ORIGINAL PAPER

Structured prior distributions for the covariance matrix in latent factor models

Sarah Elizabeth Heaps[1] · Ian Hyla Jermyn[1]

Abstract

Factor models are widely used for dimension reduction in the analysis of multivariate data. This is achieved through decomposition of a $p \times p$ covariance matrix into the sum of two components. Through a latent factor representation, they can be interpreted as a diagonal matrix of idiosyncratic variances and a shared variation matrix, that is, the product of a $p \times k$ factor loadings matrix and its transpose. If $k \ll p$, this defines a parsimonious factorisation of the covariance matrix. Historically, little attention has been paid to incorporating prior information in Bayesian analyses using factor models where, at best, the prior for the factor loadings is order invariant. In this work, a class of structured priors is developed that can encode ideas of dependence structure about the shared variation matrix. The construction allows data-informed shrinkage towards sensible parametric structures while also facilitating inference over the number of factors. Using an unconstrained reparameterisation of stationary vector autoregressions, the methodology is extended to stationary dynamic factor models. For computational inference, parameter-expanded Markov chain Monte Carlo samplers are proposed, including an efficient adaptive Gibbs sampler. Two substantive applications showcase the scope of the methodology and its inferential benefits.

Keywords Covariance matrix · Dimension reduction · Intraday gas demand · Latent factor models · Stationary dynamic factor models · Structured prior distributions

Mathematics Subject Classification 62F15 · 62J99 · 62M10

1 Introduction

Factor models are widely used as a tool for dimension reduction in explaining the covariance structure of multivariate data. Let $y_i = (y_{i1}, \ldots, y_{ip})^T$ be a p-dimensional random vector with mean μ and covariance matrix Ω for observation unit $i = 1, \ldots, n$. The classic factor model posits that the covariances between the p components of y_i can be explained by their mutual dependence on a smaller number k of unknown *common factors* $\eta_i = (\eta_{i1}, \ldots, \eta_{ik})^T$. Specifically, the model expresses y_i as a noisy affine transformation of η_i,

$$y_i = \mu + \Lambda \eta_i + \epsilon_i, \tag{1}$$

in which the $p \times k$ matrix Λ is called the *factor loadings matrix*. The components of the error term ϵ_i are often called the *specific factors* (or "uniquenesses") and their variances are often termed the *idiosyncratic variances*. In (1), we take the specific and common factors, ϵ_i and η_i, to be uncorrelated, zero-mean multivariate normal random variables, with

$$\epsilon_i \sim N_p(0, \Sigma) \quad \text{and} \quad \eta_i \sim N_k(0, I_k). \tag{2}$$

In general, the common factors are assumed to explain all the shared variation between the components of y_i and so Σ is constrained to be a diagonal matrix with positive diagonal elements, $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. It then follows from (1) that the marginal covariance matrix Ω can be expressed as

$$\Omega = \Lambda \Lambda^T + \Sigma \tag{3}$$

in which the product $\Delta = (\delta_{ij}) = \Lambda \Lambda^T$ will henceforth be termed the *shared variation matrix*. If $k \ll p$, the right-hand-

✉ Sarah Elizabeth Heaps
sarah.e.heaps@durham.ac.uk

Ian Hyla Jermyn
i.h.jermyn@durham.ac.uk

1 Department of Mathematical Sciences, Durham University, Upper Mountjoy, Stockton Road, Durham DH1 3LE, UK

side of (3) constitutes a parsimonious factorisation of the full covariance matrix.

Factor models originally found widespread use in psychology and other social sciences (e.g. Goldberg 1990), where a primary motivation was the prospect of domain insight through interpretation of the latent factors. Since their introduction, factor models have been extended in a variety of directions, for example, by allowing sparsity in the matrix of factor loadings for high-dimensional problems (Conti et al. 2014) or by modelling the temporal evolution of variances in stochastic volatility models (Lopes and Carvalho 2007) or the latent factors in dynamic factor models (Sáfadi and Peña 2008). Accordingly, the use of factor models and their extensions has spread into a diverse array of other fields, such as finance (Aguilar and West 2000) and genetics (Carvalho et al. 2008).

In many of these application areas, the modeller is likely to hold prior beliefs about the nature of the dependence between variables. For example, when the observation vector represents repeated measures data, or observations at a collection of spatial locations, it would be reasonable to expect stronger associations between measurements that are closer together in time or space. Indeed, such ideas underpin the class of functional factor models which are tailored to variables that might be better represented as functions over a continuous domain, rather than vector-valued random quantities. Some functional factor models, and their dynamic extensions, have functional factors (e.g. Castellanos et al. 2004; Taylor-Rodriguez et al. 2019) while others have functional factor loadings. In the latter case, if we regard the p observations on unit i as measurements on a function over a continuous domain, say $\tau \in \mathbb{R}$, then we replace (1) with

$$
\begin{aligned}
y_i(\tau_j) &= \mu(\tau_j) + f_i(\tau_j) + \epsilon_i(\tau_j) \\
&= \mu(\tau_j) + \sum_{m=1}^{k} \lambda_m(\tau_j)\eta_{im} + \epsilon_i(\tau_j)
\end{aligned}
$$

for $j = 1, \ldots, p$. The key idea is then to treat the factor loading curves $\lambda_m(\tau)$ as smooth unknown functions of τ. This can be achieved by regarding $f_i(\tau_j) = \sum \lambda_m(\tau_j)\eta_{im}$ as a linear combination of (orthogonal) basis functions, often modelled using splines, which then implies a particular covariance function for $f_i(\tau)$. Although some of the research in this direction uses a Bayesian approach to inference (Kowal et al. 2017; Kowal and Canale 2022), the frequentist treatment of the problem is more common, particularly in financial applications to forecasting yield curves (Hays et al. 2012; Jungbacker et al. 2014). An alternative approach in models with functional factor loadings is to assume each $\lambda_m(\tau)$ to be a realisation of a stochastic process, for instance, a Gaussian process with a chosen covariance function; see, for example, the dynamic spatial model of Lopes et al. (2008).

Beyond analyses of functional data, most of the work in the Bayesian literature on prior specification for factor models has focused on the problem of developing priors that are exchangeable with respect to the order of the variables in the observation vector (e.g. Leung and Drton 2016; Chan et al. 2018) with more recent work focusing additionally on allowing sparsity in the factor loadings matrix (Frühwirth-Schnatter et al. 2023). There has been very little work on constructing non-exchangeable prior specifications in a general setting. A notable exception is the structured increasing shrinkage process prior of Schiavon et al. (2022) under which each factor loading in an infinite factorisation model is assigned a spike-and-slab prior where variable-specific meta-covariates are incorporated in the spike probabilities. Other related work appears in the context of vector error correction models, which are another class of reduced rank models, where Koop et al. (2009) develop an informative prior for the matrix of cointegration vectors based on economic theory about the likely cointegration relationships between variables.

In this paper, we treat the factor model as a tool for providing a dimension-reduced parameterisation of a covariance matrix. Based on the insight that the shared variation matrix $\boldsymbol{\Delta} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$ in (3) is more amenable to the specification of prior beliefs than the factor loadings matrix $\boldsymbol{\Lambda}$, our main contribution is a framework for incorporating initial beliefs about $\boldsymbol{\Delta}$ in the prior for $\boldsymbol{\Lambda}$ by exploiting the algebraic relationship between the two quantities. By switching the focus from properties of the prior for $\boldsymbol{\Lambda}$ to properties of the prior for $\boldsymbol{\Delta}$ we obtain a methodology that is more flexible than alternatives described in the literature. The prior for the shared variation matrix $\boldsymbol{\Delta}$ is induced through a structured prior distribution for the factor loadings matrix $\boldsymbol{\Lambda}$ which can encode ideas of order-invariance as well as non-exchangeable structure, modelled through dependence on covariance kernels, meta-covariates, pairwise distance metrics and more, within a single principled framework. In particular, the mean of the prior for $\boldsymbol{\Delta}$ can be any positive definite matrix of relevance. For example, it could be a phylogenetic distance matrix whose entries encode distances between nodes on a tree, or even the Gram matrix arising from a Gaussian process covariance function. The interpretability of this prior expectation facilitates straightforward incorporation of meaningful domain expertise. Moreover, the flexible class of priors allows data-informed shrinkage towards a region around the mean within a framework that facilitates inference on the number of factors and hence the extent to which dimension-reduction can be achieved. New theoretical results outline the nature of this shrinkage in the rank-reduced space over which $\boldsymbol{\Delta}$ has support while shedding light on some of the identifiability issues that can hamper computational inference for factor models.

Based on ideas from the literature on infinite factorisation models and parameter-expansion algorithms, we propose an

efficient scheme for computational inference where a single Markov chain Monte Carlo (MCMC) run yields information about both the continuous model parameters and the number of factors. Using an unconstrained reparameterisation of stationary vector autoregressions, the methodology is also extended to a class of stationary dynamic factor models for which the shared variation matrix remains meaningfully defined. To the best of our knowledge, this is the first prior, with associated inferential scheme, for a general class of dynamic factor models that constrains inference to the stationary region without imposing further restrictions. This constitutes an important contribution to the Bayesian time-series literature.

The remainder of this paper is organised as follows. After introducing our parameterisation of Bayesian factor models in Sect. 2, Sect. 3 presents the general class of structured priors and its components. In Sect. 4, we extend these ideas to a class of stationary dynamic factor models. Posterior computation is discussed in Sect. 5, then Sect. 6 presents a simulation experiment and two substantive applications that illustrate the scope of the methodology and its inferential benefits.

2 Parameterisation of the Bayesian factor model

It is well known that the parameters $\mathbf{\Lambda}$ and $\mathbf{\Sigma}$ in the marginal covariance matrix $\mathbf{\Omega} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Sigma}$ of a factor model are not identifiable from the likelihood without imposition of constraints. Indeed, even with the scale of the factors fixed at $\mathrm{Var}(\boldsymbol{\eta}_i) = \mathbf{I}_k$ in (2), there remains a rotational invariance. Consider any $k \times k$ orthogonal matrix $\mathbf{Q}$. We can pre-multiply the factors in (1) by $\mathbf{Q}$ and post-multiply the factor loadings matrix by $\mathbf{Q}^{\mathrm{T}}$ and this gives the transformed factors in (2) the same distribution as the original factors and so the marginal variance in (3) remains unchanged. The factor loadings matrix can therefore only be identified up to an orthogonal transformation. This means there are $pk - k(k-1)/2 + p$ degrees of freedom determining the marginal variance $\mathbf{\Omega}$. In an unconstrained model for $\mathbf{y}_i$, the number of degrees of freedom would be $p(p+1)/2$ and so the reduction in the number of parameters is $p(p+1)/2 - \{pk - k(k-1)/2 + p\}$ which is positive if $k < \varphi(p)$ where

$$\varphi(p) = \frac{2p + 1 - \sqrt{8p+1}}{2} \qquad (4)$$

is called the Ledermann bound. Notwithstanding rotational invariance, the first identification issue is therefore whether $\mathbf{\Sigma}$ can be identified uniquely from $\mathbf{\Omega}$. Fortunately, Bekker and ten Berge (1997) proved that if $k < \varphi(p)$ then $\mathbf{\Sigma}$, and therefore $\mathbf{\Delta} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$, is almost surely *globally identifiable*.

Given identification of $\mathbf{\Delta}$, solving the rotation problem would then guarantee unique identification of the factor loadings matrix $\mathbf{\Lambda}$.

In the Bayesian literature, the most common solution to the rotation problem uses the positive diagonal, lower triangular (PLT) constraint; denoting the constrained matrix by $\tilde{\mathbf{\Lambda}}$ and corresponding factors by $\tilde{\boldsymbol{\eta}}_1, \ldots, \tilde{\boldsymbol{\eta}}_n$, this demands $\tilde{\lambda}_{ij} = 0$ for $j > i$ and $\tilde{\lambda}_{ii} > 0$ for $i = 1, \ldots, k$. Although historically this approach has been widely used in the Bayesian literature (e.g. Geweke and Zhou 1996; Lopes and West 2004), we choose not to address the rotational invariance and instead parameterise the model using the unidentified and unconstrained factor loadings matrix $\mathbf{\Lambda}$. There are two main motivations.

First, recent work provides theoretical and empirical evidence that imposition of the PLT constraint can adversely affect computational inference. Writing $\tilde{\mathbf{\Lambda}}^{(k)}$ to denote the lower triangular matrix comprising the first k rows and columns of $\tilde{\mathbf{\Lambda}}$, problems arise when $\tilde{\mathbf{\Lambda}}^{(k)}$ is near singular, which can occur when variables in $\mathbf{y}_i$ are highly correlated. When the jth diagonal element of $\tilde{\mathbf{\Lambda}}^{(k)}$ is close to zero, the signs of $(\tilde{\lambda}_{j+1,j}, \ldots, \tilde{\lambda}_{pj})$ are only very weakly identified, producing multiple modes in the posterior corresponding to the different sign flips (e.g. Aßmann et al. 2016; Man and Culpepper 2022). This can cause poor mixing in MCMC samplers and poor approximation of the posterior for the number of factors (Chan et al. 2018). These problems have been addressed successfully by parameter-expansion techniques, which target a posterior over an expanded space with fewer pathological features (e.g. Ghosh and Dunson 2009; Chan et al. 2018). Parameterising the model and framing the problem of computational inference in terms of the unidentified factor loadings matrix $\mathbf{\Lambda}$ falls into this class of methods.

Second, parameterisation in terms of the unconstrained and unidentified factor loadings matrix $\mathbf{\Lambda}$ delivers an additional benefit in terms of prior specification. Direct elicitation of a prior for the identified factor loadings in $\tilde{\mathbf{\Lambda}}$ is difficult as they quantify relationships with latent factors whose interpretation is generally unclear *a priori*. In contrast, beliefs about the shared variation $\mathbf{\Delta}$ in linear Gaussian models can be related directly to beliefs about observable quantities. As a result, $\mathbf{\Delta}$ is generally a more intuitive parameter for the quantification of covariation. For example, in a spatial problem where the elements of $\mathbf{y}_i$ correspond to observations at different locations, a modeller might reasonably structure their beliefs through a covariance matrix for which covariance decays as a function of distance. Fortunately, subject to $k < \varphi(p)$, identifiability of $\mathbf{\Lambda}$ is not necessary for identification of $\mathbf{\Delta}$. Further, under various standard distributions for an unconstrained random matrix $\mathbf{\Lambda}$, the first and second order moments of the shared variation matrix $\mathbf{\Delta} = \tilde{\mathbf{\Lambda}}\tilde{\mathbf{\Lambda}}^{\mathrm{T}} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$ can be calculated in closed form. Through careful specifi-

cation of the prior for $\boldsymbol{\Lambda}$, a modeller can therefore capture their beliefs about the moments in the prior for the shared variation matrix.

3 Structured prior distributions

3.1 Significance of prior expectation

The prior expectation of the shared variation matrix $\boldsymbol{\Delta} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$ is defined by $E(\boldsymbol{\Delta}) = \{E(\delta_{ij})\} = E(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}})$. As we detail in Sect. 3.4, $E(\boldsymbol{\Delta})$ will be chosen to reflect beliefs about the covariation amongst the p elements of $\boldsymbol{y}_i$. In general, it will be a rank p matrix.

For $k < p$, denote by $\mathcal{S}_{p,k}^+$ the set of rank k, $p \times p$ symmetric, positive semi-definite matrices and write $\mathcal{S}_p^+$ for the space of $p \times p$ symmetric, positive definite matrices. The factor loadings matrix $\boldsymbol{\Lambda}$ is rank $k < \varphi(p) < p$ and so the prior distribution for the shared variation matrix $\boldsymbol{\Delta} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$ only offers non-zero support to $\mathcal{S}_{p,k}^+$. Because $\mathcal{S}_{p,k}^+$ is not a convex set, it need not contain the prior expectation of $\boldsymbol{\Delta}$. Indeed, as stated previously, $E(\boldsymbol{\Delta})$ will generally be rank p. Therefore, although $E(\boldsymbol{\Delta})$ represents the centre of prior mass, in general there will be zero density at this point. The significance of $E(\boldsymbol{\Delta})$ is thus not completely clear. We will elucidate its meaning via an alternative, constrained expectation, as follows.

The Frobenius norm of a matrix is simply the Euclidean norm of its vectorised form. Define the (squared) *Frobenius distance* between two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ as the squared Frobenius norm of their difference:

$$d(\boldsymbol{A}, \boldsymbol{B})^2 = \|\boldsymbol{A} - \boldsymbol{B}\|_F^2 = \mathrm{tr}\left\{(\boldsymbol{A} - \boldsymbol{B})^{\mathrm{T}}(\boldsymbol{A} - \boldsymbol{B})\right\}.$$

By analogy with the classic Euclidean expectation, we now define the *constrained expectation* of $\boldsymbol{\Delta}$ as $E_F(\boldsymbol{\Delta}) = \boldsymbol{L}_0\boldsymbol{L}_0^{\mathrm{T}} \in \mathcal{S}_{p,k}^+$ where

$$\boldsymbol{L}_0 = \min_{\boldsymbol{L} \in \mathbb{R}^{p \times k}} E_{\boldsymbol{\Lambda}}\left\{d(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}, \boldsymbol{L}\boldsymbol{L}^{\mathrm{T}})^2\right\}.$$

Theorem 1 below asserts that if the kth and $(k + 1)$th eigenvalues of $E(\boldsymbol{\Delta})$ are distinct, the constrained expectation of $\boldsymbol{\Delta}$ is the point in $\mathcal{S}_{p,k}^+$ that minimises the Frobenius distance to $E(\boldsymbol{\Delta})$. However, before proving Theorem 1, we need Proposition 1.

Proposition 1 *Let the spectral decomposition of a matrix $\boldsymbol{D} \in \mathcal{S}_p^+$ be $\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathrm{T}}$, where $\boldsymbol{M} = \mathrm{diag}(m_1, \ldots, m_p)$ is a diagonal matrix of ordered eigenvalues, $m_1 \geq m_2 \geq \cdots \geq m_p > 0$, and $\boldsymbol{U}$ is an orthogonal matrix whose columns comprise the corresponding eigenvectors. Assume that $m_k \neq m_{k+1}$. Then, for $\boldsymbol{\Lambda} \in \mathbb{R}^{p \times k}$, the matrix product $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$ which minimises the Frobenius distance to $\boldsymbol{D}$ is*

$\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} = \boldsymbol{D}^{1/2}\boldsymbol{U}^{(k)}\boldsymbol{U}^{(k)\,\mathrm{T}}\boldsymbol{D}^{1/2}$ *where $\boldsymbol{U}^{(k)}$ is the sub-matrix comprising the first k columns of $\boldsymbol{U}$. Moreover, the minimum squared Frobenius distance is equal to the sum of the squares of the last $p - k$ eigenvalues of $\boldsymbol{D}$.*

The proof of Proposition 1 is provided in the Supplementary Materials.

Theorem 1 *If $\boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{\mathrm{T}}$ denotes the spectral decomposition of $E(\boldsymbol{\Delta}) \in \mathcal{S}_p^+$ and $m_k \neq m_{k+1}$, then the constrained expectation of $\boldsymbol{\Delta}$, $E_F(\boldsymbol{\Delta}) \in \mathcal{S}_{p,k}^+$, is equal to the matrix product which minimises the Frobenius distance to $E(\boldsymbol{\Delta})$. That is, $E_F(\boldsymbol{\Delta}) = E(\boldsymbol{\Delta})^{1/2}\boldsymbol{U}^{(k)}\boldsymbol{U}^{(k)\,\mathrm{T}}E(\boldsymbol{\Delta})^{1/2}$.*

The proof of Theorem 1 is given in the Supplementary Materials. Its significance lies in the suggestion that the prior for $\boldsymbol{\Delta}$ encourages shrinkage towards the closest matrix in $\mathcal{S}_{p,k}^+$ to the rank p prior expectation $E(\boldsymbol{\Delta})$, hence for a given structure for $E(\boldsymbol{\Delta})$, approximating it as closely as possible in rank-reduced form. The Supplementary Materials also consider the case $m_k = m_{k+1}$ and the implications for computational inference.

3.2 A matrix normal prior

A random matrix $\boldsymbol{\Lambda}$ has a matrix normal distribution with location matrix $\boldsymbol{M} \in \mathbb{R}^{p \times k}$, among-row scale matrix $\boldsymbol{\Phi} \in \mathcal{S}_p^+$ and among-column scale matrix $\boldsymbol{\Psi} \in \mathcal{S}_k^+$, written $\boldsymbol{\Lambda} \sim \mathrm{N}_{p,k}(\boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$, if $\mathrm{vec}(\boldsymbol{\Lambda})$ is multivariate normal and such that $\mathrm{vec}(\boldsymbol{\Lambda}) \sim \mathrm{N}_{pk}\left(\mathrm{vec}(\boldsymbol{M}), \boldsymbol{\Psi} \otimes \boldsymbol{\Phi}\right)$. Since $(\alpha\boldsymbol{\Psi}) \otimes (\alpha^{-1}\boldsymbol{\Phi}) = \boldsymbol{\Psi} \otimes \boldsymbol{\Phi}$ for any scalar $\alpha > 0$, the overall scale of either $\boldsymbol{\Phi}$ or $\boldsymbol{\Psi}$ can be fixed without loss of generality.

Suppose that we take $\boldsymbol{\Lambda} \sim \mathrm{N}_{p,k}(\boldsymbol{0}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$ as a prior for the unconstrained factor loadings matrix and fix the scale of the among-row scale matrix by taking $\mathrm{tr}(\boldsymbol{\Phi}) = p$ or $\mathrm{tr}(\boldsymbol{\Phi}^{-1}) = p$. As remarked in Sect. 2, the shared variation matrix $\boldsymbol{\Delta} = (\delta_{ij}) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$ is a quantity which is amenable to the specification of prior beliefs, and so our goal is to derive its moments. Using standard theory for the matrix normal distribution (e.g. Gupta and Nagar 2000, Chapter 2), it can be shown that the expectation of $\boldsymbol{\Delta}$ is given by

$$E(\boldsymbol{\Delta}) = \mathrm{tr}(\boldsymbol{\Psi})\boldsymbol{\Phi}, \tag{5}$$

while the covariance between δ_{ij} and $\delta_{k\ell}$ is $\mathrm{Cov}(\delta_{ij}, \delta_{k\ell}) = \mathrm{tr}(\boldsymbol{\Psi}^2)(\phi_{ik}\phi_{j\ell} + \phi_{i\ell}\phi_{jk})$. In the special case where $i = k$ and $j = \ell$, the variance of δ_{ij} is then

$$\mathrm{Var}(\delta_{ij}) = \mathrm{tr}(\boldsymbol{\Psi}^2)(\phi_{ii}\phi_{jj} + \phi_{ij}^2). \tag{6}$$

The derivations of the moments above are provided in the Supplementary Materials.

The result in (5) is significant because of the interpretation it bestows on the among-row scale matrix $\boldsymbol{\Phi}$ as a

standardised version of our prior expectation for the shared variation matrix $\boldsymbol{\Delta}$. Parametric forms are convenient models for covariance matrices because they provide a parsimonious representation of dependence and a way to model the relationships with as-yet unobserved variables. We can therefore model the among-row scale matrix $\boldsymbol{\Phi}$ using a parametric form. This encourages shrinkage of the shared variation matrix $\boldsymbol{\Delta}$ towards an area around that parametric form but without enforcing the structure as a model constraint.

Suppose that $\boldsymbol{\Phi} = \boldsymbol{R}(\boldsymbol{\vartheta})$ in which $\boldsymbol{\vartheta}$ is a low-dimensional vector of hyperparameters on which the parametric form $\boldsymbol{R}(\cdot)$ depends. We complete our prior for the factor loadings matrix by taking $\boldsymbol{\Psi} = \mathrm{diag}(\psi_1, \ldots, \psi_k)$ and assigning a prior to $\boldsymbol{\vartheta}$ and the ψ_i. Parametric forms $\boldsymbol{R}(\cdot)$ and the prior for the ψ_i are discussed in Sects. 3.4 and 3.5, respectively.

3.3 A matrix-*t* prior

The expressions for the means and variances in (5) and (6) under the matrix normal prior reveal a lack of flexibility; recalling that the scale of $\boldsymbol{\Phi}$ is fixed, once the mean $E(\boldsymbol{\Delta})$ has been chosen, it is clearly not possible to scale up or down all the variances by scaling $\boldsymbol{\Psi}$. As such, we cannot assess independently the overall scale of the mean and uncertainty around that mean. A matrix-*t* prior for $\boldsymbol{\Lambda}$ remedies this problem through the introduction of a degree of freedom parameter.

Let $\boldsymbol{S} \sim \mathrm{W}_p(\varsigma + p - 1, \boldsymbol{\Phi}^{-1})$ and $\boldsymbol{X} \sim \mathrm{N}_{p,k}(\boldsymbol{0}, \boldsymbol{I}_p, \boldsymbol{\Psi})$ be independent random matrices, where $\mathrm{W}_n(q, \boldsymbol{U})$ denotes the n-dimensional Wishart distribution with $q > 0$ degrees of freedom and scale matrix $\boldsymbol{U} \in \mathcal{S}_n^+$. If we define

$$\boldsymbol{\Lambda} = (\boldsymbol{S}^{-1/2})^{\mathrm{T}} \boldsymbol{X} + \boldsymbol{M} \tag{7}$$

where $\boldsymbol{S}^{1/2}(\boldsymbol{S}^{1/2})^{\mathrm{T}} = \boldsymbol{S}$ and $\boldsymbol{M} \in \mathbb{R}^{p \times k}$, then the random matrix $\boldsymbol{\Lambda}$ has a matrix-*t* distribution, written $\boldsymbol{\Lambda} \sim \mathrm{t}_{p,k}(\varsigma, \boldsymbol{M}, \boldsymbol{\Phi}, \boldsymbol{\Psi})$. Again, one can fix the scale of either the among-row scale matrix $\boldsymbol{\Phi} \in \mathcal{S}_p^+$ or the among-column scale matrix $\boldsymbol{\Psi} \in \mathcal{S}_k^+$ without loss of generality.

Suppose that we adopt $\boldsymbol{\Lambda} \sim \mathrm{t}_{p,k}(\varsigma, \boldsymbol{0}, \check{\boldsymbol{\Phi}}, \boldsymbol{\Psi})$ as a prior for the factor loadings matrix. Expressions for the means, variances and covariances of $\boldsymbol{\Delta} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$ are derived in the Supplementary Materials. In particular, we show that $E(\boldsymbol{\Delta}) = \mathrm{tr}(\boldsymbol{\Psi})\boldsymbol{\Phi}$ for $\varsigma > 2$ where $\boldsymbol{\Phi} = \check{\boldsymbol{\Phi}}/(\varsigma - 2)$, which is identical to the expression derived under the matrix normal prior. We also show that for a fixed value of $E(\boldsymbol{\Delta})$, the variance,

$$\mathrm{Var}(\delta_{ij}) = \frac{\left\{\mathrm{tr}(\boldsymbol{\Psi})^2 + (\varsigma - 2)\mathrm{tr}(\boldsymbol{\Psi}^2)\right\}\left\{\varsigma \phi_{ij}^2 + (\varsigma - 2)\phi_{ii}\phi_{jj}\right\}}{(\varsigma - 1)(\varsigma - 4)} \text{ for } \varsigma > 4,$$

can be increased or decreased by decreasing or increasing ς, respectively. Therefore, by treating $\varsigma > 4$ as an

unknown and assigning it a prior, we can allow the data to influence the degree of shrinkage of the shared variation matrix towards the closest rank k matrix to its mean. Writing $\check{\zeta} = 1/(\varsigma - 4) \in \mathbb{R}^+$, a matrix normal distribution for $\boldsymbol{\Lambda}$ is recovered as $\check{\zeta} \to 0$. We can therefore allow controlled relaxation of the fixed shrinkage of the matrix normal prior by specifying a prior for $\check{\zeta}$ with its mode at zero. To this end, we recommend an exponential distribution $\check{\zeta} \sim \mathrm{Exp}(a_0)$. In the special case when $\boldsymbol{\Phi} = \boldsymbol{I}_p$ and $\boldsymbol{\Psi} = \psi \boldsymbol{I}_k$, we have $\mathrm{Var}(\delta_{ij}) = s_k(\check{\zeta})E(\delta_{ij})$ where $s_k(\check{\zeta}) = (1 + 2\check{\zeta})\left\{1 + (2 + k)\check{\zeta}\right\}/(1 + 3\check{\zeta})$ is an increasing function in $\check{\zeta}$ which takes its minimum value of 1 when $\check{\zeta} = 0$. By trial and improvement, we can therefore use the quantiles of the $\mathrm{Exp}(a_0)$ distribution to choose a value for a_0 which is consistent with our prior beliefs about a chosen quantile in the distribution for the scale factor $s_k(\check{\zeta})$ for various k. This is illustrated in the application in Sect. 6.2.

3.4 Model and prior for $\boldsymbol{\Phi}$

3.4.1 General form

Under either a matrix normal or matrix-*t* prior for $\boldsymbol{\Lambda}$, $E(\boldsymbol{\Delta}|\boldsymbol{\Phi}, \boldsymbol{\Psi}) = \mathrm{tr}(\boldsymbol{\Psi})\boldsymbol{\Phi}$. As explained in Sect. 3.2, since we choose to fix $\mathrm{tr}(\boldsymbol{\Phi}) = p$ or $\mathrm{tr}(\boldsymbol{\Phi}^{-1}) = p$, we can therefore interpret $\boldsymbol{\Phi}$ as a standardised version of the prior expectation of $\boldsymbol{\Delta}$. For most of the parametric forms described in this section, it will be possible to factorise $\boldsymbol{\Phi}$ or $\boldsymbol{\Xi} = \boldsymbol{\Phi}^{-1}$ as $\boldsymbol{\Phi} = \boldsymbol{R}(\boldsymbol{\vartheta})$ or $\boldsymbol{\Xi} = \boldsymbol{R}(\boldsymbol{\vartheta})$, respectively, where $\boldsymbol{R}(\cdot)$ yields a positive definite matrix with 1 s on the diagonal. The vector $\boldsymbol{\vartheta}$ typically contains a small number of unknown hyperparameters to which we assign a prior.

3.4.2 Specific examples

The simplest structure for $\boldsymbol{\Phi}$ would take $\boldsymbol{\Phi} = \boldsymbol{I}_p$ giving $E(\boldsymbol{\Delta}|\boldsymbol{\Phi}, \boldsymbol{\Psi}) = \mathrm{tr}(\boldsymbol{\Psi})\boldsymbol{I}_p$. In the matrix normal case when, additionally, $\boldsymbol{\Psi} = \psi \boldsymbol{I}_k$, we obtain the order-invariant prior presented in Leung and Drton (2016) for the identified factor loadings matrix in which $\tilde{\lambda}_{ij} \sim \mathrm{N}(0, \psi)$ for $i \neq j$ and $\tilde{\lambda}_{ii}^2 \sim \mathrm{Gam}\{(k - i + 1)/2, 1/(2\psi)\}$ for $i = 1, \ldots, k$. Although taking $\boldsymbol{\Phi} = \boldsymbol{I}_p$ will give an order-invariant prior for $\boldsymbol{\Delta}$, a more general distribution that is exchangeable with respect to the order of the variables in $\boldsymbol{y}_i$ arises by taking $\boldsymbol{\Phi}$ to be a two-parameter exchangeable matrix. Since we constrain $\mathrm{tr}(\boldsymbol{\Phi}) = p$ this yields $\boldsymbol{\Phi} = (1-\vartheta)\boldsymbol{I}_p + \vartheta \boldsymbol{J}_p$, where $-1/(p-1) < \vartheta < 1$, $\boldsymbol{J}_p = \boldsymbol{1}_p\boldsymbol{1}_p^{\mathrm{T}}$ and $\boldsymbol{1}_p$ is a p-vector of 1 s. This is the most general form for a $p \times p$ symmetric, positive definite matrix with $\mathrm{tr}(\boldsymbol{\Phi}) = p$ which is invariant under a common permutation of the rows and columns. If a modeller has nothing in their prior beliefs to distinguish among the p elements of $\boldsymbol{y}_i$, the advantage of this specification over the order-invariant prior of Leung and Drton (2016) is that it

allows shrinkage of the $p(p-1)/2$ off-diagonal elements of the shared variation matrix $\boldsymbol{\Delta}$ towards non-zero values. This can be particularly helpful when the number of observations is small relative to the dimension of $\boldsymbol{y}_i$.

In confirmatory factor analysis, structural zeros are often introduced in the factor loadings matrix based on prior beliefs about the relationships between the k factors and the variables in $\boldsymbol{y}_i$. For example, in the prototypical two-factor model, the first p_1 variables load only on factor 1 whilst the remaining $p - p_1$ variables load only on factor 2. This would correspond to a block diagonal shared variation matrix $\boldsymbol{\Delta}$. Rather than imposing these beliefs about independent group structure as a model constraint, one could instead encourage shrinkage towards the corresponding structure through the prior. With appropriate ordering of the variables, this could be achieved by taking $\boldsymbol{\Phi} = \text{blockdiag}(\boldsymbol{\Phi}_1, \ldots, \boldsymbol{\Phi}_k)$ in which each $\boldsymbol{\Phi}_i$ block is a two-parameter exchangeable matrix, as described above.

Beyond these simple examples, there are many fields of statistics in which parametric forms are adopted for correlation matrices or their inverses. Typically, they are intended to capture the idea that observations from variables that are closer together, often in time or space, tend to be more strongly (positively) correlated. For instance, the precision matrix for a vector of longitudinal measurements might be assumed to be that of a stationary autoregressive process of order one. We could centre the prior for $\boldsymbol{\Delta}$ on a matrix of this form by taking $\boldsymbol{\Xi} = (\xi_{ij})$ to be a tridiagonal matrix with $\xi_{i,i+1} = \xi_{i+1,i} = -\vartheta$ for $i = 1, \ldots, p-1, \xi_{ii} = 1 + \vartheta^2$ for $i = 2, \ldots, p-1$ and $\xi_{11} = \xi_{pp} = 1$ where $\vartheta \in (-1, 1)$.

In spatial statistics, it is very common to use a spatial correlation function, such as one from the Matérn class. A conventional spatial correlation function will depend only on the distance, and possibly direction, between sites. However, Schmidt et al. (2011) introduced the idea of recording distances in a space of dimension greater than two, by introducing covariates into the correlation function. A special case, termed the projection model, measures the distance between sites j and k in this augmented space using the Mahalanobis metric with respect to an (unknown) $c \times c$ symmetric, positive definite matrix $\boldsymbol{\Theta}$, that is,

$$d_{C,jk} = \sqrt{\{(\boldsymbol{x}_j - \boldsymbol{x}_k)^{\text{T}} \boldsymbol{\Theta} (\boldsymbol{x}_j - \boldsymbol{x}_k)\}}. \tag{8}$$

Here $\boldsymbol{x}_j$ is a vector of coordinates for site j, comprising its two spatial coordinates and the values of $c - 2$ meta-covariates. This is then combined with an exponential spatial correlation function to give $\phi_{jk} = \exp(-d_{C,jk})$. Omitting the spatial coordinates, these ideas can be extended to non-spatial settings if there are meta-covariates associated with the p variables that are thought to influence the relationships between deviations from the mean $(y_{ij} - \mu_j)$ and $(y_{ik} - \mu_k)$ in any vector measurement. Indeed, any Gaussian process

correlation matrix could be used for $\boldsymbol{\Phi}$ which has huge scope for the incorporation of meaningful domain expertise.

More generally, if the distances between variables can be measured by some metric, say d_{jk}, this could be used to structure the prior through the specification $\phi_{jk} = \exp(-d_{jk}/\vartheta)$. An ecological example is provided in Sect. 6.2 which uses phylogenetic distances between species in a joint model for species occurrence.

3.5 Model and prior for $\boldsymbol{\Psi}$

Except in confirmatory factor analysis, an appropriate number of factors is not usually known *a priori*. Various methods have been proposed in the literature to address this source of uncertainty. This includes techniques to approximate the marginal likelihood of models with different numbers of factors, including path sampling (Dutta and Ghosh 2013) and Chib's method (Chib et al. 2006). However, these methods are very computationally expensive, requiring separate Monte Carlo simulations for models with $k = 1, \ldots, H$ factors, where H is the maximum number of factors that a modeller is prepared to entertain. Lopes and West (2004) propose a reversible jump MCMC algorithm but the proposals for transdimensional moves require pilot MCMC runs for models with $k = 1, \ldots, H$ factors which rapidly becomes untenable for high-dimensional problems. Other methods that attempt to allow simultaneous inference of the number of factors and the model parameters rely on identification of individual zeros in the factor loadings matrix (Frühwirth-Schnatter and Lopes 2010; Conti et al. 2014; Frühwirth-Schnatter et al. 2023). The approach taken here in similar in spirit, relying on the removal of columns whose contribution to $\boldsymbol{\Omega}$ is negligible.

Due to the challenges in approximating the posterior for k, a recent approach that has become popular relies conceptually on an overfitted factor model (Bhattacharya and Dunson 2011; Legramanti et al. 2020; Schiavon et al. 2022). In theory, infinitely many factors are permitted but cumulative shrinkage priors are used to increasingly shrink columns of the factor loadings matrix towards zero as the column index increases. This allows the infinite factor loadings matrix to be approximated with a finite matrix by omitting columns that have been shrunk close enough to zero. The posterior for the number of non-discarded columns k^* is then used as a proxy for the posterior for k. As such, a single MCMC sample yields information about both the model parameters and the number of latent factors.

A popular increasing shrinkage prior is the multiplicative gamma process (MGP) (Bhattacharya and Dunson 2011). Conditional on a set of hyperparameters, the factor loadings within each column are assumed to be independent normal random variables with zero mean and a column-specific precision parameter. The precisions are a cumulative prod-

uct of gamma random variables and their prior expectation increases with the column index. A factor is deemed inactive *a posteriori* if the contribution of its loadings to $\boldsymbol{\Omega}$ is negligible under a chosen truncation criterion. In order to apply these ideas to our structured prior for the shared variation matrix, we take the among-column scale matrix $\boldsymbol{\Psi}$ to be diagonal and assign the reciprocals of its elements a MGP prior

$$\psi_h^{-1} = \prod_{\ell=1}^{h} \varrho_\ell,$$
$$\varrho_1 \sim \text{Gam}(a_1, 1), \quad \varrho_\ell \sim \text{Gam}(a_2, 1), \quad \ell \geq 2 \qquad (9)$$

in which the ϱ_ℓ are independent. To ensure identifiability of the shared variation matrix $\boldsymbol{\Delta} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\text{T}}$ and the idiosyncratic variances $\boldsymbol{\Sigma}$ in the likelihood, we choose to truncate the prior at $H \leq \lceil \varphi(p) \rceil - 1$ where $\varphi(p)$ was defined in (4). Guidelines on the choice of a_1 and a_2 to ensure the $1/\psi_h$ are stochastically increasing in a reasonable neighbourhood of zero are provided in Durante (2017).

4 Stationary dynamic factor models

4.1 Bayesian dynamic factor models

In the classic vector autoregressive (VAR) model, the number of parameters grows quadratically with the dimension of the observation vector. In contrast, the number of parameters in a dynamic factor model only grows quadratically with the dimension of the latent factors. This makes them a valuable tool for dimension-reduction in multivariate time series analysis. A dynamic factor model has the same observation equation as a static model,

$$\boldsymbol{y}_t = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_t + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \text{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}). \qquad (10)$$

However, the factors in a dynamic model evolve over time according to an order m vector autoregression (VAR(m)). This yields the state equation as

$$\boldsymbol{\eta}_t = \boldsymbol{\Gamma}_1 \boldsymbol{\eta}_{t-1} + \cdots + \boldsymbol{\Gamma}_m \boldsymbol{\eta}_{t-m} + \boldsymbol{\zeta}_t, \quad \boldsymbol{\zeta}_t \sim \text{N}_k(\boldsymbol{0}, \boldsymbol{\Pi}), \ (11)$$

for $t = 1, 2, \ldots$ in which the observation and factor innovations $\boldsymbol{\epsilon}_t$ and $\boldsymbol{\zeta}_t$ are serially and mutually uncorrelated. As in the static case, it is generally assumed that the common factors explain all of the shared variation and so $\boldsymbol{\Sigma}$ is taken to be diagonal. Various extensions of this model have been presented, for example, Peña and Poncela (2006) incorporate a moving average component in state equation, Aßmann et al. (2016) introduce lagged factors in (10) and Kaufmann and Schumacher (2019) allow the components

of the idiosyncratic innovations to evolve independently as univariate stationary autoregressions.

4.2 A stationary dynamic factor model

The methodology described in this paper relies on the use of a structured prior distribution for the shared variation component of the marginal variance. The marginal variance of a dynamic factor model is only meaningfully defined when the factors evolve according to a stationary process; hereafter we focus only on situations where this is a reasonable assumption. The manner by which stationarity is enforced is discussed in Sect. 4.3.

In order to identify the scale of the loadings matrix in a dynamic factor model, the scale of the factors is normally constrained by fixing $\boldsymbol{\Pi} = \text{Var}(\boldsymbol{\zeta}_t) = \boldsymbol{I}_k$. However, to retain the interpretation of $\boldsymbol{\Delta} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\text{T}}$ as a shared variation matrix, we instead constrain the stationary covariance matrix so that $\boldsymbol{G}_0 = \text{Var}(\boldsymbol{\eta}_t) = \boldsymbol{I}_k$ and hence the marginal variance of $\boldsymbol{y}_t$ remains as $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\text{T}} + \boldsymbol{\Sigma}$. We can therefore assign a prior to the unconstrained factor loadings matrix $\boldsymbol{\Lambda}$ using precisely the ideas discussed in Sect. 3 for the static model.

To complete specification of the stationary dynamic factor model, the state equation (11) must be initialised at its stationary distribution. To this end, we augment the parameter space with m auxiliary factors at times $t = 1 - m, \ldots, 0$. Generalising the definition of the stationary variance $\boldsymbol{G}_0$ above, we define $\boldsymbol{G}_i = \text{Cov}(\boldsymbol{\eta}_t, \boldsymbol{\eta}_{t+i})$ as the ith autocovariance of $\boldsymbol{\eta}_t$. The initial distribution can then be expressed as $(\boldsymbol{\eta}_{1-m}^{\text{T}}, \ldots, \boldsymbol{\eta}_0^{\text{T}})^{\text{T}} \sim \text{N}_{km}(\boldsymbol{0}, \boldsymbol{G})$ where $\boldsymbol{G}$ is a positive definite block Toeplitz matrix with $\boldsymbol{G}_{j-i}$ as the block in rows $\{k(i-1) + 1\}$ to ki and columns $\{k(j-1) + 1\}$ to kj $(i, j = 1, \ldots, m)$ and $\boldsymbol{G}_{-\ell} = \boldsymbol{G}_\ell^{\text{T}}$ $(\ell = 1, \ldots, m-1)$.

4.3 Enforcing the stationarity constraint

Denote by B the backshift operator, $B\boldsymbol{\eta}_t = \boldsymbol{\eta}_{t-1}$. The VAR model governing evolution of the factors can then be written as $\boldsymbol{\zeta}_t = (\boldsymbol{I}_k - \boldsymbol{\Gamma}_1 B - \cdots - \boldsymbol{\Gamma}_m B^m)\boldsymbol{\eta}_t = \boldsymbol{\Gamma}(B)\boldsymbol{\eta}_t$ where $\boldsymbol{\Gamma}_i \in \mathbb{R}^{k \times k}$ for $i = 1, \ldots, m$ and $\boldsymbol{\Gamma}(u) = \boldsymbol{I}_k - \boldsymbol{\Gamma}_1 u - \cdots - \boldsymbol{\Gamma}_m u^m$, $u \in \mathbb{C}$, is termed the characteristic polynomial. The process is stable if and only if all the roots of $\det\{\boldsymbol{\Gamma}(u)\} = 0$ lie outside the unit circle. Because all stable processes are stationary and unstable stationary processes are not generally of interest, this subset of the Cartesian product space in which the $\boldsymbol{\Gamma}_i$ lie is often referred to as the stationary region, $\mathcal{C}_{m,k}$. Even for the simplest vector case of an order-1 bivariate autoregression, it has a highly complex geometry.

Various authors have considered Bayesian inference for stationary dynamic factor models (e.g. Sáfadi and Peña 2008; Lopes et al. 2008). However, due to the difficulties of designing efficient MCMC samplers with state space constrained to $\mathcal{C}_{m,k}$, the autoregressive coefficient matrices

are often assumed to be diagonal, which simplifies the stationarity condition to that of a univariate autoregression. Fortunately, recent work by Heaps (2023) presents a prior for the autoregressive coefficient matrices that is constrained to $\mathcal{C}_{m,k}$ and facilitates routine computational inference. This is based on an unconstrained reparameterisation of the model through two bijective transformations. First, the model parameters $\{(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_m), \boldsymbol{\Pi}\}$ undergo a recursive mapping which yields a new parameter set $\{(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_m), \boldsymbol{\Pi}\}$. Here $\boldsymbol{P}_{i+1}$ is the $(i+1)$th partial autocorrelation matrix, that is, a standardised version of the conditional cross-covariance between $\boldsymbol{\eta}_{t+1}$ and $\boldsymbol{\eta}_{t-i}$ given the i intervening variables $(\boldsymbol{\eta}_t, \ldots, \boldsymbol{\eta}_{t-i+1})$. Each partial autocorrelation matrix lies in the space of $k \times k$ square matrices with singular values less than one. By mapping the singular values of $\boldsymbol{P}_i$ from $[0, 1)$ to the positive real line, a second mapping then constructs an unconstrained $k \times k$ square matrix $\boldsymbol{A}_i$ through $\boldsymbol{A}_i = (\boldsymbol{I}_k - \boldsymbol{P}_i \boldsymbol{P}_i^{\mathrm{T}})^{-1/2} \boldsymbol{P}_i$, $i = 1, \ldots, m$, in which $\boldsymbol{X}^{-1/2}$ denotes the inverse of the symmetric matrix square root of $\boldsymbol{X}$. Specification of a prior for the $\boldsymbol{A}_i$, discussed in Sect. 4.4, and computational inference is now routine owing to the Euclidean geometry of the parameter space.

The inverse mapping from the intermediate reparameterisation $\{(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_m), \boldsymbol{\Pi}\}$ to the original parameters $\{(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_m), \boldsymbol{\Pi}\}$ involves two recursions. The first outputs the stationary covariance matrix $\boldsymbol{G}_0$ from $\{(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_m), \boldsymbol{\Pi}\}$ and the second outputs $(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_m)$ (and recovers $\boldsymbol{\Pi}$) from $(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_m)$ and $\boldsymbol{G}_0$. It is therefore trivial to modify the reparameterisation to accommodate the constraint that $\boldsymbol{G}_0 = \boldsymbol{I}_k$; one simply omits the first recursion, and calculates $\{(\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_m), \boldsymbol{\Pi}\}$ from $(\boldsymbol{P}_1, \ldots, \boldsymbol{P}_m)$ and $\boldsymbol{G}_0 = \boldsymbol{I}_k$ in the second recursion. We note that the autocovariance matrices $\boldsymbol{G}_1, \ldots, \boldsymbol{G}_{m-1}$ needed to characterise the initial distribution of $\boldsymbol{\eta}_{(1-m):0}$ are also by-products of this second recursion.

When $\boldsymbol{\eta}_t$ follows a VAR(1) process, there is a closed form expression for the original model parameters in terms of $\boldsymbol{P}_1$ or, equivalently, $\boldsymbol{A}_1$. Dropping the 1-subscript for brevity, we can write $\boldsymbol{\Gamma} = \boldsymbol{P} = (\boldsymbol{I}_k + \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1/2} \boldsymbol{A}$ and $\boldsymbol{\Pi} = \boldsymbol{I}_k - \boldsymbol{P}\boldsymbol{P}^{\mathrm{T}} = (\boldsymbol{I}_k + \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{-1}$.

4.4 Prior for the transformed partial autocorrelation matrices

The stationary dynamic factor model described in the previous sections is invariant under rotations of the factors when compensatory transformations are applied to the factor loadings and parameters of the state equation. In particular, marginalising over the common factors, it can readily be shown that the likelihood evaluated at $(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \boldsymbol{A}_1, \ldots, \boldsymbol{A}_m)$ is the same as the likelihood evaluated at $(\boldsymbol{\mu}, \boldsymbol{\Lambda}\boldsymbol{Q}^{\mathrm{T}}, \boldsymbol{\Sigma}, \boldsymbol{Q}\boldsymbol{A}_1\boldsymbol{Q}^{\mathrm{T}}, \ldots, \boldsymbol{Q}\boldsymbol{A}_m\boldsymbol{Q}^{\mathrm{T}})$ for any orthogonal matrix $\boldsymbol{Q} \in \mathcal{O}(k)$. In the absence of any meaningful information to distinguish $\boldsymbol{A}_i$ from $\boldsymbol{Q}\boldsymbol{A}_i\boldsymbol{Q}^{\mathrm{T}}$ a priori,

we assign a prior which is rotatable, that is $\boldsymbol{A}_i$ and $\boldsymbol{Q}\boldsymbol{A}_i\boldsymbol{Q}^{\mathrm{T}}$ have the same distribution for any orthogonal matrix $\boldsymbol{Q}$. To this end, we take the $\boldsymbol{A}_i$ to be independent with $\boldsymbol{A}_i \sim \mathrm{N}_{k,k}(\boldsymbol{0}, \boldsymbol{I}_k, \boldsymbol{I}_k)$, $i = 1, \ldots, m$.

5 Posterior computation

5.1 Samplers with fixed truncation level

Consider first the static factor model. For computational inference, we use MCMC to sample from the posterior associated with our unconstrained parameterisation of the model. When the truncation level in the factor loadings matrix is fixed at H, we propose a straightforward Gibbs sampler. The full conditional distributions for most unknowns have standard forms and can be sampled directly. However, depending on the structure chosen for the among-row scale matrix $\boldsymbol{\Phi} = \boldsymbol{R}(\boldsymbol{\vartheta})$, a semi-conjugate prior for the correlation parameter(s) in $\boldsymbol{\vartheta}$ is not generally available and so we update $\boldsymbol{\vartheta}$ in a Metropolis-Hastings step. Computational inference is slightly complicated if a matrix-t, rather than matrix normal, prior is adopted for $\boldsymbol{\Lambda}$ because it is not conjugate to the likelihood. However, sampling becomes straightforward if we make use of the representation of a matrix-t random variable in (7) and augment the state space of the sampler with the matrix $\boldsymbol{S}$. The full conditional distribution for $\boldsymbol{\Lambda}$ is then matrix-normal while the full conditional distribution for $\boldsymbol{S}$ is Wishart. The only other additional unknown is the degree of freedom parameter ς. If ς is assigned the prior from Sect. 3.3, its full conditional distribution is non-standard. Rather than updating the transformed degree of freedom parameter $\check{\varsigma}$ through its own Metropolis-Hastings step, mixing can be greatly improved by performing a joint update of $\boldsymbol{S}$ and $\check{\varsigma}$. The proposal density takes the form $q(\boldsymbol{S}^*, \check{\varsigma}^* | \boldsymbol{S}, \check{\varsigma}, \ldots) = q_1(\check{\varsigma}^* | \check{\varsigma}) q_2(\boldsymbol{S}^* | \check{\varsigma}^*, \ldots)$ in which $q_1(\cdot)$ is a random walk proposal on the log-scale and $q_2(\cdot)$ is the full conditional density for $\boldsymbol{S}$.

Extension of the Gibbs sampler to handle the dynamic factor model is straightforward. The latent factors $\boldsymbol{\eta}_{1-m}, \ldots, \boldsymbol{\eta}_n$ can be sampled efficiently from their full conditional distribution in a single block using a variety of sampling approaches (Wilkinson and Yeung 2002, 2004). In the example in Sect. 6.3 we used the forward-filtering backward-sampling algorithm (e.g. Frühwirth-Schnatter 1994) but other methods that rely on banded or sparse matrix algorithms will be more efficient in higher-dimensional applications; see, for example, Chan and Jeliazkov (2009). The full conditional distributions for the transformed partial autocorrelation matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_m$ are non-standard. Given a truncation level of H, each $\boldsymbol{A}_i$ contains H^2 parameters and so it is not generally feasible to update the whole matrix in a single Metropolis-Hastings step. We therefore propose dividing the elements of

each $\boldsymbol{A}_i$ into blocks of length b and updating the blocks one-at-a-time. In the application in Sect. 6.3, we found mixing to be greatly improved by using Metropolis-adjusted Langevin (MALA) steps rather than Gaussian random walks.

Given a large enough fixed truncation level $H \leq \lceil \varphi(p) \rceil - 1$, the MGP prior increasingly shrinks the factor loadings in columns with higher indices if the data suggest their likelihood contribution is negligible. At least conceptually, these columns can be discarded. If there are $m^{[i]}$ discardable columns on iteration i of the MCMC sampler, the effective number of factors is defined as $k^{*[i]} = H - m^{[i]}$. The posterior for the effective number of factors can then be used as a proxy for the posterior for k. There are various truncation criteria for deciding which columns can be omitted. Bhattacharya and Dunson (2011) consider as inactive any factor whose loadings are all within ϵ of zero for some chosen threshold, say $\epsilon = 10^{-4}$. Schiavon and Canale (2020) suggest a more interpretable criterion which, unlike the choice of threshold ϵ, is unaffected by the scale of the data and its dimension. Denote by $\boldsymbol{\Lambda}_{k^*}$ the factor loadings matrix obtained by discarding the columns of $\boldsymbol{\Lambda}$ from $k^* + 1$ onwards. The basic idea is that if $\boldsymbol{\Omega}_{k^*} = \boldsymbol{\Lambda}_{k^*} \boldsymbol{\Lambda}_{k^*}^{\mathrm{T}} + \boldsymbol{\Sigma}$ can explain $100 T \%$ of the total variability of the data for some $T \in (0, 1)$, then one decides there are k^* active factors. The choice of proportion T then replaces the choice of threshold ϵ. This is the truncation criterion we adopt in the applications in Sect. 6.

When the truncation level H is fixed, an alternative to Gibbs sampling is to implement off-the-shelf computational inference using standard probabilistic programming software. Code for the applications in Sect. 6, written in Stan (Carpenter et al. 2017), is given in the Supplementary Materials. Stan uses Hamiltonian Monte Carlo to perform joint updates of all unknowns and so tends to converge faster and mix better than a Gibbs sampler. The caveat is the necessity to fix H; although this can be addressed in a bespoke implementation of Gibbs sampling (see Sect. 5.2 below), Stan cannot be customised in this way.

5.2 Adaptive Gibbs sampler

Especially in applications where p is large, the number of effective factors is often considerably less than the Ledermann bound $\varphi(p)$. Therefore fixing a truncation level H which is less than, but in the vicinity of, $\varphi(p)$ can be computationally inefficient. In the literature on infinite factor models, a common pragmatic solution is to use an adaptive Gibbs sampler which tunes the truncation level H as it proceeds. Adaptation occurs at iteration i with probability $p(i) = \exp(\alpha_0 + \alpha_1 i)$ where $\alpha_0 \leq 0$ and $\alpha_1 < 0$ such that the probability of adaptation decreases over the course of the simulation. This is necessary to satisfy the diminishing adaptation condition of Roberts and Rosenthal (2007).

During an adaptation step, $k^{*[i]}$ is compared to the current value of H. The basic idea is to delete any inactive factors or to add an extra factor if all H factors are active. In the static model, implementation is straightforward. If $k^{*[i]} < H$, H is reduced to $k^{*[i]}$ and any inactive factors are deleted along with the corresponding columns of $\boldsymbol{\Lambda}$ and components of $\boldsymbol{\varrho}$ and $\boldsymbol{\Psi}$. If $k^{*[i]} = H$ and $H < \lceil \varphi(p) \rceil - 1$, H is increased by 1 then an extra factor, column of factor loadings, and component of $\boldsymbol{\varrho}$ are sampled from their priors. In the dynamic model, the adaptation step is slightly more involved because of the additional complexity in the distribution of the latent factors. Full details of the adaptive Gibbs samplers for both static and dynamic models are given in the Supplementary Materials, along with R code for the applications in Sect. 6. We also provide details on how the posterior output can be post-processed to obtain samples under a parameterisation in which the factor loadings matrix is identified by the PLT constraint. This is incredibly valuable for parameter interpretation and the assessment of MCMC diagnostics.

6 Applications

6.1 Simulation experiment

6.1.1 Simulation settings

In order to investigate posterior sensitivity to the prior specification in the idealised setting in which we know the data were generated from a factor model with known shared variation matrix $\boldsymbol{\Delta}$, we carried out a series of simulation experiments. In the simplified setting in which $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_p$, denote by β the proportion of the total variation in $\boldsymbol{\Omega} = \boldsymbol{\Delta} + \boldsymbol{\Sigma}$ that is explained by the common factors, that is, $\beta = \mathrm{tr}(\boldsymbol{\Delta})/\mathrm{tr}(\boldsymbol{\Omega})$. Then for given β and $\boldsymbol{\Delta}$, the corresponding value of σ^2 can be computed as $\sigma^2 = (1 - \beta)\mathrm{tr}(\boldsymbol{\Delta})/(k\beta)$. Three different combinations of p and k, representing different degrees of dimension reduction, were considered: $(p = 24, k = 6)$, $(p = 48, k = 9)$ and $(p = 72, k = 9)$ so that k/p was equal to 25, 18.75 and 12.5%, respectively. For each combination of p and k, we also considered three values for β, $\beta \in \{0.9, 0.95, 0.99\}$, giving nine sets of values for $\{(p, k), \beta\}$ in total. We refer to these as *simulation settings*. Note that we only set $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_p$ for the purposes of simulating the data; for inference, we still allow $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$.

Under each simulation setting, we simulated 24 data sets of size $n = 50$ based on an exponential correlation matrix for $\boldsymbol{\Phi}$, taking $\boldsymbol{\Delta}$ to be the closest rank-k matrix to $\boldsymbol{\Phi}$ (see Proposition 1). In the simulation of each data set, in order to fix the value of $\boldsymbol{\Delta}$, we simulated p pairs of coordinates $\boldsymbol{x}_j = (x_{j1}, x_{j2})^{\mathrm{T}}$, $j = 1, \ldots, p$, by sampling uniformly at random from the unit square, and then set the expected shared variation matrix equal to $\boldsymbol{\Phi}$ with (j, k) element $\phi_{jk} =$

$\exp(-\|\boldsymbol{x}_j - \boldsymbol{x}_k\|_2/\vartheta)$ where the length-scale parameter was equal to

$$\vartheta = \begin{cases} -1/\log(0.05), & \text{for data sets } 1, \dots, 6; \\ -1/\log(0.10), & \text{for data sets } 7, \dots, 12; \\ -1/\log(0.15), & \text{for data sets } 13, \dots, 18; \\ -1/\log(0.20), & \text{for data sets } 19, \dots, 24. \end{cases}$$

For a distance of 1 in the unit square, the correlations therefore ranged from 0.05 to 0.20.

6.1.2 Prior distributions

Each data set generated under each simulation setting was used to update five prior distributions for $\boldsymbol{\Lambda}$. Three are structured matrix normal priors, with different assumptions about $\boldsymbol{\Phi}$, and two are widely used priors in factor analysis which allow inference on the number of factors:

SN (expcov.)   A structured matrix normal prior in which $\boldsymbol{\Phi}$ was taken to have the same form as that used to simulate the data, that is, an exponential correlation matrix. The length-scale parameter was taken to be unknown and assigned the distribution $\log(\vartheta) \sim \mathrm{N}(0, 1)$.

SN (fixed)   A structured matrix normal prior in which $\boldsymbol{\Phi}$ was taken to have the same form as that used to simulate the data, that is, an exponential correlation matrix. The length-scale parameter was fixed at the value used to simulate the data.

SN (exch.)   A structured matrix normal prior in which $\boldsymbol{\Phi}$ was taken to have a different form to that used to simulate the data, specifically we took $\boldsymbol{\Phi} = (1 - \vartheta)\boldsymbol{I}_p + \vartheta\boldsymbol{J}_p$, $-1/(p - 1) < \vartheta < 1$, representing a two-parameter exchangeable matrix with trace fixed at p. The correlation parameter was taken to be unknown and assigned the distribution $\mathrm{logit}(\check{\vartheta}) \sim \mathrm{N}(0, 1)$ where $\check{\vartheta} = \{1 + (p - 1)\vartheta\}/p$.

MGP   The original (exchangeable) multiplicative gamma process prior described in Bhattacharya and Dunson (2011), which is a global–local prior. We took $a_1 = 1$ and $a_2 = 2$ in the multiplicative gamma process for the global precision parameters and $\nu = 3$ in the distribution for the local precision parameters.

CUSP   A cumulative shrinkage process prior (Legramanti et al. 2020), which is a slab-and-spike prior. We took $a_\theta = b_\theta = 2$ in the distribu-

tion for the slab, $\theta_\infty = 0.05$ for the position of the spike and $\alpha = 5$ in the cumulative stick-breaking construction controlling the probability assigned to the spike.

For the three structured matrix normal priors, we chose $a_1 = 1$ and $a_2 = 2$ in the multiplicative gamma process for the diagonal elements of $\boldsymbol{\Psi}$. In all cases, the idiosyncratic variances were assigned the prior $1/\sigma_j^2 \sim \gamma(1, 0.3)$ independently for $j = 1, \dots, p$.

6.1.3 Analysis and results

All analyses were run using an adaptive Gibbs sampler, allowing adaptation after 500 iterations and setting the parameters in the diminishing adaptation condition at $\alpha_0 = -1$ and $\alpha_1 = -5 \times 10^{-4}$. The chains were run for 10K iterations, following a burn-in of 10K, and thinned to retain every fifth sample to reduce computational overheads. For the three structured matrix normal priors and the original MGP prior, we applied the truncation criterion of Schiavon and Canale (2020), retaining as active enough factors to explain at least $T = 0.95$ of the total variation in the data.

The (squared) geodesic distance between two symmetric positive definite matrices $\boldsymbol{A} \in \mathcal{S}_p^+$ and $\boldsymbol{B} \in \mathcal{S}_p^+$ is defined as

$$d_G(\boldsymbol{A}, \boldsymbol{B})^2 = \left[\sum_{i=1}^{p} \log\left\{ \lambda_i\left(\boldsymbol{A}^{-1}\boldsymbol{B}\right)\right\}^2 \right]$$

in which $\lambda_i(\boldsymbol{X})$ denotes the ith eigenvalue of $\boldsymbol{X}$ (Lim and Sepulchre 2019). For simulation experiments in which $\beta = 0.99$, Table 1 reports the median and interquartile range across the 24 data sets of the posterior mean of $d_G(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$ where $\boldsymbol{\Omega}_0$ denotes the value used to simulate the data. Corresponding Tables for the simulations where $\beta = 0.95$ and $\beta = 0.9$ are presented in the Supplementary Materials. We chose the distance metric $d_G(\cdot, \cdot)$ because, by construction, $\boldsymbol{\Omega}$ is close to singular and so the geodesic distance will be better suited than a metric, such as the Frobenius distance, which ignores the geometry of the space. In particular, it will give large values whenever $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ do not approximately span the same subspace of $\mathbb{R}^p$, that is when their null spaces are not approximately the same. We also report the median and interquartile range for the posterior mean of the effective number of factors k^*.

Focusing first on the posterior mean of the geodesic distance $d_G(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$, it is immediately clear from Table 1 and the corresponding tables in the Supplementary Materials that both the medians and interquartile ranges across data sets are typically smallest when we use a prior in which $\boldsymbol{\Phi}$ is based on an exponential correlation matrix. As we would expect, this is especially true when the length-scale parameter in the

Table 1 Summaries of the simulation experiment when $\beta = 0.99$

Prior	(p, k)	$E\{d_G(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0)|\boldsymbol{y}\}$		$E(k^*|\boldsymbol{y})$	
		Median	IQR	Median	IQR
SN (expcov.)	(24, 6)	15.88	0.38	6.34	0.56
SN (fixed)	(24, 6)	15.78	0.39	6.30	0.89
SN (exch.)	(24, 6)	15.93	0.39	6.63	0.69
MGP	(24, 6)	16.02	0.45	6.27	0.86
CUSP	(24, 6)	16.25	0.36	3.13	0.57
SN (expcov.)	(48, 9)	21.25	0.58	7.54	1.95
SN (fixed)	(48, 9)	21.18	0.65	7.32	1.47
SN (exch.)	(48, 9)	22.20	1.04	5.96	2.02
MGP	(48, 9)	22.15	0.64	6.41	2.32
CUSP	(48, 9)	22.86	0.84	3.00	0.23
SN (expcov.)	(72, 9)	23.89	1.03	7.76	1.38
SN (fixed)	(72, 9)	23.69	1.37	7.85	1.55
SN (exch.)	(72, 9)	25.81	1.87	6.15	0.92
MGP	(72, 9)	24.99	0.94	7.21	1.81
CUSP	(72, 9)	27.89	1.07	3.00	0.00

prior is fixed at the value used in simulating the data. Nevertheless, the modest differences between posterior inferences under SN (expcov.) and SN (fixed) are reassuring; in the analysis of a real data set, though we may have strong prior beliefs about a suitable structure for $E(\boldsymbol{\Delta})$, we are likely to be more uncertain about the values of the hyperparameters in that structure. We also observe that the gains from assuming the correct form for $\boldsymbol{\Phi}$ in this simulation experiment are greatest when the total variation in $\boldsymbol{\Delta}$ makes up a larger proportion of the total variation in $\boldsymbol{\Omega}$ (that is, when $\beta = 0.99$ and $\beta = 0.95$) and when the degree of dimension reduction is larger. This is as expected because in both cases, the prior for $\boldsymbol{\Lambda}$, and therefore $\boldsymbol{\Delta}$, is likely to impart more influence on the posterior for $\boldsymbol{\Omega}$. It is interesting to observe that when we adopt a structured matrix normal prior but assume an incorrect structure for $\boldsymbol{\Phi}$, the posterior mean for $d_G(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0)$ still tends to be smaller than under the MGP or CUSP priors. This is likely to be because we simulated all data sets based on exponential correlation matrices for $\boldsymbol{\Phi}$ which have strictly positive diagonal elements. Under SN (exch.), we are shrinking $\boldsymbol{\Delta}$ towards a rank-reduced approximation to a two-parameter exchangeable matrix with a common (potentially positive) off-diagonal element whereas the prior expectation for $\boldsymbol{\Delta}$ under the MGP and CUSP priors is diagonal.

A discussion of the results concerning the posterior mean of the effective number of factors k^* can be found in the Supplementary Materials.

6.2 Co-occurrence of Finnish birds

6.2.1 Model and prior

As discussed in Sect. 1, most of the literature on prior specification for factor models focuses on constructing a distribution for the factor loadings matrix that is exchangeable with respect to the order of the components in the observation vector. An exception is the *structured increasing shrinkage (SIS) process* (Schiavon et al. 2022). This prior allows meta-covariates to inform the within-column sparsity structure of the factor loadings matrix in an infinite factorisation model. In this section, we re-examine the Finnish birds data set analysed in Schiavon et al. (2022) to illustrate the extra flexibility afforded by our approach.

The Finnish bird data comprise information on the co-occurrence of the 50 most common bird species in Finland from 137 sampling areas in 2014 (Lindström et al. 2015). The occurrences are arranged into a $n \times p$ binary matrix $\boldsymbol{Y} = (y_{ij})$ where $y_{ij} = 1$ if bird species j was observed in area i and $y_{ij} = 0$ otherwise. Information available to explain the variation in the mean $\boldsymbol{\mu}_i$ at each sampling area include a $n \times c$ matrix of environmental covariates $\boldsymbol{W}$ containing measurements on spring temperature and its square and a five-level habitat type (broadleaved forests, coniferous forests, open habitats, urban habitats, wetlands) so that $c = 7$. Information available to structure the prior for the marginal variance $\boldsymbol{\Omega}$ include a phylogenetic tree, indicating the evolutionary relationships amongst the 50 species, and a $p \times q$ matrix of meta-covariates $\boldsymbol{X}$ consisting of two species traits: logarithm of typical body mass and a three-level migratory strategy (short-distance migrant, resident species, long-distance migrant). Following Schiavon et al. (2022), we model species presence or absence using a multivariate probit regression model where $y_{ij} = \mathbb{I}(z_{ij} > 0)$ and we assume the latent variables $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})^{\mathrm{T}}$ are such that $\boldsymbol{z}_i = \boldsymbol{B}^{\mathrm{T}} \boldsymbol{w}_i + \boldsymbol{\Lambda} \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i$ for $i = 1, \ldots, n$. Here $\boldsymbol{w}_i^{\mathrm{T}}$ is the ith row of $\boldsymbol{W}$, $\boldsymbol{B} = (\beta_{ij})$ is a $c \times p$ matrix of regression coefficients, and the idiosyncratic variances on the diagonal of $\mathrm{Var}(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma}$ are set equal to 1 to prevent compensatory rescaling of $\boldsymbol{B}$ and $\boldsymbol{\Omega} = \boldsymbol{\Lambda} \boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Sigma}$. Our prior for the coefficients in $\boldsymbol{B}$ is identical to that of Schiavon et al. (2022) and described in the Supplementary Materials.

The SIS process prior for $\boldsymbol{\Lambda}$ is built on a series of assumptions of conditional independence. As a consequence, the expectation of the shared variation matrix $\boldsymbol{\Delta}$ is diagonal, irrespective of the priors assigned to its hyperparameters. Moreover, every factor loading λ_{ij} has a spike-and-slab distribution in which the meta-covariates only enter through the logit of the spike probability. As such, there is no mechanism for incorporating the rich information on evolutionary relationships that is expressed through the phylogenetic tree.

In contrast, using the approach outlined in Sect. 3.4 we can structure the prior for the shared variation so that it incorporates the information from the phylogenetic tree in addition to the two meta-covariates. The former can be encoded as a matrix of phylogenetic distances between the 50 species, whose (i, j)th entry, say $d_{P,ij}$, is the sum of the branch lengths along the path between the leaves for species i and j. This distance function is a metric over the set of leaves (Steel 2016, Chapter 6). Another measure of the dissimilarity between species can be obtained using the generalised distance function from the projection model in (8). As we only have binary response data, which is unlikely to be as informative as continuous data, we simplify (8) by taking the matrix Θ to be diagonal. This yields $d_{C,ij} = \left\{ \sum_{k=1}^{4} (x_{ik} - x_{jk})^2 / \vartheta_k^2 \right\}^{1/2}$ as a metric for the distance between the meta-covariates for species i and j, where x_{i1}, x_{i2} and x_{i3} are indicators for whether species i is a short-distance migrant, resident or long-distance migrant, respectively, while x_{i4} is the logarithm of typical body mass for species i. Since $d_{P,ij}$ and $d_{C,ij}$ are metrics over a set comprising the species and their associated traits, the sum is also a metric. We can therefore take as a generalised measure of distance between species i and j, $d_{ij} = d_{C,ij} + d_{P,ij}/\vartheta_5$, in which $\vartheta_i > 0$ for $i = 1, \ldots, 5$. So that all the length-scale parameters ϑ_i operate on a comparable scale, the meta-covariates and branch lengths are standardised prior to calculating $d_{C,ij}$ and $d_{P,ij}$. Finally, we relate the generalised distance d_{ij} to the among-row scale matrix Φ by using the exponential correlation function, $\phi_{ij} = \exp(-d_{ij})$. We assume the length-scale parameters to be independent in the prior and assign distributions $\log \vartheta_i \sim N(0, 10)$.

Based on available phylogenetic and ecological information, this prior encourages shrinkage towards a rank-reduced approximation of an expected shared variation matrix $E(\Delta)$ with higher correlations between types of bird that are more closely related in terms of their evolutionary history and species traits. Since we allow the length-scale parameters ϑ_i to be unknown we are also able to learn which variables are more important in explaining relationships between species. However, the structure for the expected shared variation matrix represented through Φ is conjectural and so we elect to use the matrix-t prior whose extra degree of freedom parameter ς allows the data to influence the degree of shrinkage. As discussed in Sect. 3.3, we induce a prior for $\varsigma > 4$ by specifying a distribution for $\check{\varsigma} = 1/(\varsigma - 4) \in \mathbb{R}^+$. Specifically, we take $\check{\varsigma} \sim \text{Exp}(1)$ which implies that when the number of factors k is 5, 10 and 15, the probability of the scale factor $s_k(\check{\varsigma})$ lying between 1 and x is 0.75 for $x = 7.8$, $x = 12.9$ and $x = 18.0$, respectively. For the diagonal elements in Ψ, we assign a MGP prior (9) with $a_1 = 2$, $a_2 = 6$ and a maximum truncation point H that satisfies $H \leq \lceil \varphi(50) \rceil - 1 = 40$.

6.2.2 Analysis and results

We ran two chains, initialised at different starting points, using the adaptive Gibbs sampler discussed in Sect. 5.2 and another two chains using Stan's implementation of the (non-adaptive) Hamiltonian Monte Carlo sampler with a fixed truncation level of $H = 10$. In the Gibbs sampler, adaptation was allowed after 500 iterations and, following Schiavon et al. (2022), the parameters in the diminishing adaptation condition were set at $\alpha_0 = -1$ and $\alpha_1 = -5 \times 10^{-4}$. After a burn-in of 20K iterations, each chain was run for a further 20K iterations, thinning to store every fifth draw in order to reduce computational overheads. Using the truncation criterion of Schiavon and Canale (2020), we retain as active enough factors to explain at least $T = 0.999$ of the total variation in the data. The posterior mode (and median) for the number of effective factors in each case was 5 with 95% credible interval $(3, 6)$. Comparing the output of identified parameters across the chains from both samplers using the usual graphical diagnostic checks gave no evidence of any lack of convergence.

To gauge performance against another non-exchangeable prior, the model was also fitted under the SIS process prior using an adaptive Gibbs sampler, with the hyperparameters in the prior and tuning parameters in the sampler set in line with the specification from Section 5 of Schiavon et al. (2022). In this case, more or less all the posterior mass for the number of effective factors was stacked at 4.

The marginal prior and posterior densities for the logarithms of the length-scale parameters $\vartheta_1, \ldots, \vartheta_5$ and the transformed degree of freedom parameter $\check{\varsigma} = 1/(\varsigma - 4)$ under the structured matrix-t prior are displayed in Fig. 1. Absence of a relationship between a component of the generalised distance metric and the expected shared variation between species would be indicated by an infinite value for the corresponding length-scale parameter. Therefore the concentration of all posterior densities in Fig. 1a over a finite interval near the origin highlights the role played by the generalised distance metric in explaining the structure of the shared variation matrix. It is also clear that the phylogenetic distances, whose length-scale parameter is ϑ_5, have the greatest influence. For the transformed degree of freedom parameter $\check{\varsigma}$ in Fig. 1b, the similarity between the prior and posterior indicate that the information learned from the data is modest. Nevertheless, the shift in the position of the mode from 0 in the prior to around 0.1 in the posterior suggests that the data support less shrinkage towards the expected shared variation matrix than would be implied by a matrix normal prior for which $\check{\varsigma} = 0$.

Further qualitative evidence of the importance of the phylogenetic information on the structure of the marginal covariance matrix $\Omega = (\omega_{ij})$ is demonstrated in Fig. 2 which shows the mean of the posterior for the marginal correlation

Fig. 1 Marginal posterior density for **a** the logarithms of the length scale parameters ϑ_1 (—), ϑ_2 (—), ϑ_3 (—), ϑ_4 (—), ϑ_5 (—) and **b** the transformed degree of freedom parameter ξ. Also shown are the priors (---). (color figure online)

(a)   (b)
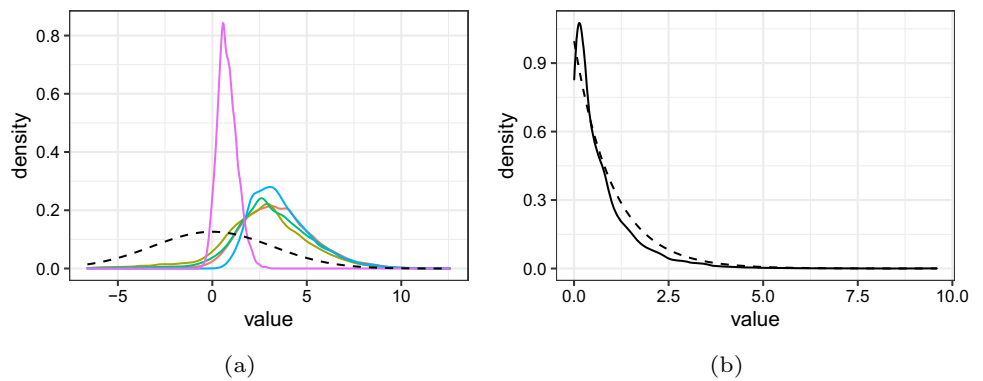
matrix $S_\Omega^{-1}\Omega S_\Omega^{-1}$, where $S_\Omega = \mathrm{diag}(\sqrt{\omega_{11}}, \ldots, \sqrt{\omega_{pp}})$. The order of the 50 bird species is based on their phylogenetic tree, which is shown in the margins of the plot. There is clear structure in the matrix with several groups of species that share a recent common ancestor displaying strong positive correlations. This includes the species in the first ten rows, which correspond to birds from the Corvoidea superfamily and the Sylviidae family; the species in rows 19 to 30, which correspond largely to birds from the Musicapoidea superfamily; and the species in rows 42 to 47, which correspond mostly to birds from the Scolopacidae superfamily. It is also clear that most of the rows and columns predominated by near-zero correlations correspond to basal species, like *Columba palumbus* and *Grus grus*. Nevertheless, the structure of the matrix is clearly not dictated by evolutionary proximity alone. Interestingly, the corresponding plot from the analysis under the SIS process prior appears to be very sparse; see the Supplementary Materials. At least in part, this can be attributed to shrinkage of Δ towards its prior mean, which is a diagonal matrix.

In order to provide a more formal comparison between the two model-prior combinations, we consider measures of goodness-of-fit and predictive performance. In the former case, we follow Schiavon et al. (2022) by computing the pseudo marginal likelihood (PML), defined as the product of conditional predictive ordinates (Gelfand and Dey 1994). In terms of predictive performance, we consider the Brier score, which is a widely used proper scoring rule for categorical variables, calculated using a 4-fold cross-validation approach (e.g. Gneiting and Raftery 2007). In our case, the score is positively oriented so that large values indicate better performance. The results are shown in Table 2 from which we can conclude that the combination of the factor model and structured matrix-*t* prior gives the best-fit to the data and, more markedly, the best predictive performance. Further details on the calculation of the PML and Brier score can be found in the Supplementary Materials.

6.3 Hourly demand for natural gas

6.3.1 Model and prior

To illustrate the use of our methodology in the context of a dynamic model, we consider an application to modelling the hourly demand for natural gas. In the UK, gas in the national transmission system is transported to individual customers through eight regional distribution networks which are responsible for maintaining a reliable gas supply to customers at all times. This requires accurate short-term forecasts of the hourly demand for gas at individual points in the network called offtakes, where gas is taken off to supply locally. We consider data from a single offtake site in a regional distribution network in the North of England. Hourly gas demand data are available for the period from May 2012 to June 2015 along with the average daily temperature at the offtake site. In addition to the weather, the other main factors that are known to influence the mean level of residential demand for gas are seasonal and calendar effects (Soldo 2012). We therefore additionally incorporate covariates that represent the day of the year, day of the week and whether or not a day is a public holiday.

The measurement of the demand for gas at an hourly resolution makes construction of a model challenging because the underlying dynamics of the process are likely to vary over a much longer time scale. For example, consider a dynamic linear model for hourly gas demand that allows for calendar and weather effects and incorporates a time-varying local level. Suppose that this local level is modelled as evolving according to a random walk or autoregression. Deviations from behaviour that would be considered typical given the calendar and weather effects would likely persist over days, rather than hours, and so hourly evolution of the local level would demand a near-zero innovation variance. Similarly, if calendar effects such as the day of the week, were regarded as dynamic, hourly evolution would demand vanishingly small innovation variances.
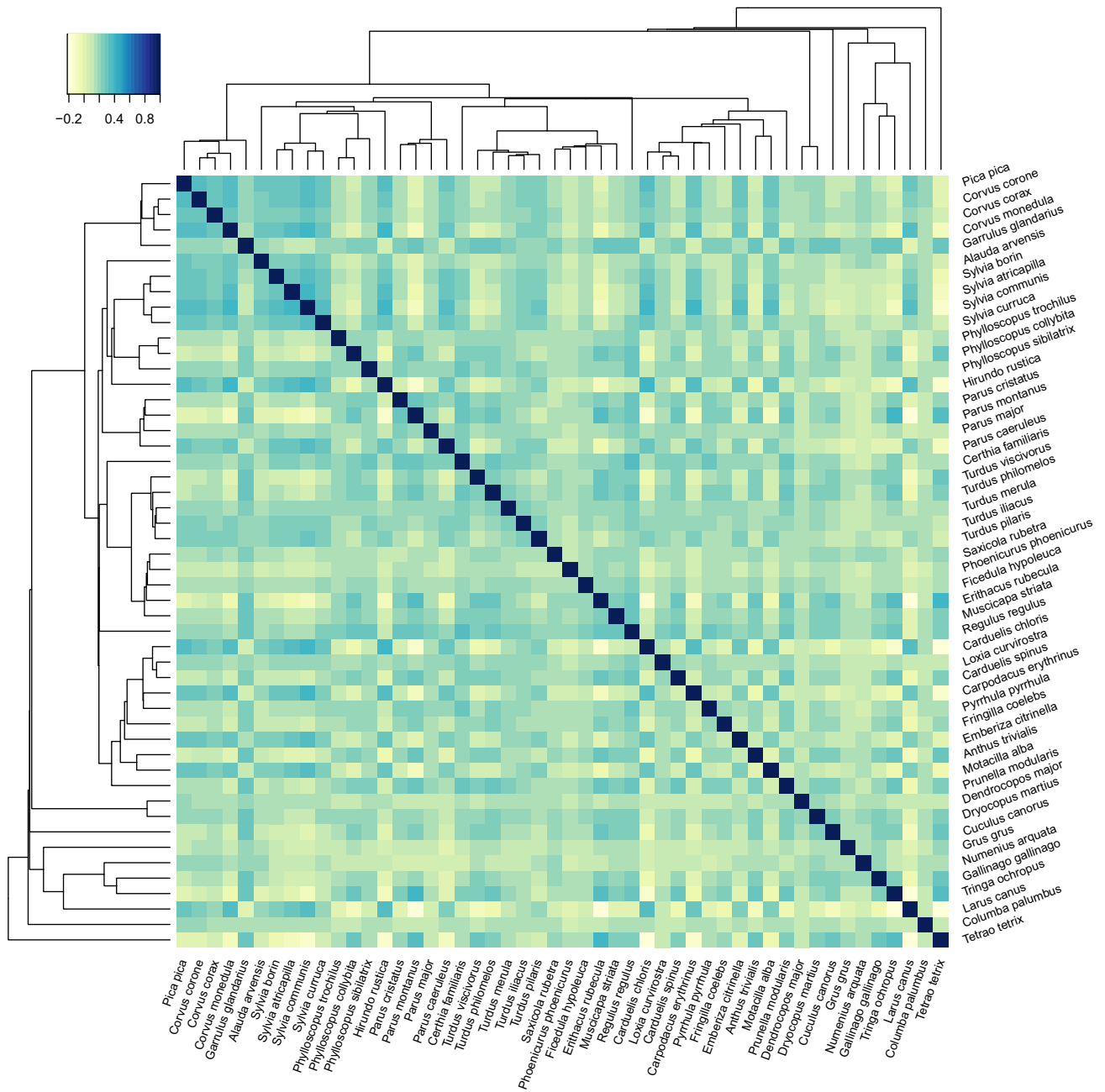
Fig. 2 Mean of the posterior for the marginal correlation matrix under the structured matrix-t prior. The species are ordered according to their phylogenetic tree which is visualised in the margins

Table 2 Goodness-of-fit and predictive performance of the two model-prior combinations

	Structured matrix-t	SIS process
Log PML	**−2831.6**	−2936.5
Brier score	**−0.29120**	−0.85396

Bold indicates the score of the model-prior combination that performs better against that metric. Both metrics are positively oriented so that the larger values indicate better performance

Motivated by these concerns, we define a *gas-day-vector* as a longitudinal series of measurements of the log-transformed demand for gas over 24 h, starting at 07:00, which is the beginning of a gas-day. Denote by $\boldsymbol{y}_t = (y_{t1}, \ldots, y_{t,24})^{\mathrm{T}}$ the gas-day-vector on day t so that $y_{t,h}$ denotes the log-transformed demand for gas at hour h of gas-day t. We elect to model the gas-day-vectors using a dynamic factor model with observation equation $\boldsymbol{y}_t = \boldsymbol{\mu}_t + \boldsymbol{\Lambda}\boldsymbol{\eta}_t + \boldsymbol{\epsilon}_t$, where $\boldsymbol{\epsilon}_t \sim \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, $t = 1, \ldots, n$. For simplicity in this illustrative application, we model evolution of the factors

using a first order vector autoregression, $\boldsymbol{\eta}_t = \boldsymbol{\Gamma}\boldsymbol{\eta}_{t-1} + \boldsymbol{\zeta}_t$, where $\boldsymbol{\zeta}_t \sim N_k(\mathbf{0}, \boldsymbol{\Pi})$. Due to the stability of the demand for gas over the time period in question, it is reasonable to assume that the departures from the time-varying mean $\boldsymbol{y}_t - \boldsymbol{\mu}_t$ follow a stationary process. We therefore constrain the stationary variance of the $\boldsymbol{\eta}_t$ to be $\boldsymbol{I}_k$ and take the initial distribution to be $\boldsymbol{\eta}_0 \sim N_k(\mathbf{0}, \boldsymbol{I}_k)$. Reparameterising the model for the factors in terms of the transformed partial autocorrelation matrices yields a single unconstrained square matrix $\boldsymbol{A} = (\alpha_{ij})$ with $\boldsymbol{\Gamma} = (\boldsymbol{I}_k + \boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})^{-1/2}\boldsymbol{A}$ and $\boldsymbol{\Pi} = (\boldsymbol{I}_k + \boldsymbol{A}\boldsymbol{A}^{\mathsf{T}})^{-1}$. As discussed in Sect. 4.4, we assign a rotatable prior to $\boldsymbol{A} = (\alpha_{ij})$, and therefore $\boldsymbol{\Gamma}$, by taking $\alpha_{ij} \sim N(0, 1)$.

The variances of the specific factors $\boldsymbol{\epsilon}_t$ are assigned the prior $1/\sigma_j^2 \sim \gamma(3.1, 2.1)$ independently for $j = 1, \ldots, p$ which ensures that the distributions have finite variance. The time-varying mean $\boldsymbol{\mu}_t$ is modelled as $\boldsymbol{\mu}_t = \boldsymbol{B}^{\mathsf{T}}\boldsymbol{w}_t$ in which $\boldsymbol{w}_t^{\mathsf{T}} = (w_{t1}, \ldots, w_{tc})$ is row t of an $n \times c$ matrix of daily covariates $\boldsymbol{W}$. The columns of $\boldsymbol{W}$ allow for an intercept, a non-linear effect of deviations from seasonal average temperature and fixed effects for the day-of-the-week, the occurrence of a public holiday, and the day-of-the-year, which is represented as a truncated Fourier series with six harmonics. Full details of the model for $\boldsymbol{\mu}_t$ are given in the Supplementary Materials along with a description of the associated priors.

In order to reflect both the persistence of above or below average demand for gas and the longitudinal but circular nature of a gas-day-vector, a sensible model for the inverse of the expected shared variation matrix $\boldsymbol{\Xi} = \boldsymbol{\Phi}^{-1}$ would be the precision matrix of a stationary circular autoregressive process of order one (Held and Rue 2010). We therefore take $\boldsymbol{\Xi} = (\xi_{ij})$ to be a tridiagonal Toeplitz matrix with corners; that is, $\xi_{i,i+1} = \xi_{i+1,i} = -\vartheta/2$ for $i = 1, \ldots, p-1$, $\xi_{1p} = \xi_{p1} = -\vartheta/2$ and $\xi_{ii} = 1$ for $i = 1, \ldots, p$ where we assume $\vartheta \in [0, 1)$ and take logit$(\vartheta) \sim N(0, 2)$. Exploratory analysis using data from another offtake site suggest that this structure is sensible based on the inverse of the sample covariance matrix. Given that the existing model is already rather complex, we chose to adopt a matrix normal prior for $\boldsymbol{\Lambda}$ in this example. The diagonal elements in the among-column scale matrix $\boldsymbol{\Psi}$ are given a MGP prior (9), with $a_1 = 2$, $a_2 = 3$ and a maximum truncation point of $H = \lceil \varphi(24) \rceil - 1 = 17$.

6.3.2 Analysis and results

We ran two chains, initialised at different starting points, using the adaptive Gibbs sampler discussed in Sect. 5.2 and another two chains using the (non-adaptive) Hamiltonian Monte Carlo sampler with a fixed truncation level of $H = \lceil \varphi(24) \rceil - 1 = 17$. After a burn-in of 50K iterations, each Gibbs chain was run for a further 200K iterations, thinning to retain every 100th draw in order to reduce computational overheads. Adaptation was allowed after 5K iterations and

the parameters in the diminishing adaptation condition were set at $\alpha_0 = -1$ and $\alpha_1 = -5 \times 10^{-5}$. The Hamiltonian Monte Carlo sampler was run for 20K iterations after a burn-in of the same length, thinning the output to retain every 10th draw. As in the Finnish birds example, we retain as active enough factors to explain at least $T = 0.999$ of the total variation in the data. The posterior mode (and median) for the number of effective factors in each case was 10 with 95% credible interval (9, 12). Comparing the output of identified parameters across the chains for the two samplers using the usual graphical diagnostic checks gave no evidence of any lack of convergence.

Let $\boldsymbol{\Upsilon} = \boldsymbol{\Omega}^{-1} = (\upsilon_{ij})$ denote the marginal precision matrix of the process and let $\boldsymbol{S}_{\Upsilon} = \text{diag}(\sqrt{\upsilon_{11}}, \ldots, \sqrt{\upsilon_{pp}})$. Figure 3 shows the posterior mean for the standardised precision matrix $\boldsymbol{S}_{\Upsilon}^{-1}\boldsymbol{\Upsilon}\boldsymbol{S}_{\Upsilon}^{-1}$. It is clear that its structure is reasonably consistent with the tridiagonal Toeplitz matrix with corners on which the prior for $\boldsymbol{\Delta}^{-1}$ is centered. However, there are some deviations, most notably another one or two bands of non-zero elements below the subdiagonal and above the supradiagonal. There is also some evidence of at least a partial band in the vicinity of $\upsilon_{i,i+12}$ for $i = 1, \ldots, 12$. This may be due to people switching their heating on twice per day, at around 7:00 in the morning and around 19:00 in the evening. This picture is reinforced by the posterior for the identified factor loadings matrix $\tilde{\boldsymbol{\Lambda}}$, shown in the Supplementary Materials, many of whose columns display a double-hump shape with smaller loadings towards the middle of the gas-day. As remarked in Sect. 4.3, in order to simplify the geometry of the stationary region for stationary dynamic factor models, the autoregressive coefficient matrices in the state equation are often assumed to be diagonal. However, this example provides considerable evidence in favour of our fully flexible approach, with the posterior probabilities $\Pr(\tilde{\gamma}_{i,jk} > 0 | \boldsymbol{y}_{1:n})$ for many off-diagonal elements being close to zero or one.

In order to assess the forecasting performance of the model, the number of factors was fixed at $k = 10$ and the model was refitted, holding back the last 25% of observations as test data. In keeping with the short-term horizon of interest, we considered forecasting $h = 24$ h ahead and, for comparative purposes, $h = 1$ hour ahead. Application of the standard forward filtering and forecasting recursions for (multivariate) dynamic linear models are not appropriate here as they do not allow within-day updates or forecasts. We therefore modify the standard forward filter so that the time-step is hours, rather than days, and at hours $2, \ldots, 24$ within each gas-day, we perform a partial update, comprising an observation step but no prediction step. The forecasting algorithm is similarly modified to allow forecasts to be issued hourly. The full algorithms are given in the Supplementary Materials. By using the draws from the posterior obtained from the fit to the training data and sampling one-for-one from the h-step
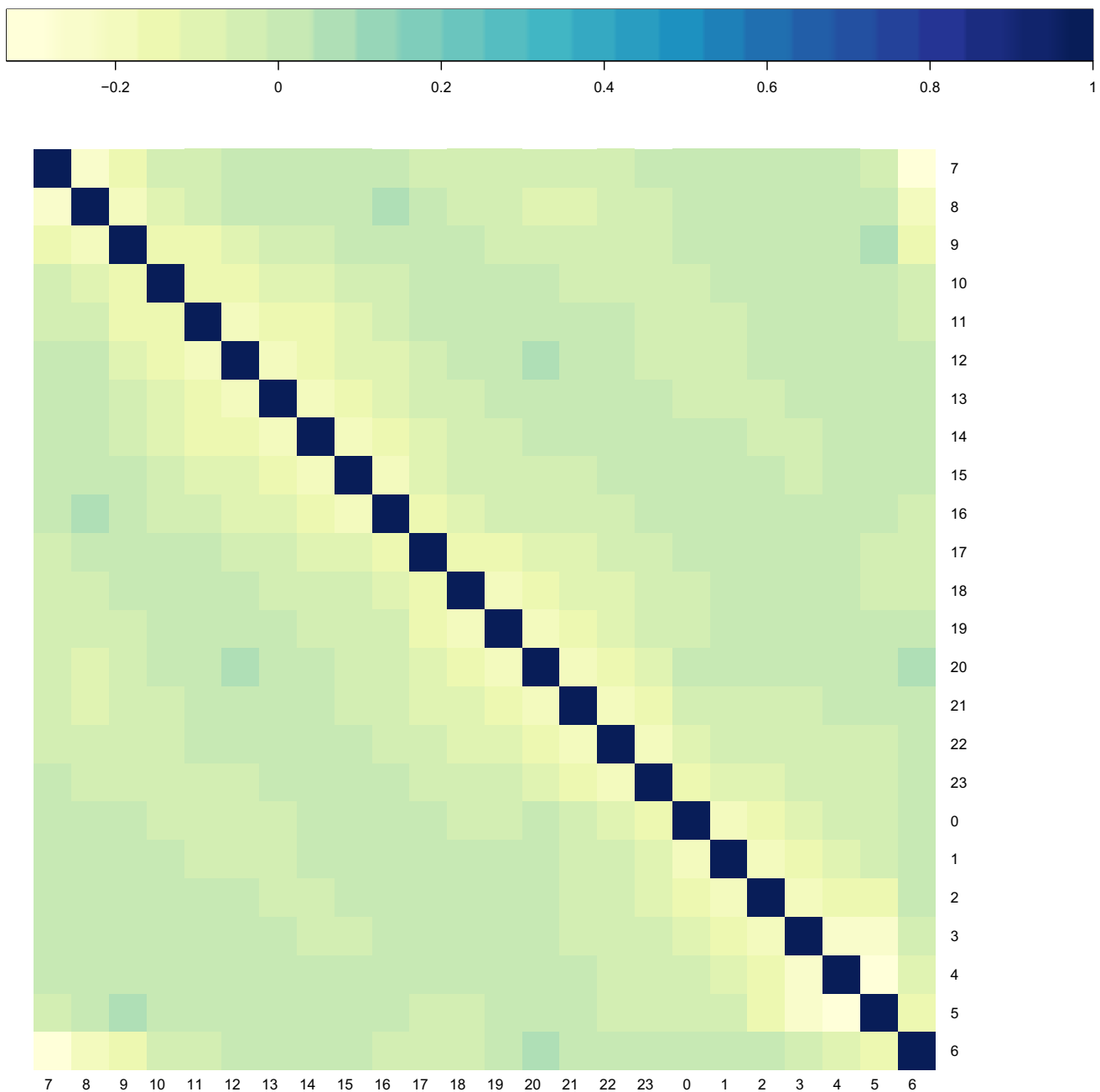
Fig. 3 Mean of the posterior for the marginal standardised precision matrix. The labels indicate hour of the day in a 24-h clock

ahead predictive distributions, we obtain samples from the h-step ahead posterior predictive distributions. For $h = 1$ and $h = 24$, these are visualised in Fig. 4 for the first 5% of times in the hold-out period, along with the test data. An analogous figure in the Supplementary Materials shows the last 5% of times. From these plots it is clear that the posterior predictive distributions are both accurate and precise over the forecast horizons of interest.

7 Discussion

We have proposed a class of structured priors for the loadings matrix of a Bayesian factor model with accompanying inferential algorithms. The novelty lies in the insight that the shared variation matrix $\mathbf{\Delta} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}}$ is much more amenable to the specification of prior beliefs than the factor loadings matrix $\mathbf{\Lambda}$. The prior is based on a matrix normal or matrix-t distribution for $\mathbf{\Lambda}$. Two important features are the choice of parametric structure for the among-row scale matrix $\mathbf{\Phi}$ and the increasing shrinkage prior for the diagonal among-
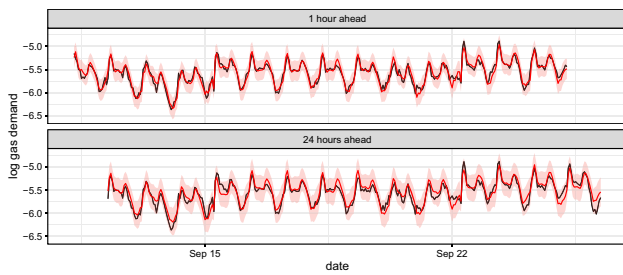
Fig. 4 For the first 5% of times in the hold-out period, means (—) and 95% equi-tailed credible intervals (▮) for the one-step ahead and 24-step ahead posterior predictive distributions. The observed data are also shown (—). (color figure online)

column scale matrix Ψ. The matrix Φ characterises the conditional prior expectation of Δ. Parametric forms are widely adopted as models for covariance matrices in longitudinal studies, spatial statistics, Gaussian processes, and a host of other areas. By adopting a parametric form, informed by domain expertise, as a model for the expectation of Δ, we benefit from shrinkage towards a rank-reduced matrix that is close to that structure. In general, the number of factors in a factor model is not known *a priori*. This is addressed through the increasing shrinkage process prior for $\psi_1, \ldots, \psi_k$ which allows columns of loadings whose likelihood contribution is negligible to be shrunk towards zero and discarded. A prior that encourages no more factors than are needed is helpful, particularly for prediction, because of the reduction in epistemic uncertainty that is afforded. At the cost of slightly more involved computational inference, the matrix-t version of the structured prior also offers a degree of freedom parameter ς which allows the data to influence the degree of shrinkage of Δ towards a rank-reduced matrix that is close to its mean.

Supplementary information

Supplementary information includes: (i) a document containing proofs and derivations of the results used in the paper, complete descriptions of the adaptive Gibbs samplers, further background and results on the simulation experiment and two applications; (ii) Stan programmes for non-adaptive Hamiltonian Monte Carlo sampling and R code implementing adaptive Gibbs sampling for the two applications.

Declarations

References

Aguilar, O., West, M.: Bayesian dynamic factor models and portfolio allocation. J. Bus. Econ. Stat. **18**(3), 338–357 (2000)

Aßmann, C., Boysen-Hogrefe, J., Pape, M.: Bayesian analysis of static and dynamic factor models: an ex-post approach towards the rotation problem. J. Economet. **192**, 190–206 (2016)

Bekker, P.A., ten Berge, J.M.F.: Generic global identification in factor analysis. Linear Algebra Appl. **264**, 255–263 (1997)

Bhattacharya, A., Dunson, D.B.: Sparse Bayesian infinite factor models. Biometrika **98**(2), 291–306 (2011)

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M.A., Guo, J., Li, P., Riddell, A.: Stan: a probabilistic programming language. J. Stat. Softw. **76**(1), 1–32 (2017)

Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q., West, M.: High-dimensional sparse factor modeling: applications in gene expression genomics. J. Am. Stat. Assoc. **103**(484), 1438–1456 (2008)

Castellanos, L., Vu, V.Q., Perel, S., Schwartz, A.B., Kass, R.E.: A multivariate Gaussian process factor model for hand shape during reach-to-grasp movements. Stat. Sinica **25**(1), 5–24 (2004)

Chan, J.C.C., Jeliazkov, I.: Efficient simulation and integrated likelihood estimation in state space models. Int. J. Math. Model. Numer. Optim. **1**, 101–120 (2009)

Chan, J., Leon-Gonzalez, R., Strachan, R.W.: Invariant inference and efficient computation in the static factor model. J. Am. Stat. Assoc. **113**(522), 819–828 (2018)

Chib, S., Nardari, F., Shephard, N.: Analysis of high dimensional multivariate stochastic volatility models. J. Economet. **134**(2), 341–371 (2006)

Conti, G., Frühwirth-Schnatter, S., Heckman, J.J., Piatek, R.: Bayesian exploratory factor analysis. J. Economet. **183**, 31–57 (2014)

Durante, D.: A note on the multiplicative gamma process. Stat. Probab. Lett. **122**, 198–204 (2017)

Dutta, R., Ghosh, J.K.: Bayes model selection with path sampling: factor models and other examples. Stat. Sci. **28**(1), 95–115 (2013)

Frühwirth-Schnatter, S.: Data augmentation and dynamic linear models. J. Time Ser. Anal. **15**, 183–202 (1994)

Frühwirth-Schnatter, S., Lopes, H.F.: Parsimonious Bayesian factor analysis when the number of factors is unknown. Technical Report, Chicago Booth (2010)

Frühwirth-Schnatter, S., Hosszejni, D., Lopes, H.F.: Sparse Bayesian factor analysis when the number of factors is unknown. arXiv:2301.06459 (2023)

Gelfand, A.E., Dey, D.K.: Bayesian model choice: asymptotics and exact calculations. J. R. Stat. Soc. B **56**(3), 501–514 (1994)

Geweke, J., Zhou, G.: Measuring the pricing error of the arbitrage pricing theory. Rev. Financial Stud. **9**(2), 557–587 (1996)

Ghosh, J., Dunson, D.B.: Default prior distributions and efficient posterior computation in Bayesian factor analysis. J. Comput. Graph. Stat. **18**(2), 306–320 (2009)

Gneiting, T., Raftery, A.E.: Strictly proper scoring rules, prediction, and estimation. J. Am. Stat. Assoc. **102**(477), 359–378 (2007)

Goldberg, L.R.: An alternative 'description of personality': the big-five factor structure. J. Pers. Soc. Psychol. **59**(6), 1216–1229 (1990)

Gupta, A.K., Nagar, D.K.: Matrix variate distributions. Chapman & Hall/CRC, Boca Raton (2000)

Hays, S., Shen, H., Huang, J.X.: Functional dynamic factor models with application to yield curve forecasting. Ann. Appl. Stat. **6**(3), 870–894 (2012)

Heaps, S.E.: Enforcing stationarity through the prior in vector autoregressions. J. Comput. Graph. Stat. **32**(1), 74–83 (2023)

Held, L., Rue, H.: Conditional and intrinsic autoregressions. In: Gelfand, A.E., Diggle, P.J., Fuentes, M., Guttorp, P. (eds.) Handbook of Spatial Statistics, 1st edn. Handbooks of Modern Statistical Methods, pp. 201–216. CRC Press, Boca Raton (2010). Chap. 13

Jungbacker, B., Koopman, S.J., Van der Wel, M.: Smooth dynamic factor analysis with application to the US term structure of interest rates. J. Appl. Economet. **29**(3), 65–90 (2014)

Kaufmann, S., Schumacher, C.: Bayesian estimation of sparse dynamic factor models with order-independent and ex-post identification. J. Economet. **210**, 116–134 (2019)

Koop, G., Leon-Gonzalez, R., Strachan, R.W.: Efficient posterior simulation for cointegrated models with priors on the cointegration space. Economet. Rev. **29**(2), 224–242 (2009)

Kowal, D.R., Canale, A.: Semiparametric functional factor models with Bayesian rank selection. arXiv:2108.02151v3 (2022)

Kowal, D.R., Matteson, D.S., Ruppert, D.: A Bayesian multivariate functional dynamic linear model. J. Am. Stat. Assoc. **112**(518), 733–744 (2017)

Legramanti, S., Durante, D., Dunson, D.B.: Bayesian cumulative shrinkage for infinite factorizations. Biometrika **107**(3), 745–752 (2020)

Leung, D., Drton, M.: Order-invariant prior specification in Bayesian factor analysis. Stat. Probab. Lett. **111**, 60–66 (2016)

Lim, L.-H., Sepulchre, R.: Geometric distance between positive definite matrices of different dimensions. IEEE Trans. Inf. Theory **65**(9), 5401–5405 (2019)

Lindström, Å., Green, M., Husby, M., Kålås, J.A., Lehikoinen, A.: Large-scale monitoring of waders on their Boreal and Arctic breeding grounds in Northern Europe. Ardea **103**(1), 3–15 (2015)

Lopes, H.F., Carvalho, C.M.: Factor stochastic volatility with time varying loadings and Markov switching regimes. J. Stat. Plan. Inf. **137**, 3082–3091 (2007)

Lopes, H.F., West, M.: Bayesian model assessment in factor analysis. Stat. Sinica **14**, 41–67 (2004)

Lopes, H.F., Salazar, E., Gamerman, D.: Spatial dynamic factor analysis. Bayesian Anal. **3**(4), 759–792 (2008)

Man, A.X., Culpepper, S.A.: A mode-jumping algorithm for Bayesian factor analysis. J. Am. Stat. Assoc. **117**(537), 277–290 (2022)

Peña, D., Poncela, P.: Nonstationary dynamic factor analysis. J. Stat. Plan. Inf. **136**, 1237–1257 (2006)

Roberts, G.O., Rosenthal, J.S.: Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. J. Appl. Probab. **44**(2), 458–475 (2007)

Sáfadi, T., Peña, D.: Bayesian analysis of dynamic factor models: an application to air pollution and mortality in São Paulo, Brazil. Environmetrics **19**(1), 95–117 (2008)

Schiavon, L., Canale, A.: On the truncation criteria in infinite factor models. Stat **9**(1), 298 (2020)

Schiavon, L., Canale, A., Dunson, D.B.: Generalized infinite factorization models. Biometrika **109**(3), 817–835 (2022)

Schmidt, A.M., Guttorp, P., O'Hagan, A.: Considering covariates in the covariance structure of spatial processes. Environmetrics **22**, 487–500 (2011)

Soldo, B.: Forecasting natural gas consumption. Appl. Energy **92**, 26–37 (2012)

Steel, M.: Phylogeny: Discrete and Random Processes in Evolution. Society for Industrial and Applied Mathematics, Philadelphia (2016)

Taylor-Rodriguez, D., Finley, A.O., Dhatta, A., Babcock, C., Andersen, H., Cook, B.D., Mortin, D.C., Banerjee, S.: Spatial factor models for high-dimensional and large spatial data. Stat. Sinica **29**(3), 1155–1180 (2019)

Wilkinson, D.J., Yeung, S.K.H.: Conditional simulation from highly structured Gaussian systems, with application to blocking-MCMC for the Bayesian analysis of very large linear models. Stat. Comput. **12**, 287–300 (2002)

Wilkinson, D.J., Yeung, S.K.H.: A sparse matrix approach to Bayesian computation in large linear models. Comput. Stat. Data Anal. **44**, 493–516 (2004)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.