Attention-Based Vector Autoregressions

Juan Antolín-Díaz*
MIT Sloan

Juan F. Rubio-Ramírez†
Emory University and
Federal Reserve Bank of Atlanta

January 2026

Abstract

We introduce attention-based vector autoregressions, a new class of nonlinear time series models in which the conditional mean and variance are modeled through the attention mechanism, the core of the Transformer architecture in language modeling. Attention leads to state-dependent, time-varying dynamics, that are deterministic functions of past data. The framework nests constant-parameter linear VARs and approximates Dynamic Factor Models, but provides dimensionality reduction with a natural economic interpretation. We develop Bayesian priors and estimation methods suitable for time series and sample sizes available in macroeconomics and finance, and characterize the model's theoretical properties. An empirical application to forecasting U.S. inflation after the pandemic demonstrates gains in point and density forecast accuracy relative to standard benchmarks.

Keywords: Vector autoregression, time-varying parameters, attention mechanism, density forecasting, conditional heteroskedasticity.

JEL Classification: C32, C45, C53, E37

*MIT Sloan School of Management, 100 Main Street, Cambridge, MA 02142. Email: jantolin@mit.edu.
†Department of Economics, Emory University, Atlanta, GA 30322; and Federal Reserve Bank of Atlanta.

1 Introduction

Transformer models (Vaswani et al., 2017) have achieved remarkable success in language modeling. Since at their core they model the joint distribution of a sequence, an application to economic time series seems natural. It is unclear, however, whether they can work well with the small samples typically available in macroeconomics and finance; there is no econometric framework for interpreting their structure or conducting inference, and little attention has been paid to well-calibrated density forecasts.

This paper addresses these gaps, introducing attention-based vector autoregressions, a class of nonlinear multivariate time series models in which the conditional mean and covariance matrix are functions of past observables, parameterized using the attention mechanism. Unlike existing non-linear and time-varying VAR extensions, attention-based VARs are scalable, do not require filtering of latent processes, and provide an agnostic benchmark when the researcher does not want to commit ex-ante neither to a time-invariant linear structure nor to a specific form of nonlinearity or time variation.

Our main theoretical result is a formal connection between the attention mechanism and VARs (Sims, 1980) and factor models (Geweke, 1977; Stock and Watson, 2002), the workhorse models of macroeconometrics. The conditional mean admits an exact time-varying VAR representation whose coefficients are closed-form functions of the observed history, nesting constant-coefficient VARs as a special case. The embedding dimension and the number of attention heads jointly control the rank of the coefficient matrices, imposing the same low-dimensional structure that makes factor models amenable to large cross sections and gives them a natural economic interpretation consistent with dynamic equilibrium models (Fernández-Villaverde et al., 2007); the same structure extends to the conditional covariance, producing a complete density model. This connection also reveals a normalization problem in the attention model analogous to the rotational indeterminacy of factor loadings, which we resolve with identification restrictions.

Unlike time-varying parameter models that treat the coefficients as latent states requiring filtering (Primiceri, 2005), all conditional moments here are deterministic functions of observables, so the likelihood is available in closed form and exact Bayesian inference is feasible even as the number of variables increases. We exploit the model's partially conjugate structure in a Gibbs sampler: the output projection and error covariance admit closed-form posterior draws, while the nonlinear attention parameters are updated via Elliptical Slice Sampling (Murray et al., 2010). The connection with VARs and factor models allows us to construct a shrinkage prior that is centered at time-invariant, low-rank dynamics, making the constant-coefficient VAR a limiting case of the attention specification. The prior echoes

the Minnesota prior (Doan et al., 1984; Litterman, 1986) in spirit, encoding the belief that a parsimonious linear model is a reasonable starting point while letting the data pull the posterior towards richer dynamics when warranted.

The post-pandemic inflation surge, which caught the Federal Reserve and professional forecasters by surprise, provides a natural test case: can the nonlinearities captured by the attention mechanism improve forecasts during such a dramatic episode? Inflation is notoriously difficult to forecast (Stock and Watson, 2007): its dynamics are well known to vary over time, cross-sectional comovements are strong, and the large cross section strains conventional TVP-VAR methods. We put the model to the test in a recursive forecasting exercise for U.S. personal consumption expenditure (PCE) price inflation, using panels of increasingly disaggregated price indices, with up to several hundred variables, to test the scalability of the approach. Even in a minimal univariate setting with a single attention head, the attention-based VAR produces more accurate point and density forecasts than a Bayesian VAR with Minnesota prior estimated using state-of-the-art methods (Banbura et al., 2010; Giannone et al., 2015) at all horizons evaluated, particularly at longer horizons.[1] The gains grow with the forecast horizon: at $h = 24$ months the RMSE falls by 19% relative to the Bayesian VAR, and the log predictive score improves from -2.33 to -2.02, indicating substantially sharper and better-calibrated predictive densities.

Related literature. This paper connects two literatures that have developed largely in parallel, separated by methodological traditions and language.

On the econometrics side, the VAR model (Sims, 1980; Hamilton, 1994; Lütkepohl, 2005) and factor models (Geweke, 1977; Stock and Watson, 2002; Bai and Ng, 2002; Bernanke et al., 2005) are the dominant frameworks for multivariate time series in macroeconomics and finance. Extensions with time-varying parameters include random walk specifications (Primiceri, 2005; Cogley and Sargent, 2005), Markov-switching (Hamilton, 1989), smooth-transition autoregressions (Tong, 1990; Teräsvirta, 1994), and dynamic shrinkage approaches (Hauzenberger et al., 2024); Arias et al. (2024) develop sign-restricted inference for TV-VARs. Unlike these models, where the form of time variation is specified ex ante, attention learns the state dependence from the data. Multivariate GARCH models (Bollerslev, 1990; Engle, 2002) capture time-varying covariances; our parameterization is related to factor GARCH (Engle et al., 1990) but ties the covariance to the same state that drives the mean. The observation-driven formulation follows Creal et al. (2013) and Harvey (2013). Bayesian shrinkage (Banbura et al., 2010; Koop and Korobilis, 2013) provides regularization; our prior

[1] The present version reports results for the single-head, univariate specification only. Preliminary results confirm the good performance of the model when extended to multiple heads and larger cross sections.

adapts this idea to the attention parameterization.

On the machine learning side, the attention mechanism (Bahdanau et al., 2015) and the Transformer (Vaswani et al., 2017) have transformed sequence modeling. Machine learning methods are increasingly used in economics and finance (Gu et al., 2020; Farrell et al., 2021), and Transformer architectures have been applied to time series forecasting (Lim et al., 2021; Zhou et al., 2021), though Zeng et al. (2023) show that simple linear models can outperform them, underscoring the need for structural restrictions in small samples. Kelly et al. (2025) embed a Transformer in the stochastic discount factor for asset pricing. Softmax attention can be viewed as a Nadaraya–Watson estimator with learned kernels (Tsai et al., 2019; Goel and Bartlett, 2024; Lu et al., 2024), and in-context learning theory shows that Transformers can implement gradient descent and ridge regression in their forward pass (Von Oswald et al., 2023; Akyürek et al., 2023). Non-identification issues related to functional equivalence and parameter symmetries are familiar in the optimization literature; for attention architectures, explicit characterization of these symmetries is more recent (Henry et al., 2024; François and Ravera, 2025). We connect this problem to the rotational indeterminacy of factor loadings and resolve it with analogous restrictions. More broadly, we go beyond existing representation results by developing a complete econometric framework with priors, posterior computation, and density forecasting.

Organization of the paper. Section 2 introduces the attention-only model, building from single-head attention and its TV-VAR representation to multi-head attention with its approximation and regularization properties. Section 5 extends the framework to time-varying covariance. Section 6 develops the nonlinear extensions that arise from the full Transformer architecture. Section 7 develops estimation and inference methods. Section 8 concludes.

2 Single-Head Attention Models

In this section we introduce single-head attention models. These models can be used to model both the conditional mean and variance of the time series. We will focus on the conditional mean. After defining basic notation we will show that single-head attention models have TV-VAR representation with a low-rank structure. We will also present an algorithm to draw from the posterior of the parameters.

2.1 Notation and definitions

We begin by establishing notation and defining key concepts.

Basic notation. Let $\{\boldsymbol{y}_t\}_{t\geq 1}$ be an n-dimensional stochastic process with $\boldsymbol{y}_t \in \mathbb{R}^n$. Let $\mathcal{F}_t = \sigma(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t)$ denote the natural filtration. For $t \geq L+1$, we denote $\boldsymbol{y}_{t-L:t-1} = (\boldsymbol{y}_{t-L}, \ldots, \boldsymbol{y}_{t-1}) \in (\mathbb{R}^n)^L$ the "context window" of length L; observations $t \leq L$ are treated as a presample. The set of $n \times n$ symmetric positive definite matrices is denoted $\mathcal{S}_{++}^n$.

Softmax function. For $\boldsymbol{a} = (a_1, \ldots, a_L)^\top \in \mathbb{R}^L$, the softmax function is

$$\operatorname{softmax}(\boldsymbol{a})_\ell = \frac{\exp(a_\ell)}{\sum_{j=1}^L \exp(a_j)}, \quad \ell = 1, \ldots, L. \tag{1}$$

Softmax is shift-invariant: for any constant $c \in \mathbb{R}$, $\operatorname{softmax}(\boldsymbol{a} + c\mathbf{1}) = \operatorname{softmax}(\boldsymbol{a})$. A numerically stable way to evaluate (1) is therefore to set $m = \max_j a_j$ and compute

$$\operatorname{softmax}(\boldsymbol{a})_\ell = \frac{\exp(a_\ell - m)}{\sum_{j=1}^L \exp(a_j - m)}.$$

The output $\operatorname{softmax}(\boldsymbol{a}) \in \mathbb{R}^L$ is a probability vector: all entries are non-negative and sum to one. The softmax function provides a smooth approximation to the argmax operation: $\operatorname{softmax}(\beta \boldsymbol{a}) \to \boldsymbol{e}_{i^\star}$ as $\beta \to \infty$, where $i^\star \in \arg\max_i a_i$ and $\boldsymbol{e}_{i^\star}$ is the corresponding standard basis vector.

2.2 Econometric environment

We model $\boldsymbol{y}_t$ given the information set $\mathcal{F}_{t-1}$ by specifying its first two conditional moments. In particular, we assume that the conditional mean and conditional covariance exist and are $\mathcal{F}_{t-1}$-measurable functions of a fixed "context window" $\boldsymbol{y}_{t-L:t-1}$:

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) \in \mathbb{R}^n, \qquad \mathbb{V}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) \in \mathcal{S}_{++}^n, \tag{2}$$

where $g_{\boldsymbol{\theta}}(\cdot)$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\cdot)$ are both $\mathcal{F}_{t-1}$-measurable functions to be specified, L is the fixed "context length" (the number of lags entering the context window), and $\boldsymbol{\theta} \in \Theta$ collects all parameters. Equivalently, we can write an innovation representation $\boldsymbol{y}_t = g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) + \boldsymbol{\varepsilon}_t$, where $\mathbb{E}[\boldsymbol{\varepsilon}_t \mid \mathcal{F}_{t-1}] = 0$ and $\operatorname{Var}(\boldsymbol{\varepsilon}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1})$.

Example: the linear VAR model. The VAR(p) model (Sims, 1980; Lütkepohl, 2005) is the special case with $L = p$, linear conditional mean, and constant conditional covariance:

$$g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-p:t-1}) = \boldsymbol{c} + \sum_{\ell=1}^{p} \mathbf{A}_\ell \boldsymbol{y}_{t-\ell}, \qquad \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-p:t-1}) = \boldsymbol{\Sigma},$$

where $\boldsymbol{\theta} = \{\boldsymbol{c}, \mathbf{A}_1, \ldots, \mathbf{A}_p, \boldsymbol{\Sigma}\}$ and $\Theta = \mathbb{R}^n \times \mathbb{R}^{n\times n} \times \ldots \times \mathbb{R}^{n\times n} \times \mathcal{S}^n_{++}$.

2.3 Single-head attention for the conditional mean

For the rest of this section, we assume a constant conditional covariance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) = \boldsymbol{\Sigma}$; Section 5 extends the framework to time-varying covariance. We parameterize the conditional mean by first mapping the context window to a d-dimensional state vector

$$\boldsymbol{h}_t = \phi_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) \in \mathbb{R}^d,$$

of which the conditional mean is an affine function:

$$g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) = \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t, \tag{3}$$

where $\mathbf{W}_\mu \in \mathbb{R}^{n\times d}$ and $\boldsymbol{b}_\mu \in \mathbb{R}^n$. All model outputs depend on the past only through $\boldsymbol{h}_t$. Unlike classical state-space models, $\boldsymbol{h}_t$ is a deterministic (though nonlinear) function of observables rather than a latent vector with its own stochastic dynamics; this is an observation-driven specification in the terminology of Creal et al. (2013). It remains to specify the state mapping $\phi_{\boldsymbol{\theta}}$. We adapt the attention mechanism of Vaswani et al. (2017) to a time series context, and describe it below in three steps:

Linear Embedding. For each lag $\ell \in \{1, \ldots, L\}$, define the d-dimensional linear "embedding"

$$\boldsymbol{x}_{t-\ell} = \boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell} \in \mathbb{R}^d, \tag{4}$$

where $\mathbf{W}_x \in \mathbb{R}^{d\times n}$ is a parameter matrix and $\boldsymbol{p}_\ell \in \mathbb{R}^d$ is a lag-specific intercept, which we call a "positional effect."[2] Let $\boldsymbol{p} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L] \in \mathbb{R}^{d\times L}$.

[2]In the machine learning literature, the $\boldsymbol{p}_\ell$ are called "positional encodings" (Vaswani et al., 2017). In natural language models, each input $\boldsymbol{y}_{t-\ell}$ is a discrete token mapped to a learned vector via a lookup table, and the positional encoding is added to distinguish identical tokens appearing at different positions in the sequence. In our setting, $\boldsymbol{y}_{t-\ell} \in \mathbb{R}^n$ is already a continuous vector, so $\mathbf{W}_x \boldsymbol{y}_{t-\ell}$ is a linear projection rather than a lookup. The positional effects $\boldsymbol{p}_\ell$ play an analogous role: they allow the model to distinguish the same observation vector appearing at different lags.

Linear Projections. Three linear projections are then applied to each embedding:

$$\boldsymbol{q}_t = \mathbf{W}_q \boldsymbol{x}_{t-1} \in \mathbb{R}^{d_k}, \ \boldsymbol{k}_{t-\ell} = \mathbf{W}_k \boldsymbol{x}_{t-\ell} \in \mathbb{R}^{d_k}, \text{ and } \boldsymbol{v}_{t-\ell} = \mathbf{W}_v \boldsymbol{x}_{t-\ell} \in \mathbb{R}^{d_v}, \tag{5}$$

for $\ell = 1, \ldots, L$, where $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d_k \times d}$ and $\mathbf{W}_v \in \mathbb{R}^{d_v \times d}$. The "query" $\boldsymbol{q}_t$ is formed from the most recent observation; the "keys" $\boldsymbol{k}_{t-\ell}$ and "values" $\boldsymbol{v}_{t-\ell}$ are formed from each lagged observation. We assume that $\mathbf{W}_s$ for $s \in \{x, q, k, v\}$ are full rank.

Attention weights. The relevance of each lag is measured by the "score"

$$s_{t,\ell} = b_\ell + \frac{\boldsymbol{q}_t^\top \boldsymbol{k}_{t-\ell}}{\sqrt{d_k}} \in \mathbb{R}, \quad \ell = 1, \ldots, L, \tag{6}$$

where $b_\ell \in \mathbb{R}$ is a lag-specific score bias and the scaling by $\sqrt{d_k}$ is a variance-stabilizing normalization.[3] Let $\boldsymbol{b} = (b_1, \ldots, b_L)^\top \in \mathbb{R}^L$. The scores are transformed into non-negative weights summing to one via the softmax function:

$$\alpha_{t,\ell} = \frac{\exp(s_{t,\ell})}{\sum_{j=1}^{L} \exp(s_{t,j})} \in \mathbb{R}, \quad \ell = 1, \ldots, L. \tag{7}$$

Substituting Equations (4) and (5) shows that the attention weights depend on $(\mathbf{W}_q, \mathbf{W}_k)$ only through the product $\mathbf{W}_q^\top \mathbf{W}_k$, a fact used below in the identification discussion. The attention weight vector $\boldsymbol{\alpha}_t = (\alpha_{t,1}, \ldots, \alpha_{t,L})^\top$ lies in the interior of the $(L-1)$-simplex and is $\mathcal{F}_{t-1}$-measurable. We will assume that $s_{t,\ell}$ are finite, hence $\alpha_{t,\ell} > 0$ for $\ell = 1, \ldots, L$.

Aggregation. The state vector $\boldsymbol{h}_t$ is the weighted average of values:

$$\boldsymbol{h}_t = \phi_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) = \sum_{\ell=1}^{L} \alpha_{t,\ell} \, \boldsymbol{v}_{t-\ell} \in \mathbb{R}^{d_v}, \tag{8}$$

so that $d = d_v$ in the single-head case. From this point onward in this section, we use only d for this common dimension. The conditional mean in Equation (3) is then

$$g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) = \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t = \boldsymbol{b}_\mu + \mathbf{W}_\mu \phi_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) = \boldsymbol{b}_\mu + \mathbf{W}_\mu \sum_{\ell=1}^{L} \alpha_{t,\ell} \, \boldsymbol{v}_{t-\ell},$$

[3] If the entries of $\boldsymbol{q}_t$ and $\boldsymbol{k}_{t-\ell}$ have unit variance and are uncorrelated, the score has variance approximately one regardless of d_k. The scaling stabilizes scores with respect to dimension but not measurement scale; in empirical work, standardizing $\boldsymbol{y}_t$ mitigates scale-driven variation in attention weights.

where $\boldsymbol{\theta} = \{\boldsymbol{b}_\mu, \mathbf{W}_\mu, \boldsymbol{p}, \mathbf{W}_x, \mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \boldsymbol{b}\}$ and $\boldsymbol{\Theta} = \mathbb{R}^n \times \mathbb{R}^{n\times d} \times \mathbb{R}^{d\times L} \times \mathbb{R}^{d\times n} \times \mathbb{R}^{d_k\times d} \times \mathbb{R}^{d_k\times d} \times \mathbb{R}^{d\times d} \times \mathbb{R}^L$.

2.4 Identification

The model defined by Equations (3)–(8) is not identified: distinct values of the parameter vector $\boldsymbol{\theta}$ can produce an identical conditional mean. The following proposition formalizes the claim.

Proposition 1 (Invariance to Reparametrizations)**.** *The conditional mean is invariant to each of the following transformations: (i)* embedding-space basis change*: $(\boldsymbol{p}, \mathbf{W}_x, \mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v)$ replaced by $(\mathbf{G}\boldsymbol{p}, \mathbf{G}\mathbf{W}_x, \mathbf{W}_q\mathbf{G}^{-1}, \mathbf{W}_k\mathbf{G}^{-1}, \mathbf{W}_v\mathbf{G}^{-1})$ for any invertible $\mathbf{G} \in \mathbb{R}^{d\times d}$; (ii)* value-space reparameterization*: $(\mathbf{W}_\mu, \mathbf{W}_v)$ replaced by $(\mathbf{W}_\mu\mathbf{S}^{-1}, \mathbf{S}\mathbf{W}_v)$ for any invertible $\mathbf{S} \in \mathbb{R}^{d\times d}$; (iii)* query-key reparameterization*: $(\mathbf{W}_q, \mathbf{W}_k)$ replaced by $(\mathbf{C}^{-\top}\mathbf{W}_q, \mathbf{C}\mathbf{W}_k)$ for any invertible $\mathbf{C} \in \mathbb{R}^{d_k\times d_k}$; (iv)* score-bias location shift*: $(b_1, \ldots, b_L)$ replaced by $(b_1+c, \ldots, b_L+c)$ for any $c \in \mathbb{R}$.*

Proof. For (i), let $\boldsymbol{x}_{t-\ell} = \boldsymbol{p}_\ell + \mathbf{W}_x\boldsymbol{y}_{t-\ell}$. Under the transformation, $\boldsymbol{x}'_{t-\ell} = \mathbf{G}\boldsymbol{x}_{t-\ell}$, so the scores satisfy

$$(\boldsymbol{x}'_{t-1})^\top(\mathbf{W}'_q)^\top\mathbf{W}'_k\boldsymbol{x}'_{t-\ell} = \boldsymbol{x}_{t-1}^\top\mathbf{G}^\top\mathbf{G}^{-\top}\mathbf{W}_q^\top\mathbf{W}_k\mathbf{G}^{-1}\mathbf{G}\boldsymbol{x}_{t-\ell} = \boldsymbol{x}_{t-1}^\top\mathbf{W}_q^\top\mathbf{W}_k\boldsymbol{x}_{t-\ell}.$$

Likewise, the value vectors are unchanged because $\mathbf{W}'_v\boldsymbol{x}'_{t-\ell} = \mathbf{W}_v\mathbf{G}^{-1}\mathbf{G}\boldsymbol{x}_{t-\ell} = \mathbf{W}_v\boldsymbol{x}_{t-\ell}$. For (ii), $\mathbf{W}'_\mu\mathbf{W}'_v = \mathbf{W}_\mu\mathbf{S}^{-1}\mathbf{S}\mathbf{W}_v = \mathbf{W}_\mu\mathbf{W}_v$, and $\mathbf{S}$ does not enter the scores. For (iii), the scores depend on $\mathbf{W}_q$ and $\mathbf{W}_k$ only through $\mathbf{W}_q^\top\mathbf{W}_k$, and $(\mathbf{C}^{-\top}\mathbf{W}_q)^\top(\mathbf{C}\mathbf{W}_k) = \mathbf{W}_q^\top\mathbf{C}^{-1}\mathbf{C}\mathbf{W}_k = \mathbf{W}_q^\top\mathbf{W}_k$. For (iv), the transformed scores satisfy $s'_{t,\ell} = s_{t,\ell} + c$ for all ℓ, hence

$$\alpha'_{t,\ell} = \frac{\exp(s_{t,\ell}+c)}{\sum_{j=1}^L \exp(s_{t,j}+c)} = \frac{\exp(c)\exp(s_{t,\ell})}{\exp(c)\sum_{j=1}^L \exp(s_{t,j})} = \alpha_{t,\ell}.$$

Since the value vectors are unchanged, the conditional mean is unchanged. □

We can impose some restriction to $\boldsymbol{\theta}$ that will allow us to identify the model. Next corollary represents such restrictions.

Corollary 1 (Exact Normalizations)**.** *The following restrictions rule out the basis changes, matrix factorizations, and common score shift in Proposition 1 without restricting the set of conditional means the model can represent: (i) require a $d \times d$ submatrix of $\mathbf{W}_x$ to be lower-triangular with positive diagonal; (ii) set $\mathbf{W}_v = \mathbf{I}_d$ and estimate $\mathbf{W}_\mu$ directly, since only the product enters the conditional mean; (iii) parameterize the score kernel $\mathbf{F} := \mathbf{W}_q^\top\mathbf{W}_k$*

directly, since the attention weights depend on $\mathbf{W}_q$ *and* $\mathbf{W}_k$ *only through* $\mathbf{F}$*; (iv) impose one location normalization on the score biases, such as* $b_1 = 0$*.*

Since $\mathbf{F} = \mathbf{W}_q^\top \mathbf{W}_k \in \mathbb{R}^{d\times d}$ with $\mathbf{W}_q, \mathbf{W}_k \in \mathbb{R}^{d_k \times d}$, its rank satisfies

$$\operatorname{rank}(\mathbf{F}) \le \min\{\operatorname{rank}(\mathbf{W}_q), \operatorname{rank}(\mathbf{W}_k)\} = \min\{d_k, d\}.$$

Thus $d_k < d$ imposes a low-rank restriction on the score kernel, while $d_k \ge d$ allows $\mathbf{F}$ to be any $d \times d$ matrix. A unified presentation treats all the cases is as follows. Let $q := \min\{d, d_k\}$ and can write the compact singular value decomposition $\mathbf{F} = \mathbf{U}_q \mathbf{D}_q \mathbf{V}_q^\top$ where $\mathbf{U}_q, \mathbf{V}_q \in \mathbb{R}^{d\times q}$ satisfy $\mathbf{U}_q^\top \mathbf{U}_q = \mathbf{V}_q^\top \mathbf{V}_q = \mathbf{I}_q$ and $\mathbf{D}_q = \operatorname{diag}(\lambda_1, \ldots, \lambda_q)$ with $\lambda_1 \ge \cdots \ge \lambda_q > 0$. A compatible pair $(\mathbf{W}_q, \mathbf{W}_k)$ is obtained by

$$\mathbf{W}_q = \begin{bmatrix} \mathbf{D}_q^{1/2} \mathbf{U}_q^\top \\ \mathbf{0}_{(d_k - q)\times d} \end{bmatrix}, \qquad \mathbf{W}_k = \begin{bmatrix} \mathbf{D}_q^{1/2} \mathbf{V}_q^\top \\ \mathbf{0}_{(d_k - q)\times d} \end{bmatrix},$$

for which $\mathbf{W}_q^\top \mathbf{W}_k = \mathbf{F}$. This nests the previous cases as special cases: if $d_k < d$, then $q = d_k$ and there is no padding; if $d_k = d$, then $q = d$ and again there is no padding; if $d_k > d$, then $q = d$ and the construction reduces to the zero-row padding argument above. The corresponding parameter count is

$$2\Big(dq - \frac{q(q+1)}{2}\Big) + q = 2dq - q^2,$$

which equals d^2 when $d_k \ge d$ and $2dd_k - d_k^2$ when $d_k < d$.

Importantly, there is a local mapping from Euclidean space into the constrained set $\mathcal{S}(d,q) \times \mathbb{R}^q_{>,+} \times \mathcal{S}(d,q)$ where

$$\mathcal{S}(d,q) := \{\mathbf{Q} \in \mathbb{R}^{d\times q} : \mathbf{Q}^\top \mathbf{Q} = \mathbf{I}_q\} \qquad \text{and} \qquad \mathbb{R}^q_{>,+} := \{(\lambda_1, \ldots, \lambda_q) : \lambda_1 > \cdots > \lambda_q > 0\}.$$

Specifically, let $\boldsymbol{\xi} = (\mathbf{A}_U, \boldsymbol{\omega}_U, \mathbf{A}_V, \boldsymbol{\omega}_V, \boldsymbol{\eta}) \in \mathbb{R}^{(d-q)q} \times \mathbb{R}^{q(q-1)/2} \times \mathbb{R}^{(d-q)q} \times \mathbb{R}^{q(q-1)/2} \times \mathbb{R}^q$. Construct skew-symmetric matrices $\boldsymbol{\Omega}_U, \boldsymbol{\Omega}_V \in \mathbb{R}^{q\times q}$ from $\boldsymbol{\omega}_U$ and $\boldsymbol{\omega}_V$, define the Cayley transforms $\mathbf{R}_U = (\mathbf{I}_q - \boldsymbol{\Omega}_U)(\mathbf{I}_q + \boldsymbol{\Omega}_U)^{-1}$ and $\mathbf{R}_V = (\mathbf{I}_q - \boldsymbol{\Omega}_V)(\mathbf{I}_q + \boldsymbol{\Omega}_V)^{-1}$, and let

$$\mathbf{G}(\mathbf{A}) := \begin{bmatrix} \mathbf{I}_q \\ \mathbf{A} \end{bmatrix} (\mathbf{I}_q + \mathbf{A}^\top \mathbf{A})^{-1/2} \in \operatorname{St}(d,q).$$

Then define $\mathbf{U}_q = \mathbf{G}(\mathbf{A}_U)\mathbf{R}_U$ and $\mathbf{V}_q = \mathbf{G}(\mathbf{A}_V)\mathbf{R}_V$. For the singular values, set

$$\lambda_q = e^{\eta_q} > 0,\ \lambda_j = \lambda_{j+1} + e^{\eta_j} > 0 \text{ for } j = 1, \ldots, q-1, \text{ and } \mathbf{D}_q = \operatorname{diag}(\lambda_1, \ldots, \lambda_q).$$

This produces the rank-q kernel $\mathbf{F} = \mathbf{U}_q \mathbf{D}_q \mathbf{V}_q^\top$. Hence, when $q = d_k$, one can consider the unrestricted Euclidean vector $\boldsymbol{\xi}$ and map it into $(\mathbf{U}_q, \mathbf{D}_q, \mathbf{V}_q)$. When $q = d$ we can directly consider $\mathbf{F}$.

2.5 A TV-VAR class

We now show that the single-head attention model admits an exact representation as a VAR with time-varying, state-dependent coefficients that depend on past observations. Hence, the single-head attention model defines a class of TV-VARs. Proposition 2 defines such a class.

Assumption 1. *The model uses single-head attention with: (i) context length L; (ii) affine conditional mean as in Equation (3); (iii) linear embeddings as in Equation (4); (iv) linear projections as in Equation (5); (v) attention weights as in Equations (6)-(7); (vi) aggregation as in Equation (8).*

Proposition 2 (TV-VAR Representation)**.** *Under Assumption 1, the conditional mean satisfies*

$$g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) = \boldsymbol{b}_t + \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell}, \tag{9}$$

where $\mathbf{A}_{t,\ell} = \alpha_{t,\ell} \mathbf{B}$ with $\mathbf{B} = \mathbf{W}_\mu \mathbf{W}_v \mathbf{W}_x \in \mathbb{R}^{n \times n}$, and

$$\boldsymbol{b}_t = \boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \mathbf{W}_v \boldsymbol{p}_\ell.$$

Proof. Substitute definitions. Since $\boldsymbol{v}_{t-\ell} = \mathbf{W}_v(\boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell})$ and $\boldsymbol{h}_t = \sum_\ell \alpha_{t,\ell} \boldsymbol{v}_{t-\ell}$,

$$\begin{aligned} g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) &= \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t = \boldsymbol{b}_\mu + \mathbf{W}_\mu \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_v(\boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}) \\ &= \boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \mathbf{W}_v \boldsymbol{p}_\ell + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \mathbf{W}_v \mathbf{W}_x \boldsymbol{y}_{t-\ell}. \end{aligned}$$

The result follows with $\mathbf{B} = \mathbf{W}_\mu \mathbf{W}_v \mathbf{W}_x$. □

Proposition 2 shows that the attention-based model has the structure of a VAR(L) with time-varying coefficients. Both $\mathbf{A}_{t,\ell}$ and $\boldsymbol{b}_t$ are $\mathcal{F}_{t-1}$-measurable, since the attention weights $\alpha_{t,\ell}$ are deterministic functions of the history. The time variation arises entirely through

these weights. Moreover, notice that the factorization

$$\mathbf{A}_{t,\ell} = \alpha_{t,\ell}\mathbf{B}, \qquad \ell = 1, \ldots, L. \tag{10}$$

provides implicit regularization: all coefficient matrices share a single base matrix $\mathbf{B}$ so they can only differ across lags through the scalar weights $\alpha_{t,\ell}$ and it is easy to show that their rank can be lower than n, in particular, we have thatrank$(\mathbf{A}_{t,\ell}) = rank(\mathbf{B}) \leq \min\{n, d\}$ for all t, ℓ. Additionally, since the attention weights $\alpha_{t,\ell}$ sum to one, we also have $\sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} = \mathbf{B}$ for all t, so the sum of coefficients is time-invariant even when individual lag coefficients vary with the state. The multi-head extension overcomes these restrictions.

Setting $\mathbf{W}_q\mathbf{W}_x = 0$ and $\mathbf{W}_k\mathbf{W}_x = 0$ makes the attention scores depend only on the positional effects, $s_{t,\ell} = \boldsymbol{p}_1^\top(\mathbf{W}_q)^\top\mathbf{W}_k\boldsymbol{p}_\ell/\sqrt{d_k}$, so the weights $\alpha_{t,\ell}$ are constant and the model reduces to a time-invariant VAR(L). However, the proportionality across lags places restrictions still holds. In particular, under Assumption 1, a time-invariant VAR(L) conditional mean with $(\mathbf{A}_1, \ldots, \mathbf{A}_L) \neq (0, \ldots, 0)$ can be obtained from the single-head attention model if and only if there exist scalars $a_1, \ldots, a_L > 0$ with $\sum_{\ell=1}^{L} a_\ell = 1$ and a matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$ such that

$$\mathbf{A}_\ell = a_\ell\mathbf{B}, \qquad \ell = 1, \ldots, L. \tag{11}$$

The proportionality constraint in Equations (10)-(11) is the key limitation of the single-head model: all lag coefficient matrices must be scalar multiples of a single $\mathbf{B}$, so every variable pair shares the same lag profile up to scale. Even in the time-invariant case, many economically meaningful VARs violate this; for instance, a bivariate VAR where variable 1 Granger-causes variable 2 at lag 1 but not at lag 2 cannot satisfy Equation (11) unless the entire lag-2 coefficient matrix is zero.

Example 1 (AR(2) with state-dependent coefficients)**.** *Consider a univariate model ($n = 1$) with $L = 2$ lags and constant conditional variance σ^2. Set identity embeddings $x_{t-\ell} = y_{t-\ell}$, i.e. $p_\ell = 0$ and $W_x = 1$. Set $W_q = W_k = \sqrt{\beta}$ (scalars, $\beta \geq 0$), $W_v = 1$, $b_\mu = 0$ and $W_\mu = \varphi$ with $|\varphi| < 1$. Then $B = W_\mu W_v W_x = \varphi$, the scores are*

$$s_{t,1} = \beta y_{t-1}^2, \qquad s_{t,2} = \beta y_{t-1}y_{t-2},$$

and the attention weights are

$$\alpha_{t,1} = \frac{\exp(\beta y_{t-1}^2)}{\exp(\beta y_{t-1}^2) + \exp(\beta y_{t-1}y_{t-2})} \text{ and } \alpha_{t,2} = 1 - \alpha_{t,1}$$

yielding the TV-AR(2):

$$y_t = \underbrace{\varphi\alpha_{t,1}}_{\phi_{1,t}} y_{t-1} + \underbrace{\varphi(1-\alpha_{t,1})}_{\phi_{2,t}} y_{t-2} + \varepsilon_t, \qquad \mathbb{E}[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0, \quad \mathrm{Var}(\varepsilon_t \mid \mathcal{F}_{t-1}) = \sigma^2.$$

This is an observation-driven, smooth-transition autoregression with constant variance. The coefficients $(\varphi\alpha_{t,1}, \varphi\alpha_{t,2})$ *are nonnegative convex combinations scaled by* φ*, so they sum to* φ*. The attention mechanism acts as soft lag selection: since* $s_{t,1} > s_{t,2}$ *if and only if* $y_{t-1}(y_{t-1} - y_{t-2}) > 0$*, the model places more weight on lag 1 when* y_{t-1} *and the recent change have the same sign. The parameter* β *controls the sharpness of this selection, with* $\beta \to 0$ *yielding uniform weights and* $\beta \to \infty$ *approaching a hard argmax.*

2.6 Algorithm for the Single Head Model

We now describe a Gibbs sampler that draws from the posterior of $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ under Gaussian errors. The sampler alternates between conjugate draws, where the conditional posterior is available in closed form, and Elliptical Slice Sampling (ESS) steps for the remaining nonlinear blocks. For $t = L+1, \ldots, T$, assume:

$$\boldsymbol{y}_t \mid \boldsymbol{y}_{t-L:t-1}, \boldsymbol{\theta} \sim \mathcal{N}\big(g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}),\, \boldsymbol{\Sigma}\big),$$

with $\boldsymbol{\Sigma} \in \mathcal{S}^n_{++}$. Conditional on $\boldsymbol{\Sigma}$, the log-likelihood (up to an additive constant) equals:

$$\log L(\boldsymbol{\theta}) = -\frac{1}{2} \sum_{t=L+1}^{T} \big(\boldsymbol{y}_t - g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1})\big)^\top \boldsymbol{\Sigma}^{-1} \big(\boldsymbol{y}_t - g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1})\big).$$

We impose the normalizations in Corollary 1: set $\mathbf{W}_v = \mathbf{I}_d$, $\boldsymbol{b}_\mu = \mathbf{0}$, parameterize the score kernel $\mathbf{F} = \mathbf{W}_q^\top \mathbf{W}_k$ directly, and set $b_1 = 0$. The parameter vector reduces to $\boldsymbol{\theta} = (\mathbf{W}_x, \boldsymbol{p}_1, \ldots, \boldsymbol{p}_L, \mathbf{W}_\mu, \mathbf{F}, b_2, \ldots, b_L)$. This direct parameterization treats $\mathbf{F}$ as the primitive $d \times d$ kernel; it is the natural choice when one wants the unrestricted case $d_k \geq d$. If instead one wishes to enforce the low-rank restriction associated with $d_k < d$, one can replace the $\mathbf{F}$ block by a factorization $\mathbf{F} = \mathbf{U}\mathbf{V}^\top$ with $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{d \times d_k}$ and sample those factors under additional identifying restrictions. Assign independent Gaussian priors $\boldsymbol{\vartheta}_b \sim \mathcal{N}(\boldsymbol{m}_b, \mathbf{V}_b)$ to each block and an inverse-Wishart prior $\boldsymbol{\Sigma} \sim \mathrm{IW}(\nu_0, \mathbf{S}_0)$ to the error covariance.

Conjugate blocks. Since $\boldsymbol{b}_\mu = \mathbf{0}$ and $\mathbf{W}_v = \mathbf{I}_d$, the conditional mean is $\boldsymbol{y}_t = \mathbf{W}_\mu \boldsymbol{h}_t + \boldsymbol{\varepsilon}_t$ with $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, which is linear in $\mathbf{W}_\mu$ given the state vectors $\boldsymbol{h}_{L+1}, \ldots, \boldsymbol{h}_T$. Stacking $\boldsymbol{\Theta} = \mathbf{W}_\mu^\top$ $(d \times n)$, $\mathbf{X} = [\boldsymbol{h}_{L+1}, \ldots, \boldsymbol{h}_T]^\top$ $(T_{\text{eff}} \times d)$, and $\mathbf{Y} = [\boldsymbol{y}_{L+1}, \ldots, \boldsymbol{y}_T]^\top$ $(T_{\text{eff}} \times n)$, the

model takes the matrix regression form $\mathbf{Y} = \mathbf{X}\boldsymbol{\Theta}+\mathbf{E}$. With prior $\boldsymbol{\Theta} \mid \boldsymbol{\Sigma} \sim \mathrm{MN}(\boldsymbol{\Theta}_0, \sigma_0^{-2}\mathbf{I}_d, \boldsymbol{\Sigma})$, the conditional posterior is

$$\boldsymbol{\Theta} \mid \boldsymbol{\Sigma}, \boldsymbol{h}_{1:T}, \boldsymbol{y} \sim \mathrm{MN}(\bar{\boldsymbol{\Theta}}, \mathbf{N}_T^{-1}, \boldsymbol{\Sigma}), \qquad \mathbf{N}_T = \sigma_0^{-2}\mathbf{I}_d + \mathbf{X}^\top\mathbf{X}, \qquad \bar{\boldsymbol{\Theta}} = \mathbf{N}_T^{-1}(\sigma_0^{-2}\boldsymbol{\Theta}_0 + \mathbf{X}^\top\mathbf{Y}). \tag{12}$$

With prior $\boldsymbol{\Sigma} \sim \mathrm{IW}(\nu_0, \mathbf{S}_0)$, the conditional posterior of the error covariance is

$$\boldsymbol{\Sigma} \mid \boldsymbol{\Theta}, \boldsymbol{y} \sim \mathrm{IW}\big(\nu_0 + T_{\text{eff}} + d,\ \mathbf{S}_0 + \mathbf{E}^\top\mathbf{E} + \sigma_0^{-2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)^\top(\boldsymbol{\Theta} - \boldsymbol{\Theta}_0)\big), \tag{13}$$

where $\mathbf{E} = \mathbf{Y} - \mathbf{X}\boldsymbol{\Theta}$ are the residuals. The second quadratic term in the scale matrix arises from the matrix-normal prior on $\boldsymbol{\Theta}$.

ESS blocks. The remaining four blocks, $\mathbf{W}_x$, $\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L\}$, $\mathbf{F}$, and $(b_2, \ldots, b_L)$, enter the likelihood nonlinearly through the attention weights and are updated by Elliptical Slice Sampling (ESS; Murray et al., 2010). Algorithm 1 describes the ESS step and Algorithm 2 describes the full Gibbs sampler. For each ESS block $\boldsymbol{\vartheta}_b$, hold $\boldsymbol{\theta}_{-\boldsymbol{\vartheta}_b}$ fixed and define $L(\boldsymbol{\vartheta}_b) = L(\boldsymbol{\vartheta}_b, \boldsymbol{\theta}_{-\boldsymbol{\vartheta}_b})$, then the posterior is $p(\boldsymbol{\vartheta}_b \mid \boldsymbol{\theta}_{-\boldsymbol{\vartheta}_b}, \boldsymbol{y}) \propto L(\boldsymbol{\vartheta}_b)\pi(\boldsymbol{\vartheta}_b)$. If we let $\ell(\boldsymbol{\vartheta}_b) = \log L(\boldsymbol{\vartheta}_b)$ and given the current value $\hat{\boldsymbol{\vartheta}}_b$ the following ESS step produces a draw from $p(\boldsymbol{\vartheta}_b \mid \boldsymbol{\theta}_{-\boldsymbol{\vartheta}_b}, \boldsymbol{y})$.

Algorithm 1 Elliptical Slice Sampling (ESS) update for one block $\boldsymbol{\vartheta}_b$

Draw $\boldsymbol{\nu} \sim \mathcal{N}(0, \mathbf{V}_b)$.
Draw $u \sim \mathrm{Unif}(0, 1)$ and set $\log \hat{y}_b \leftarrow \ell(\hat{\boldsymbol{\vartheta}}_b) + \log u$.
Draw $\phi \sim \mathrm{Unif}(0, 2\pi)$ and set $[\phi_{\min}, \phi_{\max}] \leftarrow [\phi - 2\pi, \phi]$.
loop
    Propose $\boldsymbol{\vartheta}(\phi) \leftarrow \boldsymbol{m}_b + (\hat{\boldsymbol{\vartheta}}_b - \boldsymbol{m}_b)\cos\phi + \boldsymbol{\nu}\sin\phi$.
    if $\ell(\boldsymbol{\vartheta}(\phi)) \geq \log \hat{y}_b$ **then**
        return $\boldsymbol{\vartheta}(\phi)$.
    else
        if $\phi < 0$ **then**
            $\phi_{\min} \leftarrow \phi$.
        else
            $\phi_{\max} \leftarrow \phi$.
        end if
        Draw a new $\phi \sim \mathrm{Unif}(\phi_{\min}, \phi_{\max})$ and repeat.
    end if
end loop

Using Algorithm 1, we define the following Gibbs sampler, whose stationary distribution is $p(\boldsymbol{\theta}, \boldsymbol{\Sigma} \mid \boldsymbol{y}_{L+1:T})$.[4]

[4]We initialize the sampler at a MAP estimate obtained by profile optimization: for fixed attention

Algorithm 2 Gibbs sampler for the single-head attention VAR

Require: $\mathbf{W}_v = \mathbf{I}_d$, $\boldsymbol{b}_\mu = \mathbf{0}$, $\mathbf{F} = \mathbf{W}_q^\top \mathbf{W}_k$ (Corollary 1)
Require: Initial values $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$; prior hyperparameters $(\boldsymbol{m}_b, \mathbf{V}_b)_{b=1}^B$, ν_0, $\mathbf{S}_0$
for $m = 1, \dots, M$ **do**
  ▷ Conjugate blocks
  Draw $\mathbf{W}_\mu \mid \boldsymbol{h}_{1:T}, \boldsymbol{\Sigma}, \boldsymbol{y}$ from the matrix-normal posterior
  Draw $\boldsymbol{\Sigma} \mid \mathbf{W}_\mu, \boldsymbol{h}_{1:T}, \boldsymbol{y}$ from the inverse-Wishart posterior
  ▷ ESS blocks
  for $b \in \{\mathbf{W}_x,\ \{\boldsymbol{p}_1, \dots, \boldsymbol{p}_L\},\ \mathbf{F}\}$ **do**
    Update $\boldsymbol{\vartheta}_b$ using Algorithm 1
  end for
end for

Each ESS block updates one group of parameters while holding all others fixed. The target for each block is the full conditional $p(\boldsymbol{\vartheta}_b \mid \boldsymbol{\theta}_{-\boldsymbol{\vartheta}_b}, \boldsymbol{\Sigma}, \boldsymbol{y}) \propto L(\boldsymbol{\vartheta}_b)\pi(\boldsymbol{\vartheta}_b)$, where the prior $\pi(\boldsymbol{\vartheta}_b)$ is Gaussian by construction and the likelihood enters only through the whitened residuals. During the ESS sweep, $\boldsymbol{\Sigma}$ and $\mathbf{W}_\mu$ are held fixed, so the Cholesky factor $\mathbf{L}$ of $\boldsymbol{\Sigma}$ and the whitened output matrix $\mathbf{L}^{-1}\mathbf{W}_\mu$ can be computed once and reused across all ESS evaluations. The whitened log-likelihood (up to a constant) reduces to

$$\ell(\boldsymbol{\vartheta}_b) = -\tfrac{1}{2} \sum_{t=L+1}^{T} \left\| \mathbf{L}^{-1}(\boldsymbol{y}_t - \mathbf{W}_\mu \boldsymbol{h}_t) \right\|^2,$$

where $\boldsymbol{h}_t$ depends on $\boldsymbol{\vartheta}_b$ through the attention mechanism: updating $\mathbf{W}_x$ changes the embedded context vectors $\boldsymbol{x}_{t-\ell}$ and hence the attention weights and aggregated state; updating the positional embeddings $\{\boldsymbol{p}_\ell\}$ shifts all context vectors by a lag-specific constant; and updating the score kernel $\mathbf{F}$ changes the pairwise scores and therefore the attention weights directly. Each of these modifications propagates to $\boldsymbol{h}_t$ and thus to the residuals, but the dependence is nonlinear, which is why ESS rather than a conjugate draw is required.

The lower-triangular constraint on $\mathbf{W}_x$ (Corollary 1(i)) is enforced via a selection matrix that maps a reduced vector of free parameters to $\mathrm{vec}(\mathbf{W}_x)$: above-diagonal elements are fixed at zero, and diagonal elements are sampled in log-space to ensure positivity.

2.7 Application: Univariate forecasting with a single head

We illustrate Algorithm 2 in a pseudo real-time out-of-sample forecasting exercise for U.S. aggregate monthly PCE inflation. The sample runs from January 1960 through January

parameters $(\mathbf{W}_x, \{\boldsymbol{p}_\ell\}, \mathbf{F})$, the optimal $\mathbf{W}_\mu$ and $\boldsymbol{\Sigma}$ are available in closed form from (12)–(13), so gradient-based optimization operates only on the lower-dimensional nonlinear block.

2026. We use the January 2026 vintage, abstracting from revisions. We use the minimal single-head specification ($H = 1$, $d = d_k = d_v = 1$, $L = 12$) with homoskedastic errors. The model admits a representation as a time-varying AR(12). The benchmark is a (univariate) Bayesian VAR with the same lag length ($p = L = 12$) and Minnesota prior.[5] Both models are recursively re-estimated over expanding windows from December 2019 to January 2026. At each date we generate predictive densities for horizons $h = 1, \ldots, 24$ months; the forecast target is annualized cumulative inflation $(12/h) \sum_{j=1}^{h} y_{t_0+j}$.

We first assess the convergence properties of the algorithm. Figure 1 reports trace plots from four independent chains initialized at dispersed starting values. The traces of the reduced-form coefficient $B_{1,1}$, the error variance $\Sigma_{1,1}$, the posterior mean attended lag, and the score bias b_2 all mix rapidly and converge to the same stationary distribution, with no visible dependence on initial conditions after the burn-in period.

The state dependence of the attention weights allows the effective autoregressive coefficients to vary with the recent history of inflation. Figure 2 illustrates this by plotting frozen-coefficient impulse response functions at the two dates whose profiles differ most across the sample. The impulse responses are computed by fixing the attention weights at their values for a given date and iterating the resulting VAR(L) coefficients forward. The figure reveals modest time variation in persistence: shocks decay somewhat faster at the 2008 date than at the 1980 date, but the difference is small in magnitude. This is a direct consequence of the single-head restriction (Proposition 2): with $H = 1$, all coefficient matrices $A_{t,\ell}$ are proportional to a single base matrix B, and the only source of time variation is the reweighting of lags through the scalar attention weights $\alpha_{t,\ell}$. The model can shift which lags receive more weight, but it cannot change the shape of the response to each lag independently. Richer time variation in impulse response profiles requires the multi-head specification of Section 3, where $H > 1$ heads allow the coefficient matrices to move in a higher-dimensional subspace.

Despite these limitations, the single-head model produces sizable forecast improvements. Table 1 reports root mean squared errors and average log predictive scores (Gneiting and Raftery, 2007) at selected horizons. The attention-based VAR matches the Bayesian VAR at short horizons and increasingly dominates at longer ones, with RMSE improvements of 12% at $h = 12$ and 19% at $h = 24$. Log predictive scores favor the attention-based VAR at every horizon. Figure 3 illustrates these results. Panel (a) plots posterior median twelve-month-ahead forecasts against realized values. The two models produce broadly similar forecasts, but the attention-based VAR forecasts higher annualized inflation during the 2020–2021

[5]The Minnesota prior overall tightness, lag decay, and the sum-of-coefficients and dummy-initial-observation hyperparameters are chosen by maximizing the marginal likelihood following Giannone et al. (2015).

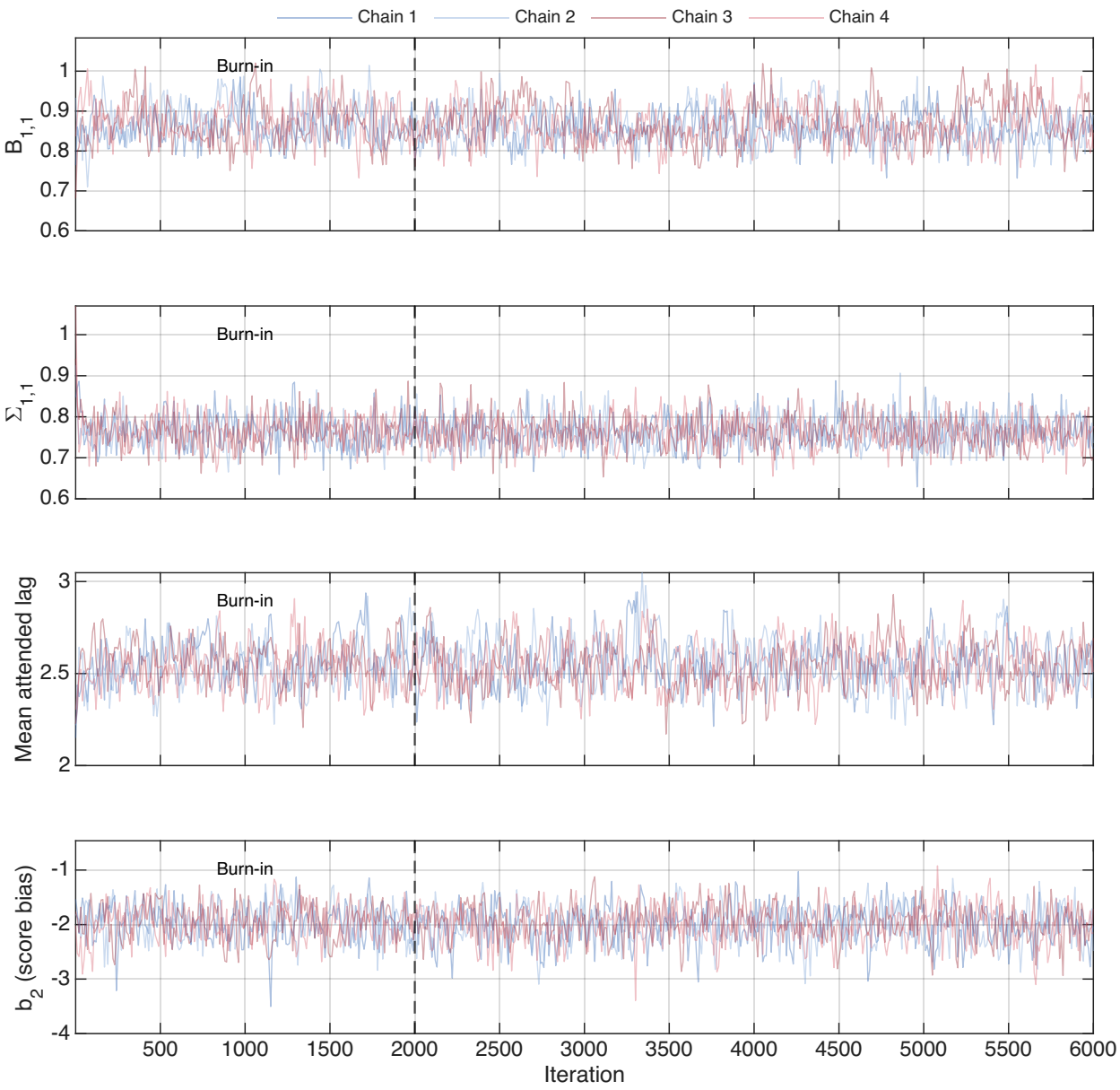
Table 1: Out-of-sample forecast comparison: PCE inflation

	RMSE			Log predictive score	
h	Benchmark	BVAR	ATVAR	BVAR	ATVAR
1	2.33	2.18	2.16	−2.42	−2.21
3	1.91	1.64	1.61	−1.92	−1.91
6	1.93	1.54	1.45	−1.84	−1.82
12	2.32	1.79	1.58	−2.03	−1.89
24	2.89	2.21	1.80	−2.33	−2.02

Notes: ATVAR denotes the single-head attention-based VAR ($H = 1$, $d = d_k = d_v = 1$, $L = 12$). BVAR is a Bayesian VAR(12) with Minnesota prior. Benchmark is the trailing twelve-month average. The forecast target is annualized cumulative inflation $(12/h)\sum_{j=1}^{h} y_{t_0+j}$. Expanding windows from 2019M12 to 2026M01. Log predictive scores are averages across evaluation dates; higher (less negative) is better.

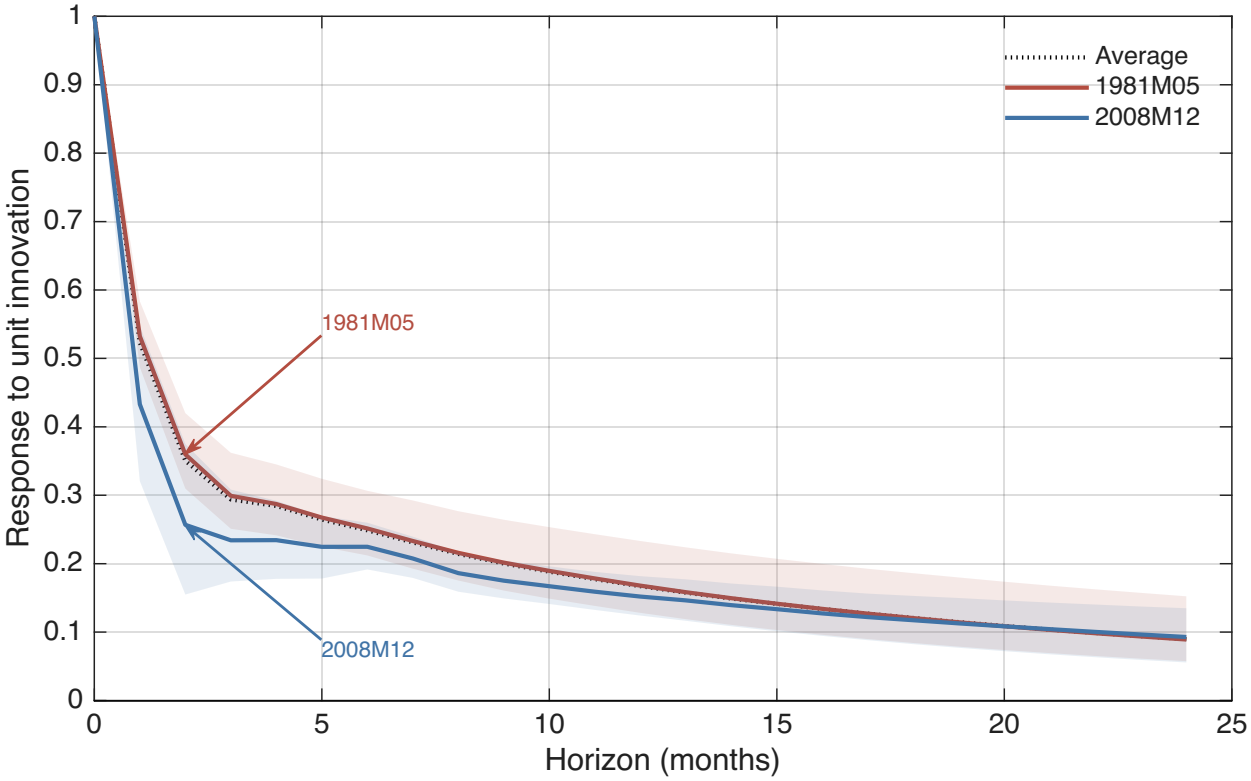
run-up (e.g., 3.2% vs. 2.7% from 2020M12, against a realized 6.1%), and predicts a faster disinflation through mid-2022 (e.g., 4.3% vs. 5.4% from 2022M07, against a realized 3.4%). Panels (b) and (c) compare the predictive densities as of January 2026, where the attention-based VAR predicts a faster acceleration of inflation.

Figure 1: Convergence diagnostics for the attention VAR sampler



Notes: Each panel shows the trace of an identified quantity across four independent chains initialized at different seeds. From top to bottom: the scalar coefficient B, the variance $\Sigma_{1,1}$, the posterior mean attended lag $\bar{\ell}_t = \sum_\ell \ell\, \alpha_{t,\ell}$, and the score bias b_2. The dashed vertical line marks the end of the burn-in period. All chains mix over the same range after burn-in, indicating convergence.

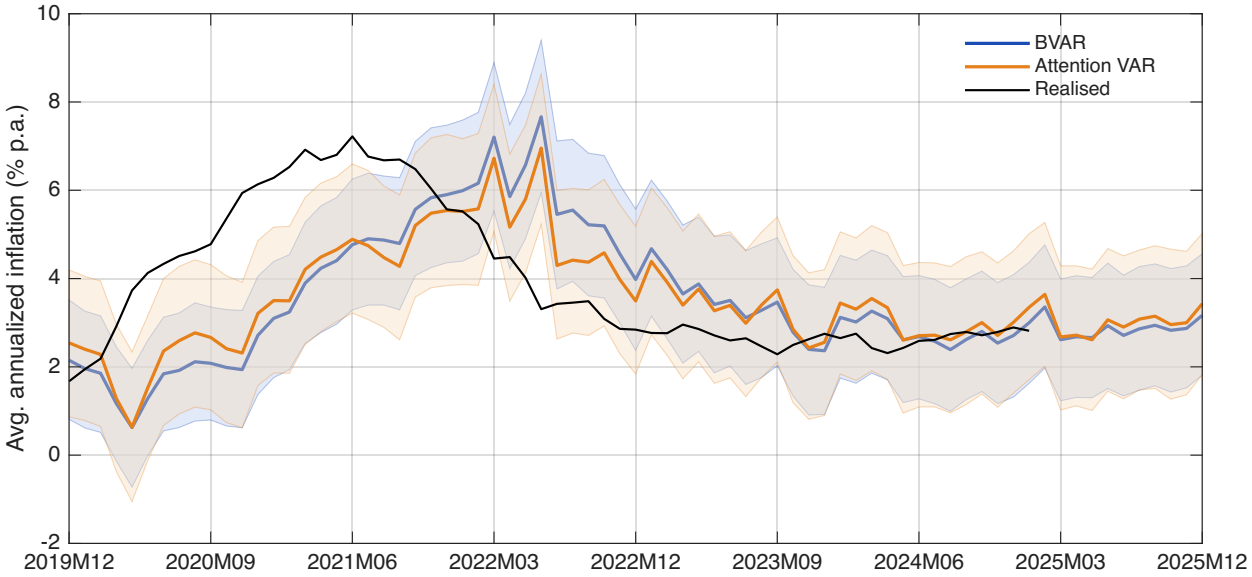
Figure 2: Time-varying impulse responses to an inflation innovation



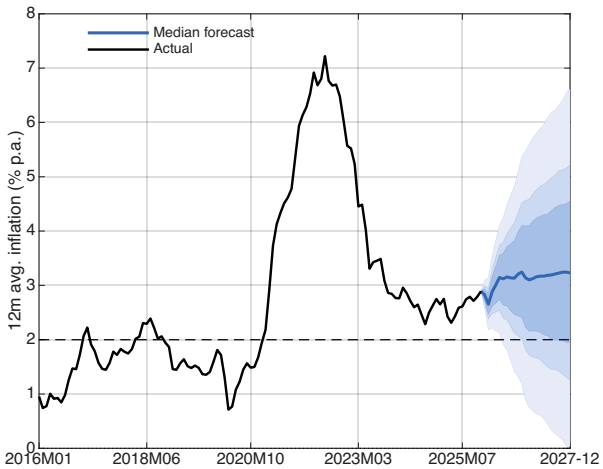
Notes: Impulse response functions to a one-standard-deviation inflation innovation at selected dates, computed from the posterior of the attention VAR parameters. The red and blue lines show the response at the two dates with the most and least persistent dynamics, respectively, as measured by the frozen-coefficient impulse response. The dashed black line shows the average impulse response across all time periods. Shaded bands show the 68% posterior credible interval at each date. The attention mechanism allows the persistence of inflation shocks to vary with economic conditions without imposing a specific law of motion for the time variation.

Figure 3: Out-of-sample forecasting: PCE inflation

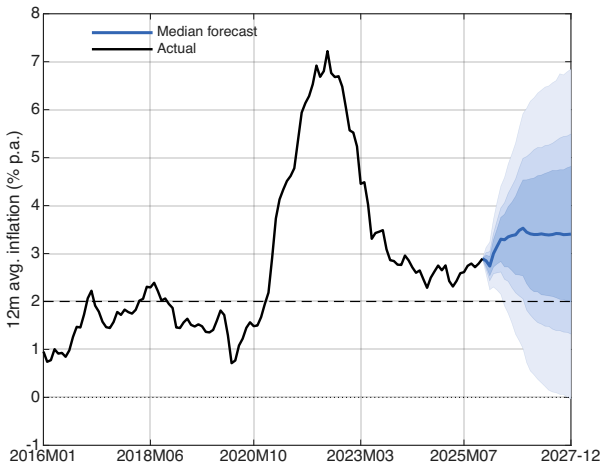
(a) Forecast versus realized: $h = 12$ months



(b) Bayesian VAR



(c) Attention-based VAR



Notes: Panel (a): twelve-month-ahead cumulative inflation forecasts from the attention VAR and the Bayesian VAR, evaluated over a recursive out-of-sample exercise from December 2019 to the end of the sample. Shaded bands show the 68% posterior predictive interval; solid lines show the posterior median forecast; the black line shows realized annualized twelve-month average PCE inflation. Panel (b): posterior predictive density of annualized cumulative PCE inflation from the Bayesian VAR, estimated at the last sample vintage. Shaded regions show the 68% and 90% posterior predictive intervals at horizons one through twenty-four months; the solid line shows the posterior median forecast; the dashed line shows historical annualized inflation. Panel (c): same as panel (b) for the attention VAR.

3 Multi-head attention

The proportionality constraint and the related time-invariant coefficient sum show that a single attention head imposes severe restrictions. To overcome these limitations, Vaswani et al. (2017) propose to linearly combine H attention heads, each with separate parameters that follow the single-head attention structure (4)–(8).

In other words, for each head $m \in \{1, \dots, H\}$, define projection matrices $\mathbf{W}_q^m, \mathbf{W}_k^m \in \mathbb{R}^{d_k \times d}$ and $\mathbf{W}_v^m \in \mathbb{R}^{d_v \times d}$. Using a common set of embeddings as in (4), each head computes its own queries, keys, values, scores, weights, and output:

$$\boldsymbol{q}_t^m = \mathbf{W}_q^m \boldsymbol{x}_{t-1}, \quad \boldsymbol{k}_{t-\ell}^m = \mathbf{W}_k^m \boldsymbol{x}_{t-\ell}, \quad \boldsymbol{v}_{t-\ell}^m = \mathbf{W}_v^m \boldsymbol{x}_{t-\ell}, \tag{14}$$

$$s_{t,\ell}^m = \frac{(\boldsymbol{q}_t^m)^\top \boldsymbol{k}_{t-\ell}^m}{\sqrt{d_k}}, \quad \alpha_t^m = \mathrm{softmax}(s_t^m), \tag{15}$$

$$\boldsymbol{c}_t^m = \sum_{\ell=1}^{L} \alpha_{t,\ell}^m \, \boldsymbol{v}_{t-\ell}^m \in \mathbb{R}^{d_v}, \tag{16}$$

for head $m = 1, \dots, H$ and $\ell = 1, \dots, L$. Then, the head outputs are combined via

$$\boldsymbol{h}_t = \mathbf{W}^O \begin{pmatrix} \boldsymbol{c}_t^1 \\ \vdots \\ \boldsymbol{c}_t^H \end{pmatrix} = \sum_{m=1}^{H} \mathbf{W}^{O,m} \boldsymbol{c}_t^m \in \mathbb{R}^d, \tag{17}$$

where $\mathbf{W}^O = [\mathbf{W}^{O,1} \cdots \mathbf{W}^{O,H}]$ with $\mathbf{W}^{O,m} \in \mathbb{R}^{d \times d_v}$. The conditional mean Equation (3) becomes

$$g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) = \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t = \boldsymbol{b}_\mu + \mathbf{W}_\mu \sum_{m=1}^{H} \mathbf{W}^{O,m} \sum_{\ell=1}^{L} \alpha_{t,\ell}^m \, \boldsymbol{v}_{t-\ell}^m, \tag{18}$$

where $\boldsymbol{\theta} = \{\boldsymbol{b}_\mu, \mathbf{W}_\mu, \boldsymbol{p}, \mathbf{W}_x, \mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v, \mathbf{W}^O\}$ and
$\Theta = \mathbb{R}^n \times \mathbb{R}^{n \times d} \times \mathbb{R}^{Ld} \times \mathbb{R}^{d \times n} \times \mathbb{R}^{d_k \times Hd} \times \mathbb{R}^{d_k \times Hd} \times \mathbb{R}^{d_v \times Hd} \times \mathbb{R}^{d \times Hd_v}$.

Assumption 2. *The model uses H-head attention with: (i) context length L and H heads; (ii) affine conditional mean as in Equation (18); (iii) linear embeddings as in Equation (4); (iv) linear projections as in Equation (14); (v) attention weights as in Equation (15); (vi) aggregation as in Equations (16)–(17).*

Proposition 3 (TV-VAR Representation under H-head attention)**.** *Under Assumption 2, the conditional mean satisfies*

$$g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}) = \boldsymbol{b}_t + \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \, \boldsymbol{y}_{t-\ell}, \tag{19}$$

where

$$\mathbf{A}_{t,\ell} = \sum_{m=1}^{H} \alpha_{t,\ell}^{m} \mathbf{B}^{m}, \qquad \mathbf{B}^{m} = \mathbf{W}_{\mu} \mathbf{W}^{O,m} \mathbf{W}_{v}^{m} \mathbf{W}_{x} \in \mathbb{R}^{n \times n}, \tag{20}$$

with $\operatorname{rank}(\mathbf{B}^{m}) \leq \min\{d_v, d, n\}$*, and* $\boldsymbol{b}_t = \boldsymbol{b}_{\mu} + \sum_{m=1}^{H} \sum_{\ell=1}^{L} \alpha_{t,\ell}^{m} \mathbf{W}_{\mu} \mathbf{W}^{O,m} \mathbf{W}_{v}^{m} \boldsymbol{p}_{\ell}$*. Each* $\mathbf{A}_{t,\ell}$ *lies in the linear span of* $\{\mathbf{B}^{m}\}_{m=1}^{H}$*, a subspace of* $\mathbb{R}^{n \times n}$ *with dimension at most* H*.*

Proof. The proof follows by substituting the aggregation Equation (17) into the conditional mean equation and collecting terms. □

We now derive the conditions under which the multi-head model can represent an arbitrary constant parameter VAR. For the model to represent a linear VAR with constant coefficients, attention weights must not depend on the data.[6] The score for head m at lag ℓ is

$$s_{t,\ell}^{m} = \frac{(\mathbf{W}_{q}^{m} \boldsymbol{x}_{t-1})^{\top} (\mathbf{W}_{k}^{m} \boldsymbol{x}_{t-\ell})}{\sqrt{d_k}}.$$

To make scores not depend on the data, one imposes the following restrictions:

$$\mathbf{W}_{q}^{m} \mathbf{W}_{x} = 0 \quad \text{and} \quad \mathbf{W}_{k}^{m} \mathbf{W}_{x} = 0 \quad \text{for all } m. \tag{21}$$

Under Equation (21), the scores are constant in t. The value projections $\boldsymbol{v}_{t-\ell}^{m} = \mathbf{W}_{v}^{m}(\boldsymbol{p}_{\ell} + \mathbf{W}_{x} \boldsymbol{y}_{t-\ell})$ remain a function of lagged data, preserving the VAR structure.

Proposition 4 (Representation of time-invariant VARs)**.** *Under Assumption 2, any time-invariant VAR(L) conditional mean can be obtained from the H-head attention model with context length L and deterministic attention weights* $\boldsymbol{\alpha}^{m} = (\alpha_1^{m}, \ldots, \alpha_L^{m})^{\top} \in \mathbb{R}^{L}$ *satisfying Equation (21). Suppose:*

(i) $H \geq L$ *and the* $L \times H$ *mixing matrix*

$$\boldsymbol{\mathcal{M}} = \begin{pmatrix} \alpha_1^1 & \cdots & \alpha_1^H \\ \vdots & \ddots & \vdots \\ \alpha_L^1 & \cdots & \alpha_L^H \end{pmatrix} \in \mathbb{R}^{L \times H}$$

has full row rank L*;*

(ii) $d, d_v \geq n$*, and let* $\mathbf{W}_{\mu} \in \mathbb{R}^{n \times d}$ *and* $\mathbf{W}_{x} \in \mathbb{R}^{d \times n}$ *be any matrices with* $\operatorname{rank}(\mathbf{W}_{\mu}) = \operatorname{rank}(\mathbf{W}_{x}) = n$*.*

[6] Since $\operatorname{softmax}(s)_{\ell} > 0$ for all $s \in \mathbb{R}^{L}$, every lag receives strictly positive weight, so a VAR(p) with $p < L$ (which requires zero weight on lags $\ell > p$) can only be approximated, not represented exactly. However, by driving the positional scores for lags $\ell > p$ to $-\infty$, the weights on those lags can be made arbitrarily small, so the approximation is as tight as desired. For a VAR(L) that uses all L lags, the issue does not arise.

Then for any target VAR coefficients $(\mathbf{A}_1, \ldots, \mathbf{A}_L)$ *with* $\mathbf{A}_\ell \in \mathbb{R}^{n\times n}$*, one can construct matrices*

$$\{\mathbf{W}^{O,m}\}_{m=1}^{H} \text{ and } \{\mathbf{W}_v^m\}_{m=1}^{H}$$

such that

$$\mathbf{A}_\ell = \sum_{m=1}^{H} \alpha_\ell^m \mathbf{B}^m \text{ where } \mathbf{B}^m = \mathbf{W}_\mu \mathbf{W}^{O,m} \mathbf{W}_v^m \mathbf{W}_x \text{ for } \ell = 1, \ldots, L$$

and for each (i, j)*, we have that*

$$\boldsymbol{b}_{ij} = \boldsymbol{\mathcal{M}}^\top (\boldsymbol{\mathcal{M}}\boldsymbol{\mathcal{M}}^\top)^{-1} \boldsymbol{a}_{ij}, \tag{22}$$

where $\boldsymbol{a}_{ij} = ([\mathbf{A}_1]_{ij}, \ldots, [\mathbf{A}_L]_{ij})^\top \in \mathbb{R}^L$ *and* $\boldsymbol{b}_{ij} = ([\mathbf{B}^1]_{ij}, \ldots, [\mathbf{B}^H]_{ij})^\top \in \mathbb{R}^H$*. When* $H > L$*, the resulting* $\{\mathbf{B}^m\}_{m=1}^{H}$ *are not unique.*

Proof. For each (i, j), define $\boldsymbol{a}_{ij} = ([\mathbf{A}_1]_{ij}, \ldots, [\mathbf{A}_L]_{ij})^\top$. Since

$$[\mathbf{A}_\ell]_{ij} = \sum_{m=1}^{H} \alpha_\ell^m [\mathbf{B}^m]_{ij},$$

we have the linear system $\boldsymbol{a}_{ij} = \boldsymbol{\mathcal{M}}\boldsymbol{b}_{ij}$ with $\boldsymbol{b}_{ij} = ([\mathbf{B}^1]_{ij}, \ldots, [\mathbf{B}^H]_{ij})^\top$. Because $\boldsymbol{\mathcal{M}}$ has full row rank, this system is solvable for every $\boldsymbol{a}_{ij}$, and Equation (22) gives the minimum-norm solution. Stacking all entries (i, j) defines matrices $\mathbf{B}^1, \ldots, \mathbf{B}^H$ satisfying $\mathbf{A}_\ell = \sum_{m=1}^{H} \alpha_\ell^m \mathbf{B}^m$ for all ℓ.

It remains to construct $\mathbf{W}^{O,m}$ and $\mathbf{W}_v^m$ for each m. Since $\operatorname{rank}(\mathbf{W}_\mu) = n$, pick any right inverse $\mathbf{R}_\mu \in \mathbb{R}^{d\times n}$ such that $\mathbf{W}_\mu \mathbf{R}_\mu = \mathbf{I}_n$. Since $\operatorname{rank}(\mathbf{W}_x) = n$, pick any left inverse $\boldsymbol{L}_e \in \mathbb{R}^{n\times d}$ such that $\boldsymbol{L}_e \mathbf{W}_x = \mathbf{I}_n$. Define

$$\widetilde{\mathbf{M}}_m = \mathbf{R}_\mu \mathbf{B}^m \boldsymbol{L}_e \in \mathbb{R}^{d\times d}.$$

Then $\operatorname{rank}(\widetilde{\mathbf{M}}_m) \le \operatorname{rank}(\mathbf{B}^m) \le n \le d_v$. Hence there exist matrices $\mathbf{W}^{O,m} \in \mathbb{R}^{d\times d_v}$ and $\mathbf{W}_v^m \in \mathbb{R}^{d_v\times d}$ such that $\widetilde{\mathbf{M}}_m = \mathbf{W}^{O,m}\mathbf{W}_v^m$. Therefore,

$$\mathbf{W}_\mu \mathbf{W}^{O,m} \mathbf{W}_v^m \mathbf{W}_x = \mathbf{W}_\mu \widetilde{\mathbf{M}}_m \mathbf{W}_x = \mathbf{W}_\mu \mathbf{R}_\mu \mathbf{B}^m \boldsymbol{L}_e \mathbf{W}_x = \mathbf{B}^m.$$

Thus the required matrices exist. Non-uniqueness follows because when $H > L$, $\boldsymbol{a}_{ij} = \boldsymbol{\mathcal{M}}\boldsymbol{b}_{ij}$ has infinitely many solutions; moreover, the choices of inverses and matrix factorization are also non-unique. □

Example 2 (Two-Head Representation of VAR(2))**.** *Consider a bivariate VAR(2) with arbitrary target coefficients* $(\mathbf{A}_1, \mathbf{A}_2)$*. With* $H = L = 2$ *heads, let the deterministic attention weights be* $\boldsymbol{\alpha}^1 = (a, 1-a)^\top$ *and* $\boldsymbol{\alpha}^2 = (b, 1-b)^\top$ *with* $a \neq b$ *and* $a, b \in (0,1)$*. The mixing matrix is*

$$\boldsymbol{\mathcal{M}} = \begin{pmatrix} a & b \\ 1-a & 1-b \end{pmatrix}, \qquad \det(\boldsymbol{\mathcal{M}}) = a(1-b) - b(1-a) = a - b \neq 0.$$

Since $\boldsymbol{\mathcal{M}}$ *is invertible, the unique solution for* $(\mathbf{B}^1, \mathbf{B}^2)$ *is:*

$$\mathbf{B}^1 = \frac{(1-b)\mathbf{A}_1 - b\mathbf{A}_2}{a-b}, \qquad \mathbf{B}^2 = \frac{-(1-a)\mathbf{A}_1 + a\mathbf{A}_2}{a-b}. \tag{23}$$

Direct substitution verifies $\mathbf{A}_1 = a\mathbf{B}^1 + b\mathbf{B}^2 = \mathbf{A}_1$ *and* $\mathbf{A}_2 = (1-a)\mathbf{B}^1 + (1-b)\mathbf{B}^2 = \mathbf{A}_2$*. Thus any VAR(2), regardless of the structure of* $(\mathbf{A}_1, \mathbf{A}_2)$*, can be exactly represented by a two-head model with appropriately chosen base matrices, provided the architectural conditions on* (d, d_v) *are satisfied.*

An H-head attention model with $H \geq L$ can therefore represent any linear VAR(L), provided that $d, d_v \geq n$ and $\text{rank}(\mathbf{W}_\mu) = \text{rank}(\mathbf{W}_x) = n$.

3.1 Sum-identity and gain relaxation

As in the single-head case, softmax normalization implies $\sum_{\ell=1}^{L} \alpha_{t,\ell}^m = 1$ for each head m. Summing (20) over lags gives

$$\sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} = \sum_{m=1}^{H} \mathbf{B}^m, \tag{24}$$

which is constant in t even when individual attention weights vary over time. Under stability, the sum-identity fixes the instantaneous long-run multiplier $(\mathbf{I}_n - \sum_\ell \mathbf{A}_{t,\ell})^{-1}$: the model can redistribute weight across lags but cannot change the cumulative feedback matrix.

We introduce a generalization to relax this constraint while preserving the TV-VAR structure. Introduce head-specific, state-dependent gains

$$g_t^m = \exp(\boldsymbol{u}_m^\top \boldsymbol{h}_t^0 + \kappa_m) > 0, \qquad \boldsymbol{h}_t^0 = \sum_{m=1}^{H} \mathbf{W}^{O,m} \boldsymbol{c}_t^m, \tag{25}$$

where $\boldsymbol{u}_m \in \mathbb{R}^d$ and $\kappa_m \in \mathbb{R}$ are learned parameters, adding $H(d+1)$ parameters total. The state $\boldsymbol{h}_t^0$ is the multi-head output before the mean projection, so the gains are observation-driven functions of the same history $\boldsymbol{y}_{t-L:t-1}$ that determines the attention weights. Replac-

ing $\boldsymbol{c}_t^m$ by the gain-scaled output $g_t^m \boldsymbol{c}_t^m$ in the mixing step (17) yields modified coefficient matrices

$$\widetilde{\mathbf{A}}_{t,\ell} = \sum_{m=1}^{H} g_t^m \alpha_{t,\ell}^m \mathbf{B}^m,$$

so that $\sum_\ell \widetilde{\mathbf{A}}_{t,\ell} = \sum_m g_t^m \mathbf{B}^m$ now varies with t. The cross-lag rank structure is preserved: $\operatorname{rank}(\widetilde{\boldsymbol{\mathcal{A}}}_t) \le \min\{H, L\}$. Each head acts as a mode with its own lag shape α_t^m (the relative weights across lags) and gain g_t^m (the overall amplitude).[7]

4 Regularization and economic interpretation

The preceding sections established that multi-head attention can represent any linear VAR when $H \ge L$, $d \ge n$, and $d_v \ge n$ (Proposition 4). When these conditions are not met, the factored coefficient structure $\mathbf{A}_{t,\ell} = \sum_m \alpha_{t,\ell}^m \mathbf{B}^m$ imposes two types of rank restriction on the model's dynamics.

Proposition 5 (Cross-lag and within-lag rank restrictions)**.** *Consider an H-head attention model. For any fixed t, define*

$$\begin{aligned}
\boldsymbol{\mathcal{A}}_t &= \big[\operatorname{vec}(\mathbf{A}_{t,1}) \ \cdots \ \operatorname{vec}(\mathbf{A}_{t,L})\big] \in \mathbb{R}^{n^2 \times L}, \\
\boldsymbol{\mathcal{M}}_t &= [\boldsymbol{\alpha}_t^1 \ \cdots \ \boldsymbol{\alpha}_t^H] \in \mathbb{R}^{L \times H}, \\
\boldsymbol{\mathcal{B}} &= \big[\operatorname{vec}(\mathbf{B}^1) \ \cdots \ \operatorname{vec}(\mathbf{B}^H)\big] \in \mathbb{R}^{n^2 \times H}.
\end{aligned}$$

Then $\boldsymbol{\mathcal{A}}_t = \boldsymbol{\mathcal{B}} \boldsymbol{\mathcal{M}}_t^\top$, hence $\operatorname{rank}(\boldsymbol{\mathcal{A}}_t) \le \min\{H, L\}$. Each $\mathbf{B}^m = \mathbf{W}_\mu \mathbf{W}^{O,m} \mathbf{W}_v^m \mathbf{W}_x$ satisfies $\operatorname{rank}(\mathbf{B}^m) \le \min\{n, d, d_v\}$, and in particular $\operatorname{rank}(\mathbf{B}^m) \le d_v$ when $d_v < n$.

Proof. Immediate from $\mathbf{A}_{t,\ell} = \sum_{m=1}^H \alpha_{t,\ell}^m \mathbf{B}^m$ and standard rank inequalities. □

The first restriction is within-lag reduced rank. Because the same $\mathbf{W}_\mu$ and $\mathbf{W}_x$ appear in every head and every lag, we can factor

$$\mathbf{A}_{t,\ell} = \mathbf{W}_\mu \Big(\sum_{m=1}^{H} \alpha_{t,\ell}^m \mathbf{W}^{O,m} \mathbf{W}_v^m \Big) \mathbf{W}_x,$$

so that $\operatorname{col}(\mathbf{A}_{t,\ell}) \subseteq \operatorname{col}(\mathbf{W}_\mu)$ and $\operatorname{row}(\mathbf{A}_{t,\ell}) \subseteq \operatorname{row}(\mathbf{W}_x)$ for all t and ℓ, with $\operatorname{rank}(\mathbf{A}_{t,\ell}) \le \min\{\operatorname{rank}(\mathbf{W}_\mu), \operatorname{rank}(\mathbf{W}_x)\} \le d$. All coefficient matrices share the same row and column

[7] Unbounded gains can generate explosive dynamics. In practice, one may bound gains via a sigmoid function or regularize the gain parameters $(\boldsymbol{u}_m, \kappa_m)$ during estimation; see Appendix A.2 for the detailed TV-VAR representation with gains.

spaces, determined by the embedding and output projections; the within-head dimension d_v imposes a secondary restriction on individual $\mathbf{B}^m$.

The second restriction is cross-lag reduced rank: the lag-stacked coefficient matrix $\boldsymbol{\mathcal{A}}_t$ has rank at most H, so for each (i,j) the vector of lag coefficients $\boldsymbol{a}_t^{ij} = ([\mathbf{A}_{t,1}]_{ij}, \ldots, [\mathbf{A}_{t,L}]_{ij})^\top$ lies in the span of H common attention weight vectors:

$$\boldsymbol{a}_t^{ij} = \sum_{m=1}^{H} [\mathbf{B}^m]_{ij}\, \boldsymbol{\alpha}_t^m, \qquad \boldsymbol{\alpha}_t^m = (\alpha_{t,1}^m, \ldots, \alpha_{t,L}^m)^\top. \tag{26}$$

When $H \ll L$, the $n^2 L$ lag coefficients in $\boldsymbol{\mathcal{A}}_t$ are governed by only H shared attention weight vectors $\boldsymbol{\alpha}_t^m$ and $n^2 H$ mixing weights $[\mathbf{B}^m]_{ij}$, so the model can accommodate many lags without a proportional increase in the effective number of free parameters. Thus, the attention architecture achieves dimensionality reduction. In particular, for fixed (d, H, d_k, d_v, L), the conditional mean requires

$$P_{\text{Att}} = 2dn + n + \underbrace{dL + H(2d_k + d_v)d + Hd \cdot d_v}_{\text{independent of } n} = O(n) \tag{27}$$

parameters, compared with $P_{\text{VAR}} = n + n^2 L = O(n^2)$ for an unrestricted VAR(L).[8] For moderate or large n, the attention-based VAR has fewer parameters than a corresponding time-invariant VAR. The dependence on L is also mild: P_{Att} grows as dL through the positional encodings, compared with $n^2 L$ for the unrestricted VAR. Since d is typically much smaller than n^2, this opens the door to specifying models with very long lag structures.

"No free lunch": Regularization vs. representability. The $O(n)$ parameter scaling in (27) requires holding d and d_v fixed as n grows. Under the conditions of Proposition 4, which require $d \geq n$ and $d_v \geq n$, the parameter count becomes $O(n^2)$ and the advantage over an unrestricted VAR disappears. One cannot simultaneously have unrestricted VAR representation and achieve dimensionality reduction. The researcher controls this tradeoff through three architectural choices: d (and d_v) governs the dimension of the cross-sectional propagation subspace; H governs the cross-lag rank, i.e., the dimension of the lag-profile subspace; and L governs the context length. Using $H \ll L$ is a deliberate regularization choice that keeps the cost of extending the context window low. Unlike Minnesota-style shrinkage (Doan et al., 1984), which penalizes coefficient magnitudes directly, the rank restriction constrains lag coefficient vectors to lie in a low-dimensional subspace.

[8] The terms scaling with n are $\mathbf{W}_x$ (dn), $\mathbf{W}_\mu$ (dn), and $\boldsymbol{b}_\mu$ (n). The remaining terms are the positional encodings $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L$ (dL), the per-head projection matrices $\mathbf{W}_q^m$, $\mathbf{W}_k^m$, $\mathbf{W}_v^m$ $(H(2d_k + d_v)d)$, and the output mixing matrices $\mathbf{W}^{O,m}$ $(Hd \cdot d_v)$. See Table A.1 in the Internet Appendix.

4.1 Connection to dynamic factor models

A natural question is whether these rank restrictions are an arbitrary feature of the attention architecture or can be given an economic interpretation. It turns out that an analogous low-rank coefficient structure arises in dynamic factor models (DFMs; Sargent and Sims, 1977; Bernanke et al., 2005; Stock and Watson, 2005) and linearized DSGE models (Fernández-Villaverde et al., 2007), where a low-dimensional state generates comovement among many observables. Consider first the case of a DFM without idiosyncratic components.

Proposition 6 (Representation of factor-VAR reduced forms)**.** *Consider a DFM without idiosyncratic components: r factors $\boldsymbol{f}_t \in \mathbb{R}^r$ satisfy $\boldsymbol{y}_t = \boldsymbol{\Lambda}\boldsymbol{f}_t$ and $\boldsymbol{f}_t = \sum_{\ell=1}^{L} \boldsymbol{\Phi}_\ell \boldsymbol{f}_{t-\ell} + \boldsymbol{\eta}_t$, where $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times r}$ has full column rank and $\boldsymbol{\Lambda}^+$ denotes any left inverse. The reduced-form VAR(L) is*

$$\boldsymbol{y}_t = \sum_{\ell=1}^{L} \mathbf{A}_\ell^\star \, \boldsymbol{y}_{t-\ell} + \boldsymbol{u}_t, \qquad \mathbf{A}_\ell^\star = \boldsymbol{\Lambda}\boldsymbol{\Phi}_\ell\boldsymbol{\Lambda}^+, \qquad \boldsymbol{u}_t = \boldsymbol{\Lambda}\boldsymbol{\eta}_t. \tag{28}$$

The factorization through $\mathbb{R}^r$ implies:

(i) $\operatorname{col}(\mathbf{A}_\ell^\star) \subseteq \operatorname{col}(\boldsymbol{\Lambda})$ *and* $\operatorname{row}(\mathbf{A}_\ell^\star) \subseteq \operatorname{row}(\boldsymbol{\Lambda}^+)$ *for every ℓ, with* $\operatorname{rank}(\mathbf{A}_\ell^\star) \le r$*;*

(ii) $\operatorname{rank}(\boldsymbol{\mathcal{A}}^\star) \le \min\{L, r^2\}$ *for the lag-stacked matrix* $\boldsymbol{\mathcal{A}}^\star = [\operatorname{vec}(\mathbf{A}_1^\star) \ \cdots \ \operatorname{vec}(\mathbf{A}_L^\star)]$.

An H-head attention model with $d \ge r$, $d_v \ge r$, and $H \ge L$ can represent any such reduced form by setting $\mathbf{W}_\mu = \boldsymbol{\Lambda}$, $\mathbf{W}_x = \boldsymbol{\Lambda}^+$, and choosing deterministic attention weights $\{\alpha_\ell^m\}$ satisfying Equation (21) with the $L \times H$ mixing matrix $\boldsymbol{\mathcal{M}}$ of full row rank.

Proof. The system $\boldsymbol{\Phi}_\ell = \sum_{m=1}^{H} \alpha_\ell^m \mathbf{G}^m$ is the element-wise linear system $\operatorname{vec}(\boldsymbol{\Phi}_\ell) = \sum_m \alpha_\ell^m \operatorname{vec}(\mathbf{G}^m)$ for each ℓ. Since $\boldsymbol{\mathcal{M}}$ has full row rank L, this system has a solution by the same argument as Proposition 4, applied in $r \times r$ factor space. With $\mathbf{W}_\mu = \boldsymbol{\Lambda}$ and $\mathbf{W}_x = \boldsymbol{\Lambda}^+$, we have $\mathbf{B}^m = \mathbf{W}_\mu \mathbf{W}^{O,m} \mathbf{W}_v^m \mathbf{W}_x = \boldsymbol{\Lambda}\mathbf{G}^m\boldsymbol{\Lambda}^+$, where $\mathbf{W}^{O,m}$ and $\mathbf{W}_v^m$ are constructed as in the proof of Proposition 4 with $d \ge r$ and $d_v \ge r$. Then $\mathbf{A}_\ell = \sum_m \alpha_\ell^m \mathbf{B}^m = \boldsymbol{\Lambda}(\sum_m \alpha_\ell^m \mathbf{G}^m)\boldsymbol{\Lambda}^+ = \boldsymbol{\Lambda}\boldsymbol{\Phi}_\ell\boldsymbol{\Lambda}^+ = \mathbf{A}_\ell^\star$. □

In the representation, $\mathbf{W}_\mu$ is analogous to the factor loading matrix $\boldsymbol{\Lambda}$, $\mathbf{W}_x$ to the factor extraction matrix $\boldsymbol{\Lambda}^+$, d to the factor dimension r, and H to the cross-lag rank bound r^2 (since the $r \times r$ matrices $\boldsymbol{\Phi}_\ell$ span an at most r^2-dimensional space).[9] These parallels provide guidance for model specification and prior construction. Choosing d small is analogous to assuming few factors drive the dynamics. The number of heads H controls the cross-lag

[9] The internal decomposition of $\mathbf{B}^m$ into $\mathbf{W}_\mu$, $\mathbf{W}^{O,m}$, $\mathbf{W}_v^m$, $\mathbf{W}_x$ is not identified, analogous to rotational indeterminacy in DFMs (Bai and Wang, 2015); only the products $\mathbf{B}^m$ and the resulting coefficients $\mathbf{A}_{t,\ell}$ are identified.

complexity: $H = r^2$ suffices to capture the full factor dynamics, $H < r^2$ imposes additional regularization. In Bayesian estimation, these parallels suggest shrinking toward small d and H to favor parsimonious factor-like dynamics, and toward $\mathbf{W}_x\mathbf{W}_\mu = \mathbf{I}_d$ to tie the encoding and decoding subspaces together as in the DFM. The connection also suggests a specific modelling of the innovation covariance: in the DFM, $\mathrm{Var}(\boldsymbol{u}_t) = \boldsymbol{\Lambda}\boldsymbol{\Omega}\boldsymbol{\Lambda}'$ has rank at most r, so the covariance should share the same low-dimensional structure as the conditional mean. Section 5 develops such a parameterization.

Example 3 (Two structural shocks with different persistence)**.** *Consider n macroeconomic variables driven by two independent AR(1) factors:*

$$f_{mt} = \rho_m f_{m,t-1} + \eta_{mt}, \qquad m = 1, 2, \qquad |\rho_1| > |\rho_2|,$$

with $\boldsymbol{y}_t = \boldsymbol{\lambda}_1 f_{1t} + \boldsymbol{\lambda}_2 f_{2t}$, where $\boldsymbol{\lambda}_m \in \mathbb{R}^n$ are loading vectors. Since each factor is scalar, $\boldsymbol{\omega}_m^\top \boldsymbol{\lambda}_m = 1$ defines the left inverse $\boldsymbol{\omega}_m$, and the reduced form is a VAR(1) with

$$\mathbf{A}_1^\star = \rho_1 \boldsymbol{\lambda}_1 \boldsymbol{\omega}_1^\top + \rho_2 \boldsymbol{\lambda}_2 \boldsymbol{\omega}_2^\top,$$

where $\mathrm{rank}(\mathbf{A}_1^\star) \le 2$*. By Proposition 6, a two-head attention model ($H = 2$, $L = 1$, $d_v \ge 1$) represents this with $\mathbf{B}^m = \rho_m \boldsymbol{\lambda}_m \boldsymbol{\omega}_m^\top$: each head captures one structural shock, with the attention weight on the single lag equal to one.*

More generally, with measurement error or idiosyncratic shocks, the DFM reduced form is VARMA rather than VAR, equivalently a VAR(∞) under invertibility (Stock and Watson, 2005). A finite VAR(L) cannot represent this exactly, but the attention model can approximate it by increasing L.

Proposition 7 (Approximation of DFM reduced forms)**.** *Consider a DFM with r factors, fixed loadings $\boldsymbol{\Lambda}$, stable factor dynamics, and idiosyncratic shocks, so that the reduced form is a VAR(∞) with coefficients $\mathbf{A}_\ell^\star = \boldsymbol{\Lambda}\boldsymbol{\Phi}_\ell^{(\infty)}\boldsymbol{\Lambda}^+$ satisfying Assumption B.1. For any L, an H-head attention model with $d \ge r$, $d_v \ge r$, and $H \ge L$ can represent the truncated VAR(L) exactly. The within-lag rank of $\mathbf{A}_\ell^\star$ is at most r for every ℓ, and d and H need not grow with L. As $L \to \infty$, the truncation error vanishes at rate $O(\varrho^L)$ (Lemma B.1 in Appendix B), so the attention model approximates the full DFM conditional mean arbitrarily well.*

Proof. For each $\ell \le L$, $\mathbf{A}_\ell^\star = \boldsymbol{\Lambda}\boldsymbol{\Phi}_\ell^{(\infty)}\boldsymbol{\Lambda}^+$ factors through $\mathbb{R}^r$, so Proposition 6 applies to the truncated coefficient sequence $(\mathbf{A}_1^\star, \ldots, \mathbf{A}_L^\star)$. The truncation bound follows from Lemma B.1. □

The cross-lag rank restriction makes increasing L feasible: the estimation cost grows as dL rather than n^2L, so the model can use a long context window to reduce truncation bias without the curse of dimensionality that would plague an unrestricted VAR. Appendix B formalizes this approximation-estimation tradeoff.

5 Attention for the Conditional Variance

The baseline model assumes a constant conditional covariance $\boldsymbol{\Sigma}$. This section parameterizes $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1})$ as a function of the same state vector $\boldsymbol{h}_t$ that determines the conditional mean. The parameterization ties the covariance structure to the same architectural bottlenecks that constrain the conditional mean, with a single dimension parameter q controlling the tradeoff between regularization and flexibility.

5.1 Parameterization

Section 4 established that $\mathrm{col}(\mathbf{A}_{t,\ell}) \subseteq \mathrm{col}(\mathbf{W}_\mu)$ for all t and ℓ: the conditional mean propagates shocks through a d-dimensional subspace determined by $\mathbf{W}_\mu$, analogous to the loading space $\mathrm{col}(\boldsymbol{\Lambda})$ in a dynamic factor model. We impose the same low-dimensional structure on the innovation covariance.

Write the innovation as

$$\boldsymbol{\varepsilon}_t = \boldsymbol{\Lambda}_\Sigma \boldsymbol{\xi}_t + \mathbf{D}_t \boldsymbol{z}_t, \qquad \boldsymbol{\xi}_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(0, \boldsymbol{\Omega}_t), \qquad \boldsymbol{z}_t \mid \mathcal{F}_{t-1} \sim \mathcal{N}(0, \mathbf{I}_n), \tag{29}$$

with $\boldsymbol{\xi}_t \perp \boldsymbol{z}_t$, where $\boldsymbol{\Lambda}_\Sigma \in \mathbb{R}^{n \times q}$ is a time-invariant loading matrix, $\boldsymbol{\xi}_t \in \mathbb{R}^q$ is a vector of common shocks, $\boldsymbol{z}_t \in \mathbb{R}^n$ captures idiosyncratic variation, and $\mathbf{D}_t = \mathrm{diag}(\sigma_{1,t}, \ldots, \sigma_{n,t})$. The implied covariance is

$$\boldsymbol{\Sigma}_t = \boldsymbol{\Lambda}_\Sigma \, \boldsymbol{\Omega}_t \, \boldsymbol{\Lambda}_\Sigma^\top + \mathbf{D}_t^2, \tag{30}$$

which is positive definite whenever $\mathbf{D}_t^2 \succ 0$. This is the "low rank plus diagonal" structure of approximate factor models (Chamberlain and Rothschild, 1983; Stock and Watson, 2002).

The architectural link is in the choice of $\boldsymbol{\Lambda}_\Sigma$. Since $\mathbf{W}_\mu \in \mathbb{R}^{n \times d}$ already determines the column space through which the conditional mean operates, we parameterize

$$\boldsymbol{\Lambda}_\Sigma = \mathbf{W}_\mu \mathbf{W}_\Sigma, \qquad \mathbf{W}_\Sigma \in \mathbb{R}^{d \times q}, \tag{31}$$

where $\mathbf{W}_\Sigma$ is a learned parameter matrix that selects which q directions within the d-dimensional embedding space carry common shocks into the observables. The product enforces $\mathrm{col}(\boldsymbol{\Lambda}_\Sigma) \subseteq \mathrm{col}(\mathbf{W}_\mu)$: the covariance factor loadings live in the same subspace

that generates the conditional mean. Since $\operatorname{rank}(\mathbf{\Lambda}_\Sigma) \le \operatorname{rank}(\mathbf{W}_\mu) \le d$, only $q_{\text{eff}} = \min\{q, \operatorname{rank}(\mathbf{W}_\mu)\}$ distinct common-shock directions can be loaded; setting $q > d$ adds no expressiveness. In the no-measurement-error DFM of Section 4, $\mathbf{W}_\mu$ plays the role of $\mathbf{\Lambda}$, so $\mathbf{\Lambda}_\Sigma = \mathbf{W}_\mu \mathbf{W}_\Sigma$ is the reduced-form counterpart of the statement that innovations load on the same factor space as the levels.

The idiosyncratic log-variances are affine functions of the state:

$$\log \sigma_{i,t}^2 = b_{\delta,i} + \boldsymbol{w}_{\delta,i}^\top \boldsymbol{h}_t, \qquad i = 1, \dots, n. \tag{32}$$

The common-shock covariance $\mathbf{\Omega}_t \in \mathcal{S}_{++}^q$ is diagonal:

$$\mathbf{\Omega}_t = \operatorname{diag}\big(\exp(\lambda_{1,t}), \dots, \exp(\lambda_{q,t})\big), \qquad \boldsymbol{\lambda}_t = \boldsymbol{b}_\lambda + \mathbf{W}_\lambda \boldsymbol{h}_t, \tag{33}$$

where $\mathbf{W}_\lambda \in \mathbb{R}^{q \times d}$ and $\boldsymbol{b}_\lambda \in \mathbb{R}^q$. All cross-sectional correlation comes through the loadings $\mathbf{\Lambda}_\Sigma$; the common shocks are conditionally uncorrelated but heteroskedastic.

Computation. The low-rank-plus-diagonal structure allows likelihood evaluation in $O(nq^2 + q^3)$ rather than $O(n^3)$. The matrix determinant lemma and the Woodbury identity give

$$|\mathbf{\Sigma}_t| = |\mathbf{D}_t^2|\,|\mathbf{\Omega}_t|\,\big|\mathbf{\Omega}_t^{-1} + \mathbf{\Lambda}_\Sigma^\top \mathbf{D}_t^{-2} \mathbf{\Lambda}_\Sigma\big|, \tag{34}$$

$$\mathbf{\Sigma}_t^{-1} = \mathbf{D}_t^{-2} - \mathbf{D}_t^{-2} \mathbf{\Lambda}_\Sigma \big(\mathbf{\Omega}_t^{-1} + \mathbf{\Lambda}_\Sigma^\top \mathbf{D}_t^{-2} \mathbf{\Lambda}_\Sigma\big)^{-1} \mathbf{\Lambda}_\Sigma^\top \mathbf{D}_t^{-2}, \tag{35}$$

so the dominant cost per observation is the $q \times q$ inversion in the Woodbury term.

5.2 Special cases

The specification nests several standard covariance models through restrictions on its parameters.

Constant covariance (time-invariant VAR). When $\mathbf{W}_\delta = 0$ and $\mathbf{W}_\lambda = 0$, neither the idiosyncratic nor the common-shock variances depend on $\boldsymbol{h}_t$, giving $\mathbf{D}_t \equiv \mathbf{D} = \operatorname{diag}(\exp(\boldsymbol{b}_\delta/2))$ and $\mathbf{\Omega}_t \equiv \mathbf{\Omega} = \operatorname{diag}(\exp(\boldsymbol{b}_\lambda))$. The covariance reduces to

$$\mathbf{\Sigma} = \mathbf{\Lambda}_\Sigma\, \mathbf{\Omega}\, \mathbf{\Lambda}_\Sigma^\top + \mathbf{D}^2, \tag{36}$$

which is a constant positive definite matrix. Combined with the time-invariant conditional mean obtained when $\mathbf{W}_q^m \mathbf{W}_x = 0$ and $\mathbf{W}_k^m \mathbf{W}_x = 0$ (Proposition 4), this recovers the standard linear VAR with Gaussian innovations. If $\operatorname{rank}(\mathbf{W}_\mu) = n$ (hence $d \ge n$) and $q \ge n$,

the constant-covariance subclass is saturated: by appropriate choice of $(\mathbf{W}_\Sigma, \boldsymbol{b}_\lambda, \boldsymbol{b}_\delta)$ it can match any $\boldsymbol{\Sigma} \in \mathcal{S}^n_{++}$. In the regularized regime $\text{rank}(\mathbf{W}_\mu) < n$, the model is restricted to a low-rank-plus-diagonal family.

Factor-model covariance. In the no-measurement-error DFM of Section 4, $\boldsymbol{\varepsilon}_t = \boldsymbol{\Lambda}\boldsymbol{\eta}_t$ and $\text{Var}(\boldsymbol{\varepsilon}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{\Lambda}\mathbf{Q}\boldsymbol{\Lambda}^\top$ has rank at most $r < n$. Setting $\boldsymbol{\Lambda}_\Sigma = \boldsymbol{\Lambda}$, $\boldsymbol{\Omega}_t \equiv \boldsymbol{\Omega}$, and $\mathbf{D}_t = 0$ recovers this singular covariance exactly, but the density is not well defined in $\mathbb{R}^n$. With $\mathbf{D}_t^2 \succ 0$, the covariance becomes $\boldsymbol{\Sigma}_t = \boldsymbol{\Lambda}\boldsymbol{\Omega}\boldsymbol{\Lambda}^\top + \mathbf{D}_t^2$, the structure of approximate factor models and FAVAR specifications (Bernanke et al., 2005). The idiosyncratic variances regularize the covariance to full rank while preserving the factor interpretation.

5.3 Regularization, identification, and the role of q

The dimension q controls the complexity of the covariance specification, paralleling the role of (d, H) for the conditional mean. Small q imposes strong regularization: all cross-sectional dependence is channeled through q common-shock directions, and the covariance parameters scale as $O(nq)$ rather than $O(n^2)$. When q is large (up to $\text{rank}(\mathbf{W}_\mu)$), the common component can approximate a wide class of covariance matrices, but with fixed $\boldsymbol{\Lambda}_\Sigma$ and diagonal $\boldsymbol{\Omega}_t$ the time variation in cross-sectional dependence is restricted to q volatility factors plus the n idiosyncratic variances.

Because $\boldsymbol{\Omega}_t$ is diagonal, $(\boldsymbol{\Lambda}_\Sigma, \boldsymbol{\lambda}_t)$ are not separately identified up to column-wise rescaling and permutation. The identified object is $\boldsymbol{\Sigma}_t$. For interpretability one may normalize $\|\boldsymbol{\Lambda}_{\Sigma,\cdot j}\| = 1$ for each j, so that $\exp(\lambda_{j,t})$ is the variance of common shock j.

5.4 Illustrative examples

Example 4 (Univariate model with ARCH-type volatility)**.** *Consider $n = 1$ with constant conditional mean $\mu_t = c$ and time-varying variance. Augment the input to include squared values: $\tilde{y}_{t-\ell} = (y_{t-\ell}, y^2_{t-\ell})^\top \in \mathbb{R}^2$. With $W_x = (0, 1) \in \mathbb{R}^{1\times 2}$ and uniform attention $\alpha_{t,\ell} = 1/L$, the state becomes $h_t = L^{-1}\sum_{\ell=1}^L y^2_{t-\ell}$, and*

$$\log \sigma_t^2 = b_\delta + w_\delta h_t = b_\delta + \frac{w_\delta}{L}\sum_{\ell=1}^{L} y^2_{t-\ell}.$$

With $w_\delta > 0$ and $c = 0$ (so $y^2_{t-\ell} \approx \varepsilon^2_{t-\ell}$), this is a log-linear ARCH(L) model. When attention weights are learned rather than uniform, the model becomes $\log \sigma_t^2 = b_\delta + w_\delta \sum_\ell \alpha_{t,\ell} y^2_{t-\ell}$, with data-dependent weights on recent versus distant squared values. Setting $w_\delta = 0$ reduces to

homoskedasticity.

Example 5 (Bivariate model with state-dependent correlation)**.** *Consider $n = 2$ with constant conditional mean and $q = 1$. Write $\boldsymbol{\Lambda}_\Sigma = (\ell_1, \ell_2)^\top \in \mathbb{R}^2$. The covariance is $\boldsymbol{\Sigma}_t = \boldsymbol{\Lambda}_\Sigma \boldsymbol{\Lambda}_\Sigma^\top \exp(\lambda_t) + \mathbf{D}_t^2$, so the off-diagonal element is $[\boldsymbol{\Sigma}_t]_{12} = \ell_1 \ell_2 \exp(\lambda_t)$ and the conditional correlation is*

$$\rho_t = \frac{\ell_1 \ell_2 \exp(\lambda_t)}{\sqrt{(\ell_1^2 \exp(\lambda_t) + \sigma_{1,t}^2)(\ell_2^2 \exp(\lambda_t) + \sigma_{2,t}^2)}},$$

where $\lambda_t = b_\lambda + \boldsymbol{w}_\lambda^\top \boldsymbol{h}_t$. When $\ell_1 \ell_2 > 0$, an increase in $\exp(\lambda_t)$ raises both the numerator and the common-shock share of each diagonal element, pushing $|\rho_t|$ toward one. If the attention mechanism learns that $\boldsymbol{h}_t$ increases during periods of joint large moves, correlations rise in volatile states, consistent with the empirical finding that cross-asset correlations increase during market stress (Longin and Solnik, 2001). Setting $w_\delta = 0$ and $w_\lambda = 0$ gives a constant correlation.

6 Nonlinear Extensions

The baseline attention-only specification (Sections 2–5) yields a conditional mean that is linear in lagged observables: $\boldsymbol{\mu}_t = \boldsymbol{b}_t + \sum_\ell \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell}$. The standard Transformer architecture (Vaswani et al., 2017) includes three additional components that break this linearity: a feedforward sublayer, residual connections, and layer normalization. This section describes each component, shows how stacking multiple layers composes these nonlinearities, and characterizes how the resulting model relates to the TV-VAR representation.

6.1 Feedforward sublayer

The feedforward sublayer applies a two-layer network position-wise to each state vector:

$$\mathrm{FFN}(\boldsymbol{u}) = \boldsymbol{b}_2 + \mathbf{W}_2 \phi(\boldsymbol{b}_1 + \mathbf{W}_1 \boldsymbol{u}),$$

where $\mathbf{W}_1 \in \mathbb{R}^{d_{\text{ff}} \times d}$, $\mathbf{W}_2 \in \mathbb{R}^{d \times d_{\text{ff}}}$, and ϕ is an elementwise nonlinearity such as the rectified linear unit $\phi(z) = \max\{0, z\}$. Applied after the attention aggregation step, the feedforward sublayer transforms the state $\boldsymbol{h}_t$ before it enters the conditional mean projection. This breaks the linearity of Proposition 2: $\boldsymbol{\mu}_t = \boldsymbol{b}_\mu + \mathbf{W}_\mu \mathrm{FFN}(\boldsymbol{h}_t)$ is no longer an affine function of lagged observables. The conditional mean can now capture threshold effects, saturation, and other nonlinear responses to the history.

6.2 Residual connections and layer normalization

Residual connections (He et al., 2016) add the input of each sublayer to its output:

$$\tilde{\boldsymbol{h}}_t^{(k)} = \boldsymbol{h}_t^{(k)} + \mathrm{Attn}^{(k)}(\mathrm{LN}(\boldsymbol{h}_\cdot^{(k)}))_t,$$
$$\boldsymbol{h}_t^{(k+1)} = \tilde{\boldsymbol{h}}_t^{(k)} + \mathrm{FFN}^{(k)}(\mathrm{LN}(\tilde{\boldsymbol{h}}_t^{(k)})),$$

where $\boldsymbol{h}_t^{(k)} \in \mathbb{R}^d$ is the hidden state at position t in layer k. When the attention and feedforward contributions are small relative to the identity term, the model adds nonlinear corrections on top of a near-linear baseline. The identity mapping operates in state space $\mathbb{R}^d$, not observation space $\mathbb{R}^n$.

Layer normalization (Ba et al., 2016) standardizes each state vector across its coordinates:

$$\mathrm{LN}(\boldsymbol{u}) = \boldsymbol{\gamma} \odot \frac{\boldsymbol{u} - \bar{\boldsymbol{u}}}{\sqrt{\widehat{\mathrm{Var}}(\boldsymbol{u}) + \epsilon}} + \boldsymbol{\beta},$$

where $\bar{\boldsymbol{u}}$ and $\widehat{\mathrm{Var}}(\boldsymbol{u})$ are the coordinate mean and variance, $\boldsymbol{\gamma}, \boldsymbol{\beta} \in \mathbb{R}^d$ are learned parameters, and $\odot$ denotes elementwise multiplication. Layer normalization stabilizes gradient magnitudes during estimation and makes the model invariant to the overall scale of the state vector, but it introduces an additional nonlinearity through the division by the coordinate standard deviation.

6.3 Multi-layer models and local linearization

With K layers, each combining attention, feedforward, residual, and layer normalization sublayers, the model composes K nonlinear transformations. The TV-VAR form (9) holds exactly in state space $\mathbb{R}^d$ within each layer (with hidden states replacing observations), but not in observation space $\mathbb{R}^n$. The overall mapping from $\boldsymbol{y}_{t-L:t-1}$ to $\boldsymbol{\mu}_t$ is a composition of K attention-feedforward blocks followed by the output projection, which is generally a nonlinear function of the lagged observables.

To recover interpretable coefficient matrices, define effective VAR coefficients as Jacobians:

$$\mathbf{J}_{t,\ell} = \frac{\partial \boldsymbol{\mu}_t}{\partial \boldsymbol{y}_{t-\ell}} \in \mathbb{R}^{n \times n}.$$

These provide local impulse response interpretations: $\mathbf{J}_{t,\ell}$ measures the marginal effect of $\boldsymbol{y}_{t-\ell}$ on $\boldsymbol{\mu}_t$ at the realized history. For structural analysis, these Jacobians play the role of the autoregressive coefficients in a standard VAR.[10] This local linearization reduces to the

[10] Because the Jacobians $\mathbf{J}_{t,\ell}$ depend on the full history, impulse responses in multi-layer attention-based

exact TV-VAR representation of Proposition 2 when $K = 1$ and the feedforward sublayer is omitted, so the baseline model is a special case. Increasing K or adding feedforward layers trades the closed-form coefficient structure for greater expressive power.

7 Estimation

7.1 Identification

Proposition 1 establishes that the primitive parameters are not identified. We distinguish between exact restrictions that eliminate specific symmetries and softer anchoring through the prior.

Exact normalizations. In the single-head case with $d_v = d$, only the product $\mathbf{M} := \mathbf{W}_\mu \mathbf{W}_v$ enters the conditional mean, so setting $\mathbf{W}_v = \mathbf{I}_d$ and estimating $\mathbf{M}$ directly eliminates the value-space rotation with no loss of generality. For the query-key block, the scores depend only on the product $\mathbf{F} := \mathbf{W}_q^\top \mathbf{W}_k$, so parameterizing $\mathbf{F}$ directly eliminates the query-key rotation. When $d_k < d/2$, retaining the factored form $(\mathbf{W}_q, \mathbf{W}_k)$ is more parsimonious ($2d_k d$ vs. d^2 free parameters), but requires a normalization such as $\mathbf{W}_k \mathbf{W}_k^\top = \mathbf{I}_{d_k}$. For the intercept decomposition, since $\sum_\ell \alpha_{t,\ell} = 1$, one normalization such as $\mathbf{r}_1 = \mathbf{0}$ (where $\mathbf{r}_\ell = \mathbf{W}_\mu \mathbf{W}_v \boldsymbol{p}_\ell$) resolves the non-uniqueness between $\boldsymbol{b}_\mu$ and the positional terms.

Multi-head case. With $H > 1$ heads, two additional symmetries arise. First, the model is invariant to permutations of the head index: relabeling heads m and m' together with their associated projections $(\mathbf{W}_q^m, \mathbf{W}_k^m, \mathbf{W}_v^m, \mathbf{W}^{O,m})$ leaves all reduced-form objects unchanged. Second, each head has its own value-space rotation: only the product $\mathbf{W}^{O,m}\mathbf{W}_v^m$ enters the conditional mean, so replacing $(\mathbf{W}_v^m, \mathbf{W}^{O,m})$ with $(\mathbf{S}_m \mathbf{W}_v^m, \mathbf{W}^{O,m}\mathbf{S}_m^\top)$ for any orthogonal $\mathbf{S}_m$ is observationally equivalent. Unlike the single-head case, one cannot collapse $\mathbf{W}_v^m$ and $\mathbf{W}^{O,m}$ into a single matrix without losing the low-rank structure of each $\mathbf{B}^m = \mathbf{W}_\mu \mathbf{W}^{O,m} \mathbf{W}_v^m \mathbf{W}_x$. The embedding-space and query-key rotations carry over from Proposition 1, now acting on all heads simultaneously for the embedding and independently per head for the query-key block. The head permutation symmetry is discrete and can be handled by ordering conventions (e.g., by decreasing $\|\mathbf{B}^m\|_F$); the per-head value-space rotations are continuous

VARs are nonlinear functions of the coefficient matrices, not linear combinations of head-specific responses. One cannot decompose the impulse response at horizon h into a sum of H head-specific impulse responses, since the iterated composition $\mathbf{J}_{t+h,1} \cdots \mathbf{J}_{t+1,1}$ involves cross-products of attention weights from different heads and time periods.

and require either per-head triangular constraints on $\mathbf{W}_v^m$ or soft anchoring through the prior.

Soft anchoring. The embedding-space rotation is the most consequential symmetry because it acts on $(\mathbf{W}_x, \boldsymbol{p}_\ell, \mathbf{W}_q^m, \mathbf{W}_k^m, \mathbf{W}_v^m)$ simultaneously across all heads. Hard constraints such as requiring a $d \times d$ submatrix of $\mathbf{W}_x$ to be lower-triangular with positive diagonal would fix the latent basis but complicate the sampler, which relies on unconstrained Gaussian blocks and Elliptical Slice Sampling. An effective alternative is to anchor the basis through informative priors: centering $(\mathbf{W}_\mu, \mathbf{W}_v^m, \mathbf{W}_x)$ at an SVD factorization of an OLS estimate of $\mathbf{B}$ defines a preferred coordinate system that reduces posterior drift across equivalent parameterizations without imposing hard constraints.

7.2 A shrinkage prior centered at time-invariant dynamics

The attention parameterization can generate arbitrarily nonlinear, state-dependent dynamics, so regularization toward parsimony is essential. This subsection constructs a prior that is centered at a constant-coefficient VAR but remains within the attention architecture, making "time invariance" a shrinkage limit rather than a separate model.

Reparameterization. The source of time variation in the attention-based VAR is the dependence of scores on lagged data through the query and key projections. We isolate this source by decomposing each head's projections into a positional component and a data-dependent perturbation:

$$\mathbf{W}_q^m = \mathbf{W}_{q,0}^m + \lambda_m \widetilde{\mathbf{W}}_q^m, \qquad \mathbf{W}_k^m = \mathbf{W}_{k,0}^m + \lambda_m \widetilde{\mathbf{W}}_k^m, \tag{37}$$

where $\mathbf{W}_{q,0}^m \mathbf{W}_x = 0$ and $\mathbf{W}_{k,0}^m \mathbf{W}_x = 0$, so that the "base" projections $(\mathbf{W}_{q,0}^m, \mathbf{W}_{k,0}^m)$ satisfy the deterministic weights condition (21). The scalar $\lambda_m \in \mathbb{R}$ controls the degree of state dependence for head m: when $\lambda_m = 0$, the scores reduce to

$$s_{t,\ell}^m = \frac{\left(\mathbf{W}_{q,0}^m \boldsymbol{p}_1\right)^\top \left(\mathbf{W}_{k,0}^m \boldsymbol{p}_\ell\right)}{\sqrt{d_k}},$$

which depend only on the positional effects. Attention weights $\boldsymbol{\alpha}_t^m$ are then constant in t, and the conditional mean is a time-invariant VAR(L) with cross-lag rank at most H (Proposition 4). As $|\lambda_m|$ increases, scores become functions of the lagged data and the coefficients become time-varying.

Hierarchical shrinkage prior. The reparameterization (37) reduces the question of "how much time variation" to the magnitudes of $\lambda_1, \ldots, \lambda_H$. A natural prior specification places a global-local shrinkage prior on these scales:

$$\lambda_m \mid \tau_\lambda, \nu_m \sim \mathcal{N}(0, \tau_\lambda^2 \nu_m^2), \qquad \nu_m \sim \text{half-Cauchy}(1), \qquad \tau_\lambda \sim \text{half-Cauchy}(1). \tag{38}$$

The global scale τ_λ controls overall shrinkage toward time invariance, while the local scales ν_m allow individual heads to escape if the data support state-dependent dynamics. This is the attention-architecture analogue of centering a TV-VAR prior at constant coefficients: the prior concentrates near the constant-coefficient submodel but has heavy tails that accommodate departures when warranted by the likelihood.

Nesting of constant-coefficient VARs. Setting $\lambda_m = 0$ for all m reduces the model exactly to a constant-coefficient VAR(L) with the cross-lag rank restriction of Proposition 5. The prior (38) therefore nests the linear VAR as a limiting case ($\lambda \to 0$), with the posterior concentrating near this limit when the data are consistent with time-invariant dynamics. The remaining parameters $(\mathbf{W}_{q,0}^m, \mathbf{W}_{k,0}^m, \mathbf{W}_v^m, \mathbf{W}_m^O, \mathbf{W}_\mu, \mathbf{W}_x, \{\boldsymbol{p}_\ell\})$ receive conventional priors; details are provided below.

Posterior inference on the degree of nonlinearity. The posterior distribution of λ_m directly quantifies the degree of state dependence contributed by head m: small $|\lambda_m|$ means head m operates as a nearly constant-coefficient component, while large $|\lambda_m|$ indicates that its lag weights vary substantially with the state of the economy. The following result formalizes this relationship.

Lemma 1 (Local control of time variation by λ_m)**.** *Fix a head m and assume there exists $S_{\max} < \infty$ such that $\sup_{t,\ell} |s_{t,\ell}^m| \le S_{\max}$ almost surely under the parameterization (37). Then there exists a constant $C = C(S_{\max}, L)$ such that for all t,*

$$\|\boldsymbol{\alpha}_t^m - \boldsymbol{\alpha}_0^m\|_1 \le C\,|\lambda_m| \max_{\ell \le L} \|\boldsymbol{y}_{t-\ell}\|,$$

where $\boldsymbol{\alpha}_0^m$ denotes the (time-invariant) weights at $\lambda_m = 0$. Consequently,

$$\|\mathbf{A}_{t,\ell} - \mathbf{A}_{0,\ell}\|_F \le C\,|\lambda_m|\,\|\mathbf{B}^m\|_F \max_{j \le L} \|\boldsymbol{y}_{t-j}\| + \sum_{m' \ne m} \|\mathbf{A}_{t,\ell}^{(m')} - \mathbf{A}_{0,\ell}^{(m')}\|_F.$$

Proof sketch. Under (37), $s_{t,\ell}^m$ is differentiable in λ_m with derivative linear in $\boldsymbol{y}_{t-\ell}$ and $\boldsymbol{y}_{t-1}$. On bounded score sets, softmax is Lipschitz (bounded Jacobian), implying $\|\boldsymbol{\alpha}_t^m - \boldsymbol{\alpha}_0^m\|_1 \lesssim$

$\|\boldsymbol{s}_t^m - \boldsymbol{s}_0^m\|_1$. The score difference is $O(|\lambda_m| \max_{\ell \le L} \|\boldsymbol{y}_{t-\ell}\|)$ by linearity of the projections. The coefficient bound follows from $\mathbf{A}_{t,\ell}^m = \alpha_{t,\ell}^m \mathbf{B}^m$. □

Lemma 1 implies that the posterior of $\{\lambda_1, \ldots, \lambda_H\}$ provides a head-by-head decomposition of the degree of nonlinearity in the estimated dynamics. If the posterior of λ_m concentrates near zero for most heads, the data favor approximately linear, time-invariant dynamics; if λ_m is large for some heads, those heads capture state-dependent behavior. Combined with the head-specific scales ω_m discussed in Appendix B, which control each head's overall contribution, the practitioner obtains a two-dimensional summary per head: ω_m measures "how important" and λ_m measures "how nonlinear."

7.3 Likelihood

Under the conditional Gaussian specification (2), the log-likelihood given the first L observations is

$$\ell(\boldsymbol{\theta}, \boldsymbol{\Sigma}) = -\frac{(T-L)n}{2} \log(2\pi) - \frac{T-L}{2} \log|\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{t=L+1}^{T} (\boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_t - \boldsymbol{\mu}_t(\boldsymbol{\theta})), \quad (39)$$

where $\boldsymbol{\mu}_t(\boldsymbol{\theta}) = g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1})$ is the conditional mean from Equation (3) and $\boldsymbol{\Sigma}$ is the error covariance. When the covariance is time-varying (Section 5), the middle term becomes $-\frac{1}{2}\sum_{t=L+1}^{T} \log|\boldsymbol{\Sigma}_t(\boldsymbol{\theta})|$ and $\boldsymbol{\Sigma}$ in the quadratic form is replaced by $\boldsymbol{\Sigma}_t(\boldsymbol{\theta})$. The log-likelihood conditions on the initial L observations and treats them as given.

7.4 Posterior computation

The posterior is explored via the Gibbs sampler of Algorithm 2. The conjugate conditional posteriors for $\mathbf{W}_\mu$ and $\boldsymbol{\Sigma}$ are given in Equations (12)–(13); the nonlinear attention parameters are updated by Elliptical Slice Sampling (Algorithm 1).[11]

8 Conclusion

This paper develops a theoretical framework connecting Transformer architectures to classical econometric models for multivariate time series. The main contribution is an exact representation result: for attention-only specifications, the conditional mean is a VAR with

[11] The conjugate profiling structure also admits maximum likelihood estimation: for fixed attention parameters, the optimal $\mathbf{W}_\mu$ and $\boldsymbol{\Sigma}$ are available in closed form, so gradient-based optimization of the profile log-likelihood operates only on the lower-dimensional nonlinear block.

time-varying, state-dependent coefficients that arise endogenously from the attention mechanism. This connection provides interpretability (attention weights as lag importance), reveals implicit regularization (low-rank, parameter-tied coefficients), and offers theoretical grounding for applying Transformer models to economic time series.

We extend the framework to model conditional heteroskedasticity using a factor-coherent covariance parameterization that ties the innovation loading space to the same low-dimensional subspace that constrains the conditional mean. The resulting "low rank plus diagonal" covariance structure nests constant covariance and approximate factor-model covariance as special cases, with all time variation driven by the same observation-driven state. For estimation, we develop a Gibbs sampler that exploits the conjugate structure of the output projection and error covariance while using Elliptical Slice Sampling for the nonlinear attention parameters, with a shrinkage prior that nests the constant-coefficient VAR as a limiting case. In a recursive forecasting exercise for PCE inflation, even the minimal single-head specification produces more accurate point and density forecasts than a Bayesian VAR at horizons from one to twenty-four months.

Several directions merit further investigation. First, the stability analysis for general multi-layer architectures requires stronger theoretical foundations; results from the observation-driven model literature (Creal et al., 2013; Straumann and Mikosch, 2006) may provide a starting point. Second, the framework could be extended to structural analysis by identifying structural shocks through restrictions on the contemporaneous impact matrix, analogous to identified VARs; the methods of Arias et al. (2024) for inference in time-varying SVARs with sign restrictions provide a natural complement. Third, the asymptotic properties of the maximum likelihood estimator and associated inference procedures remain to be developed. Finally, the connection between attention mechanisms and kernel methods suggests potential links to Gaussian process models that warrant exploration.

References

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. (2023). What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations (ICLR)*.

Arias, J. E., Rubio-Ramírez, J. F., Shin, M., and Waggoner, D. F. (2024). Inference based on time-varying SVARs identified with sign restrictions. Working Paper 24-18, Federal Reserve Bank of Philadelphia.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Wang, P. (2015). Identification and Bayesian estimation of dynamic factor models. *Journal of Business & Economic Statistics*, 33(2):221–240.

Banbura, M., Giannone, D., and Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1):71–92.

Bernanke, B. S., Boivin, J., and Eliasz, P. (2005). Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach. *Quarterly Journal of Economics*, 120(1):387–422.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: A multivariate generalized ARCH model. *Review of Economics and Statistics*, 72(3):498–505.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

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

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Doan, T., Litterman, R., and Sims, C. (1984). Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3(1):1–100.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate GARCH models. *Journal of Business & Economic Statistics*, 20(3):339–350.

Engle, R. F., Ng, V. K., and Rothschild, M. (1990). Asset pricing with a factor-ARCH covariance structure: Empirical estimates for treasury bills. *Journal of Econometrics*, 45(1–2):213–237.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Fernández-Villaverde, J., Rubio-Ramírez, J. F., Sargent, T. J., and Watson, M. W. (2007). ABCs (and Ds) of understanding VARs. *American Economic Review*, 97(3):1021–1026.

François, J. and Ravera, L. (2025). Toward manifest relationality in Transformers via symmetry reduction. *arXiv preprint arXiv:2602.18948*.

Geweke, J. (1977). The dynamic factor analysis of economic time series. In Aigner, D. J. and Goldberger, A. S., editors, *Latent Variables in Socio-Economic Models*, pages 365–383. North-Holland, Amsterdam.

Giannone, D., Lenza, M., and Primiceri, G. E. (2015). Prior selection for vector autoregressions. *Review of Economics and Statistics*, 97(2):436–451.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Goel, G. and Bartlett, P. (2024). Can a Transformer represent a Kalman filter? In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference (L4DC)*, volume 242 of *PMLR*.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273.

Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57(2):357–384.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails*. Econometric Society Monographs. Cambridge University Press.

Hauzenberger, N., Huber, F., and Koop, G. (2024). Dynamic shrinkage priors for large time-varying parameter regressions using scalable Markov chain Monte Carlo methods. *Studies in Nonlinear Dynamics & Econometrics*, 28(2):201–225.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Henry, N. W., Marchetti, G. L., and Kohn, K. (2024). Geometry of lightning self-attention: Identifiability and dimension. *arXiv preprint arXiv:2408.17221*.

Kelly, B. T., Kuznetsov, B., Malamud, S., and Xu, T. A. (2025). Artificial intelligence asset pricing models. *NBER Working Paper No. 33351*.

Koop, G. and Korobilis, D. (2013). Large time-varying parameter VARs. *Journal of Econometrics*, 177(2):185–198.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4):1748–1764.

Litterman, R. B. (1986). Forecasting with Bayesian vector autoregressions: Five years of experience. *Journal of Business & Economic Statistics*, 4(1):25–38.

Longin, F. and Solnik, B. (2001). Extreme correlation of international equity markets. *Journal of Finance*, 56(2):649–676.

Lu, S., Bigdeli, S. A., and Föllmer, H. (2024). Elliptical attention. In *Advances in Neural Information Processing Systems*, volume 37.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

Murray, I., Adams, R. P., and MacKay, D. J. C. (2010). Elliptical slice sampling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 541–548.

Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *Review of Economic Studies*, 72(3):821–852.

Sargent, T. J. and Sims, C. A. (1977). Business cycle modeling without pretending to have too much a priori economic theory. In *New Methods in Business Cycle Research: Proceedings from a Conference*, pages 45–109. Federal Reserve Bank of Minneapolis.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica*, 48(1):1–48.

Stock, J. H. and Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179.

Stock, J. H. and Watson, M. W. (2005). Implications of dynamic factor models for VAR analysis. In *NBER Working Paper No. 11467*.

Stock, J. H. and Watson, M. W. (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39(s1):3–33.

Straumann, D. and Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *Annals of Statistics*, 34(5):2449–2495.

Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89(425):208–218.

Tong, H. (1990). *Non-linear Time Series: A Dynamical System Approach*. Oxford University Press.

Tsai, Y.-H. H., Bai, S., Thattai, M., and Salakhutdinov, R. (2019). Transformer dissection: An unified understanding for Transformer's attention via the lens of kernel. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 4344–4353.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. (2023). Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

Zeng, A., Chen, M., Zhang, L., and Xu, Q. (2023). Are Transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128.

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of AAAI Conference on Artificial Intelligence*, pages 11106–11115.

Internet Appendix

beginequation*12pt] **Attention-Based Vector Autoregressions**

beginequation*8pt] Juan Antolín-Díaz Juan F. Rubio-Ramírez

beginequation*20pt]

Table of Contents

A Additional architectural details **2**

A.1 Causal masking 2

A.2 Gain relaxation 2

A.3 Parameter summary 3

B Regularization, approximation, and Bayesian adaptivity **3**

B.1 Targets, truncation bias, and approximation error 4

B.2 Architectural rank regularization and complexity scaling in L 5

B.3 Soft regularization via priors and posterior concentration on low-dimensional sets 6

B.4 Oracle-type bounds for Bayesian predictive performance 7

A Additional architectural details

A.1 Causal masking

In implementations processing all positions simultaneously, a causal mask enforces autoregressive structure. Define $\mathbf{M} \in \mathbb{R}^{T\times T}$ with $M_{ij} = 0$ if $j \le i$ and $M_{ij} = -\infty$ otherwise. Masked scores are $\boldsymbol{s} + \mathbf{M}$, so positions $j > i$ receive zero weight after softmax. The standard convention uses the hidden state at position i (incorporating $\boldsymbol{y}_i$) to predict $\boldsymbol{y}_{i+1}$.

A.2 Gain relaxation

The sum-identity (24) fixes the long-run multiplier over time. A minimal modification breaks this constraint without altering the linear-in-lags structure. Define the pre-gain state

$$\boldsymbol{h}_t^0 = \sum_{m=1}^{H} \mathbf{W}^{O,m} \boldsymbol{c}_t^m, \tag{A.1}$$

and introduce head-specific positive gains

$$g_t^m = \exp(\boldsymbol{u}_m^\top \boldsymbol{h}_t^0 + \kappa_m) > 0, \tag{A.2}$$

where $\boldsymbol{u}_m \in \mathbb{R}^d$ and $\kappa_m \in \mathbb{R}$ are learned parameters (adding $H(d+1)$ parameters total). Replace $\boldsymbol{c}_t^m$ by the gain-scaled output $\widetilde{\boldsymbol{c}}_t^m = g_t^m \boldsymbol{c}_t^m$ in the output mixing step, yielding

$$\boldsymbol{h}_t = \sum_{m=1}^{H} \mathbf{W}^{O,m} \widetilde{\boldsymbol{c}}_t^m, \qquad \boldsymbol{\mu}_t = \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t.$$

This construction is observation-driven: $\boldsymbol{h}_t^0$ depends only on the history $\boldsymbol{y}_{t-L:t-1}$, the gains g_t^m are deterministic functions of $\boldsymbol{h}_t^0$, and there is no fixed-point computation.

The conditional mean retains the TV-VAR form

$$\boldsymbol{\mu}_t = \widetilde{\boldsymbol{b}}_t + \sum_{\ell=1}^{L} \widetilde{\mathbf{A}}_{t,\ell} \boldsymbol{y}_{t-\ell}, \qquad \widetilde{\mathbf{A}}_{t,\ell} = \sum_{m=1}^{H} g_t^m \alpha_{t,\ell}^m \mathbf{B}^m, \tag{A.3}$$

with intercept $\widetilde{\boldsymbol{b}}_t = \boldsymbol{b}_\mu + \sum_{m=1}^{H} g_t^m \sum_{\ell=1}^{L} \alpha_{t,\ell}^m \mathbf{W}_\mu \mathbf{W}^{O,m} \mathbf{W}^{V,m} \boldsymbol{p}_\ell$. The cross-lag factorization is preserved: letting $\widetilde{\boldsymbol{\mathcal{M}}}_t = [g_t^1 \boldsymbol{\alpha}_t^1 \ \cdots \ g_t^H \boldsymbol{\alpha}_t^H]$,

$$\widetilde{\boldsymbol{\mathcal{A}}}_t = \boldsymbol{\mathcal{B}} \widetilde{\boldsymbol{\mathcal{M}}}_t^\top, \qquad \text{so} \qquad \operatorname{rank}(\widetilde{\boldsymbol{\mathcal{A}}}_t) \le \min\{H, L\} \text{ for every } t.$$

The sum-identity becomes $\sum_{\ell=1}^{L} \widetilde{\mathbf{A}}_{t,\ell} = \sum_{m=1}^{H} g_t^m \mathbf{B}^m$, which varies with t whenever gains vary, allowing regime-dependent amplification or attenuation of each dynamic mode.

In this architecture, $\alpha_{t,\ell}^m$ determines the within-head distribution over lags (a "lag shape"), while g_t^m controls the overall contribution of head m (a "mode gain"). If the state $\boldsymbol{h}_t^0$ captures volatility or crisis indicators, gains can adapt the long-run multiplier to the economic regime without requiring discrete regime specification. Unbounded gains can generate explosive dynamics: if $\sum_m g_t^m \mathbf{B}^m$ has spectral radius exceeding unity, the implied VAR is locally unstable. In practice, one may bound gains via $g_t^m = \bar{g} \cdot \sigma(\boldsymbol{u}_m^\top \boldsymbol{h}_t^0 + \kappa_m)$ with sigmoid σ and maximum gain $\bar{g}$, or regularize the gain parameters $(\boldsymbol{u}_m, \kappa_m)$ during estimation. This connects to the stability analysis in Section **??**.

A.3 Parameter summary

Table A.1: Parameter matrices and vectors (constant covariance specification)

Component	Parameter	Dimension	Count
Embedding	$\mathbf{W}_x$	$d \times n$	dn
	$\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L$	$d \times 1$ each	dL
Attention (per head)	$\mathbf{W}_q^m$, $\mathbf{W}_k^m$	$d_k \times d$ each	$2d_k d$
	$\mathbf{W}_v^m$	$d_v \times d$	$d_v d$
Output mixing	$\mathbf{W}_1^O, \ldots, \mathbf{W}_H^O$	$d \times d_v$ each	$Hd \cdot d_v$
Conditional mean	$\mathbf{W}_\mu$	$n \times d$	nd
	$\boldsymbol{b}_\mu$	$n \times 1$	n
Covariance (Cholesky)	$\boldsymbol{\mathcal{L}}$	$n \times n$ (lower tri.)	$n(n+1)/2$

Notes: Tuning parameters (L, H, d, d_k, d_v) are fixed by the researcher. Total parameter count is $O(Hd^2 + nd)$ for the attention and mean components. Additional parameters for time-varying covariance are summarized in Section 5.

B Regularization, approximation, and Bayesian adaptivity

This section formalizes the tradeoff between (i) architectural regularization induced by attention factorization and rank constraints, (ii) prior-driven shrinkage within a more flexible parameterization, and (iii) the choice of context length L as a bias-variance (approximation-estimation) lever. The key point is that rank-based architectural regularization changes the

way statistical complexity scales with L, making it feasible to increase L to reduce truncation bias in settings where unrestricted VAR estimation would be ill-posed.

B.1 Targets, truncation bias, and approximation error

Let T denote the sample size, and treat $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_L)$ as presample values. Consider a generic target conditional mean of the form

$$\boldsymbol{\mu}_t^\star = \boldsymbol{c}^\star + \sum_{\ell=1}^{\infty} \mathbf{A}_\ell^\star \boldsymbol{y}_{t-\ell}, \qquad t \geq L+1, \tag{B.1}$$

with $\mathbb{E}\|\boldsymbol{y}_t\|^2 < \infty$. Representations of the form (B.1) arise from stable VARMA and state-space models via their AR(∞) expansions (Lütkepohl, 2005; Hamilton, 1994). We use (B.1) only as an approximation benchmark: the true DGP need not be linear, but (B.1) provides a canonical notion of "long memory in lags."

For a fixed context length L, define the truncation residual

$$\boldsymbol{r}_{t,L}^\star = \sum_{\ell>L} \mathbf{A}_\ell^\star \boldsymbol{y}_{t-\ell}.$$

The truncation bias is controlled by tail decay of $\{\mathbf{A}_\ell^\star\}$.

Assumption B.1 (Geometric lag decay)**.** *There exist constants $C_A > 0$ and $\varrho \in (0,1)$ such that $\|\mathbf{A}_\ell^\star\|_{\mathrm{op}} \leq C_A \varrho^\ell$ for all $\ell \geq 1$, and $\sup_t \mathbb{E}\|\boldsymbol{y}_t\|^2 \leq M_2 < \infty$.*

Assumption B.1 is satisfied by broad classes of stable VARMA and linear state-space models (after reparameterization) and is a convenient sufficient condition for exponentially decaying truncation error.

Lemma B.1 (Truncation bound)**.** *Under Assumption B.1,*

$$\mathbb{E}\|\boldsymbol{r}_{t,L}^\star\|^2 \leq M_2 \left(\sum_{\ell>L} \|\mathbf{A}_\ell^\star\|_{\mathrm{op}} \right)^2 \leq M_2 \left(\frac{C_A \varrho^{L+1}}{1-\varrho} \right)^2.$$

Proof. By the triangle inequality and Cauchy–Schwarz,

$$\|\boldsymbol{r}_{t,L}^\star\| \leq \sum_{\ell>L} \|\mathbf{A}_\ell^\star\|_{\mathrm{op}} \, \|\boldsymbol{y}_{t-\ell}\| \leq \left(\sum_{\ell>L} \|\mathbf{A}_\ell^\star\|_{\mathrm{op}} \right) \sup_{\ell>L} \|\boldsymbol{y}_{t-\ell}\|.$$

Taking expectations and using $\mathbb{E}\|\boldsymbol{y}_t\|^2 \leq M_2$ yields the first inequality; the second is the geometric tail bound. □

Lemma B.1 isolates the benefit of increasing L: for geometrically decaying AR(∞) coefficients, truncation error falls exponentially in L. The cost of increasing L is statistical: in an unrestricted VAR(L), the coefficient dimension grows like n^2L, so estimation error typically dominates quickly.

B.2 Architectural rank regularization and complexity scaling in L

Write the (possibly time-varying) lag-stacked coefficient matrix implied by the attention architecture as

$$\boldsymbol{\mathcal{A}}_t(\boldsymbol{\theta}) = \big[\mathrm{vec}(\mathbf{A}_{t,1}(\boldsymbol{\theta})) \ \cdots \ \mathrm{vec}(\mathbf{A}_{t,L}(\boldsymbol{\theta}))\big] \in \mathbb{R}^{n^2 \times L}.$$

Under H-head attention, Proposition 5 implies the cross-lag rank restriction

$$\mathrm{rank}(\boldsymbol{\mathcal{A}}_t(\boldsymbol{\theta})) \le \min\{H, L\} \qquad \text{for every } t. \tag{B.2}$$

This rank bound is a hard (architectural) restriction. It reduces the intrinsic dimension of admissible coefficient objects from n^2L (unrestricted VAR) to $O(n^2H + HL)$.

To make this precise, consider the constant-coefficient case $\mathbf{A}_{t,\ell} \equiv \mathbf{A}_\ell$ and define

$$\boldsymbol{\mathcal{A}} = \big[\mathrm{vec}(\mathbf{A}_1) \ \cdots \ \mathrm{vec}(\mathbf{A}_L)\big].$$

Any $\boldsymbol{\mathcal{A}}$ with $\mathrm{rank}(\boldsymbol{\mathcal{A}}) \le H$ admits a factorization $\boldsymbol{\mathcal{A}} = \mathbf{U}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{n^2 \times H}$ and $\mathbf{V} \in \mathbb{R}^{L \times H}$. Hence the coefficient object can be parameterized with at most $H(n^2 + L)$ degrees of freedom (up to non-identification), instead of n^2L.

This change in scaling is central for long-memory approximation: when H is small relative to L, the estimation cost of increasing L is linear rather than quadratic in n.

Remark B.1 (Approximating VARMA with large L under rank constraints)**.** *Under Assumption B.1, the truncation error from replacing* (B.1) *by a finite VAR(L) mean is* $O(\varrho^{2L})$ *in mean square (Lemma B.1). In an unrestricted VAR(L), the estimation error typically grows with* n^2L*, forcing small* L *in macroeconomic samples. Under the cross-lag rank restriction* (B.2)*, a large-L approximation is statistically feasible when* $H \ll L$*, because the intrinsic dimension grows like* $n^2H + HL$*. This is the sense in which rank-based architectural regularization permits increasing* L *to reduce truncation bias while controlling estimation error.*

B.3 Soft regularization via priors and posterior concentration on low-dimensional sets

Architectural restrictions control complexity by shrinking the model class. A complementary strategy is to enlarge the class and regularize with a prior that concentrates on low-dimensional subsets. This provides an explicit notion of posterior concentration on low-dimensional structure and allows the data to "opt out" of the low-dimensional approximation when warranted.

Two low-dimensional structures are particularly salient in attention-based VARs.

Few dynamic modes (effective number of heads). Introduce head-specific scales $\omega_1, \ldots, \omega_H$ and write

$$\mathbf{B}^m = \omega_m \widetilde{\mathbf{B}}^m, \qquad \omega_m \geq 0, \tag{B.3}$$

with a prior that shrinks many ω_m toward zero (global-local shrinkage, spike-and-slab, etc.). Global-local priors such as the horseshoe (Carvalho et al., 2010) are convenient because they are continuous (no reversible-jump machinery) while still producing aggressive shrinkage.

Given (B.3), define the posterior "effective head count" at threshold $\varepsilon > 0$:

$$H_{\text{eff}}(\varepsilon; \boldsymbol{\theta}) = \sum_{m=1}^{H} \mathbf{1}\{\omega_m > \varepsilon\}.$$

Posterior draws of $\{\omega_m\}$ yield posterior probabilities $\Pi(H_{\text{eff}}(\varepsilon) \leq r \mid \boldsymbol{y}_{1:T})$, which provide direct inference on whether the posterior concentrates on an approximately r-dimensional cross-lag subspace.

Time invariance (degree of state dependence). Within the attention architecture, time variation in coefficients is induced by the dependence of scores on lagged data. This can be parameterized with explicit scale parameters and centered at a time-invariant submodel (Section 7.2). Posterior concentration near that submodel is then measured by the posterior of those scales.

More generally, for any low-dimensional subset $\mathcal{S} \subset \Theta$ (e.g., "at most r active heads," "time-invariant weights," "small gain variation"), define its ε-neighborhood under a discrepancy $d(\cdot, \cdot)$:

$$\mathcal{S}_\varepsilon = \{\boldsymbol{\theta} \in \Theta : \ d(\boldsymbol{\theta}, \mathcal{S}) \leq \varepsilon\}.$$

A conceptually clean posterior concentration diagnostic is $\Pi(\mathcal{S}_\varepsilon \mid \boldsymbol{y}_{1:T})$ as $\varepsilon \downarrow 0$, with d chosen to reflect the estimand of interest (e.g., predictive mean discrepancy, coefficient-stack

discrepancy, or Jacobian discrepancy for multilayer models).

B.4 Oracle-type bounds for Bayesian predictive performance

We now state an oracle inequality for Bayesian predictive log loss. The result is most naturally expressed in terms of the Bayes marginal likelihood (equivalently, the sequential Bayes predictive density), and it holds without independence assumptions because it is a variational identity for mixtures of conditional densities. Related "resolvability" and PAC-Bayes bounds appear in Barron and Cover (1991); Barron (1999); Catoni (2004); McAllester (1998); Seeger (2002).

For each parameter value $\boldsymbol{\theta}$, let $p_{\boldsymbol{\theta}}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})$ denote the model-implied conditional density. Let Π be a prior on $\boldsymbol{\theta}$, and define the Bayes mixture conditional density

$$p_{\Pi}(\boldsymbol{y}_{L+1:T} \mid \boldsymbol{y}_{1:L}) = \int \prod_{t=L+1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})\, \Pi(d\boldsymbol{\theta}).$$

Let $\Pi(\cdot \mid \boldsymbol{y}_{1:T})$ denote the corresponding posterior, and let Q denote an arbitrary probability measure on Θ.

Proposition B.1 (Variational oracle inequality for log loss)**.** *For any realized sample* $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$*,*

$$-\log p_{\Pi}(\boldsymbol{y}_{L+1:T} \mid \boldsymbol{y}_{1:L}) \le \inf_{Q} \left\{ \mathbb{E}_Q\left[-\sum_{t=L+1}^{T} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) \right] + \mathrm{KL}(Q \| \Pi) \right\}. \tag{B.4}$$

Equality holds when Q is the posterior $\Pi(\cdot \mid \boldsymbol{y}_{1:T})$.

Proof. This is the Gibbs variational identity. Let $f(\boldsymbol{\theta}) = -\sum_{t=L+1}^{T} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})$. Then

$$-\log \int e^{-f(\boldsymbol{\theta})} \Pi(d\boldsymbol{\theta}) \le \inf_{Q} \left\{ \mathbb{E}_Q[f(\boldsymbol{\theta})] + \mathrm{KL}(Q \| \Pi) \right\},$$

with equality at the Gibbs measure proportional to $e^{-f(\boldsymbol{\theta})}\Pi(d\boldsymbol{\theta})$, i.e., the posterior. □

Proposition B.1 implies that the Bayes mixture predictive density achieves, up to a complexity penalty $\mathrm{KL}(Q\|\Pi)$, the performance of the best randomized predictor Q in the model class. Taking expectations under the true DGP yields an oracle inequality in expected conditional KL divergence. Let $p^\star(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})$ denote the true conditional density and define the cumulative conditional KL risk

$$\mathcal{K}_T(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{t=L+1}^{T} \log \frac{p^\star(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})}{p_{\boldsymbol{\theta}}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})} \right], \qquad \mathcal{K}_T(\Pi) = \mathbb{E}\left[\sum_{t=L+1}^{T} \log \frac{p^\star(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})}{p_{\Pi}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}, \boldsymbol{y}_{1:t-1})} \right],$$

where $p_\Pi(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}, \boldsymbol{y}_{1:t-1})$ denotes the sequential Bayes predictive density (the usual prequential interpretation of $p_\Pi(\boldsymbol{y}_{L+1:T} \mid \boldsymbol{y}_{1:L})$).

Corollary B.1 (Expected KL oracle inequality)**.** *Under integrability of the log likelihood ratios,*

$$\frac{1}{T-L}\,\mathcal{K}_T(\Pi) \le \inf_Q \left\{ \frac{1}{T-L}\,\mathbb{E}_Q[\mathcal{K}_T(\boldsymbol{\theta})] + \frac{1}{T-L}\,\mathrm{KL}(Q\|\Pi) \right\}.$$

Proof. Apply Proposition B.1, subtract $\mathbb{E}[-\log p^\star(\boldsymbol{y}_{L+1:T} \mid \boldsymbol{y}_{1:L})]$ from both sides, and use the tower property to rewrite expected log likelihood differences as cumulative conditional KL divergences. □

Specialization to the conditional Gaussian model. If both $p^\star(\cdot \mid \mathcal{F}_{t-1})$ and $p_{\boldsymbol{\theta}}(\cdot \mid \mathcal{F}_{t-1})$ are Gaussian with common covariance $\boldsymbol{\Sigma}$ and means $\boldsymbol{\mu}_t^\star$ and $\boldsymbol{\mu}_t(\boldsymbol{\theta})$, then

$$\log \frac{p^\star(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})}{p_{\boldsymbol{\theta}}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1})} = \frac{1}{2}\,\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_t^\star - \boldsymbol{\mu}_t(\boldsymbol{\theta}))\|^2 + \text{martingale difference},$$

and the conditional KL risk reduces to a quadratic prediction risk.

Corollary B.2 (Quadratic oracle bound under homoskedastic Gaussian likelihood)**.** *Assume* $p^\star(\cdot \mid \mathcal{F}_{t-1}) = \mathcal{N}(\boldsymbol{\mu}_t^\star, \boldsymbol{\Sigma})$ *and* $p_{\boldsymbol{\theta}}(\cdot \mid \mathcal{F}_{t-1}) = \mathcal{N}(\boldsymbol{\mu}_t(\boldsymbol{\theta}), \boldsymbol{\Sigma})$ *with the same* $\boldsymbol{\Sigma} \succ 0$*. Then*

$$\frac{1}{T-L}\,\mathbb{E}\left[\sum_{t=L+1}^{\top} \frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_t^\star - \boldsymbol{\mu}_t(\Pi))\|^2\right] \le \inf_Q \left\{ \frac{1}{T-L}\,\mathbb{E}_Q\left[\sum_{t=L+1}^{\top} \frac{1}{2}\|\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{\mu}_t^\star - \boldsymbol{\mu}_t(\boldsymbol{\theta}))\|^2\right] + \frac{1}{T-L}\,\mathrm{KL}(Q\|\Pi) \right\},$$

where $\boldsymbol{\mu}_t(\Pi)$ *denotes the Bayes predictive mean.*

Implication for choosing (H, L)**: approximation vs. complexity.** Corollary B.2 reduces selection of regularization strength to balancing an approximation term and a complexity term. Under rank-based architectural regularization, the complexity penalty grows roughly like $n^2H + HL$ (coefficient object) or, under the explicit attention parameterization, like the number of free parameters P_{Tr} in (27), which scales linearly in L rather than n^2L. Combined with Lemma B.1, this yields the economically relevant conclusion: for stable long-memory targets, L can increase with T to reduce truncation bias, because the estimation penalty need not grow like n^2L when $H \ll L$.

A simple corollary illustrates the canonical scaling.

Corollary B.3 (Logarithmic growth of the optimal context length under geometric tails)**.** *Assume the target mean admits the AR(∞) representation* (B.1) *and Assumption B.1 holds. Suppose the estimation complexity penalty for the chosen regularization scheme is of the form*

$$\mathrm{pen}(L) = \frac{C_0 + C_1 L}{T - L}$$

for constants $C_0, C_1 > 0$ (e.g. C_1 proportional to an intrinsic dimension such as H or d). Then any choice L_T satisfying

$$L_T \asymp \log T$$

balances the truncation bound $O(\varrho^{2L})$ and the estimation penalty $\mathrm{pen}(L)$ *up to multiplicative constants.*

Proof sketch. Lemma B.1 gives a truncation bound of order ϱ^{2L}. Balancing ϱ^{2L} with L/T yields $L \asymp \log T$. □

Corollary B.3 is intentionally stated at a high level: the precise form of the penalty depends on whether regularization is imposed through architectural constraints ((H, d_v) fixed) or via priors that induce posterior concentration on low-dimensional sets (shrinking ω_m and λ_m). The substantive implication is that rank-based architectural regularization (or an equivalent low-dimensional prior) permits L to grow with T, a key advantage when approximating VARMA/state-space dynamics by long-lag VAR representations.

References

Barron, A. R. (1999). Information-theoretic characterization of Bayes performance and the choice of priors in parametric and nonparametric problems. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 6: Proceedings of the Sixth Valencia International Meeting*, pages 27–52. Oxford University Press.

Barron, A. R. and Cover, T. M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory*, 37(4):1034–1054.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Catoni, O. (2004). *Statistical Learning Theory and Stochastic Optimization*, volume 1851 of *Lecture Notes in Mathematics*. Springer.

Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.

McAllester, D. A. (1998). Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT)*, pages 230–234. ACM.

Seeger, M. (2002). PAC-Bayesian generalisation error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3:233–269.