UNIVERSITY OF TEXAS ARLINGTON

Bayesian Hierarchical Dynamic Factor Models

Anthony M. Thomas, Jr.
anthony.thomas3@mavs.uta.edu

Department of Mathematics
The University of Texas at Arlington

September 5, 2020

Contents

Section 1

Bayesian Inference

Bayesian Inference can be described by two parts:

1. Build a model based on data $\boldsymbol{X}$ and parameters $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2\}$
   - Likelihood: $p(\boldsymbol{X}|\boldsymbol{\Theta})$
   - Prior: $p(\boldsymbol{\Theta})$
2. Compute the posterior
   - Posterior:
     $$p(\boldsymbol{\Theta}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})}{p(\boldsymbol{X})}$$
   - Report summaries, e.g. posterior expectations
     $$\mathbb{E}[h(\boldsymbol{\Theta})|\boldsymbol{X}]$$
     or marginal posterior expectations
     $$\mathbb{E}[h(\boldsymbol{\Theta}_i)|\boldsymbol{X}]$$

Factor Analysis

- **Factor Analysis** (FA) is a method that assumes that the covariance structure of a set of cross-sectional observations can be described in terms of a linear combination of latent variables called factors
- A sample of P observations are related to set of factors through the equation

$$\boldsymbol{X}_i = \boldsymbol{\Lambda} \boldsymbol{F}_i + \boldsymbol{e}_i, \qquad i = 1, \ldots, P \tag{1}$$

where

  - $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{Ni})^\top$ denotes a vector of observations for variable i
  - $\boldsymbol{F}_i = (F_{1i}, \ldots, F_{Ki})^\top$ denotes a vector of factors for variable i
  - $\boldsymbol{e}_i = (e_{1i}, \ldots, e_{Ni})^\top$ denotes a vector of measurement errors and idiosyncratic (unique) factors for variable i
  - $\boldsymbol{\Lambda} = [\lambda_{nk}]_{N \times K}$ denotes a matrix of factor loadings

- The following assumptions are typically made in FA:
  1. $\text{rank}(\boldsymbol{\Lambda}) = K$
  2. $\mathbb{E}[\boldsymbol{X}_i] = \mathbb{E}[\boldsymbol{e}_i] = \mathbf{0}_N$ and $\mathbb{E}[\boldsymbol{F}_i] = \mathbf{0}_K \quad \forall i$
  3. $\text{Var}(\boldsymbol{F}_i) = \mathbf{I}_K$ and $\text{Var}(\boldsymbol{e}_i) = \boldsymbol{\Sigma}$ where $\boldsymbol{\Sigma} = \text{diag}\left(\sigma_1^2, \ldots, \sigma_N^2\right) \quad \forall i$
  4. $\text{Cov}(\boldsymbol{F}_i, \boldsymbol{e}_i) = \mathbf{0}_{K \times N} \quad \forall i$
- Under these assumptions it follows that

$$\text{Var}(\boldsymbol{X}_i) = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Sigma} \quad \forall i$$

- Typical uses of FA:
  1. Dimension reduction: explain the covariation between N variables using $K < N$ factors
  2. Data interpretation: find factors that explain the covariation
  3. Theory testing: test whether a hypothesized factor structure fits observed data

Time Series Factor Analysis

- In [1] FA is extended to time series data as **time series factor analysis** (TSFA)
- A sample of T time series observations are related to the factors through the equation

$$\boldsymbol{X}_t = \boldsymbol{\Lambda}\boldsymbol{F}_t + \boldsymbol{e}_t \qquad t = 1, \ldots, T \tag{2}$$

  where
  - $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{Nt})^\top$ denotes a vector of observations at time t
  - $\boldsymbol{F}_t = (F_{1t}, \ldots, F_{Kt})^\top$ denotes a vector of factors at time t
  - $\boldsymbol{e}_t$ is a vector of measurement errors and idiosyncratic factors at time t
  - $\boldsymbol{\Lambda} = [\lambda_{nk}]_{N\times K}$ denotes a matrix of factor loadings

Dynamic Factor Analysis

- In **dynamic factor analysis** (DFA) the factors are assumed to not only affect the observations contemporaneously, but affect them through their lags as well:

$$X_{nt} = \boldsymbol{\lambda}^n(L)\boldsymbol{F}_t + \boldsymbol{e}_t \quad n = 1, \ldots, N \tag{3}$$

where

$$\boldsymbol{\lambda}^n(L) = \boldsymbol{\lambda}_0^n + \boldsymbol{\lambda}_1^n L + \cdots + \boldsymbol{\lambda}_q^n L^q$$

$$L^s \boldsymbol{F}_t = \boldsymbol{F}_{t-s} \quad \forall s \geq 0$$

is a distributed lag polynomial of factor loadings in the lag operator L for the nth series

Dynamic Factor Analysis

- In DFA the factors are modeled as a time series process
- The time series process is commonly taken to be a vector autoregressive process, i.e.

$$\boldsymbol{\Psi}(L)\boldsymbol{F}_t = \boldsymbol{\varepsilon}_t \tag{4}$$

where

$$\boldsymbol{\Psi}(L) = \mathbf{I}_K - \boldsymbol{\Psi}_1 L - \cdots - \boldsymbol{\Psi}_p L^p$$

is a matrix polynomial of autoregressive coefficients in the lag operator L

Section 2

Hierarchical Dynamic Factor Model

- In [2], the authors generalize the dynamic factor model by positing that for each t, the nth series in a given block b, denoted by X_{bnt}, has three sources of variation:
    1. idiosyncratic
    2. block-specific
    3. common

Hierarchical Dynamic Factor Model

- A three-level representation of the data for $b = 1, \ldots, B$ is given as

$$X_{bnt} = \boldsymbol{\lambda}^{n}_{G.b}(L)\boldsymbol{G}_{bt} + e_{Xbnt} \tag{5}$$

$$\boldsymbol{G}_{bt} = \boldsymbol{\Lambda}_{F.b}(L)\boldsymbol{F}_t + \boldsymbol{e}_{Gbt} \tag{6}$$

$$\boldsymbol{\Psi}_F(L)\boldsymbol{F}_t = \boldsymbol{\epsilon}_{Ft}, \tag{7}$$

  where
  - $\boldsymbol{\lambda}^{n}_{G.b}(L)$ denotes a distributed lag polynomial of block-level factor loadings
  - $\boldsymbol{\Lambda}_{F.b}(L)$ denotes a distributed lag matrix polynomial of common factor loadings
  - $\boldsymbol{G}_{bt} = (G_{b1t}, \ldots, G_{bK_{Gb}t})^{\top}$ denotes the block-level factors
  - $\boldsymbol{F}_t = (F_{1t}, \ldots, F_{K_F t})^{\top}$ denotes the common factors

Hierarchical Dynamic Factor Model

- For some blocks, it may be appropriate to break up the data into subblocks, which adds another source of variation
- Let Z_{bsnt} be the nth series in subblock s of block b
- A four-level representation of the subblock data is given as

$$\begin{aligned} Z_{bsnt} &= \boldsymbol{\lambda}^{n}_{H.bs}(L)\boldsymbol{H}_{bst} + e_{Zbsnt} \\ \boldsymbol{H}_{bst} &= \boldsymbol{\Lambda}_{G.bs}(L)\boldsymbol{G}_{bt} + \boldsymbol{e}_{Hbst} \\ \boldsymbol{G}_{bt} &= \boldsymbol{\Lambda}_{F.b}(L)\boldsymbol{F}_{t} + \boldsymbol{e}_{Gbt} \\ \boldsymbol{\Psi}_{F}(L)\boldsymbol{F}_{t} &= \boldsymbol{\epsilon}_{Ft} \end{aligned}$$

where

- $\boldsymbol{\lambda}^{n}_{H.bs}(L)$ denotes a distributed lag polynomial of subblock-level factor loadings
- $\boldsymbol{\Lambda}_{G.bs}(L)$ denotes a distributed lag matrix polynomial of block-level factor loadings
- $\boldsymbol{H}_{bst} = (H_{bs1t}, \ldots, H_{bsK_{Hbs}t})^{\top}$ denotes the subblock-level factors

Hierarchical Dynamic Factor Model

- The idiosyncratic components, the subblock-specific, block-specific, and common factors are assumed to be stationary, Gaussian autoregressive processes of orders q_{Zbsn}, q_{Xbn}, q_{Hbsi}, q_{Gbj}, and q_{Fk}, respectively, i.e.

$$\begin{aligned}
\psi_{Z.bsn}(L)e_{Zbsnt} &= \epsilon_{Zbsnt}, & \epsilon_{Zbsnt} &\sim \mathcal{N}\left(0, \sigma^2_{Zbsn}\right) & n &= 1, \ldots, N_{bs} \\
\psi_{X.bn}(L)e_{Xbnt} &= \epsilon_{Xbnt}, & \epsilon_{Xbnt} &\sim \mathcal{N}\left(0, \sigma^2_{Xbn}\right) & n &= 1, \ldots, N_{b} \\
\psi_{H.bsi}(L)e_{Hbsit} &= \epsilon_{Hbsit}, & \epsilon_{Hbsi} &\sim \mathcal{N}\left(0, \sigma^2_{Hbsi}\right) & i &= 1, \ldots, K_{Hbs} \\
\psi_{G.bj}(L)e_{Gbjt} &= \epsilon_{Gbjt}, & \epsilon_{Gbjt} &\sim \mathcal{N}\left(0, \sigma^2_{Gbj}\right) & j &= 1, \ldots, K_{Gb} \\
\psi_{F.k}(L)F_{kt} &= \epsilon_{Fkt}, & \epsilon_{Fkt} &\sim \mathcal{N}\left(0, \sigma^2_{Fk}\right) & k &= 1, \ldots, K_{F}
\end{aligned}$$

Hierarchical Dynamic Factor Model

- Not all series need to belong to blocks and subblocks, hence the data used in a four-level model are a mixture of Z_{bsnt}, X_{bnt}, and X_{nt}
- For brevity, we work with only the three-level model for which all data belongs to a block, therefore data will consist of X_{bnt} for $b = 1, \ldots, B$

Hierarchical Dynamic Factor Model

- In [2] they develop a framework to estimate the posterior distribution of the factors and model parameters
- The main steps are outlined:

Algorithm 1: Gibbs Sampler for a Hierarchical DFM

Input: Data $\{\boldsymbol{X}_t\}$

Output: Sample from posterior distribution over factors and model parameters

Initialize: $\{\boldsymbol{F}_t\}^{(0)}, \{\boldsymbol{G}_t\}^{(0)}, \{\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}\}^{(0)}$

for $i = 1, \ldots, M$ **do**

$$\{\boldsymbol{G}_t\}^{(i)} \sim \{\boldsymbol{G}_t\} | \{\boldsymbol{F}_t\}^{(i-1)}, \{\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}\}^{(i-1)}, \{\boldsymbol{X}_t\}$$

$$\{\boldsymbol{F}_t\}^{(i)} \sim \{\boldsymbol{F}_t\} | \{\boldsymbol{G}_t\}^{(i)}, \{\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}\}^{(i-1)}, \{\boldsymbol{X}_t\}$$

$$\{\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}\}^{(i)} \sim \boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Sigma} | \{\boldsymbol{F}_t\}^{(i)}, \{\boldsymbol{G}_t\}^{(i)}, \{\boldsymbol{X}_t\}$$

end

return $\left\{ \{\boldsymbol{F}_t\}^{(i)}, \{\boldsymbol{G}_t\}^{(i)}, \{\boldsymbol{\Lambda}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}\}^{(i)} \right\}_{i=1}^{M}$

Section 3

Why Variational Bayesian Inference?

- Consider a model with data $\boldsymbol{X}$ and latent variables $\boldsymbol{Z}$
- The goal is to compute the joint posterior of the latent variables given the data

$$p(\boldsymbol{Z}|\boldsymbol{X}) = \frac{p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z})}{p(\boldsymbol{X})} \tag{8}$$

  - Likelihood: $p(\boldsymbol{X}|\boldsymbol{Z})$
  - Prior: $p(\boldsymbol{Z})$
  - Evidence: $p(\boldsymbol{X})$

$$p(\boldsymbol{X}) = \int p(\boldsymbol{X}, \boldsymbol{Z})\, d\boldsymbol{Z} \tag{9}$$

Why Variational Bayesian Inference?

- For complex models (8) and (9) either have no closed-form or require high-dimensional integration which causes the “inference problem”
- As a result, the joint posterior has to be approximated
- Markov chain Monte Carlo (MCMC) has been the gold standard to solve this problem
  1. Construct an ergodic Markov chain on $\boldsymbol{Z}$ whose stationary distribution is the joint posterior $p(\boldsymbol{Z}|\boldsymbol{X})$
  2. Sample from the chain to collect samples from the stationary distribution
  3. Approximate the posterior with an empirical estimate constructed from a subset of the collected samples
  4. Use the subset of collected samples to estimate expectations of interest

Why Variational Bayesian Inference?

- A major disadvantage for MCMC is that, while it is eventually accurate, it often takes a long time to obtain results
- Variational Bayes (VB) typical obtains results much faster
- While MCMC uses sampling to solve the inference problem, VB instead uses optimization

Variational Bayesian Inference

1. Posit a family of "nice" approximate densities $\mathcal{Q}$
2. Find a member of that family that is "closest" to the exact posterior

$$q^{\star}(\boldsymbol{Z}) = \underset{q(\boldsymbol{Z}) \in \mathcal{Q}}{\arg\min}\, \mathrm{KL}(q(\boldsymbol{Z}) \,||\, p(\boldsymbol{Z}|\boldsymbol{X})) \tag{10}$$

where

$$\mathrm{KL}(q(\boldsymbol{Z}) \,||\, p(\boldsymbol{Z}|\boldsymbol{X})) = \mathbb{E}_q[\log q(\boldsymbol{Z})] - \mathbb{E}_q[\log p(\boldsymbol{Z}|\boldsymbol{X})] \tag{11}$$

Variational Bayesian Inference

It follows that

$$\mathrm{KL}(q(\boldsymbol{Z}) \,||\, p(\boldsymbol{Z}|\boldsymbol{X})) = \mathbb{E}_q[\log q(\boldsymbol{Z})] - \mathbb{E}_q[\log p(\boldsymbol{X}, \boldsymbol{Z})] + \log p(\boldsymbol{X})$$

This reveals that the objective in (10) depends on the evidence, thus it cannot be computed directly.

Variational Bayesian Inference

Instead, we optimize an alternative objective that is equivalent to (10) up to an added constant called the **evidence lower bound** (ELBO)

$$\mathrm{ELBO}(q) = \mathbb{E}_q[\log p(\boldsymbol{X}, \boldsymbol{Z})] - \mathbb{E}_q[\log q(\boldsymbol{Z})] \tag{12}$$

It can be shown that

$$\underset{q(\boldsymbol{Z})\in\mathcal{Q}}{\arg\min}\, \mathrm{KL}(q(\boldsymbol{Z}) \,||\, p(\boldsymbol{Z}|\boldsymbol{X})) = \underset{q(\boldsymbol{Z})\in\mathcal{Q}}{\arg\max}\, \mathrm{ELBO}(q)$$

Variational Bayesian Inference

The VB framework is now

1. Posit a family of “nice” approximate densities $\mathcal{Q}$
2. Find a member of that family that is “closest” to the exact posterior, i.e.

$$q^{\star}(\boldsymbol{Z}) = \underset{q(\boldsymbol{Z}) \in \mathcal{Q}}{\arg\max}\, \mathrm{ELBO}(q) \tag{13}$$

Solving this optimization problem is still difficult in general

- Using the mean-field assumption can make it easier

Mean-Field Assumption

1. Partition the latent variables into M groups, say $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M$
2. Assume that the distributions in $\mathcal{Q}$ factorize across the groups, i.e.

$$\mathcal{Q} = \left\{ q : q(\boldsymbol{Z}) = \prod_{m=1}^{M} q_m(\boldsymbol{Z}_m) \right\}$$

3. Learning the optimal q now reduces to learning the optimal $q_1, \ldots, q_M$
4. Straightforward to optimize via coordinate ascent
5. This is **NOT** a modeling assumption

Mean-Field Assumption

Interestingly, under the mean-field assumption, the optimization problem for a single q_m has the solution:

$$q_m(\boldsymbol{Z}_m) = \frac{\exp\{\mathbb{E}_{-m}[\log p(\boldsymbol{X}, \boldsymbol{Z})]\}}{\int \exp\{\mathbb{E}_{-m}[\log p(\boldsymbol{X}, \boldsymbol{Z})]\}\, d\boldsymbol{Z}_m} \tag{14}$$

This establishes what is called the **coordinate ascent variational inference** algorithm.

Coordinate Ascent Variational Inference

Algorithm 2: Coordinate ascent variational inference

Input: Model $p(\boldsymbol{X}, \boldsymbol{Z})$, Data $\boldsymbol{X}$,

Output: Variational density $q(\boldsymbol{Z}) = \prod_{m=1}^{M} q_m(\boldsymbol{Z}_m)$

Initialize: Variational densities $q_m(\boldsymbol{Z}_m)$

while *the ELBO has not converged* **do**

  for $m \in \{1, 2, \ldots, M\}$ **do**

    Set $q_m(\boldsymbol{Z}_m) \propto \exp\{\mathbb{E}_{-m}[\log p(\boldsymbol{X}, \boldsymbol{Z})]\}$

  end

  Compute ELBO(q)

end

return $q(\boldsymbol{Z})$

The Orthogonal Factor Model

Here we present a VB framework for the orthogonal time series factor analysis model. The orthogonal FA model assumes the form

$$\boldsymbol{X}_t = \boldsymbol{\Lambda}\boldsymbol{F}_t + \boldsymbol{e}_t$$

with the following assumptions:

1. $\boldsymbol{F}_t \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_K, \mathbf{I_K})$
2. $\boldsymbol{e}_t \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}_N, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma} = \text{diag}\left(\sigma_1^2, \ldots, \sigma_N^2\right)$
3. $\boldsymbol{F}_t$ and $\boldsymbol{e}_s$ are independent for every pair t, s

where $\mathbf{0}_K$ and $\mathbf{0}_N$ are zero-vectors of lengths K and N, respectively, and $\mathbf{I}_K$ is the $K \times K$ identity matrix.

Under the assumptions in the previous slide the likelihood is

$$p(\mathbf{X}_{1:T}|\mathbf{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) \propto \det(\boldsymbol{\Sigma})^{-\frac{T}{2}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} (\mathbf{X}_t - \boldsymbol{\Lambda}\mathbf{F}_t)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{X}_t - \boldsymbol{\Lambda}\mathbf{F}_t) \right\}$$

We assign the following prior distributions:

$$p(\boldsymbol{\Lambda}|\boldsymbol{\Sigma}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{\lambda}_n \,|\, \mathbf{0}_K, \sigma_n^2 \mathbf{I}_K\right)$$

$$p(\boldsymbol{\Sigma}) = \prod_{n=1}^{N} \text{Scale-inv-}\chi^2\left(\sigma_n^2 \,|\, \nu_0, \tau_0^2\right)$$

The goal here is to find a variational approximation to the posterior distribution over using CAVI, i.e. to find

$$q^\star(\boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) \approx p(\boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma} | \boldsymbol{X}_{1:T})$$

such that

$$q^\star(\boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) = \underset{q(\boldsymbol{F}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) \in \mathcal{Q}}{\arg\max} \ \mathrm{ELBO}(q)$$

where

$$\begin{aligned} \mathcal{Q} &= \{q : q(\boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) = q(\boldsymbol{F}_{1:T}) q(\boldsymbol{\Lambda}) q(\boldsymbol{\Sigma})\} \\ &= \left\{ q : q(\boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) = \prod_{t=1}^{T} q(\boldsymbol{F}_t) \prod_{n=1}^{N} q(\boldsymbol{\lambda}_n) q\left(\sigma_n^2\right) \right\} \end{aligned}$$

- With the mean-field assumption we have

$$
\begin{aligned}
\log q(\boldsymbol{F}_{1:T}) &= \mathbb{E}_{-\boldsymbol{F}}[\log p(\boldsymbol{X}_{1:T}, \boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma})] + \text{const} \\
\log q(\boldsymbol{\Lambda}) &= \mathbb{E}_{-\boldsymbol{\Lambda}}[\log p(\boldsymbol{X}_{1:T}, \boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma})] + \text{const} \\
\log q(\boldsymbol{\Sigma}) &= \mathbb{E}_{-\boldsymbol{\Sigma}}[\log p(\boldsymbol{X}_{1:T}, \boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma})] + \text{const}
\end{aligned}
$$

- In this case the log-joint $\log p(\boldsymbol{X}_{1:T}, \boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma})$ expands as

$$
\log p(\boldsymbol{X}_{1:T} | \boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) + \log p(\boldsymbol{F}_{1:T}) + \log p(\boldsymbol{\Lambda} | \boldsymbol{\Sigma}) + \log p(\boldsymbol{\Sigma})
$$

- Therefore,

$$
\begin{aligned}
\log q(\boldsymbol{F}_{1:T}) &= \mathbb{E}_{-\boldsymbol{F}}[\log p(\boldsymbol{X}_{1:T} | \boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) + \log p(\boldsymbol{F}_{1:T})] + \text{const} \\
\log q(\boldsymbol{\Lambda}) &= \mathbb{E}_{-\boldsymbol{\Lambda}}[\log p(\boldsymbol{X}_{1:T} | \boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) + \log p(\boldsymbol{\Lambda} | \boldsymbol{\Sigma})] + \text{const} \\
\log q(\boldsymbol{\Sigma}) &= \mathbb{E}_{-\boldsymbol{\Sigma}}[\log p(\boldsymbol{X}_{1:T} | \boldsymbol{F}_{1:T}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) + \log p(\boldsymbol{\Lambda} | \boldsymbol{\Sigma}) + \log p(\boldsymbol{\Sigma})] + \text{const}
\end{aligned}
$$

- Substituting in the expressions for the likelihood and prior for $\boldsymbol{F}_{1:T}$ we get

$$\log q(\boldsymbol{F}_{1:T}) = -\frac{1}{2}\sum_{t=1}^{T} \boldsymbol{F}_t^\top \left(\mathbf{I}_K + \mathbb{E}_{-\boldsymbol{F}}\left[\Lambda^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right]\right)\boldsymbol{F}_t + \sum_{t=1}^{T} \mathbf{X}_t^\top \mathbb{E}_{-\boldsymbol{F}}\left[\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}\right]\boldsymbol{F}_t + \text{const}$$

- The equation above is Gaussian wrt $\boldsymbol{F}_t$. Thus,

$$q(\boldsymbol{F}_{1:T}) = \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{F}_t \,|\, \mathbf{m}_{\boldsymbol{F}t}, \mathbf{P}_{\boldsymbol{F}})$$

where

$$\mathbf{P}_{\boldsymbol{F}}^{-1} = \mathbf{I}_K + \sum_{n=1}^{N} \mathbb{E}_{\boldsymbol{\Sigma}}\left[\frac{1}{\sigma_n^2}\right]\mathbb{E}_{\boldsymbol{\Lambda}}\left[\boldsymbol{\lambda}_n\boldsymbol{\lambda}_n^\top\right]$$
$$\mathbf{m}_{\boldsymbol{F}t} = \mathbf{P}_{\boldsymbol{F}}\,\mathbb{E}_{\boldsymbol{\Sigma}}\left[\boldsymbol{\Sigma}^{-1}\right]\mathbb{E}_{\boldsymbol{\Lambda}}[\boldsymbol{\Lambda}]\boldsymbol{X}_t$$

- Proceeding in a similar way, we have that

$$q(\boldsymbol{\Lambda}) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{\lambda}_n \,|\, \mathbf{m}_{\boldsymbol{\lambda} n}, \mathbf{P}_{\boldsymbol{\lambda} n})$$

$$q(\boldsymbol{\Sigma}) = \prod_{n=1}^{N} \text{Scale-inv-}\chi^2(\sigma_n^2 \,|\, \nu_\sigma, \tau_n^2)$$

where

$$\mathbf{P}_{\boldsymbol{\lambda} n}^{-1} = \mathbb{E}_{\boldsymbol{\Sigma}}\left[\frac{1}{\sigma_n^2}\right]\left(T\,\mathbf{I}_K + \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{F}}\left[\boldsymbol{F}_t \boldsymbol{F}_t^\top\right]\right)$$

$$\mathbf{m}_{\boldsymbol{\lambda} n} = \mathbf{P}_{\boldsymbol{\lambda} n} \mathbb{E}_{\boldsymbol{\Sigma}}\left[\frac{1}{\sigma_n^2}\right] \sum_{t=1}^{T} X_{nt} \mathbb{E}_{\boldsymbol{F}}[\boldsymbol{F}_t]$$

$$\nu_\sigma = T + \nu_0$$

$$\nu_\sigma \tau_n^2 = \nu_0 \tau_0^2 + \mathbb{E}_{\boldsymbol{\Lambda}}\left[\boldsymbol{\lambda}_n^\top \boldsymbol{\lambda}_n\right] + \sum_{t=1}^{T}\left[X_{nt}^2 - 2X_{nt}\mathbb{E}_{\boldsymbol{\Lambda}}[\boldsymbol{\lambda}_n]^\top \mathbb{E}_{\boldsymbol{F}}[\boldsymbol{F}_t] + \operatorname{tr}\left[\mathbb{E}_{\boldsymbol{\Lambda}}\left[\boldsymbol{\lambda}_n \boldsymbol{\lambda}_n^\top\right] \mathbb{E}_{\boldsymbol{F}}\left[\boldsymbol{F}_t \boldsymbol{F}_t^\top\right]\right]\right]$$

Section 4

Future Work

The main goal of the project is to develop a variational Bayesian framework to handle the four-level model from [2].

Section 5

References I

Paul D Gilbert, Erik Meijer, et al. "Time series factor analysis with an application to measuring money". In: *University of Groningen, Research School SOM Research Report 05F10* (2005).

Emanuel Moench, Serena Ng, and Simon Potter. "Dynamic Hierarchical Factor Models". en. In: *Review of Economics and Statistics* 95.5 (Dec. 2013), pp. 1811–1817. ISSN: 0034-6535, 1530-9142. DOI: 10.1162/REST_a_00359.