \frac{1}{\sqrt{2\pi}s}\exp(-\frac{1}{2s}(x^*/d)^2)\pi(s \mid \lambda)ds$$
$$= 2\exp(-\lambda|x^*/D|) = 2\exp(-\lambda x^*/D) \leq (x^*/D)^{-\lambda}.$$

The above implies that $1 - (x^*/D)^{-\lambda} \leq P_1^*\left(a^2(k,l) \leq \frac{(x^*)^2}{D^2}\right)$ and further that $(1 - (x^*/D)^{-\lambda})^{D^2} \leq P_1^*\left(a^2(k,l) \leq \frac{(x^*)^2}{D^2}, \text{for all } 1 \leq k,l \leq D\right) \leq P_1^*(||vec(A_k)|| \leq x^*)$. This implies that $P_1^*(||vec(A_k)|| > x^*) > 1 - (1 - (x^*/D)^{-\lambda})^{D^2}$. For large x^* and choosing λ large enough, one can use the binomial expansion to write (when D is even)

$$1 - (1 - (x^*/D)^{-\lambda})^{D^2} = 1 - \left\{(1 - D^2(x^*/D)^{-\lambda}) + \frac{D^2(D^2-1)}{2}(x^*/D)^{-2\lambda}\left(1 - \frac{D^2-2}{3}(x^*/D)^{-\lambda}\right) + \right.$$
$$\frac{D^2(D^2-1)(D^2-2)(D^2-3)}{4!}(x^*/D)^{-4\lambda}\left(1 - \frac{D^2-4}{5}(x^*/D)^{-\lambda}\right) + \ldots +$$
$$\left.\frac{D^2(D^2-1)\ldots\{D^2-(D^2-1)\}}{(D^2)!}(x^*/D)^{-(D^2)\lambda}\right\} = D^2(x^*/D)^{-\lambda} - \kappa^* \lesssim (x^*)^{-\lambda},$$

for large x^* and λ such that $1 - D^2(x^*/D)^{-\lambda} > 0$ and for some positive constant κ^*. Similar calculations hold for odd D.

Further, when $P_1^*(vec(A_k)) = N_{D^2}(vec(A_k); \boldsymbol{\mu}, \Lambda)$, $\Lambda \sim IW(\Lambda_0, \nu_\lambda)$, the resulting distribution follows a multivariate t-distribution. Hence one can write $P_1^*(||vec(A_k)|| > x^*) \leq (x^*)^{(\nu_\lambda - D^2+1)/2}$, using arguments similar to those in Canale and De Blasi (2017).

Appendix B. Additional Lemmas

Lemma 4: *The distance between densities f_{P_1} and f_{P_2} under the PDPM-VAR can be expressed as*

$$||f_{P_1} - f_{P_2}||_1 \leq 2\epsilon^2 + \sum_{h_1, h_\sigma < H_n} |\pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} - \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}| + 4\epsilon$$
$$+ \sum_{h_1, h_\sigma \leq H_n} \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma}\left\|\prod_{t=1}^{T}\phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K}A^{(1)}_{k,h_1}\mathbf{x}_{t-k}\big) - \prod_{t=1}^{T}\phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K}A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big)\right\|_1.$$

Proof: $||f_{P_1} - f_{P_2}||_1 =$

$$
\begin{aligned}
& || \sum_{h_1,h_\sigma \geq 1} \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1}\mathbf{x}_{t-k}\big) - \sum_{h_1,h_\sigma \geq 1} \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma} \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big)||_1 \\
= & \Big\| \sum_{h_1,h_\sigma > H_n} \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1}\mathbf{x}_{t-k}\big) - \sum_{h_1,h_\sigma > H_n} \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma} \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big) \\
+ & \sum_{h_1,h_\sigma \leq H_n} \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} \big\{ \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1}\mathbf{x}_{t-k}\big) - \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big) \big\} \\
+ & \sum_{h_1,h_\sigma < H_n} \big(\pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} - \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}\big) \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big) - \sum_{j'=1,2} \Big\{ \sum_{h_1 \leq H_n} \sum_{h_\sigma > H_n} \pi^{(j')}_{h_1}\pi^{(j')}_{\sigma,h_\sigma} \\
& \prod_{t=1}^{T} \phi_{\Sigma^{(j')}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(j')}_{k,h_1}\mathbf{x}_{t-k}\big) + \sum_{h_1 > H_n} \sum_{h_\sigma \leq H_n} \pi^{(j')}_{h_1}\pi^{(j')}_{\sigma,h_\sigma} \prod_{t=1}^{T} \phi_{\Sigma^{(j')}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(j')}_{k,h_1}\mathbf{x}_{t-k}\big) \Big\} \Big\|_1 .
\end{aligned}
$$

The upper bound for the right hand side of the above equation may be further written as $\sum_{j'=1,2} \Big\{ \sum_{h_1,h_\sigma > H_n} \pi^{(j')}_{h_1}\pi^{(j')}_{\sigma,h_\sigma} \Big\| \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(j')}_{k,h_1}\mathbf{x}_{t-k}\big) \Big\|_1 \Big\} + \sum_{h_1,h_\sigma \leq H_n} \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} \times \Big\| \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1}\mathbf{x}_{t-k}\big) - \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big) \Big\|_1 + \sum_{h_1,h_\sigma < H_n} \Big(\pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} - \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma} \Big) \Big\| \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big) \Big\|_1 + \sum_{j'=1,2} \Big\{ \sum_{h_1 \leq H_n} \sum_{h_\sigma > H_n} \pi^{(j')}_{h_1}\pi^{(j')}_{\sigma,h_\sigma} \times \Big\| \prod_{t=1}^{T} \phi_{\Sigma^{(j')}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(j')}_{k,h_1}\mathbf{x}_{t-k}\big) \Big\|_1 + \sum_{h_1 > H_n} \sum_{h_\sigma \leq H_n} \pi^{(j')}_{h_1}\pi^{(j')}_{\sigma,h_\sigma} \Big\| \prod_{t=1}^{T} \phi_{\Sigma^{(j')}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(j')}_{k,h_1}\mathbf{x}_{t-k}\big) \Big\|_1 \Big\}.$

The right hand side of the above can be further bounded as $\sum_{h_1,h_\sigma \leq H_n} \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} \Big\| \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1}\mathbf{x}_{t-k}\big) - \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big) \Big\|_1 + \Big| \sum_{h_1,h_\sigma > H_n} \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} + \sum_{h_1,h_\sigma > H_n} \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma} \Big| + \sum_{h_1,h_\sigma < H_n} \Big| \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} - \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma} \Big| + \sum_{j'=1,2} \Big\{ \sum_{h_1 \leq H_n} \sum_{h_\sigma > H_n} \pi^{(j')}_{h_1}\pi^{(j')}_{\sigma,h_\sigma} + \sum_{h_1 > H_n} \sum_{h_\sigma \leq H_n} \pi^{(j')}_{h_1}\pi^{(j')}_{\sigma,h_\sigma} \Big\}$
$\leq \sum_{h_1,h_\sigma \leq H_n} \pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} \Big\| \prod_{t=1}^{T} \phi_{\Sigma^{(1)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(1)}_{k,h_1}\mathbf{x}_{t-k}\big) - \prod_{t=1}^{T} \phi_{\Sigma^{(2)}_{h_\sigma}}\big(\mathbf{x}_t - \sum_{k=1}^{K} A^{(2)}_{k,h_1}\mathbf{x}_{t-k}\big) \Big\|_1$
$+2\epsilon^2 + \sum_{h_1,h_\sigma < H_n} |\pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} - \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}| + 4\epsilon$, using the fact that $\sum_{h_1 \leq H_n} \sum_{h_\sigma > H_n} \pi^{(j')}_{h_1}\pi^{(j')}_{\sigma,h_\sigma} = (\sum_{h_1 \leq H_n} \pi^{(j')}_{h_1})(\sum_{h_\sigma > H_n} \pi^{(j')}_{\sigma,h_\sigma}) \leq \epsilon$, since $\sum_{h_1 > H_n} \pi_{h_1} < \epsilon, \sum_{h_\sigma > H_n} \pi_{\sigma,h_\sigma} < \epsilon$.

Lemma 5: *The upper bound for* $\sum_{h_1,h_\sigma < H_n} |\pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma} - \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}|$ *is given by* $\sum_{h_1,h_\sigma < H_n} \Big| \tilde{\pi}^{(1)}_{h_1}\tilde{\pi}^{(1)}_{\sigma,h_\sigma} - \pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma} \Big| + \Big| 1 - (1-\epsilon)^2 \Big|$, *where* $\tilde{\pi} = \frac{\pi_h}{(1 - \sum_{h > H} \pi_h)}$.

Proof: Note that $\sum_{h_1,h_\sigma<H_n}|\pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma}-\pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}|\leq$

$$\begin{aligned}
&\sum_{h_1,h_\sigma<H_n}\left|\pi^{(1)}_{h_1}\pi^{(1)}_{\sigma,h_\sigma}-\Big(1-\sum_{h_1>H}\pi^{(1)}_{h_1}\Big)\Big(1-\sum_{h_\sigma>H}\pi^{(2)}_{\sigma,h_\sigma}\Big)\pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}\right| \\
+\;&\sum_{h_1,h_\sigma<H_n}\left|\Big(1-\sum_{h_1>H}\pi^{(1)}_{h_1}\Big)\Big(1-\sum_{h_\sigma>H}\pi^{(2)}_{\sigma,h_\sigma}\Big)\pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}-\pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}\right| \\
=\;&\Big(1-\sum_{h_1>H}\pi^{(1)}_{h_1}\Big)\Big(1-\sum_{h_\sigma>H}\pi^{(2)}_{\sigma,h_\sigma}\Big)\sum_{h_1,h_\sigma<H_n}\left|\tilde\pi^{(1)}_{h_1}\tilde\pi^{(1)}_{h_\sigma}-\pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}\right| \\
+\;&\left|\Big(1-\sum_{h_1>H}\pi^{(1)}_{h_1}\Big)\Big(1-\sum_{h_\sigma>H}\pi^{(2)}_{\sigma,h_\sigma}\Big)-1\right|\Big(\sum_{h_1,h_\sigma<H_n}\pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}\Big) \\
\leq\;&\sum_{h_1,h_\sigma<H_n}\left|\tilde\pi^{(1)}_{h_1}\tilde\pi^{(1)}_{\sigma,h_\sigma}-\pi^{(2)}_{h_1}\pi^{(2)}_{\sigma,h_\sigma}\right|+\Big|1-(1-\epsilon)^2\Big|,\ \text{where } \tilde\pi=\frac{\pi_h}{(1-\sum_{h>H}\pi_h)}. \quad (36)
\end{aligned}$$

Lemma 6: *For the PDPM-VAR the prior tail condition (2A) in Theorem 2 is satisfied.*

Proof: For the PDPM-VAR, the prior probability on the sieve $\mathcal{F}_n$ defined in (30) satisfies

$$\begin{aligned}
\Pi(\mathcal{F}_n^c)\;\leq\;&Pr\Big(\sum_{h_\sigma>H_n}\pi_{\sigma,h_\sigma}>\epsilon\Big)+Pr\Big(\sum_{h_1>H_n}\pi_{h_1}>\epsilon\Big)+H_nP_2^*\Big(\lambda_D(\Sigma)\leq\underline{\sigma}_n^2\Big)+\\
&H_nP_2^*\Big(\lambda_1(\Sigma)>\underline{\sigma}_n^2(1+\epsilon/\sqrt{D})^{M_n}\Big)+H_nP_2^*\Big(\frac{\lambda_1(\Sigma_{h_\sigma})}{\lambda_D(\Sigma_{h_\sigma})}>n^{H_n}\Big),
\end{aligned}$$

using the fact that $P[(A\cap B\cap C)^c]=P[A^c\cup B^c\cup C^c]\leq P(A^c)+P(B^c)+P(C^c)$. Using the stick-breaking representation for DP for the first term and the prior tail conditions, and following similar steps as in the proof of Proposition 2 in Shen et al. (2013), one has $\Pi(\mathcal{F}_n^c)$

$$\begin{aligned}
\lesssim\;&\sum_{m=1,2}\big\{\frac{e\alpha_m}{H_n}\log(1/\epsilon)\big\}^{H_n}+H_n\big\{e^{-c_1\underline{\sigma}_n^{-2c_2}}+\underline{\sigma}_n^{-2c_3}\big(1+\epsilon/\sqrt{D}\big)^{-c_3M_n}+(n^{\frac{1}{\log n}})^{-\kappa Cn\epsilon^2}\big\}\\
\lesssim\;&2\big(Cn\epsilon^2/\log(n)\big)^{-Cn\epsilon^2/\log(n)}+\big(Cn\epsilon^2/\log(n)\big)\big(e^{-c_1n}+n^{c_3/c_2}(1+\epsilon/\sqrt{D})^{-c_3n}+e^{-\kappa Cn\epsilon^2}\big)\lesssim e^{-bn}
\end{aligned}$$

since $n^{\frac{1}{\log n}}=e$ and due to the fact that $\big(Cn\epsilon^2/\log(n)\big)\log\{-Cn\epsilon^2/\log(n)\}>Cn\epsilon^2$ for large n, where $0<b<\min\{C\epsilon^2/2,c_1,c_3\log(1+\epsilon/\sqrt{D}),\kappa C\epsilon^2\}$. Hence the first condition $(2A)$ in Theorem 2 is satisfied.

Lemma 7: For the PDPM-VAR, we have $\sum_{j_{h_1}\in N}\sum_{1\leq l_{h_\sigma}\leq H_n}\sqrt{N(\epsilon,\mathcal{F}_{n,j_{h_1}l_{h_\sigma}})\Pi(\mathcal{F}_{n,j_{h_1}l_{h_\sigma}})}e^{-(4-c)n\epsilon^2}\to 0$ as $n\to\infty$ for a suitable choice of constants.

Proof: Hence we can write $\sqrt{N(\epsilon, \mathcal{F}_{n,\mathbf{jl}}, ||\cdot||_1)\Pi(\mathcal{F}_{n,\mathbf{jl}})}$

$$
\begin{aligned}
&\lesssim \sqrt{\mathcal{C}_1^{H_n}\big(\frac{1}{\epsilon}\big)^{C_1 H_n} \times \bigg\{\mathcal{C}\bigg(1+o_n(1)\bigg)^{KD^2}\bigg\}^{H_n} \times \big(\frac{1}{\epsilon}\big)^{H_n KD^2 + D(D-1)}} \\
&\times \exp\bigg\{\frac{1}{2}\big(DH_n\log(M) + H_n\log\frac{C_1}{\epsilon}\big)\bigg\} \times n^{H_n^3(KD^2(T-2)+\frac{1}{2}KD^2)+\frac{H_n}{c_2}\left(KD^2(T-1)+TKD^3\right)} \\
&\times \bigg\{n^{-H_n^3(r+1)K}\prod_{h_1\le H_n}(j_{h_1}-1)^{-1_{(j_{h_1}\ge 2)}K(r+1)}\bigg\} \times \bigg\{\prod_{h_1\le H_n}(j_{h1})^{2KD^2(T-2)+K(D^2-1)}\bigg\} \\
&\times \bigg\{\prod_{h_\sigma\le H_n}(n^{l_{h_\sigma}})^{\frac{D(D-1)}{2}+KD^2H_n(T-1)\frac{D-1}{2}}\bigg\} \times \bigg\{\prod_{h_\sigma\le H_n}(n^{\frac{\kappa}{2}(l_{h_\sigma}-1)})^{-1_{(l_{h_\sigma}\ge 1)}}\bigg\}.
\end{aligned}
$$

The above can be simplified further as

$$
\begin{aligned}
&\sqrt{\mathcal{C}_1^{H_n}}\exp\bigg\{H_n\log(\frac{C_1}{\epsilon}) + \frac{H_n}{2}\big(KD^2+D(D-1)\big)\log(\frac{1}{\epsilon})\bigg\} \\
&\times \exp\bigg\{\frac{1}{2}\big(DH_n\log(M) + \frac{H_n}{c_2}\big(KD^2(T-1)+TKD^3\big)\log(n)\big)\bigg\} \times n^{KH_n^3\left(D^2(T-2)+\frac{1}{2}D^2-r\right)-KH_n^3} \\
&\times \bigg\{\prod_{h_1\le H_n}(j_{h1})^{KD^2(T-2)+\frac{K}{2}(D^2-1)}\bigg\}\bigg\{\prod_{h_1\le H_n}(j_{h_1}-1)^{-1_{(j_{h_1}\ge 2)}K(r+1)}\bigg\} \\
&\times \prod_{h_\sigma\le H_n}\bigg\{n^{l_{h_\sigma}}\bigg\}^{\frac{D(D-1)}{4}+KD^2H_n(T-1)\frac{D-1}{4}} \times \bigg\{\prod_{h_\sigma\le H_n}(n^{\kappa(l_{h_\sigma}-1)/2})^{-1_{(l_{h_\sigma}\ge 1)}}\bigg\}, \qquad (37)
\end{aligned}
$$

Note that $n^{KH_n^3\left(D^2(T-2)+\frac{1}{2}D^2-r\right)}$ is bounded when $r > D^2(T-2)+\frac{1}{2}D^2$. Further, looking at the terms involving j in the last line of (37), one can sum over j (for a fixed h_1) to have

$$
\sum_{j_{h_1}\ge 2}\big\{(j_{h1})^{KD^2(T-2)+\frac{K}{2}(D^2-1)}(j_{h_1}-1)^{-1_{(j_{h_1}\ge 2)}K(r+1)}\big\} \approx \sum_{j_{h_1}\ge 2}(j_{h1})^{KD^2(T-2)+\frac{K}{2}(D^2-1)-K(r+1)} = \bigg(1+\mathcal{B}\bigg),
$$

where $\mathcal{B}$ is a suitable finite constant that does not depend on n when r is large enough such that $r > D^2(T-2)+\frac{1}{2}D^2$, and where the approximation holds for n large enough.

Similarly, looking at the terms involving l, one can sum over l (for a fixed h_σ) to have $\sum_{l_{h_\sigma}\ge 1}\bigg\{(n^{l_{h_\sigma}})^{\frac{D(D-1)}{4}-(\kappa/2)}\bigg\} \times \bigg\{(n^{l_{h_\sigma}})^{KD^2H_n(T-1)\frac{D-1}{4}}\bigg\}$
$\le \sqrt{\sum_{l_{h_\sigma}\ge 1}\bigg\{(n^{l_{h_\sigma}})^{\frac{D(D-1)}{2}-(\kappa/2)}\bigg\}} \times \sqrt{\sum_{l_{h_\sigma}\ge 1}\bigg\{(n^{l_{h_\sigma}})^{KD^2H_n(T-1)\frac{D-1}{2}}\bigg\}}$ where the inequality is using Cauchy-Schwartz, and the first term in the upper bound (denoted by $\mathcal{B}_1$) is finite when $\kappa > D(D-1)/2$.

Hence the upper bound on the combined terms in the last two lines in (37) is given by
$\big(1+\mathcal{B}\big)^{H_n}\big(1+\mathcal{B}_1\big)^{H_n/2}n^{KH_n^3\left(D^2(T-2)+\frac{1}{2}D^2-r\right)-KH_n^3} \times \prod_{h_\sigma\le H_n}\sqrt{\sum_{l_{2\le h_\sigma\le H_n}}(n^{l_{h_\sigma}})^{KD^2H_n(T-1)\frac{D-1}{2}}}$
$= \big(1+\mathcal{B}\big)^{H_n}\big(1+\mathcal{B}_1\big)^{H_n/2}n^{KH_n^3\left(D^2(T-2)+\frac{1}{2}D^2-r\right)} \times \prod_{h_\sigma\le H_n}\sqrt{\sum_{l_{2\le h_\sigma\le H_n}}\bigg\{\frac{(n^{l_{h_\sigma}})^{KD^2H_n(T-1)\frac{D-1}{2}}}{n^{2KH_n^2}}\bigg\}}.$

The sum within the square root can be simplified as $\sum_{l_{2\leq h_\sigma\leq H_n}}\left\{\frac{(n^{l_{h_\sigma}})^{KD^2H_n(T-1)\frac{D-1}{2}}}{n^{2KH_n^2}}\right\}=\sum_{h_\sigma\leq H_n}\left(n^{KD^2H_n(T-1)\frac{D-1}{2}-2KH_n^2}\right)^{l_{h_\sigma}}=\left(\frac{1-(r^*)^{H_n}}{1-r^*}\right)<1$, where the equality is obtained by summing H_n terms in geometric progression, and $r^*=n^{KD^2H_n(T-1)\frac{D-1}{2}-2KH_n^2}<1$ since $KD^2H_n(T-1)\frac{D-1}{2}-2KH_n^2<0$ when n is large enough, and recalling that $1\leq l_{h_\sigma}=h_\sigma\leq H_n$. Combining the above expressions, $\sum_{j_{h_1}\in N}\sum_{1\leq l_{h_\sigma\leq H_n}}\sqrt{N(\epsilon,\mathcal{F}_{n,j_{h_1}l_{h_\sigma}})\Pi(\mathcal{F}_{n,j_{h_1}l_{h_\sigma}})}$

$$\lesssim\left(1+\max\{\mathcal{B},\mathcal{B}_1\}\right)^{H_n}n^{KH_n^3\left(D^2(T-2)+\frac{1}{2}D^2-r\right)}\left(\frac{1-(r^*)^{H_n}}{1-r^*}\right)^{H_n}$$
$$\times\exp\left\{H_n\log(\frac{C_1}{\epsilon})+\frac{H_n}{2}(KD^2+D(D-1))\log(\frac{1}{\epsilon})\right\}$$
$$\times\exp\left\{\frac{Cn\epsilon^2}{2}\big(D+\frac{1}{2c_2}KD^2(T-1+TD)\big)\right\}\lesssim\left(\frac{\mathcal{K}^*}{n^{KH_n^2\left(r-D^2(T-2)+\frac{1}{2}D^2\right)}}\right)^{H_n}$$
$$\times\exp\left\{\frac{Cn\epsilon^2}{2}\big\{\big(D+\frac{1}{2c_2}KD^2(T-1)\big)+\log(\frac{C_1}{\epsilon})+\frac{1}{2}(KD^2+D(D-1))\log(\frac{1}{\epsilon}))\big\}\right\}\quad(38)$$

where $r^*<1$, $r-D^2(T-2)+\frac{1}{2}D^2>0$, $H_n\log(M_n)=Cn\epsilon^2$, and $\mathcal{K}^*>0$ is some finite constant that is a function of K,D,ϵ, and other constants but does not depend on n.

Therefore, $\sum_{j_{h_1}\in N}\sum_{1\leq l_{h_\sigma\leq H_n}}\sqrt{N(\epsilon,\mathcal{F}_{n,j_{h_1}l_{h_\sigma}})\Pi(\mathcal{F}_{n,j_{h_1}l_{h_\sigma}})}e^{-(4-c)n\epsilon^2}\to 0$ as $n\to\infty$ for a suitable choice of C such that $\frac{1}{2}\big(D+\frac{1}{2c_2}KD^2(T-1)\big)C<1$. Hence the condition $(2B)$ in Theorem 2 is satisfied and the strong posterior consistency result is proved corresponding to the PDPM-VAR model.

Appendix C. Posterior Computation Steps

C.1 Residual Covariance Updates:

Under the low rank representation, we impose DP mixture priors on $(\mathbf{\Gamma}_i^*,\ \mathbf{\Xi}_i,\ \mathbf{\Psi}_i)$ leading to a mixture prior on Σ_i. This corresponds to the prior $\Sigma_i\sim\sum_{h_\sigma=1}^{\infty}\pi_{\sigma,h_\sigma}\delta_{(\mathbf{\Gamma}^*_{h_\sigma},\mathbf{\Xi}_{h_\sigma},\mathbf{\Psi}_{h_\sigma})}$, where $(\mathbf{\Gamma}^*_{h_\sigma},\mathbf{\Xi}_{h_\sigma},\mathbf{\Psi}_{h_\sigma})\sim P_2^*\equiv P_{\mathbf{\Gamma}^*}\times P_{\mathbf{\Xi}}\times P_{\mathbf{\Psi}}$. Here $P_{\mathbf{\Gamma}^*}$ is a product of independent standard normal distributions, $P_{\mathbf{\Xi}}$ is a product of independent $Gamma(1/2,1/2)$ distributions yielding a half-Cauchy prior on the diagonal elements of $\mathbf{\Gamma}$ and a Cauchy prior on the lower-off-diagonal elements of $\mathbf{\Gamma}$ as in Ghosh and Dunson (2009), and the inverse of the diagonal elements of $\mathbf{\Psi}$ have independent $Gamma(\alpha_\sigma,\beta_\sigma)$ priors. Note that here $\mathbf{\Gamma}_i^*$ is not a square matrix, and by diagonal elements we refer to elements $\Gamma_{i,1,1},\ldots,\Gamma_{i,B,B}$.

Under the stick-breaking representation of Sethuraman (1994), we can write $\pi_{\sigma,h_\sigma}=\nu_{\sigma,h_\sigma}\prod_{l_\sigma=1}^{h_\sigma-1}(1-\nu_{\sigma,l_\sigma})$, $\nu_{l_\sigma}\sim Beta(1,\alpha_2)$. We use the slice sampling approach of Walker (2007) to facilitate sampling. This approach introduces a cluster membership indicator, V, with $V_i=h_\sigma$ when subject i belongs to cluster h_σ, and let $\mathcal{V}_{h_\sigma}=\{i:V_i=h_\sigma\}$ be the indices of all subjects belonging to covariance cluster h_σ and let n_{σ,h_σ} be the cardinality of this set. Let g_i be a uniformly distributed latent variable used to reduce the stick-breaking representation of the DPM to a finite sum. Our sampler updates ν_{σ,h_σ}, and g_i as: $\nu_{\sigma,h_\sigma}|\{V_1,\ldots,V_N\}\sim Beta\left(1+n_{\sigma,h_\sigma},\alpha_2+\sum_{i=1}^n I_{(V_i>\nu_{\sigma,h_\sigma})}\right)$, $g_i|V_i\sim U(0,\pi_{\sigma,V_i})$.

The cluster membership indicators V_i are then sampled from a multinomial distribution with posterior probabilities $\big(p(V_i = 1|-), \ldots, p(V_i = h^*_\sigma|-)\big)$ expressed as, $p(V_i = h_\sigma|-) \sim \frac{I_{(g_i < \pi_{\sigma,h_\sigma})} \prod_{t=1}^{T_i} \phi_{\Sigma_{h_\sigma}}(x_{i,t}; A_{i1}, \ldots, A_{iK})}{\sum_{h'_\sigma=1}^{h^*_\sigma} \left\{ I_{(g_i < \pi_{\sigma,h'_\sigma})} \prod_{t=1}^{T_i} \phi_{\Sigma_{h'_\sigma}}(x_{i,t}; A_{i1}, \ldots, A_{iK}) \right\}}$, where $h^*_\sigma = \min\{h_\sigma : g_i > 1 - \sum_{h'_\sigma=1}^{h_\sigma} \pi_{\sigma,h_\sigma}, \text{for all } i\}$. Conditioned on the cluster memberships, it is straightforward to update the variables in the low rank representation of Σ using similar steps as in Ghosh and Dunson (2009). We start by sampling the elements of $\mathbf{\Gamma}_{h^*_\sigma}$ one row at a time from their full conditionals: $\mathbf{\Gamma}^*_{h_\sigma,d'}|- \sim N\left(\mu_{\mathbf{\Gamma}^*_{h_\sigma,d'}}, \Sigma_{\mathbf{\Gamma}^*_{h_\sigma,d'}}\right)$, where $\Sigma_{\mathbf{\Gamma}^*_{h_\sigma,d'}} = \left(\sigma^{-2}_{h_\sigma,d'} \sum_{i \in \mathcal{V}_{h_\sigma}} \sum_{t=1}^{T_i} (\mathcal{E}^*_{\eta,itd'})'(\mathcal{E}^*_{\eta,itd'}) + I_{\min\{d',B\}}\right)^{-1}$, $\mu_{\mathbf{\Gamma}^*_{h_\sigma,d'}} = \Sigma_{\mathbf{\Gamma}^*_{h_\sigma,d'}} \left(\sigma^{-2}_{h_\sigma,d'} \sum_{i \in \mathcal{V}_{h_\sigma}} \sum_{t=1}^{T_i} \mathcal{E}^*_{\eta,itd'} \left[x^{(d')}_{i,t} - A_{i,d',\bullet}\mathbf{z}_{i,t}\right]\right)$, $\mathcal{E}^*_{\eta,itd'} = \left(\ \eta^*_{i,t,1}, \ldots,\ \eta^*_{i,t,\min\{d',B\}}\right)'$, $x^{(d)}_{i,t}$ is the response for the ith subject at the d'th node and tth time point, and $A'_{ik,d'\bullet}$ is the transpose of the d'th row of A_{ik}, $A_{i,d'\bullet}$ is the $DK \times 1$ vector formed by stacking the $A'_{ik,d'\bullet}$ across all lags, and $\mathbf{z}_{i,t} = [\mathbf{x}'_{i,t-1}, \ldots, \mathbf{x}'_{i,t-K}]'$ is the $DK \times 1$ vector of previous outcomes used to predict the outcome at time t, padded with zeros for the case that $t - k < 1$.

The conditionals for the remaining terms in the low rank representation for Σ_i are: $\eta^*_{i,t}|- \sim N\left(\mu\ \eta^*_{i,t}, \Sigma\ \eta^*_{i,t}\right)$, $\xi_{h_\sigma,b}|- \sim Gamma\left(\frac{1+N_{h_\sigma}}{2}, \frac{1}{2}\left[1 + \sum_{i \in \mathcal{V}_{h_\sigma}} \sum_{t=1}^{T_i}\ \eta^{*2}_{i,t,b}\right]\right)$, $\sigma^{-2}_{h_\sigma,d'}|- \sim Gamma\left(a_\sigma + \frac{N_{h_\sigma}}{2}, b_\sigma + \frac{1}{2}\sum_{i \in \mathcal{V}_{h_\sigma}} \sum_{t=1}^{T_i} \left[x^{(d')}_{i,t} - A_{i,d',\bullet}\mathbf{z}_{i,t} - \mathbf{\Gamma}^*_{i,d'}\boldsymbol{\eta}^*_{i,t}\right]^2\right)$ where N_{h_σ} is equal to the total number of time points across all subjects in covariance cluster h_σ, $\Sigma\ \eta^*_{i,t} = \left(\mathbf{\Xi}^{-1}_{V_i} + \mathbf{\Gamma}'_{V_i}\mathbf{\Psi}_{V_i}\mathbf{\Gamma}_{V_i}\right)^{-1}$, and $\mu\ \eta^*_{i,t} = \Sigma\ \eta^*_{i,t}\mathbf{\Gamma}^{*\prime}_{V_i}\mathbf{\Psi}_{V_i}\left[x^{(d')}_{i,t} - A_{i,d',\bullet}\mathbf{z}_{i,t}\right]$.

C.2 Autocovariance Parameter Updates

C.2.1 Computation Steps for PDPM-VAR

As with the covariance terms, we use the stick-breaking representation (Sethuraman, 1994) of the Dirichlet process to enable posterior computation under the DPM priors. For the autocovariance terms, we can express the prior as, $A_i|P_\Theta \sim P_\Theta$, $P_\Theta = \sum_{h_1=1}^{\infty} \pi_{h_1}\delta_{A_h}$, where $\pi_{h_1} = \nu_{h_1}\prod_{l_1<\nu_{h_1}}(1-\nu_{l_1})$, $\nu_{h_1} \sim Beta(1, \alpha_1)$, and $A_h \sim P^*_1$, where P^*_1 is a multivariate normal distribution with mean $\mathbf{0}$ and variance $\text{diag}\{\boldsymbol{\tau}^2\}$. The prior for the individual τ^2 terms is given by $p(\tau^2_{k,d'}) = \frac{\lambda^2}{2}\exp\{-\frac{1}{2}\lambda^2\tau^2_{k,d'}\}$, which implies a double exponential base measure for modeling the autocovariance terms (Park and Casella, 2008). Furthermore, we place a conjugate $Gamma(r, \delta)$ hyperprior on λ^2 to facilitate Gibbs sampling. Throughout our applications we fix $r = 1.0$ and $\delta = 2.0$, which yield good performance in a wide range of settings. As with the residual covariance, we use the slice sampling approach of Walker (2007) to facilitate sampling from this infinite mixture. Let H_1 be a cluster membership indicator, where $H_{1,i} = h_1$ if subject i belongs to the h_1th autocovariance matrix cluster. Let $\mathcal{H}_{h_1} = \{i : H_{1,i} = h_1\}$ be the indices of all subjects belonging to autocovariance cluster h_1, with n_{h_1} being the cardinality of this set. The sampler proceeds by introducing a latent uniform variable u_i, relating the cluster memberships to the stick breaking representation of the DPM. The sampler proceeds by iteratively sampling ν_{h_1} and $u_{1,i}$ from their full conditionals, $\nu_{h_1}|\{H_{1,i}\} \sim Beta\left(1 + n_{h_1}, \alpha_1 + \sum_{i=1}^{n} I_{(H_{1,i}>h_1)}\right)$, $u_{1,i}|\nu_{h_1}, H_{1,i} \sim U(0, \pi_{H_{1,i}})$. The cluster memberships, $H_{1,i}$, are then sampled from a multinomial distribution with posterior

probabilities $(P(H_{1,i}=1|-),\ldots,P(H_{1,i}={h_1}^*|-))$ given by:
$P(H_{1,i}=h_1|-)=\frac{I_{(u_{1,i}<\pi_{h_1})}\prod_{t=1}^{T_i}\phi_{\Sigma_{V_i}}\left(\mathbf{x}_{i,t}-\sum_{k=1}^{K}A_{k,h_1}\mathbf{x}_{i,t-k}\right)}{\sum_{{h_1}'=1}^{{h_1}^*}\left\{I_{(u_{1,i}<\pi_{h_1}')}\prod_{t=1}^{T_i}\phi_{\Sigma_{V_i}}\left(\mathbf{x}_{i,t}-\sum_{k=1}^{K}A_{k,{h_1}'}\mathbf{x}_{i,t-k}\right)\right\}}$ where ${h_1}^*=\min\{h_1:$ $u_{1,i}>1-\sum_{{h_1}'=1}^{h_1}\pi_{A,{h_1}'}, \text{for all } i\}$.

Conditioned on the cluster memberships, we sample the autocovariance matrices across all lags one outcome at a time. The full conditional for $A_{h_{1,d'}\bullet}$ is given by $A_{h_{1,d'}\bullet}|- \sim N(\boldsymbol{\mu}^*_{A_{h_1},d'},\boldsymbol{\Sigma}^*_{A_{h_1},d'})$ with variance and mean: $\boldsymbol{\Sigma}^*_{A_{h_1},d'}=\left(\sum_{i\in\mathcal{H}_{h_1}}\sum_{t=1}^{T_i}\sigma^{-2}_{i,d'}\mathbf{z}_{i,t}\mathbf{z}'_{i,t}+\boldsymbol{\Lambda}_D^{-1}\right)^{-1}$, $\boldsymbol{\mu}^*_{A_{h_1},d'}=\boldsymbol{\Sigma}^*_{A_{h_1},d'}\left\{\sum_{i\in\mathcal{H}_{h_1}}\sum_{t=1}^{T_i}\left[\sigma^{-2}_{i,d'}\mathbf{z}_{i,t}\left(x^{(d')}_{i,t}-\boldsymbol{\Gamma}^*_{i,d'}\boldsymbol{\eta}^*_{i,t}\right)\right]\right\}$, respectively.

Finally, the parameters of the double exponential base measure can be updated using the approach outlined in Park and Casella (2008). The variance term in the base measure can be sampled using $\tau^{-2}_{k,d'}\sim InverseGaussian\left(\sqrt{\frac{\lambda^2}{CA^2_{k,d'}}},\lambda^2\right)$, for $k=1,\ldots,K$ and $d'=1,\ldots,D^2$, where C is the number of clusters. The posterior distribution for the lasso parameter is a gamma distribution, $\lambda^2|\boldsymbol{\tau}^2\sim Gamma(KD^2+r,\delta+\sum_{k=1}^{K}\sum_{d'=1}^{D^2}\frac{\tau^2_{k,d'}}{2})$.

C.2.2 Computation Steps for rgPDPM-VAR

The rgDPM-VAR requires some modification to the slice sampling approach. In particular, the sampler for the rgDPM-VAR extends the latent terms in the slice sampler along the outcome dimension. Let $H_{1d'}$ be the vector of autocovariance cluster indices for outcome d', with $H_{1d',i}=h_{1,d'}$ when subject i belongs to outcome d' cluster $h_{1,d'}$, and let $\mathcal{H}_{h_{1,d'}}=\{i:H_{1d',i}=h_{1,d'}\}$ be the indices of all subjects belonging to outcome d' cluster $h_{1,d'}$, with $n_{h_{1,d'}}$ being the cardinality of this set. Then we have the following full conditionals: $\nu_{h_{1,d'}}|\{H_{1d',i}\}\sim Beta\left(1+n_{d',h},\alpha_1+\sum_{i=1}^{n}I_{(H_{1d',i}>h_{1,d'})}\right)$, $u_{1d',i}|\nu_{h_{1,d'}},H_{1d',i}\sim U(0,\pi_{d',H_{1d',i}})$. The cluster memberships, $H_{1d',i}$, are then sampled from a multinomial distribution with posterior probabilities $\left(P(H_{1d',i}=1|-),\ldots,P(H_{1d',i}=h_{1,d'}{}^*|-)\right)$ given by:
$P\left(H_{1d',i}=h_{1,d'}|-\right)=\frac{I_{(u_{1d',i}<\pi_{d',h_{1,d'}})}\prod_{t=1}^{T_i}\phi_{\Sigma_{V_i}}\left(x^{(d')}_{i,t}-\sum_{k=1}^{K}A_{k,h_{1d'}}\mathbf{x}_{t-k}\right)}{\sum_{h_{1,d'}'=1}^{h_{1,d'}{}^*}\left\{I_{(u_{1d',i}<\pi_{d',h_{1,d'}'})}\prod_{t=1}^{T_i}\phi_{\Sigma_{V_i}}\left(x^{(d')}_{i,t}-\sum_{k=1}^{K}A_{k,h_{1,d'}'}\mathbf{x}_{t-k}\right)\right\}}$, where $h_{1,d'}{}^*=$ $\min\{h:u_{1d',i}>1-\sum_{h_{1,d'}'=1}^{h}\pi_{d',h_{1,d'}'},\text{for all } i\}$. Conditioned on the cluster memberships, the autocovariance terms can be updated in an identical manner to the PDPM-VAR. When updating the parameters of the double exponential base measure the variance terms can be sampled from inverse Gaussian distributions: $\tau^{-2}_{k,d',d^*}\sim InverseGaussian\left(\sqrt{\frac{\lambda^2_{d'}}{C_{d'}\bar{A}^2_{k,d',d^*}}},\lambda^2_{d'}\right)$ for $k=1,\ldots,K$, $d^*=1,\ldots,D$ and $d'=1,\ldots,D$, where $C_{d'}$ is the number of autocovariance clusters for outcome d' and τ^{-2}_{k,d',d^*} is the variance term corresponding to the d^*th element of the d'th row of A_k, and $\bar{A}_{k,d',d^*}$ is the average of element d^* of the d'th row of A_k across the $C_{d'}$ clusters. The outcome-specific lasso parameters have gamma posteriors: $\lambda^2_{d'}|\boldsymbol{\tau}^2_{\boldsymbol{k,d',d^*}}\sim Ga(DK+r,\delta+\sum_{k=1}^{K}\sum_{d^*=1}^{D}\frac{\tau^2_{k,d',d^*}}{2})$ for $d'=1,\ldots,D$.

D	PDPM-VAR	lgPDPM-VAR	rgPDPM-VAR
50	1736	3160	2850
100	2329	3072	3011

Table 3: Average effective sample size (ESS) for elements of the autocovariance matrix for the three proposed methods for varying number of nodes (D). All ESS terms are based on 1500 burnin iterations followed by 3500 MCMC iterations.

C.2.3 Computation Steps for lgPDPM-VAR

The sampling steps under the lgPDPM-VAR model proceeds in a similar manner as the other variants outlined in the manuscript, and are omitted here for space constraints.

Appendix D. Robustness and Convergence

We assessed the mixing of the chains for the simulation experiment. We evaluated the effective sample size (ESS) and visually inspected trace plots. Table 3 displays the average ESS, providing a picture of how well the sampler does in general. The table clearly displays that we can achieve good mixing. Trace plots for some selected elements of the autocovariance matrix are displayed in Figure 6. The plots do not show any evidence of poor mixing.

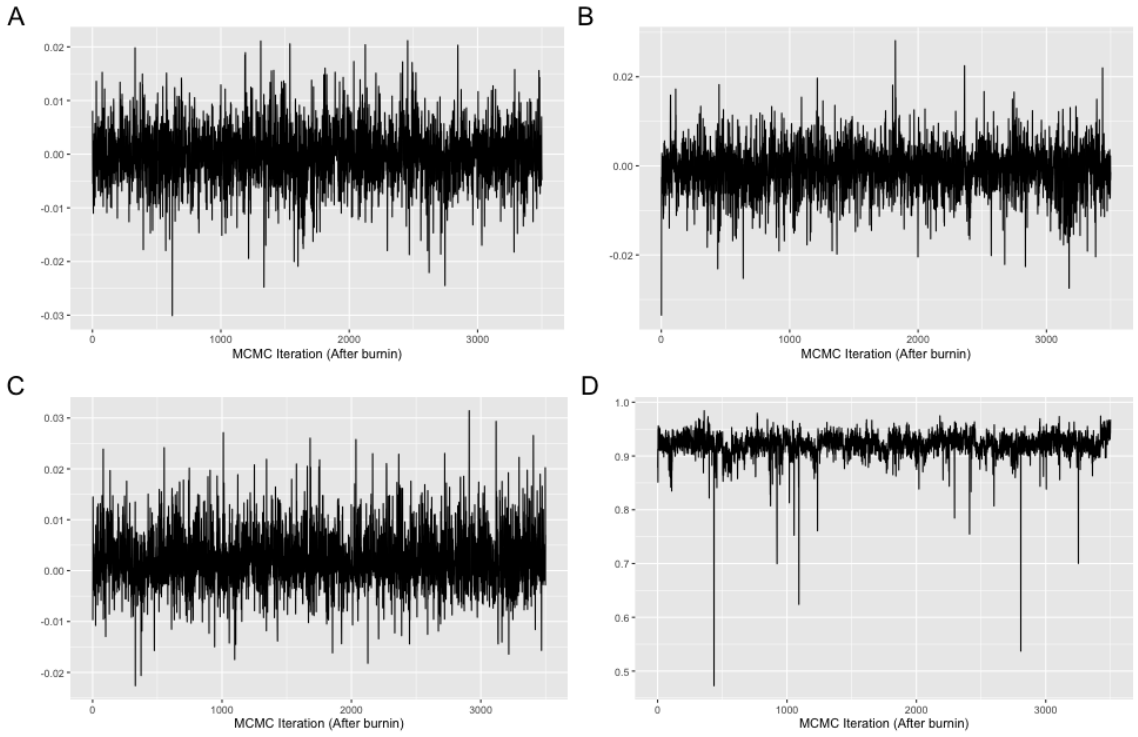

Figure 6: Trace plots for 4 selected elements of the subject-specific autocovariance matrices. (A) subject 34, node 49 lag 2 effect on node 25; (B) subject 38, node 23 lag 1 effect on node 86; (C) subject 31, node 70 lag 1 effect on node 77; (D) subject 38, node 23 lag 2 effect on node 45.

References

Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. Automatic subspace clustering of high dimensional data. *Data Mining and Knowledge Discovery*, 11(1):5–33, 2005.

Teddy J Akiki and Chadi G Abdallah. Determining the hierarchical architecture of the human brain using subject-level clustering of functional networks. *Scientific reports*, 9 (1):1–15, 2019.

Monica Billio, Roberto Casarin, and Luca Rossini. Bayesian nonparametric sparse VAR models. *Journal of Econometrics*, 212(1):97–115, 2019.

Antonio Canale and Pierpaolo De Blasi. Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation. *Bernoulli*, 23(1):379–404, 2017.

Sharon Chiang, Michele Guindani, Hsiang J Yeh, Zulfi Haneef, John M Stern, and Marina Vannucci. Bayesian vector autoregressive model for multi-subject effective connectivity inference using multi-modal neuroimaging data. *Human Brain Mapping*, 38(3):1311–1332, 2017.

Robert H Cramer and Robert B Miller. Multivariate time series analysis of bank financial behavior. *Journal of Financial and Quantitative Analysis*, 13(5):1003–1017, 1978.

Gopikrishna Deshpande, Stephan LaConte, George Andrew James, Scott Peltier, and Xiaoping Hu. Multivariate Granger causality analysis of fMRI data. *Human brain mapping*, 30(4):1361–1373, 2009.

Thomas Doan, Robert Litterman, and Christopher Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric reviews*, 3(1):1–100, 1984.

Daniele Durante, David Dunson, and Joshua Vogelstein. Nonparametric Bayes modeling of populations of networks. *Journal of the American Statistical Association*, 112(520): 1516–1530, 2017.

Alan Edelman. Eigenvalues and condition numbers of random matrices. *SIAM journal on matrix analysis and applications*, 9(4):543–560, 1988.

Robert Engle and Mark Watson. A one-factor multivariate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, 76(376):774–781, 1981.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

Subhashis Ghosal and Aad Van Der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. *The Annals of Statistics*, 35(2):697–723, 2007.

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

Satyajit Ghosh, Kshitij Khare, and George Michailidis. High-dimensional posterior consistency in Bayesian vector autoregressive models. *Journal of the American Statistical Association*, 114(526):735–748, 2018.

Cristina Gorrostieta, Mark Fiecas, Hernando Ombao, Erin Burke, and Steven Cramer. Hierarchical vector auto-regressive models and their applications to multi-subject effective connectivity. *Frontiers in computational neuroscience*, 7:159, 2013.

Hajar Hajmohammadi and Benjamin Heydecker. Multivariate time series modelling for urban air quality. *Urban Climate*, 37:100834, 2021.

Ivan Jeliazkov. Nonparametric vector autoregressions: Specification, estimation, and inference. In *VAR Models in Macroeconomics–New Developments and Applications: Essays in Honor of Christopher A. Sims*. Emerald Group Publishing Limited, 2013.

Maria Kalli and Jim E Griffin. Bayesian nonparametric vector autoregressive models. *Journal of Econometrics*, 203(2):267–282, 2018.

Min Jeong Kim, Hongbin Liu, Jeong Tai Kim, and Chang Kyoo Yoo. Sensor fault identification and reconstruction of indoor air quality (IAQ) data using a multivariate non-Gaussian model in underground building space. *Energy and buildings*, 66:384–394, 2013.

Jeong Hwan Kook, Kelly A Vaughn, Dana M DeMaster, Linda Ewing-Cobbs, and Marina Vannucci. Bvar-connect: A variational Bayes approach to multi-subject vector autoregressive models for inference on brain connectivity networks. *Neuroinformatics*, 19:39–56, 2021.

Dimitris Korobilis. VAR forecasting using Bayesian variable selection. *Journal of Applied Econometrics*, 28(2):204–230, 2013.

Suprateek Kundu and Benjamin B Risk. Scalable Bayesian matrix normal graphical models for brain functional networks. *Biometrics*, 77(2):439–450, 2021.

Javier López-de Lacalle. *tsoutliers: Detection of Outliers in Time Series*, 2024. URL `https://CRAN.R-project.org/package=tsoutliers`.

Feihan Lu, Yao Zheng, Harrington Cleveland, Chris Burton, and David Madigan. Bayesian hierarchical vector autoregressive models for patient-level predictive modeling. *PloS one*, 13(12):e0208082, 2018.

Helmut Lütkepohl. *New introduction to multiple time series analysis*. Springer Science & Business Media, 2005.

Pritthijit Nath, Pratik Saha, Asif Iqbal Middya, and Sarbani Roy. Long-term time-series pollution forecast using statistical and deep learning methods. *Neural Computing and Applications*, 33(19):12551–12570, 2021.

Trevor Park and George Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

Debdeep Pati, David B Dunson, and Surya T Tokdar. Posterior consistency in conditional distribution estimation. *Journal of multivariate analysis*, 116:456–472, 2013.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

Emiliano Santarnecchi, Alexandra Emmendorfer, Sayedhedayatollah Tadayon, Simone Rossi, Alessandro Rossi, and Alvaro Pascual-Leone. Network connectivity correlates of variability in fluid intelligence performance. *Intelligence*, 65:35–47, 2017.

Lorraine Schwartz. On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4(1):10–26, 1965.

Catia Scricciolo. Posterior rates of convergence for Dirichlet mixtures of exponential power densities. *Electronic Journal of Statistics*, 5(none):270 – 308, 2011. doi: 10.1214/11-EJS604. URL `https://doi.org/10.1214/11-EJS604`.

Jayaram Sethuraman. A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650, 1994.

Patrick Sevestre and Alain Trognon. Dynamic linear models. *The econometrics of panel data: A handbook of the theory with applications*, pages 120–144, 1996.

Weining Shen, Surya T Tokdar, and Subhashis Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika*, 100(3):623–640, 2013.

Stephen M Smith, Christian F Beckmann, Jesper Andersson, Edward J Auerbach, Janine Bijsterbosch, Gwenaëlle Douaud, Eugene Duff, David A Feinberg, Ludovica Griffanti, Michael P Harms, et al. Resting-state fmri in the human connectome project. *Neuroimage*, 80:144–168, 2013.

Surya T Tokdar. Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics*, 69(1): 90–110, 2006.

Stephen G Walker. Sampling the Dirichlet mixture model with slices. *Communications in Statistics−Simulation and Computation*, 36(1):45–54, 2007.

Charles L Weise. The asymmetric effects of monetary policy: A nonlinear vector autoregression approach. *Journal of Money, Credit and Banking*, 31(1):85–108, 1999.

Yuefeng Wu and Subhashis Ghosal. Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, 2:298–331, 2008.

Yuefeng Wu and Subhashis Ghosal. The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *Journal of Multivariate Analysis*, 101(10):2411–2419, 2010.