Attention-Based Vector Autoregressions

Juan Antolin-Diaz
MIT Sloan

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

NBER Summer Institute Workshop on Methods and Applications for Dynamic Equilibrium Models. July 15, 2026.

Motivation

- The **Transformer** model (Vaswani et al., 2017) has revolutionized language modeling by using the **Attention mechanism** (Bahdanau et al., 2015) to flexibly model probability distributions over sequences of text.
- Given the sequential nature of language, an application to time series data appears natural.
- **The promise:** A general, flexible and scalable framework to model multivariate data, featuring **unknown forms of non-nonlinearity and state dependence**.
- **What we want:** a formal econometric framework for interpreting its structure and conducting inference in the time series domain; appropriate priors for economic time series; an investigation into its performance with sample sizes and datasets common in economics and finance;

Motivation

- The **Transformer** model (Vaswani et al., 2017) has revolutionized language modeling by using the **Attention mechanism** (Bahdanau et al., 2015) to flexibly model probability distributions over sequences of text.
- Given the sequential nature of language, an application to time series data appears natural.
- **The promise:** A general, flexible and scalable framework to model multivariate data, featuring **unknown forms of non-nonlinearity and state dependence**.
- **What we want:** a formal econometric framework for interpreting its structure and conducting inference in the time series domain; appropriate priors for economic time series; an investigation into its performance with sample sizes and datasets common in economics and finance;

Motivation

- The **Transformer** model (Vaswani et al., 2017) has revolutionized language modeling by using the **Attention mechanism** (Bahdanau et al., 2015) to flexibly model probability distributions over sequences of text.
- Given the sequential nature of language, an application to time series data appears natural.
- **The promise:** A general, flexible and scalable framework to model multivariate data, featuring **unknown forms of non-nonlinearity and state dependence**.
- **What we want:** a formal econometric framework for interpreting its structure and conducting inference in the time series domain; appropriate priors for economic time series; an investigation into its performance with sample sizes and datasets common in economics and finance;

Motivation

- The **Transformer** model (Vaswani et al., 2017) has revolutionized language modeling by using the **Attention mechanism** (Bahdanau et al., 2015) to flexibly model probability distributions over sequences of text.
- Given the sequential nature of language, an application to time series data appears natural.
- **The promise:** A general, flexible and scalable framework to model multivariate data, featuring **unknown forms of non-nonlinearity and state dependence**.
- **What we want:** a formal econometric framework for interpreting its structure and conducting inference in the time series domain; appropriate priors for economic time series; an investigation into its performance with sample sizes and datasets common in economics and finance;

This paper: preview of results

- We formalize the connection between the Attention mechanism and classic time series models:
  - It turns out ***Attention is just a VAR...!***
  - A TVP-VAR in which the time variation is a parametrized function of the data...
  - ...and a set of dimension-reduction restrictions that are closely connected to those imposed by dynamic factor models and DSGE models.
  - **Attention-Based Vector Autoregressions.**
- Develop Bayesian algorithms for inference that do not require filtering of latent parameters, and propose a **shrinkage prior centered around time-invariant structures**.
- Extend the attention model to conditional covariance and illustrate the framework with several applications.

This paper: preview of results

- We formalize the connection between the Attention mechanism and classic time series models:
  - It turns out ***Attention is just a VAR...!***
  - A TVP-VAR in which the time variation is a parametrized function of the data...
  - ...and a set of dimension-reduction restrictions that are closely connected to those imposed by dynamic factor models and DSGE models.
  - **Attention-Based Vector Autoregressions.**
- Develop Bayesian algorithms for inference that do not require filtering of latent parameters, and propose a **shrinkage prior centered around time-invariant structures**.
- Extend the attention model to conditional covariance and illustrate the framework with several applications.

This paper: preview of results

- We formalize the connection between the Attention mechanism and classic time series models:
  - It turns out ***Attention is just a VAR...!***
  - A TVP-VAR in which the time variation is a parametrized function of the data...
  - ...and a set of dimension-reduction restrictions that are closely connected to those imposed by dynamic factor models and DSGE models.
  - **Attention-Based Vector Autoregressions.**
- Develop Bayesian algorithms for inference that do not require filtering of latent parameters, and propose a **shrinkage prior centered around time-invariant structures**.
- Extend the attention model to conditional covariance and illustrate the framework with several applications.

This paper: preview of results

- We formalize the connection between the Attention mechanism and classic time series models:
  - It turns out ***Attention is just a VAR...!***
  - A TVP-VAR in which the time variation is a parametrized function of the data...
  - ...and a set of dimension-reduction restrictions that are closely connected to those imposed by dynamic factor models and DSGE models.
  - **Attention-Based Vector Autoregressions.**
- Develop Bayesian algorithms for inference that do not require filtering of latent parameters, and propose a **shrinkage prior centered around time-invariant structures**.
- Extend the attention model to conditional covariance and illustrate the framework with several applications.

This paper: preview of results

- We formalize the connection between the Attention mechanism and classic time series models:
  - It turns out ***Attention is just a VAR...!***
  - A TVP-VAR in which the time variation is a parametrized function of the data...
  - ...and a set of dimension-reduction restrictions that are closely connected to those imposed by dynamic factor models and DSGE models.
  - **Attention-Based Vector Autoregressions.**
- Develop Bayesian algorithms for inference that do not require filtering of latent parameters, and propose a **shrinkage prior centered around time-invariant structures**.
- Extend the attention model to conditional covariance and illustrate the framework with several applications.

This paper: preview of results

- We formalize the connection between the Attention mechanism and classic time series models:
  - It turns out ***Attention is just a VAR...!***
  - A TVP-VAR in which the time variation is a parametrized function of the data...
  - ...and a set of dimension-reduction restrictions that are closely connected to those imposed by dynamic factor models and DSGE models.
  - **Attention-Based Vector Autoregressions.**
- Develop Bayesian algorithms for inference that do not require filtering of latent parameters, and propose a **shrinkage prior centered around time-invariant structures**.
- Extend the attention model to conditional covariance and illustrate the framework with several applications.

This paper: preview of results

- We formalize the connection between the Attention mechanism and classic time series models:
  - It turns out ***Attention is just a VAR...!***
  - A TVP-VAR in which the time variation is a parametrized function of the data...
  - ...and a set of dimension-reduction restrictions that are closely connected to those imposed by dynamic factor models and DSGE models.
  - **Attention-Based Vector Autoregressions.**
- Develop Bayesian algorithms for inference that do not require filtering of latent parameters, and propose a **shrinkage prior centered around time-invariant structures**.
- Extend the attention model to conditional covariance and illustrate the framework with several applications.

Related literature

We bridge (and translate across) two very separate literatures

► **Econometrics:** VARs, DFMs, and Nonlinear/TVP models

VARs and factor models are the workhorse multivariate time-series tools (Sims, 1980; Hamilton, 1994; Lütkepohl, 2005; Geweke, 1977; Stock and Watson, 2002; Bai and Ng, 2002). Reduced-rank VARs (Velu et al., 1986; Lütkepohl, 1991; Reinsel et al., 2022). Time-varying coefficients, regime changes, and volatility are modeled through TVP-VARs, Markov-switching, smooth transitions, dynamic shrinkage, and GARCH-type specifications (Primiceri, 2005; Cogley and Sargent, 2005; Hamilton, 1989; Tong, 1990; Teräsvirta, 1994; Hauzenberger et al., 2024; Bollerslev, 1990; Engle, 2002; Engle et al., 1990; Creal et al., 2013; Harvey, 2013).

We introduce the Attention-Based VAR as a reduced-dimension TVP-VAR without involving latent states.

► **Machine learning:** Transformers for language and time series

Attention and Transformers provide flexible sequence models (Bahdanau et al., 2015; Vaswani et al., 2017); related time-series applications include Lim et al. (2021), Zhou et al. (2021), and Zeng et al. (2023). Recent theory connects attention to VARs, kernels, in-context regression, and identification (Lu and Yang, 2025; Tsai et al., 2019; Goel and Bartlett, 2024; Lu et al., 2024; Von Oswald et al., 2023; Akyürek et al., 2023; Henry et al., 2024; François and Ravera, 2025).

We provide **theoretical foundations**, develop appropriate priors, and demonstrate that **small-sample inference** is feasible for this class of models.

Structure of the Presentation

- Theory of Attention VARs
- Economic Interpretation
- Priors and Estimation Algorithms
- Illustrative Applications
- Conclusions

Structure of the Presentation

- **Theory of Attention VARs**
- Economic Interpretation
- Priors and Estimation Algorithms
- Illustrative Applications
- Conclusions

Econometric environment

Let $\{\boldsymbol{y}_t\}_{t\geq 1}$ be an n-dimensional vector of time series. Specify conditional moments:

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

- *Context window*: $\boldsymbol{y}_{t-L:t-1} = (\boldsymbol{y}_{t-L}, \boldsymbol{y}_{t-L+1}, \cdots, \boldsymbol{y}_{t-1})$
- *Context length*: L, the number of lags in the window
- $\boldsymbol{\theta} \in \Theta$ collects all parameters
- Example: a VAR(p) model

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{c} + \sum_{\ell=1}^{p} \mathbf{A}_\ell \boldsymbol{y}_{t-\ell}, \qquad \mathrm{Var}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{\Sigma}$$

Econometric environment

Let $\{\boldsymbol{y}_t\}_{t\geq 1}$ be an n-dimensional vector of time series. Specify conditional moments:

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

- ***Context window***: $\boldsymbol{y}_{t-L:t-1} = (\boldsymbol{y}_{t-L}, \boldsymbol{y}_{t-L+1}, \ldots, \boldsymbol{y}_{t-1})$
- ***Context length***: L, the number of lags in the window
- $\theta \in \Theta$ collects all parameters
- Example: a VAR(p) model

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{c} + \sum_{\ell=1}^{p} \mathbf{A}_\ell \boldsymbol{y}_{t-\ell}, \qquad \mathrm{Var}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{\Sigma}$$

Econometric environment

Let $\{\boldsymbol{y}_t\}_{t\geq 1}$ be an n-dimensional vector of time series. Specify conditional moments:

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

- ***Context window***: $\boldsymbol{y}_{t-L:t-1} = (\boldsymbol{y}_{t-L}, \boldsymbol{y}_{t-L+1}, \ldots, \boldsymbol{y}_{t-1})$
- ***Context length***: L, the number of lags in the window
- $\theta \in \Theta$ collects all parameters
- Example: a VAR(p) model

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{c} + \sum_{\ell=1}^{p} \mathbf{A}_\ell \boldsymbol{y}_{t-\ell}, \qquad \mathrm{Var}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{\Sigma}$$

Econometric environment

Let $\{\boldsymbol{y}_t\}_{t\geq 1}$ be an n-dimensional vector of time series. Specify conditional moments:

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

- ***Context window***: $\boldsymbol{y}_{t-L:t-1} = (\boldsymbol{y}_{t-L}, \boldsymbol{y}_{t-L+1}, \ldots, \boldsymbol{y}_{t-1})$
- ***Context length***: L, the number of lags in the window
- $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ collects all parameters
- Example: a VAR(p) model

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{c} + \sum_{\ell=1}^{p} \mathbf{A}_\ell \boldsymbol{y}_{t-\ell}, \qquad \mathrm{Var}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{\Sigma}$$

Econometric environment

Let $\{\boldsymbol{y}_t\}_{t\geq 1}$ be an n-dimensional vector of time series. Specify conditional moments:

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

- ***Context window***: $\boldsymbol{y}_{t-L:t-1} = (\boldsymbol{y}_{t-L}, \boldsymbol{y}_{t-L+1}, \ldots, \boldsymbol{y}_{t-1})$
- ***Context length***: L, the number of lags in the window
- $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ collects all parameters
- Example: a VAR(p) model

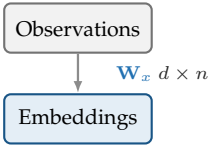
$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{c} + \sum_{\ell=1}^{p} \mathbf{A}_\ell \boldsymbol{y}_{t-\ell}, \qquad \mathrm{Var}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{\Sigma}$$

The Attention mechanism for the conditional mean: Single-head case

Observations

$$\boldsymbol{y}_{t-1}, \cdots, \boldsymbol{y}_{t-\ell}, \cdots, \boldsymbol{y}_{t-L}$$

The Attention mechanism for the conditional mean: Single-head case

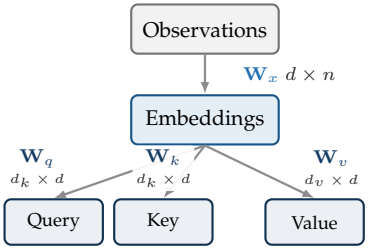


$\boldsymbol{y}_{t-1}, \cdots, \boldsymbol{y}_{t-\ell}, \cdots, \boldsymbol{y}_{t-L}$

$\boldsymbol{x}_{t-\ell} = \boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}, \quad$ for every $\ell = 1, \ldots, L$

maps the data into a d-dimensional space, $d \leq n$

The Attention mechanism for the conditional mean: Single-head case



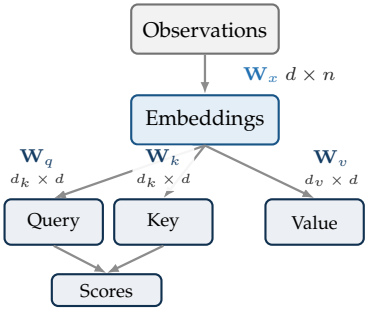
$$\boldsymbol{y}_{t-1}, \cdots, \boldsymbol{y}_{t-\ell}, \cdots, \boldsymbol{y}_{t-L}$$

$$\boldsymbol{x}_{t-\ell} = \boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}, \quad \text{for every } \ell = 1, \ldots, L$$

maps the data into a d-dimensional space, $d \leq n$

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

The Attention mechanism for the conditional mean: Single-head case



$\boldsymbol{y}_{t-1}, \cdots, \boldsymbol{y}_{t-\ell}, \cdots, \boldsymbol{y}_{t-L}$

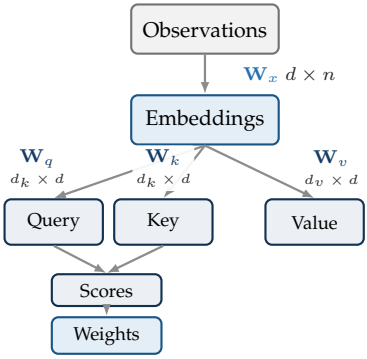
$\boldsymbol{x}_{t-\ell} = \boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}, \quad$ for every $\ell = 1, \ldots, L$

maps the data into a d-dimensional space, $d \leq n$

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

$s_{t,\ell} = b_\ell + \boldsymbol{q}_t^\top \boldsymbol{k}_{t-\ell} / \sqrt{d_k}$

The Attention mechanism for the conditional mean: Single-head case



$$\boldsymbol{y}_{t-1}, \cdots, \boldsymbol{y}_{t-\ell}, \cdots, \boldsymbol{y}_{t-L}$$

$$\boldsymbol{x}_{t-\ell} = \boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}, \quad \text{for every } \ell = 1, \ldots, L$$

maps the data into a d-dimensional space, $d \leq n$

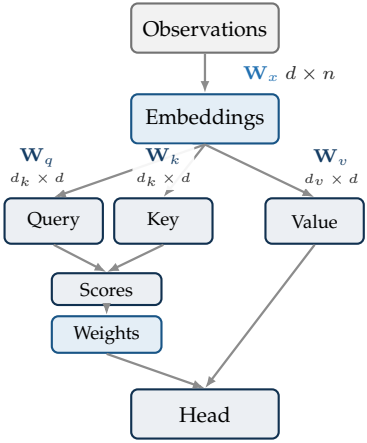
$$\boldsymbol{q}_t = \mathbf{W}_q \boldsymbol{x}_{t-1}, \quad \boldsymbol{k}_{t-\ell} = \mathbf{W}_k \boldsymbol{x}_{t-\ell}, \quad \boldsymbol{v}_{t-\ell} = \mathbf{W}_v \boldsymbol{x}_{t-\ell}$$

$$s_{t,\ell} = b_\ell + \boldsymbol{q}_t^\top \boldsymbol{k}_{t-\ell} / \sqrt{d_k}$$

$$\alpha_{t,\ell} = \text{softmax}_\ell(s_{t,\ell}) \equiv \frac{\exp(s_{t,\ell})}{\sum_{j=1}^{L} \exp(s_{t,j})}$$

key-query "similarity" mapped into weights $\alpha > 0$, $\sum \alpha = 1$

The Attention mechanism for the conditional mean: Single-head case



$\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-\ell}, \ldots, \boldsymbol{y}_{t-L}$

$\boldsymbol{x}_{t-\ell} = \boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}, \quad$ for every $\ell = 1, \ldots, L$

maps the data into a d-dimensional space, $d \leq n$

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

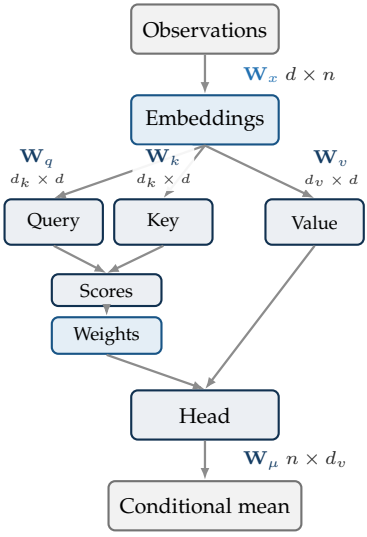
$s_{t,\ell} = b_\ell + \boldsymbol{q}_t^\top \boldsymbol{k}_{t-\ell} / \sqrt{d_k}$

$\alpha_{t,\ell} = \mathrm{softmax}_\ell(s_{t,\ell}) \equiv \frac{\exp(s_{t,\ell})}{\sum_{j=1}^{L} \exp(s_{t,j})}$

key-query "similarity" mapped into weights $\alpha > 0$, $\sum \alpha = 1$

$\boldsymbol{h}_t = \sum_{\ell=1}^{L} \alpha_{t,\ell} \, \boldsymbol{v}_{t-\ell}$

The Attention mechanism for the conditional mean: Single-head case



$$\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-\ell}, \ldots, \boldsymbol{y}_{t-L}$$

$$\boldsymbol{x}_{t-\ell} = \boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}, \quad \text{for every } \ell = 1, \ldots, L$$

maps the data into a d-dimensional space, $d \leq n$

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

$$s_{t,\ell} = b_\ell + \boldsymbol{q}_t^\top \boldsymbol{k}_{t-\ell} / \sqrt{d_k}$$

$$\alpha_{t,\ell} = \mathrm{softmax}_\ell(s_{t,\ell}) \equiv \frac{\exp(s_{t,\ell})}{\sum_{j=1}^{L} \exp(s_{t,j})}$$

key-query "similarity" mapped into weights $\alpha > 0$, $\sum \alpha = 1$

$$\boldsymbol{h}_t = \sum_{\ell=1}^{L} \alpha_{t,\ell} \, \boldsymbol{v}_{t-\ell}$$

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t$$

Some Identification Results

- The attention model is not identified [**Proposition 1**]: different values of $\mathbf{W}_\mu, \mathbf{W}_x, \mathbf{W}_v$, $\mathbf{W}_q, \mathbf{W}_k$ and $\mathbf{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L], \boldsymbol{b} = (b_1, \ldots, b_L)$ imply the same conditional mean.

  For real invertible matrices $\mathbf{G}, \mathbf{S}, \mathbf{C}$, and scalar c of appropriate dimensions:

$$\mathbf{W}_\mu^\star = \mathbf{W}_\mu \mathbf{S}^{-1}, \quad \mathbf{P}^\star = \mathbf{G}\mathbf{P}, \quad \mathbf{W}_x^\star = \mathbf{G}\mathbf{W}_x, \quad \mathbf{W}_v^\star = \mathbf{S}\mathbf{W}_v\mathbf{G}^{-1},$$
$$\mathbf{W}_q^\star = \mathbf{W}_q\mathbf{G}^{-1} \text{ or } (\mathbf{C}^{-1})^\top \mathbf{W}_q, \quad \mathbf{W}_k^\star = \mathbf{W}_k\mathbf{G}^{-1} \text{ or } \mathbf{C}\mathbf{W}_k, \quad \boldsymbol{b}^\star = \boldsymbol{b} + c\mathbf{1}_L.$$

  Each transformation leaves $g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1})$ unchanged.

- Let $\mathbf{W}_x = [\widetilde{\mathbf{W}}_x \ \mathbf{W}_{x,f}]$, assume $\widetilde{\mathbf{W}}_x$, formed by the first d columns of $\mathbf{W}_x$, is invertible. The following normalizations [**Corollary 1**] rule out the invariances above:

$$\mathbf{W}_v = \mathbf{I}_d, \qquad \widetilde{\mathbf{W}}_x = \mathbf{I}_d, \qquad \mathbf{F} = \mathbf{W}_q^\top \mathbf{W}_k, \quad \operatorname{rank}(\mathbf{F}) = d_k, \qquad b_1 = 0.$$

Some Identification Results

- The attention model is not identified [**Proposition 1**]: different values of $\mathbf{W}_\mu, \mathbf{W}_x, \mathbf{W}_v$, $\mathbf{W}_q, \mathbf{W}_k$ and $\mathbf{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L], \boldsymbol{b} = (b_1, \ldots, b_L)$ imply the same conditional mean.

  For real invertible matrices $\mathbf{G}, \mathbf{S}, \mathbf{C}$, and scalar c of appropriate dimensions:

$$\mathbf{W}_\mu^\star = \mathbf{W}_\mu \mathbf{S}^{-1}, \quad \mathbf{P}^\star = \mathbf{G}\mathbf{P}, \quad \mathbf{W}_x^\star = \mathbf{G}\mathbf{W}_x, \quad \mathbf{W}_v^\star = \mathbf{S}\mathbf{W}_v\mathbf{G}^{-1},$$
$$\mathbf{W}_q^\star = \mathbf{W}_q\mathbf{G}^{-1} \text{ or } (\mathbf{C}^{-1})^\top \mathbf{W}_q, \quad \mathbf{W}_k^\star = \mathbf{W}_k\mathbf{G}^{-1} \text{ or } \mathbf{C}\mathbf{W}_k, \quad \boldsymbol{b}^\star = \boldsymbol{b} + c\mathbf{1}_L.$$

  Each transformation leaves $g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1})$ unchanged.

- Let $\mathbf{W}_x = [\widetilde{\mathbf{W}}_x \ \mathbf{W}_{x,f}]$, assume $\widetilde{\mathbf{W}}_x$, formed by the first d columns of $\mathbf{W}_x$, is invertible. The following normalizations [**Corollary 1**] rule out the invariances above:

$$\mathbf{W}_v = \mathbf{I}_d, \qquad \widetilde{\mathbf{W}}_x = \mathbf{I}_d, \qquad \mathbf{F} = \mathbf{W}_q^\top \mathbf{W}_k, \quad \text{rank}(\mathbf{F}) = d_k, \qquad b_1 = 0.$$

Some Identification Results

- The attention model is not identified [**Proposition 1**]: different values of $\mathbf{W}_\mu, \mathbf{W}_x, \mathbf{W}_v$, $\mathbf{W}_q, \mathbf{W}_k$ and $\mathbf{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_L], \boldsymbol{b} = (b_1, \ldots, b_L)$ imply the same conditional mean.

  For real invertible matrices $\mathbf{G}, \mathbf{S}, \mathbf{C}$, and scalar c of appropriate dimensions:

$$\mathbf{W}_\mu^\star = \mathbf{W}_\mu \mathbf{S}^{-1}, \quad \mathbf{P}^\star = \mathbf{G}\mathbf{P}, \quad \mathbf{W}_x^\star = \mathbf{G}\mathbf{W}_x, \quad \mathbf{W}_v^\star = \mathbf{S}\mathbf{W}_v\mathbf{G}^{-1},$$
$$\mathbf{W}_q^\star = \mathbf{W}_q\mathbf{G}^{-1} \text{ or } (\mathbf{C}^{-1})^\top \mathbf{W}_q, \quad \mathbf{W}_k^\star = \mathbf{W}_k\mathbf{G}^{-1} \text{ or } \mathbf{C}\mathbf{W}_k, \quad \boldsymbol{b}^\star = \boldsymbol{b} + c\mathbf{1}_L.$$

  Each transformation leaves $g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1})$ unchanged.

- Let $\mathbf{W}_x = [\widetilde{\mathbf{W}}_x \ \mathbf{W}_{x,f}]$, assume $\widetilde{\mathbf{W}}_x$, formed by the first d columns of $\mathbf{W}_x$, is invertible. The following normalizations [**Corollary 1**] rule out the invariances above:

$$\mathbf{W}_v = \mathbf{I}_d, \qquad \widetilde{\mathbf{W}}_x = \mathbf{I}_d, \qquad \mathbf{F} = \mathbf{W}_q^\top \mathbf{W}_k, \quad \operatorname{rank}(\mathbf{F}) = d_k, \qquad b_1 = 0.$$

TVP-VAR representation

- Using the normalizations:

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] &= \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t \\ &= \boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu (\boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}) \\ &= \underbrace{\boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \boldsymbol{p}_\ell}_{\boldsymbol{b}_t} + \underbrace{\sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \mathbf{W}_x \boldsymbol{y}_{t-\ell}}_{\sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell}}.\end{aligned}$$

- The conditional mean can be re-written as a TVP-VAR model [**Proposition 3**]

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b}_t + \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell},$$

$$\text{with} \quad \mathbf{A}_{t,\ell} = \alpha_{t,\ell} \mathbf{B}, \qquad \mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x, \qquad \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} = \mathbf{B}.$$

TVP-VAR representation

- Using the normalizations:

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] &= \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t \\ &= \boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu (\boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}) \\ &= \underbrace{\boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \boldsymbol{p}_\ell}_{\boldsymbol{b}_t} + \underbrace{\sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \mathbf{W}_x \boldsymbol{y}_{t-\ell}}_{\sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell}}.\end{aligned}$$

- The conditional mean can be re-written as a TVP-VAR model [**Proposition 3**]

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b}_t + \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell},$$

$$\text{with} \quad \mathbf{A}_{t,\ell} = \alpha_{t,\ell} \mathbf{B}, \qquad \mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x, \qquad \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} = \mathbf{B}.$$

TVP-VAR representation

- Using the normalizations:

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] &= \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t \\ &= \boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu (\boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}) \\ &= \underbrace{\boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \boldsymbol{p}_\ell}_{\boldsymbol{b}_t} + \underbrace{\sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \mathbf{W}_x \boldsymbol{y}_{t-\ell}}_{\sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell}}.\end{aligned}$$

- The conditional mean can be re-written as a TVP-VAR model [**Proposition 3**]

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b}_t + \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell},$$

$$\text{with} \quad \mathbf{A}_{t,\ell} = \alpha_{t,\ell} \mathbf{B}, \qquad \mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x, \qquad \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} = \mathbf{B}.$$

TVP-VAR representation

- Using the normalizations:

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] &= \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t \\ &= \boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu (\boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}) \\ &= \underbrace{\boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \boldsymbol{p}_\ell}_{\boldsymbol{b}_t} + \underbrace{\sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \mathbf{W}_x \boldsymbol{y}_{t-\ell}}_{\sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell}}.\end{aligned}$$

- The conditional mean can be re-written as a TVP-VAR model [**Proposition 3**]

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b}_t + \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell},$$

$$\text{with} \quad \mathbf{A}_{t,\ell} = \alpha_{t,\ell} \mathbf{B}, \qquad \mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x, \qquad \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} = \mathbf{B}.$$

TVP-VAR representation

- Using the normalizations:

$$\begin{aligned}\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] &= \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t \\ &= \boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu (\boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}) \\ &= \underbrace{\boldsymbol{b}_\mu + \sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \boldsymbol{p}_\ell}_{\boldsymbol{b}_t} + \underbrace{\sum_{\ell=1}^{L} \alpha_{t,\ell} \mathbf{W}_\mu \mathbf{W}_x \boldsymbol{y}_{t-\ell}}_{\sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell}}.\end{aligned}$$

- The conditional mean can be re-written as a TVP-VAR model [**Proposition 3**]

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b}_t + \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell},$$

$$\text{with} \quad \mathbf{A}_{t,\ell} = \alpha_{t,\ell} \mathbf{B}, \qquad \mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x, \qquad \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} = \mathbf{B}.$$

Example: AR(2) with state-dependent coefficients

Univariate case ($n = d = d_k = 1$), $L = 2$, with the normalizations:

$$\mathbf{W}_x = 1, \qquad \mathbf{F} = \beta \geq 0, \qquad \boldsymbol{b} = (0, \delta), \qquad \delta \neq 0,$$
$$\boldsymbol{p}_1 = \boldsymbol{p}_2 = 0, \qquad b_\mu = 0, \qquad \mathbf{W}_\mu = \varphi, \qquad |\varphi| < 1.$$

Then $x_{t-\ell} = y_{t-\ell}$, $\mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x = \varphi$, and the scores are

$$s_{t,1} = \beta y_{t-1}^2, \qquad s_{t,2} = \delta + \beta y_{t-1} y_{t-2}, \qquad \alpha_{t,\ell} = \mathrm{softmax}_\ell(s_{t,\ell}) \equiv \frac{\exp(s_{t,\ell})}{\sum_{j=1}^{L} \exp(s_{t,j})}.$$

$$y_t = \underbrace{\varphi \alpha_{t,1}}_{\phi_{1,t}} y_{t-1} + \underbrace{\varphi(1 - \alpha_{t,1})}_{\phi_{2,t}} y_{t-2} + \varepsilon_t$$

- Smooth-transition autoregression with **data-driven** lag selection
- $s_{t,1} > s_{t,2}$ iff $\beta y_{t-1}(y_{t-1} - y_{t-2}) > \delta$: lag 1 receives more weight. $\beta \to 0$: constant weights.

Example: AR(2) with state-dependent coefficients

Univariate case ($n = d = d_k = 1$), $L = 2$, with the normalizations:

$$\mathbf{W}_x = 1, \qquad \mathbf{F} = \beta \geq 0, \qquad \boldsymbol{b} = (0, \delta), \qquad \delta \neq 0,$$
$$\boldsymbol{p}_1 = \boldsymbol{p}_2 = 0, \qquad b_\mu = 0, \qquad \mathbf{W}_\mu = \varphi, \qquad |\varphi| < 1.$$

Then $x_{t-\ell} = y_{t-\ell}$, $\mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x = \varphi$, and the scores are

$$s_{t,1} = \beta y_{t-1}^2, \qquad s_{t,2} = \delta + \beta y_{t-1} y_{t-2}, \qquad \alpha_{t,\ell} = \mathrm{softmax}_\ell(s_{t,\ell}) \equiv \frac{\exp(s_{t,\ell})}{\sum_{j=1}^{L} \exp(s_{t,j})}.$$

$$y_t = \underbrace{\varphi \alpha_{t,1}}_{\phi_{1,t}} y_{t-1} + \underbrace{\varphi(1 - \alpha_{t,1})}_{\phi_{2,t}} y_{t-2} + \varepsilon_t$$

- Smooth-transition autoregression with **data-driven** lag selection
- $s_{t,1} > s_{t,2}$ iff $\beta y_{t-1}(y_{t-1} - y_{t-2}) > \delta$: lag 1 receives more weight. $\beta \to 0$: constant weights.

Example: AR(2) with state-dependent coefficients

Univariate case ($n = d = d_k = 1$), $L = 2$, with the normalizations:

$$\mathbf{W}_x = 1, \qquad \mathbf{F} = \beta \geq 0, \qquad \boldsymbol{b} = (0, \delta), \qquad \delta \neq 0,$$
$$\boldsymbol{p}_1 = \boldsymbol{p}_2 = 0, \qquad b_\mu = 0, \qquad \mathbf{W}_\mu = \varphi, \qquad |\varphi| < 1.$$

Then $x_{t-\ell} = y_{t-\ell}$, $\mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x = \varphi$, and the scores are

$$s_{t,1} = \beta y_{t-1}^2, \qquad s_{t,2} = \delta + \beta y_{t-1} y_{t-2}, \qquad \alpha_{t,\ell} = \mathrm{softmax}_\ell(s_{t,\ell}) \equiv \frac{\exp(s_{t,\ell})}{\sum_{j=1}^{L} \exp(s_{t,j})}.$$

$$y_t = \underbrace{\varphi \alpha_{t,1}}_{\phi_{1,t}} y_{t-1} + \underbrace{\varphi(1 - \alpha_{t,1})}_{\phi_{2,t}} y_{t-2} + \varepsilon_t$$

- Smooth-transition autoregression with **data-driven** lag selection
- $s_{t,1} > s_{t,2}$ iff $\beta y_{t-1}(y_{t-1} - y_{t-2}) > \delta$: lag 1 receives more weight. $\beta \to 0$: constant weights.

Example: AR(2) with state-dependent coefficients

Univariate case ($n = d = d_k = 1$), $L = 2$, with the normalizations:

$$\mathbf{W}_x = 1, \qquad \mathbf{F} = \beta \geq 0, \qquad \boldsymbol{b} = (0, \delta), \qquad \delta \neq 0,$$
$$\boldsymbol{p}_1 = \boldsymbol{p}_2 = 0, \qquad b_\mu = 0, \qquad \mathbf{W}_\mu = \varphi, \qquad |\varphi| < 1.$$

Then $x_{t-\ell} = y_{t-\ell}$, $\mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x = \varphi$, and the scores are

$$s_{t,1} = \beta y_{t-1}^2, \qquad s_{t,2} = \delta + \beta y_{t-1} y_{t-2}, \qquad \alpha_{t,\ell} = \mathrm{softmax}_\ell(s_{t,\ell}) \equiv \frac{\exp(s_{t,\ell})}{\sum_{j=1}^{L} \exp(s_{t,j})}.$$

$$y_t = \underbrace{\varphi \alpha_{t,1}}_{\phi_{1,t}} y_{t-1} + \underbrace{\varphi(1 - \alpha_{t,1})}_{\phi_{2,t}} y_{t-2} + \varepsilon_t$$

- Smooth-transition autoregression with **data-driven** lag selection
- $s_{t,1} > s_{t,2}$ iff $\beta y_{t-1}(y_{t-1} - y_{t-2}) > \delta$: lag 1 receives more weight. $\beta \to 0$: constant weights.

Example: AR(2) with state-dependent coefficients

Univariate case ($n = d = d_k = 1$), $L = 2$, with the normalizations:

$$\mathbf{W}_x = 1, \qquad \mathbf{F} = \beta \geq 0, \qquad \boldsymbol{b} = (0, \delta), \qquad \delta \neq 0,$$
$$\boldsymbol{p}_1 = \boldsymbol{p}_2 = 0, \qquad b_\mu = 0, \qquad \mathbf{W}_\mu = \varphi, \qquad |\varphi| < 1.$$

Then $x_{t-\ell} = y_{t-\ell}$, $\mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x = \varphi$, and the scores are

$$s_{t,1} = \beta y_{t-1}^2, \qquad s_{t,2} = \delta + \beta y_{t-1} y_{t-2}, \qquad \alpha_{t,\ell} = \mathrm{softmax}_\ell(s_{t,\ell}) \equiv \frac{\exp(s_{t,\ell})}{\sum_{j=1}^{L} \exp(s_{t,j})}.$$

$$y_t = \underbrace{\varphi \alpha_{t,1}}_{\phi_{1,t}} y_{t-1} + \underbrace{\varphi(1 - \alpha_{t,1})}_{\phi_{2,t}} y_{t-2} + \varepsilon_t$$

- Smooth-transition autoregression with **data-driven** lag selection
- $s_{t,1} > s_{t,2}$ iff $\beta y_{t-1}(y_{t-1} - y_{t-2}) > \delta$: lag 1 receives more weight. $\beta \to 0$: constant weights.

Useful special case: a constant VAR class

- Time variation comes only through the attention weights $\alpha_{t,\ell} = \mathrm{softmax}(s_{t,\ell})$: intuition

$$s_{t,\ell} = b_\ell + \big(\boldsymbol{p}_1^\top \mathbf{F}\, \boldsymbol{p}_\ell + \boldsymbol{p}_1^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell} + \boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \boldsymbol{p}_\ell + \boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell}\big)/\sqrt{d_k}.$$

- In the special case in which $\mathbf{F} \equiv \mathbf{W}_q^\top \mathbf{W}_k = 0$, the attention weights are fixed:

$$\alpha_{t,\ell} = \alpha_\ell = \frac{\exp(b_\ell)}{\sum_{j=1}^{L} \exp(b_j)}, \qquad \sum_{\ell=1}^{L} \alpha_\ell = 1.$$

- The conditional mean becomes a constant VAR:

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b} + \sum_{\ell=1}^{L} \mathbf{A}_\ell \boldsymbol{y}_{t-\ell}, \qquad \mathbf{A}_\ell = \alpha_\ell \mathbf{B}, \qquad \mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x.$$

Useful special case: a constant VAR class

- Time variation comes only through the attention weights $\alpha_{t,\ell} = \mathrm{softmax}(s_{t,\ell})$: intuition

$$s_{t,\ell} = b_\ell + \big(\boldsymbol{p}_1^\top \mathbf{F}\, \boldsymbol{p}_\ell + \boldsymbol{p}_1^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell} + \boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \boldsymbol{p}_\ell + \boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell}\big)/\sqrt{d_k}.$$

- In the special case in which $\mathbf{F} \equiv \mathbf{W}_q^\top \mathbf{W}_k = 0$, the attention weights are fixed:

$$\alpha_{t,\ell} = \alpha_\ell = \frac{\exp(b_\ell)}{\sum_{j=1}^{L} \exp(b_j)}, \qquad \sum_{\ell=1}^{L} \alpha_\ell = 1.$$

- The conditional mean becomes a constant VAR:

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b} + \sum_{\ell=1}^{L} \mathbf{A}_\ell \boldsymbol{y}_{t-\ell}, \qquad \mathbf{A}_\ell = \alpha_\ell \mathbf{B}, \qquad \mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x.$$

Useful special case: a constant VAR class

- Time variation comes only through the attention weights $\alpha_{t,\ell} = \mathrm{softmax}(s_{t,\ell})$: intuition

$$s_{t,\ell} = b_\ell + \big(\boldsymbol{p}_1^\top \mathbf{F}\, \boldsymbol{p}_\ell + \boldsymbol{p}_1^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell} + \boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \boldsymbol{p}_\ell + \boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell}\big)/\sqrt{d_k}.$$

- In the special case in which $\mathbf{F} \equiv \mathbf{W}_q^\top \mathbf{W}_k = 0$, the attention weights are fixed:

$$\alpha_{t,\ell} = \alpha_\ell = \frac{\exp(b_\ell)}{\sum_{j=1}^{L} \exp(b_j)}, \qquad \sum_{\ell=1}^{L} \alpha_\ell = 1.$$

- The conditional mean becomes a constant VAR:

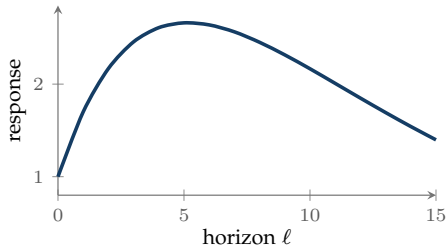
$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b} + \sum_{\ell=1}^{L} \mathbf{A}_\ell \boldsymbol{y}_{t-\ell}, \qquad \mathbf{A}_\ell = \alpha_\ell \mathbf{B}, \qquad \mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x.$$

TVP-VAR Representation: Interpretation

- The Attention model defines a (highly) restricted class of TVP-VAR models.
- We can recover an (also restricted) constant special case by setting $\mathbf{F} = \mathbf{0}$.
- $\mathbf{A}_{t,\ell} = \alpha_{t,\ell}\mathbf{B}$ for all ℓ, so coefficient matrices are proportional and $\alpha_{t,\ell} > 0 \; \forall \ell$.
- Moreover, because the embedding matrix $\mathbf{W}_x$ is $d \times n$, $\text{rank}(\mathbf{B}) = d \leq n$, and $\text{rank}(\mathbf{A}_{t,\ell}) = d$ for $\ell = 1, \ldots, L$.
- The sum of coefficients is time invariant: $\sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} = \mathbf{B}$ for all t.

Limitation of the Single-Head Model

- Consider a hump-shaped IRF



$$y_t = 1.70\, y_{t-1} - 0.7225\, y_{t-2} + \varepsilon_t$$

A single head with $L = 2$ implies

$$A_\ell = \alpha_\ell B, \qquad \alpha_1, \alpha_2 > 0, \qquad \alpha_1 + \alpha_2 = 1.$$

$$A_1 A_2 = \alpha_1 \alpha_2 B^2 \geq 0.$$

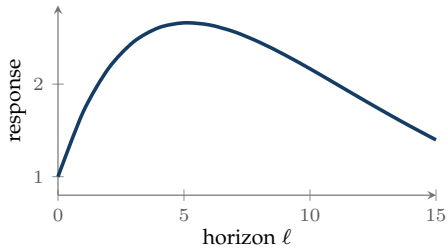
But the AR(2) above has

$$A_1 A_2 = 1.70(-0.7225) < 0.$$

- A single head cannot represent this AR(2).
- This motivates multi-head attention (Vaswani et al., 2017).

Limitation of the Single-Head Model

- Consider a hump-shaped IRF



$$y_t = 1.70\, y_{t-1} - 0.7225\, y_{t-2} + \varepsilon_t$$

A single head with $L = 2$ implies

$$A_\ell = \alpha_\ell B, \qquad \alpha_1, \alpha_2 > 0, \qquad \alpha_1 + \alpha_2 = 1.$$

$$A_1 A_2 = \alpha_1 \alpha_2 B^2 \geq 0.$$

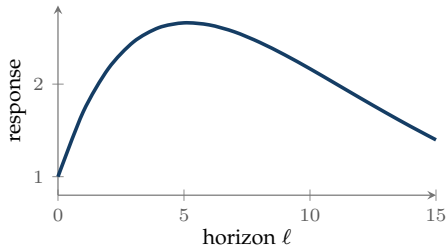
But the AR(2) above has

$$A_1 A_2 = 1.70(-0.7225) < 0.$$

- A single head cannot represent this AR(2).
- This motivates multi-head attention (Vaswani et al., 2017).

Limitation of the Single-Head Model

- Consider a hump-shaped IRF



$$y_t = 1.70\, y_{t-1} - 0.7225\, y_{t-2} + \varepsilon_t$$

A single head with $L = 2$ implies

$$A_\ell = \alpha_\ell B, \qquad \alpha_1, \alpha_2 > 0, \qquad \alpha_1 + \alpha_2 = 1.$$

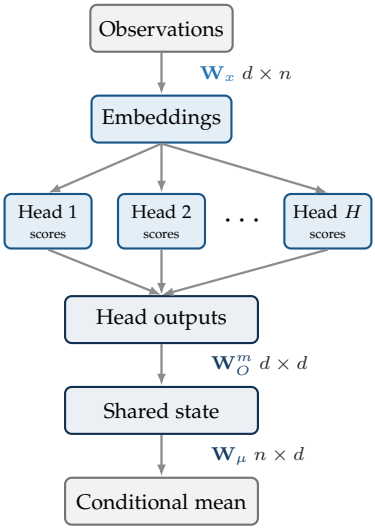
$$A_1 A_2 = \alpha_1 \alpha_2 B^2 \geq 0.$$

But the AR(2) above has

$$A_1 A_2 = 1.70(-0.7225) < 0.$$

- A single head cannot represent this AR(2).
- This motivates multi-head attention (Vaswani et al., 2017).

The Attention mechanism for the conditional mean: Multi-head case



$$\boldsymbol{y}_{t-1}, \ldots, \boldsymbol{y}_{t-\ell}, \ldots, \boldsymbol{y}_{t-L}$$

$$\boldsymbol{x}_{t-\ell} = \boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell}, \quad \ell = 1, \ldots, L$$

$$\mathbf{F}^m = (\mathbf{W}_q^m)^\top \mathbf{W}_k^m, \qquad \text{values are } \boldsymbol{x}_{t-\ell}$$

for each head $m = 1, \ldots, H$

$$s_{t,\ell}^m = b_\ell^m + \boldsymbol{x}_{t-1}^\top \mathbf{F}^m \boldsymbol{x}_{t-\ell} / \sqrt{d_k}, \quad \alpha_{t,\ell}^m = \mathrm{softmax}_\ell(s_{t,\ell}^m)$$

$$\boldsymbol{c}_t^m = \sum_{\ell=1}^L \alpha_{t,\ell}^m \boldsymbol{x}_{t-\ell}$$

$$\boldsymbol{h}_t = \sum_{m=1}^H \mathbf{W}_O^m \boldsymbol{c}_t^m$$

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b}_\mu + \mathbf{W}_\mu \boldsymbol{h}_t$$

TVP-VAR Representation: Multi-Head Case

- Substituting the H head outputs into the conditional mean:

$$\mathbb{E}[\boldsymbol{y}_t \mid \mathcal{F}_{t-1}] = \boldsymbol{b}_\mu + \mathbf{W}_\mu \sum_{m=1}^{H} \mathbf{W}_O^m \sum_{\ell=1}^{L} \alpha_{t,\ell}^m (\boldsymbol{p}_\ell + \mathbf{W}_x \boldsymbol{y}_{t-\ell})$$

$$= \boldsymbol{b}_t + \sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} \boldsymbol{y}_{t-\ell}.$$

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

$$\mathbf{A}_{t,\ell} = \sum_{m=1}^{H} \alpha_{t,\ell}^m \mathbf{B}^m, \qquad \mathbf{B}^m = \mathbf{W}_\mu \mathbf{W}_O^m \mathbf{W}_x.$$

- Multi-head attention is a TVP-VAR with fixed base matrices and state-dependent lag weights.
- Softmax weights imply $\sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} = \sum_{m=1}^{H} \mathbf{B}^m$ for all t.

Representing any Constant VAR

- To recover a constant VAR, set $\mathbf{F}^m = 0$ so each head has fixed weights. Define

$$\boldsymbol{\alpha}^m = (\alpha_1^m, \ldots, \alpha_L^m)^\top, \qquad \boldsymbol{\mathcal{M}} = \left[\boldsymbol{\alpha}^1 \ \cdots \ \boldsymbol{\alpha}^H\right], \qquad \boldsymbol{\mathcal{B}} = \left[\mathrm{vec}(\mathbf{B}^1) \ \cdots \ \mathrm{vec}(\mathbf{B}^H)\right].$$

Proposition 7: Constant-Parameter VARs

Assume $\boldsymbol{\mathcal{M}}$ is invertible.

Any arbitrary VAR(L) with coefficient matrix $\boldsymbol{\mathcal{A}} = \left[\mathrm{vec}(\mathbf{A}_1) \ \cdots \ \mathrm{vec}(\mathbf{A}_L)\right]$ can be represented by an attention model as long as $d = n$ and $H = L$ and $\mathbf{W}_\mu, \mathbf{W}_x \in \mathbb{R}^{n \times n}$ are invertible.

The base matrices are uniquely determined by $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{M}}^{-\top}$ and represented by $\mathbf{W}_O^m = \mathbf{W}_\mu^{-1}\mathbf{B}^m\mathbf{W}_x^{-1}$.

Proof

Representing any Constant VAR

- To recover a constant VAR, set $\mathbf{F}^m = 0$ so each head has fixed weights. Define

$$\boldsymbol{\alpha}^m = (\alpha_1^m, \ldots, \alpha_L^m)^\top, \qquad \boldsymbol{\mathcal{M}} = \left[\boldsymbol{\alpha}^1 \ \cdots \ \boldsymbol{\alpha}^H\right], \qquad \boldsymbol{\mathcal{B}} = \left[\mathrm{vec}(\mathbf{B}^1) \ \cdots \ \mathrm{vec}(\mathbf{B}^H)\right].$$

Proposition 7: Constant-Parameter VARs

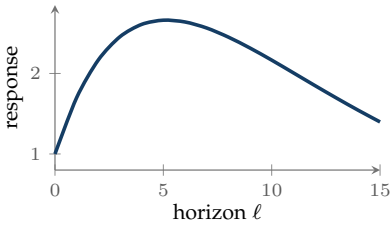
Assume $\boldsymbol{\mathcal{M}}$ is invertible.

Any arbitrary VAR(L) with coefficient matrix $\boldsymbol{\mathcal{A}} = \left[\mathrm{vec}(\mathbf{A}_1) \ \cdots \ \mathrm{vec}(\mathbf{A}_L)\right]$ can be represented by an attention model as long as $d = n$ and $H = L$ and $\mathbf{W}_\mu, \mathbf{W}_x \in \mathbb{R}^{n \times n}$ are invertible.

The base matrices are uniquely determined by $\boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{M}}^{-\top}$ and represented by $\mathbf{W}_O^m = \mathbf{W}_\mu^{-1}\mathbf{B}^m\mathbf{W}_x^{-1}$.

Proof

Example: Representing the Hump-Shaped AR(2)



$$y_t = 1.70\, y_{t-1} - 0.7225\, y_{t-2} + \varepsilon_t$$

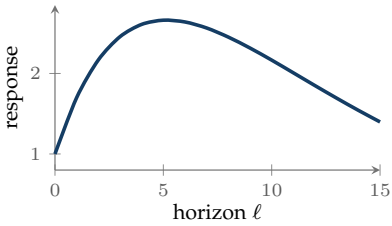
$$\boldsymbol{\alpha}^1 = \begin{pmatrix} 0.80 \\ 0.20 \end{pmatrix}, \qquad \boldsymbol{\alpha}^2 = \begin{pmatrix} 0.20 \\ 0.80 \end{pmatrix}$$

$$B^1 = 2.5075, \qquad B^2 = -1.53.$$

$$A_1 = 0.80B^1 + 0.20B^2 = 1.70,$$
$$A_2 = 0.20B^1 + 0.80B^2 = -0.7225.$$

- Two positive attention profiles reproduce the opposite-signed AR coefficients exactly.

Example: Representing the Hump-Shaped AR(2)



$$y_t = 1.70\, y_{t-1} - 0.7225\, y_{t-2} + \varepsilon_t$$

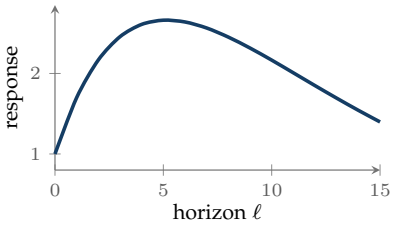
$$\boldsymbol{\alpha}^1 = \begin{pmatrix} 0.80 \\ 0.20 \end{pmatrix}, \qquad \boldsymbol{\alpha}^2 = \begin{pmatrix} 0.20 \\ 0.80 \end{pmatrix}$$

$$B^1 = 2.5075, \qquad B^2 = -1.53.$$

$$A_1 = 0.80B^1 + 0.20B^2 = 1.70,$$
$$A_2 = 0.20B^1 + 0.80B^2 = -0.7225.$$

- Two positive attention profiles reproduce the opposite-signed AR coefficients exactly.

Example: Representing the Hump-Shaped AR(2)



$$y_t = 1.70\, y_{t-1} - 0.7225\, y_{t-2} + \varepsilon_t$$

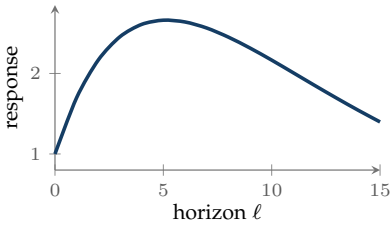
$$\boldsymbol{\alpha}^1 = \begin{pmatrix} 0.80 \\ 0.20 \end{pmatrix}, \qquad \boldsymbol{\alpha}^2 = \begin{pmatrix} 0.20 \\ 0.80 \end{pmatrix}$$

$$B^1 = 2.5075, \qquad B^2 = -1.53.$$

$$A_1 = 0.80B^1 + 0.20B^2 = 1.70,$$
$$A_2 = 0.20B^1 + 0.80B^2 = -0.7225.$$

- Two positive attention profiles reproduce the opposite-signed AR coefficients exactly.

Example: Representing the Hump-Shaped AR(2)



$$y_t = 1.70\, y_{t-1} - 0.7225\, y_{t-2} + \varepsilon_t$$

$$\boldsymbol{\alpha}^1 = \begin{pmatrix} 0.80 \\ 0.20 \end{pmatrix}, \qquad \boldsymbol{\alpha}^2 = \begin{pmatrix} 0.20 \\ 0.80 \end{pmatrix}$$

$$B^1 = 2.5075, \qquad B^2 = -1.53.$$

$$A_1 = 0.80B^1 + 0.20B^2 = 1.70,$$
$$A_2 = 0.20B^1 + 0.80B^2 = -0.7225.$$

- Two positive attention profiles reproduce the opposite-signed AR coefficients exactly.

Structure of the Presentation

- Theory of Attention VARs
- Economic Interpretation
- Priors and Estimation Algorithms
- Illustrative Applications
- Conclusions

Structure of the Presentation

- Theory of Attention VARs
- **Economic Interpretation**
- Priors and Estimation Algorithms
- Illustrative Applications
- Conclusions

Multi-head model: economic interpretation

- Proposition 7 states that we can approximate any arbitrary VAR as long as $d = n$ and $H = L$.
- The case where $d < n$ and/or $H < L$ is a restricted class, but one with a natural economic interpretation.
- Consider a **Dynamic Factor Model** with r factors and r variables measured without errors:

$$\boldsymbol{f}_t = \sum_{\ell=1}^{L} \boldsymbol{\Phi}_\ell \boldsymbol{f}_{t-\ell} + \boldsymbol{\eta}_t, \qquad \boldsymbol{y}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\xi}_t, \qquad \boldsymbol{\xi}_t = \begin{bmatrix} \mathbf{0}_r \\ \boldsymbol{\xi}_{1t} \end{bmatrix},$$
$$\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_0^\top \ \boldsymbol{\Lambda}_1^\top]^\top, \quad \det(\boldsymbol{\Lambda}_0) \neq 0, \quad \mathbf{S} = [\boldsymbol{\Lambda}_0^{-1} \ \mathbf{0}], \quad \mathbf{S}\boldsymbol{\Lambda} = \mathbf{I}_r, \quad \mathbf{S}\boldsymbol{\xi}_t = \mathbf{0}.$$

- Since $\boldsymbol{f}_t = \mathbf{S}\boldsymbol{y}_t$, the observed conditional mean has an exact restricted VAR representation:

$$\mathbb{E}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \sum_{\ell=1}^{L} \mathbf{A}_\ell^\star \boldsymbol{y}_{t-\ell}, \qquad \mathbf{A}_\ell^\star = \boldsymbol{\Lambda}\boldsymbol{\Phi}_\ell\mathbf{S}, \qquad \operatorname{rank}(\boldsymbol{\mathcal{A}}^\star) \leq \min\{L, r^2\}.$$

Multi-head model: economic interpretation

- Proposition 7 states that we can approximate any arbitrary VAR as long as $d = n$ and $H = L$.
- The case where $d < n$ and/or $H < L$ is a restricted class, but one with a natural economic interpretation.
- Consider a **Dynamic Factor Model** with r factors and r variables measured without errors:

$$\boldsymbol{f}_t = \sum_{\ell=1}^{L} \boldsymbol{\Phi}_\ell \boldsymbol{f}_{t-\ell} + \boldsymbol{\eta}_t, \qquad \boldsymbol{y}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\xi}_t, \qquad \boldsymbol{\xi}_t = \begin{bmatrix} \mathbf{0}_r \\ \boldsymbol{\xi}_{1t} \end{bmatrix},$$
$$\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_0^\top \ \boldsymbol{\Lambda}_1^\top]^\top, \quad \det(\boldsymbol{\Lambda}_0) \neq 0, \quad \mathbf{S} = [\boldsymbol{\Lambda}_0^{-1} \ \mathbf{0}], \quad \mathbf{S}\boldsymbol{\Lambda} = \mathbf{I}_r, \quad \mathbf{S}\boldsymbol{\xi}_t = \mathbf{0}.$$

- Since $\boldsymbol{f}_t = \mathbf{S}\boldsymbol{y}_t$, the observed conditional mean has an exact restricted VAR representation:

$$\mathbb{E}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \sum_{\ell=1}^{L} \mathbf{A}_\ell^\star \boldsymbol{y}_{t-\ell}, \qquad \mathbf{A}_\ell^\star = \boldsymbol{\Lambda}\boldsymbol{\Phi}_\ell\mathbf{S}, \qquad \operatorname{rank}(\boldsymbol{\mathcal{A}}^\star) \leq \min\{L, r^2\}.$$

Multi-head model: economic interpretation

- Proposition 7 states that we can approximate any arbitrary VAR as long as $d = n$ and $H = L$.
- The case where $d < n$ and/or $H < L$ is a restricted class, but one with a natural economic interpretation.
- Consider a **Dynamic Factor Model** with r factors and r variables measured without errors:

$$\boldsymbol{f}_t = \sum_{\ell=1}^{L} \boldsymbol{\Phi}_\ell \boldsymbol{f}_{t-\ell} + \boldsymbol{\eta}_t, \qquad \boldsymbol{y}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\xi}_t, \qquad \boldsymbol{\xi}_t = \begin{bmatrix} \mathbf{0}_r \\ \boldsymbol{\xi}_{1t} \end{bmatrix},$$
$$\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_0^\top \ \boldsymbol{\Lambda}_1^\top]^\top, \quad \det(\boldsymbol{\Lambda}_0) \neq 0, \quad \mathbf{S} = [\boldsymbol{\Lambda}_0^{-1} \ \mathbf{0}], \quad \mathbf{S}\boldsymbol{\Lambda} = \mathbf{I}_r, \quad \mathbf{S}\boldsymbol{\xi}_t = \mathbf{0}.$$

- Since $\boldsymbol{f}_t = \mathbf{S}\boldsymbol{y}_t$, the observed conditional mean has an exact restricted VAR representation:

$$\mathbb{E}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \sum_{\ell=1}^{L} \mathbf{A}_\ell^\star \boldsymbol{y}_{t-\ell}, \qquad \mathbf{A}_\ell^\star = \boldsymbol{\Lambda} \boldsymbol{\Phi}_\ell \mathbf{S}, \qquad \operatorname{rank}(\boldsymbol{\mathcal{A}}^\star) \leq \min\{L, r^2\}.$$

Multi-head model: economic interpretation

- Proposition 7 states that we can approximate any arbitrary VAR as long as $d = n$ and $H = L$.
- The case where $d < n$ and/or $H < L$ is a restricted class, but one with a natural economic interpretation.
- Consider a **Dynamic Factor Model** with r factors and r variables measured without errors:

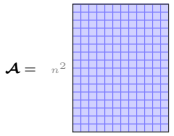
$$\boldsymbol{f}_t = \sum_{\ell=1}^{L} \boldsymbol{\Phi}_\ell \boldsymbol{f}_{t-\ell} + \boldsymbol{\eta}_t, \qquad \boldsymbol{y}_t = \boldsymbol{\Lambda} \boldsymbol{f}_t + \boldsymbol{\xi}_t, \qquad \boldsymbol{\xi}_t = \begin{bmatrix} \mathbf{0}_r \\ \boldsymbol{\xi}_{1t} \end{bmatrix},$$
$$\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_0^\top \ \boldsymbol{\Lambda}_1^\top]^\top, \quad \det(\boldsymbol{\Lambda}_0) \neq 0, \quad \mathbf{S} = [\boldsymbol{\Lambda}_0^{-1} \ \mathbf{0}], \quad \mathbf{S}\boldsymbol{\Lambda} = \mathbf{I}_r, \quad \mathbf{S}\boldsymbol{\xi}_t = \mathbf{0}.$$

- Since $\boldsymbol{f}_t = \mathbf{S}\boldsymbol{y}_t$, the observed conditional mean has an exact restricted VAR representation:

$$\mathbb{E}(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \sum_{\ell=1}^{L} \mathbf{A}_\ell^\star \boldsymbol{y}_{t-\ell}, \qquad \mathbf{A}_\ell^\star = \boldsymbol{\Lambda}\boldsymbol{\Phi}_\ell\mathbf{S}, \qquad \operatorname{rank}(\boldsymbol{\mathcal{A}}^\star) \leq \min\{L, r^2\}.$$

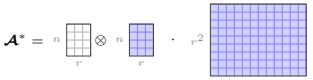
Multi-head model: dimensionality reduction

(a) Unrestricted VAR(L)



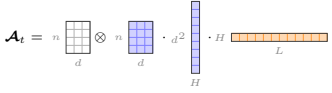
$\mathcal{A}$

(b) DFM (r factors)



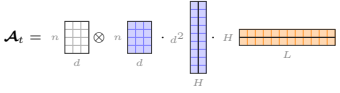
$(\mathbf{S}^\top \otimes \mathbf{\Lambda}) \cdot \mathbf{\Gamma}_\Phi$

(c) Single-Head Attention (H=1)



$(\mathbf{W}_x^\top \otimes \mathbf{W}_\mu) \cdot \widetilde{\mathbf{Q}} \cdot \boldsymbol{\alpha}_t^\top$

(d) Multi-Head Attention ($H = 2$)



$(\mathbf{W}_x^\top \otimes \mathbf{W}_\mu) \cdot \widetilde{\mathbf{Q}} \cdot \boldsymbol{\mathcal{M}}_t^\top$

- Therefore, in the attention model, the dimensions and matrices correspond to:

$$d \leftrightarrow r, \qquad \mathbf{W}_\mu \leftrightarrow \mathbf{\Lambda}, \qquad \mathbf{W}_x \leftrightarrow \mathbf{S}, \qquad H \leftrightarrow m_\Phi = \operatorname{rank}(\mathbf{\Gamma}_\Phi), \text{restrictive whenever } H < \min\{L, r^2\}.$$

Attention for the conditional mean: Summary

- Multi-head attention gives a parsimonious, observation-driven model of time variation in autoregressive coefficients.
- It also gives dimensionality reduction; parameters grow slowly with n and L. Count
- The embedding dimension $d < n$ is closely related to the number of factors r in dynamic factor models.
- The number of heads gives an additional reduction in the rank of the lag polynomial when $H < \min\{r^2, L\}$.

Attention for the covariance

- The machine learning literature gives little guidance on conditional covariance, because most work targets point forecasts.
- Predictability in macro and financial time series is low, so covariance modeling matters for density forecasts and uncertainty bands.
- Symmetric attention mechanism applied over squared residuals (no latent variables).
- Rank-reducing decomposition into q common and n idiosyncratic shocks, potentially $q \neq d$.

$$\boldsymbol{\varepsilon}_t = \mathbf{A}_{t,0}\boldsymbol{\eta}_t + \boldsymbol{e}_t, \qquad \boldsymbol{\Sigma}_t = \mathbf{A}_{t,0}\mathbf{A}_{t,0}^\top + \mathbf{D}_t.$$

$$\mathbf{A}_{t,0} = \tau_t^{1/2} \sum_{r=1}^{R} \pi_t^r \, \mathbf{W}_\varepsilon \mathbf{W}_\eta^r, \qquad \mathbf{D}_t = e^{\delta_t} \boldsymbol{\Psi}$$

Details

Structure of the Presentation

- Theory of Attention VARs
- Economic Interpretation
- Priors and Estimation Algorithms
- Illustrative Applications
- Conclusions

Structure of the Presentation

- Theory of Attention VARs
- Economic Interpretation
- **Priors and Estimation Algorithms**
- Illustrative Applications
- Conclusions

Prior specification: motivation

- Macroeconomic time series have low predictability, so flexible specifications need shrinkage.
- Persistence, unit roots, and cointegration make recent lags and long-run restrictions natural prior information.
- The Minnesota prior was designed for economic time series: shrink toward persistence and put tighter priors on more distant lags (Doan et al., 1984; Litterman, 1986).
- The mapping to time-invariant VARs and DFMs gives a natural shrinkage point; the posterior can depart from the prior if the data favors it.

Prior specification: score bias

- For exposition, consider the normalized **single-head** model. Multi-head prior
- Recall the score. The highlighted terms vary with the observations:

$$s_{t,\ell} = b_\ell + \big(\boldsymbol{p}_1^\top \mathbf{F}\, \boldsymbol{p}_\ell + \boldsymbol{p}_1^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell} + \boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \boldsymbol{p}_\ell + \boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell}\big)/\sqrt{d_k}.$$

- The prior shrinks state dependence toward zero:

$$\boldsymbol{p}_\ell \sim \mathcal{N}(\mathbf{0}, \tau_p^2 \mathbf{I}_d), \qquad \mathrm{vec}(\mathbf{F}) \sim \mathcal{N}(\mathbf{0}, \tau_F^2 \mathbf{I}_{d^2}).$$

- At this center, scores reduce to biases. Without b_ℓ, attention is uniform:

$$s_{t,\ell} = b_\ell, \qquad b_\ell = 0\ \forall \ell \quad \Longrightarrow \quad \alpha_{t,\ell} = \frac{1}{L}.$$

Lag decay and persistence

- Score biases turn the neutral attention prior into a Litterman lag-decay profile. Figure

$$b_1 = 0, \qquad b_\ell \sim \mathcal{N}(-\eta \log \ell, \tau_b^2), \qquad \ell = 2, \ldots, L.$$

- Total persistence is pinned by the cumulative feedback matrix:

$$\mathbf{B} = \mathbf{W}_\mu \mathbf{W}_x.$$

- The cumulative feedback can be shrunk toward $\mathbf{D}_\rho = \mathrm{diag}(\rho_1, \ldots, \rho_n)$ via artificial observations:

$$\mathbf{D}_\rho = \mathbf{B}^\top + \mathbf{U}_B, \qquad \mathrm{vec}(\mathbf{U}_B) \mid \mathbf{\Sigma} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma} \otimes \mathbf{\Omega}_B^{-1}), \qquad \mathbf{\Omega}_B = (\pi_1(\eta)^2/\lambda^2)\mathbf{I}_n.$$

- The two hyperparameters control different properties of the prior: η distributes persistence across lags, while λ controls shrinkage toward $\mathbf{D}_\rho$.

Bayesian estimation

- We sample from the full posterior $p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto L(\boldsymbol{\theta}) p(\boldsymbol{\theta})$.
- The state $\boldsymbol{h}_t$ is deterministic given past observations and parameters, so posterior evaluation uses the likelihood directly; no Kalman or particle filtering.
- The sampler is partially conjugate: the intercept $\boldsymbol{b}_\mu$, output map $\mathbf{W}_\mu$, and covariance $\boldsymbol{\Sigma}$ have closed-form draws.
- For the attention parameters entering nonlinearly, we use ESS is a Metropolis-Hastings update for Gaussian-prior blocks: proposals stay on a prior ellipse and are accepted by a likelihood slice.

$$\boldsymbol{\vartheta}' = \boldsymbol{m} + (\boldsymbol{\vartheta} - \boldsymbol{m}) \cos\phi + \boldsymbol{\nu} \sin\phi, \qquad \boldsymbol{\nu} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}), \qquad \ell(\boldsymbol{\vartheta}') \geq \ell(\boldsymbol{\vartheta}) + \log u.$$

Details Sampler

Timing experiments: monthly PCE at different levels of aggregation

Table: Time to obtain 5000 draws, minutes

n	$p = L$	d	H	BVAR (min.)	TVP-BVAR (min.)	ATVAR (min.)	ATVAR/BVAR	TVP/BVAR
1	12	1	2	0.0030	1.5	0.58	216.8	577.8
4	12	4	2	0.0070	30	1.7	246.4	4333.7
15	12	4	2	0.064	1.8×10^4	2.9	45.5	2.87×10^5
57	12	8	4	0.33	–	23	72.0	–
199	12	14	4	2.3	–	77	34.1	–

Structure of the Presentation

- Theory of Attention VARs
- Economic Interpretation
- Priors and Estimation Algorithms
- Illustrative Applications
- Conclusions

Structure of the Presentation

- Theory of Attention VARs
- Economic Interpretation
- Priors and Estimation Algorithms
- **Illustrative Applications**
- Conclusions

Wide range of potential applications

- Modeling inflation using many disaggregated components
- State-dependent fiscal multipliers à la Auerbach-Gorodnichenko

- **TODAY: Time variation or Rank Reduction? What features are important to describe macro data?**
  - Smets-Wouters 7-variable macroeconomic dataset
  - Fix $L = 12$. Choose different specifications for embedding dimension d, heads H, prior-tightness for time-variation in both the conditional mean and the variance.
  - Compare out-of sample RMSE performance 2010-Q1 to 2026:Q1, for forecast horizons $h = 1, \ldots, 8$ (average across horizons) against time-invariant VAR.

Wide range of potential applications

- Modeling inflation using many disaggregated components
- State-dependent fiscal multipliers à la Auerbach-Gorodnichenko

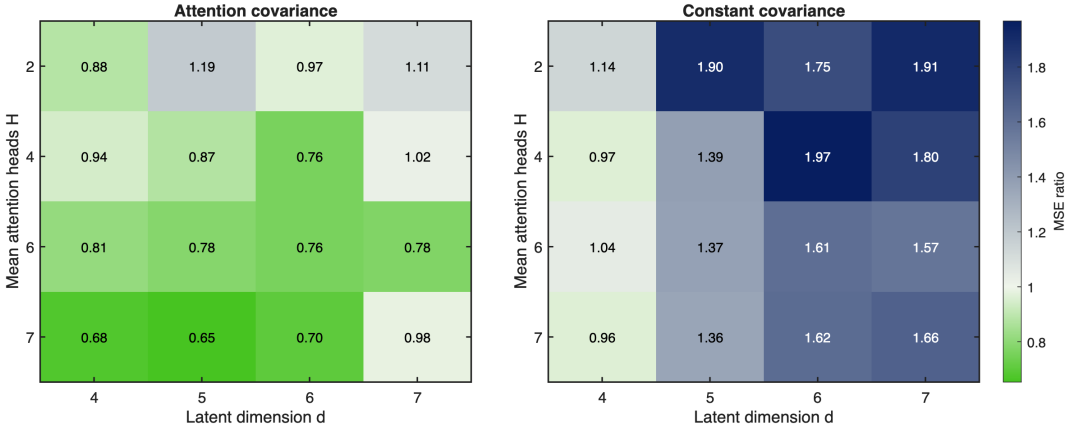
- **TODAY: Time variation or Rank Reduction? What features are important to describe macro data?**
  - Smets-Wouters 7-variable macroeconomic dataset
  - Fix $L = 12$. Choose different specifications for embedding dimension d, heads H, prior-tightness for time-variation in both the conditional mean and the variance.
  - Compare out-of sample RMSE performance 2010-Q1 to 2026:Q1, for forecast horizons $h = 1, \ldots, 8$ (average across horizons) against time-invariant VAR.

Wide range of potential applications

- Modeling inflation using many disaggregated components
- State-dependent fiscal multipliers à la Auerbach-Gorodnichenko

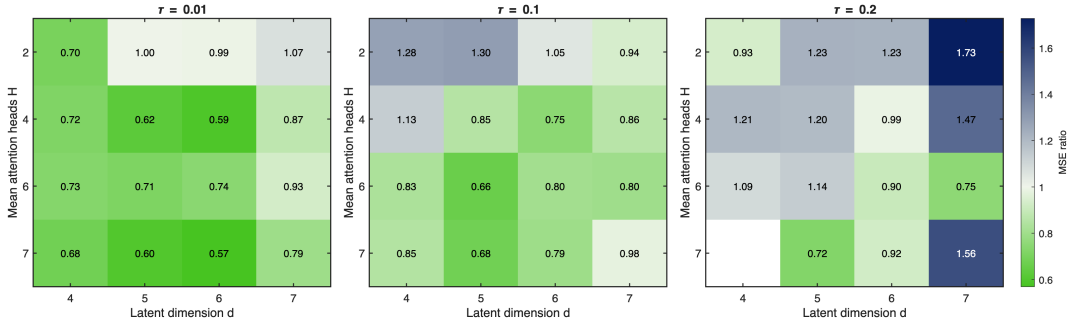
- **TODAY: Time variation or Rank Reduction? What features are important to describe macro data?**
  - Smets-Wouters 7-variable macroeconomic dataset
  - Fix $L = 12$. Choose different specifications for embedding dimension d, heads H, prior-tightness for time-variation in both the conditional mean and the variance.
  - Compare out-of sample RMSE performance 2010-Q1 to 2026:Q1, for forecast horizons $h = 1, \ldots, 8$ (average across horizons) against time-invariant VAR.

Smets–Wouters: out-of-sample GDP RMSE against BVAR



- Models with reduced d dimension, less than 12 heads, and time-varying volatility outperform.

Smets–Wouters: relative RMSE by mean prior tightness



- Conditional on time-varying volatility, tighter priors on time-variation of conditional mean do better.

Structure of the Presentation

- Theory of Attention VARs
- Economic Interpretation
- Priors and Estimation Algorithms
- Illustrative Applications
- Conclusions

Structure of the Presentation

- Theory of Attention VARs
- Economic Interpretation
- Priors and Estimation Algorithms
- Illustrative Applications
- **Conclusions**

Conclusions

- The Attention model can be understood as **a TVP-VAR with rank-reducing restrictions**.
- These restrictions are closely connected to those in factor models and have a natural **economic interpretation**.
- Bayesian inference is feasible and efficient for relevant sample sizes: no filtering is required, and Minnesota-style priors can be **centered on the linear time-invariant special case**.
- This class of models is a promising avenue to flexibly model **time variation and state dependence** in macroeconomics.

Conclusions

- The Attention model can be understood as **a TVP-VAR with rank-reducing restrictions**.
- These restrictions are closely connected to those in factor models and have a natural **economic interpretation**.
- Bayesian inference is feasible and efficient for relevant sample sizes: no filtering is required, and Minnesota-style priors can be **centered on the linear time-invariant special case**.
- This class of models is a promising avenue to flexibly model **time variation and state dependence** in macroeconomics.

Conclusions

- The Attention model can be understood as **a TVP-VAR with rank-reducing restrictions**.
- These restrictions are closely connected to those in factor models and have a natural **economic interpretation**.
- Bayesian inference is feasible and efficient for relevant sample sizes: no filtering is required, and Minnesota-style priors can be **centered on the linear time-invariant special case**.
- This class of models is a promising avenue to flexibly model **time variation and state dependence** in macroeconomics.

Conclusions

- The Attention model can be understood as **a TVP-VAR with rank-reducing restrictions**.
- These restrictions are closely connected to those in factor models and have a natural **economic interpretation**.
- Bayesian inference is feasible and efficient for relevant sample sizes: no filtering is required, and Minnesota-style priors can be **centered on the linear time-invariant special case**.
- This class of models is a promising avenue to flexibly model **time variation and state dependence** in macroeconomics.

Conclusions

- The Attention model can be understood as **a TVP-VAR with rank-reducing restrictions**.
- These restrictions are closely connected to those in factor models and have a natural **economic interpretation**.
- Bayesian inference is feasible and efficient for relevant sample sizes: no filtering is required, and Minnesota-style priors can be **centered on the linear time-invariant special case**.
- This class of models is a promising avenue to flexibly model **time variation and state dependence** in macroeconomics.

References I

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

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

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

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

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

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

References II

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

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

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

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

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

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

References III

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

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

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

N. Hauzenberger, F. Huber, and G. Koop. 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, 2024.

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

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

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

References IV

J. Lu and S. Yang. Linear transformers as VAR models: Aligning autoregressive attention mechanisms with autoregressive forecasting. *arXiv preprint arXiv:2502.07244*, 2025.

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

H. Lütkepohl. *Introduction to Multiple Time Series Analysis*. Springer-Verlag, 1991.

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

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

G. C. Reinsel, R. P. Velu, and K. Chen. *Multivariate Reduced-Rank Regression: Theory, Methods and Applications*. Springer, 2 edition, 2022.

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

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

References V

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

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

Y.-H. H. Tsai, S. Bai, M. Thattai, and R. Salakhutdinov. 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, 2019.

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

R. P. Velu, G. C. Reinsel, and D. W. Wichern. Reduced rank models for multiple time series. *Biometrika*, 73(1):105–118, 1986.

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

References VI

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

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

Conditional mean: parameter count

- After normalizations, the multi-head conditional mean has:

$$P_{\text{Att}} = 2d(n-d) + n + (dL - r_{\mathcal{U}}) + H\left(2dd_k - d_k^2 + d^2 + L - 1\right),$$
$$P_{\text{VAR}} = n + n^2 L.$$

- With fixed architectural dimensions, $P_{\text{Att}} = O(n) + O(L)$ while $P_{\text{VAR}} = O(n^2 L)$.
- Generically, $r_{\mathcal{U}} = \max\{d - Hd_k, 0\}$.

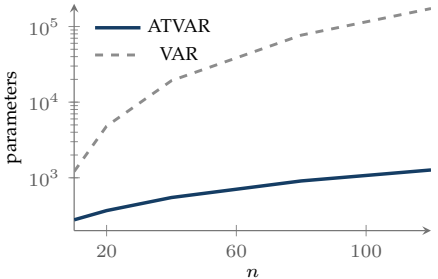


Illustration: $L = 12$, $d = d_k = H = 4$.

Back

Score terms: intuition

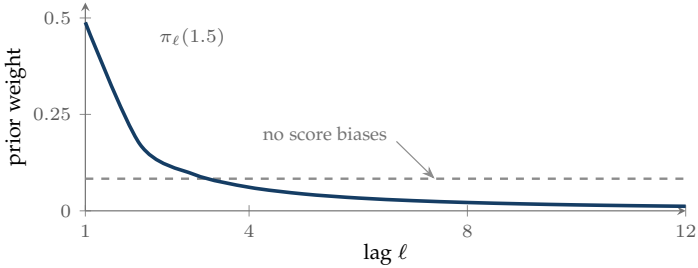
$$s_{t,\ell} = b_\ell + (\boldsymbol{p}_1^\top \mathbf{F}\, \boldsymbol{p}_\ell + \boldsymbol{p}_1^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell} + \boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \boldsymbol{p}_\ell + \boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell}) / \sqrt{d_k}.$$

- b_ℓ sets the baseline preference for lag ℓ; when $\mathbf{F} = 0$, the biases alone determine fixed weights.
- $\boldsymbol{p}_1^\top \mathbf{F}\, \boldsymbol{p}_\ell$ is a fixed position match between the current query position and lag ℓ.
- $\boldsymbol{p}_1^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell}$ and $\boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \boldsymbol{p}_\ell$ let the candidate lag and current state shift the score separately.
- $\boldsymbol{y}_{t-1}^\top \mathbf{W}_x^\top \mathbf{F}\, \mathbf{W}_x \boldsymbol{y}_{t-\ell}$ compares the current embedded state with the lagged embedded state.

Back

Lag-decay profile

$$\alpha_\ell^0 = \mathrm{softmax}(\boldsymbol{b}^0)_\ell = \frac{\ell^{-\eta}}{\sum_{j=1}^{L} j^{-\eta}} \equiv \pi_\ell(\eta).$$



Back

Attention for the covariance: Details

- The inputs are transformed squared mean residuals:

$$z_{i,t} = \varphi(\varepsilon_{i,t}^2/\psi_i), \qquad \varphi(x) = \log(x + c) - \log(1 + c).$$

- A separate attention block maps residual signals into a volatility state:

$$\begin{aligned}
\boldsymbol{x}_{t-\ell}^{\sigma} &= \boldsymbol{p}_{\ell}^{\sigma} + \mathbf{W}_{x}^{\sigma} \boldsymbol{z}_{t-\ell}, \\
s_{t,\ell}^{\sigma,m} &= b_{\ell}^{\sigma,m} + (\boldsymbol{x}_{t-1}^{\sigma})^{\top} \mathbf{F}^{\sigma,m} \boldsymbol{x}_{t-\ell}^{\sigma} / \sqrt{d_{k,\sigma}}, \\
\boldsymbol{h}_{t}^{\sigma} &= \sum_{m=1}^{H_\sigma} \mathbf{W}_{O}^{\sigma,m} \sum_{\ell=1}^{L_\sigma} \alpha_{t,\ell}^{\sigma,m} \boldsymbol{x}_{t-\ell}^{\sigma}.
\end{aligned}$$

- The state controls the size and composition of common shocks:

$$\mathbf{A}_{t,0} = \tau_t^{1/2} \sum_{r=1}^{R} \pi_t^r \, \mathbf{W}_{\varepsilon} \mathbf{W}_{\eta}^{r}, \qquad \mathbf{D}_t = e^{\delta_t} \boldsymbol{\Psi}.$$

- Positive definiteness follows from the positive diagonal $\mathbf{D}_t$; Woodbury gives likelihood evaluation at $O(nq^2)$ per period.
- Priors center the block at constant covariance; common-shock augmentation and ESS update the covariance parameters.

Back

Bayesian Estimation: Details

- The posterior uses the observation-driven likelihood:

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto \prod_{t=L+1}^{T} \mathcal{N}(\boldsymbol{y}_t; g_{\boldsymbol{\theta}}(\boldsymbol{y}_{t-L:t-1}), \boldsymbol{\Sigma})\, p(\boldsymbol{\theta}).$$

- Conditional on the nonlinear attention blocks, the mean equation is a matrix regression:

$$\mathbf{Y} = \mathbf{X}[\, \boldsymbol{b}_\mu \; \mathbf{W}_\mu \,]^\top + \mathbf{E},$$

$$\operatorname{vec}\Big([\, \boldsymbol{b}_\mu \; \mathbf{W}_\mu \,]^\top\Big) \mid \boldsymbol{\Sigma}, \cdots \sim \mathcal{N}\Big(\operatorname{vec}(\widetilde{\mathbf{M}}), \boldsymbol{\Sigma} \otimes \widetilde{\mathbf{N}}^{-1}\Big),$$

$$\boldsymbol{\Sigma} \mid \boldsymbol{b}_\mu, \mathbf{W}_\mu, \cdots \sim \mathrm{IW}(\widetilde{\nu}, \widetilde{\mathbf{S}}).$$

- ESS updates each nonlinear block $\boldsymbol{\vartheta} \in \{\mathbf{P}, \mathbf{W}_{x,f}, \mathbf{F}, \boldsymbol{b}_{2:L}\}$ with the Gaussian prior as the ellipse.
- Its log-target is the conditional log posterior excluding that Gaussian prior, e.g. $\ell(\boldsymbol{\vartheta}) = \log L(\boldsymbol{\theta}(\boldsymbol{\vartheta})) + \ell_s(\boldsymbol{z}(\boldsymbol{\vartheta})) + \cdots$.

Back

Single-head sampler with inverse-Wishart covariance

1. Given attention blocks $(\mathbf{P}, \mathbf{W}_{x,f}, \mathbf{F}, \boldsymbol{b}_{2:L})$, compute states $\boldsymbol{h}_t$ and stack the regression.
2. Draw the intercept and output map from their vector-normal conditional:
$$\operatorname{vec}\Big([\,\boldsymbol{b}_\mu \;\; \mathbf{W}_\mu\,]^\top\Big) \mid \boldsymbol{\Sigma}, \cdots \sim \mathcal{N}\Big(\operatorname{vec}(\widetilde{\mathbf{M}}), \boldsymbol{\Sigma} \otimes \widetilde{\mathbf{N}}^{-1}\Big)\,.$$
3. Draw the constant covariance from its inverse-Wishart conditional:
$$\boldsymbol{\Sigma} \mid \boldsymbol{b}_\mu, \mathbf{W}_\mu, \cdots \sim \mathrm{IW}(\widetilde{\nu}, \widetilde{\mathbf{S}}).$$
4. Update $\mathbf{P}$, $\mathbf{W}_{x,f}$, $\mathbf{F}$, and $\boldsymbol{b}_{2:L}$ block by block with ESS; repeat and retain posterior draws.

Back

Multi-head Prior

- Shared parameters $(\boldsymbol{b}_\mu, \mathbf{W}_{x,f}, \mathbf{P}, \boldsymbol{\Sigma})$ retain the single-head priors.
- Head-specific blocks are exchangeable a priori, with shared tightness parameters across heads:

$$
\begin{aligned}
b_\ell^m &\sim \mathcal{N}(-\eta \log \ell, \tau_b^2), \\
\mathrm{vec}(\mathbf{F}^m) &\sim \mathcal{N}(\mathbf{0}, \tau_F^2 \mathbf{I}_{d^2}), \\
\mathrm{vec}(\mathbf{W}_O^m) &\sim \mathcal{N}(\mathbf{0}, \tau_O^2 \mathbf{I}).
\end{aligned}
$$

- The Minnesota target is imposed on aggregate persistence, not on each head:

$$
\sum_{\ell=1}^{L} \mathbf{A}_{t,\ell} = \sum_{m=1}^{H} \mathbf{B}^m = \mathbf{W}_\mu \boldsymbol{\Pi} \mathbf{W}_x, \qquad \boldsymbol{\Pi} = \sum_{m=1}^{H} \mathbf{W}_O^m,
$$

$$
\mathbf{D}_\rho = \widetilde{\mathbf{C}}(\boldsymbol{\Pi}, \mathbf{W}_{x,f})^\top [\mathbf{I}_d \ \mathbf{W}_{\mu,f}^\top] + \mathbf{U}_B, \qquad \widetilde{\mathbf{C}}(\boldsymbol{\Pi}, \mathbf{W}_{x,f}) = \boldsymbol{\Pi}[\mathbf{I}_d \ \mathbf{W}_{x,f}].
$$

- This keeps the Minnesota persistence target shared across heads, consistent with head-label non-identification.

Back

Identification: Multi-head Case

- The model is not identified [**Proposition: Invariance to reparametrizations in the multi-head model**].
- For invertible $\mathbf{G}, \mathbf{T}, \mathbf{S}_m, \mathbf{C}_m$ and scalars c_m, the conditional mean is unchanged under:

$$\begin{aligned}
\mathbf{W}_\mu^\star &= \mathbf{W}_\mu \mathbf{T}^{-1}, \quad \mathbf{P}^\star = \mathbf{G}\mathbf{P}, \quad \mathbf{W}_x^\star = \mathbf{G}\mathbf{W}_x, \\
(\mathbf{W}_v^m)^\star &= \mathbf{S}_m \mathbf{W}_v^m \mathbf{G}^{-1}, \quad (\mathbf{W}_O^m)^\star = \mathbf{T}\mathbf{W}_O^m \mathbf{S}_m^{-1}, \\
(\mathbf{W}_q^m)^\star &= \mathbf{W}_q^m \mathbf{G}^{-1} \text{ or } (\mathbf{C}_m^{-1})^\top \mathbf{W}_q^m, \quad (\mathbf{W}_k^m)^\star = \mathbf{W}_k^m \mathbf{G}^{-1} \text{ or } \mathbf{C}_m \mathbf{W}_k^m, \\
(\boldsymbol{b}^m)^\star &= \boldsymbol{b}^m + c_m \mathbf{1}_L, \qquad m = 1, \ldots, H.
\end{aligned}$$

- Head labels are also unidentified: any permutation of the head-specific blocks leaves the conditional mean unchanged.
- The following normalizations [**Corollary: Local normalizations in the original multi-head parameterization**] rule out the continuous invariances:

$$\widetilde{\mathbf{W}}_x = \mathbf{I}_d, \quad \widetilde{\mathbf{W}}_\mu = \mathbf{I}_d, \quad \widetilde{\mathbf{W}}_{v|x}^m = \mathbf{I}_{d_v}, \quad \mathbf{F}^m = (\mathbf{W}_q^m)^\top \mathbf{W}_k^m, \quad \operatorname{rank}(\mathbf{F}^m) = d_k, \quad b_1^m = 0.$$

- Order heads by $\|\mathbf{B}^1\|_F \geq \cdots \geq \|\mathbf{B}^H\|_F$; residual common-position shifts disappear when $[\mathbf{F}^1 \ \cdots \ \mathbf{F}^H]$ has full row rank.

Back

Proof: Constant VAR Representation

For each entry (i, j), define

$$\boldsymbol{a}_{ij} = \left([\mathbf{A}_1]_{ij}, \ldots, [\mathbf{A}_L]_{ij}\right)^\top, \qquad \boldsymbol{b}_{ij} = \left([\mathbf{B}^1]_{ij}, \ldots, [\mathbf{B}^H]_{ij}\right)^\top.$$

The representation condition is a linear system:

$$[\mathbf{A}_\ell]_{ij} = \sum_{m=1}^{H} \alpha_\ell^m [\mathbf{B}^m]_{ij} \qquad \Longleftrightarrow \qquad \boldsymbol{a}_{ij} = \boldsymbol{\mathcal{M}} \boldsymbol{b}_{ij}.$$

Since $H = L$ and $\boldsymbol{\mathcal{M}}$ is invertible,

$$\boldsymbol{b}_{ij} = \boldsymbol{\mathcal{M}}^{-1} \boldsymbol{a}_{ij} \qquad \Longrightarrow \qquad \boldsymbol{\mathcal{B}} = \boldsymbol{\mathcal{A}} \boldsymbol{\mathcal{M}}^{-\top}.$$

This uniquely determines $\mathbf{B}^1, \ldots, \mathbf{B}^H$ and gives

$$\mathbf{A}_\ell = \sum_{m=1}^{H} \alpha_\ell^m \mathbf{B}^m, \qquad \ell = 1, \ldots, L.$$

Since $\mathbf{W}_\mu$ and $\mathbf{W}_x$ are invertible, set

$$\mathbf{W}_O^m = \mathbf{W}_\mu^{-1} \mathbf{B}^m \mathbf{W}_x^{-1}.$$

$$\mathbf{W}_\mu \mathbf{W}_O^m \mathbf{W}_x = \mathbf{B}^m.$$

Back