**FEM21038 and FEM21044**

# Assignment

### Instructions

This assignment covers two important topics from this course: state space models and Markov switching models. You can work in small teams of up to four students to code and analyse the data. The results that you obtain will allow you to answer the multiple choice questions in a Pop Quiz on Canvas. Therefore, I recommend to save your outputs in a form of a table or a separate pdf file. Note that the choices in the quiz may be rounded numbers; choose the number closest to your result. The Pop Quiz will have to be completed individually. If you do not complete the Pop Quiz on your own you will not receive any points.

You can start working on the problems as of today. The Pop Quiz will open on Monday, 20 November, 10.00, and close on Friday, 24 November, 23.59. You can complete the Pop Quiz only once during this period.

Next, you will need to submit the code that you used to analyze the data in a zip file. However, each group needs to submit only one zip file. Please name the zip file as follows: `YYYYYY-YYYYYY-YYYYYY-YYYYYY.zip` where `YYYYYY` is the ERNA-number for each student in the group (so between one and four ERNA numbers). Once the Pop Quiz opens, you will also see an Assignment on Canvas where you can upload the file.

Feel free to use the Matlab demo codes provided during the course as your starting point. While these are provided for your convenience, note that they will not be sufficient to answer all questions. Naturally, you are free to use other computer languages. However, no programming support beyond the Matlab demo code will be available.

### Data description

This assignment considers the following data for the United States, sampled at a quarterly frequency and obtained from the Federal Reserve Bank of St. Louis:

1. Personal consumption (PCE)
2. Employment (PAYEMS)
3. Industrial production (IPMAN).

The series from the St. Louis Fed are in levels and we compute annualized percentage changes, $y_{i,t} := 4 \times 100 \times (Y_{i,t}/Y_{i,t-1} - 1)$, where $Y_{i,t}$ denotes the level, $y_{i,t}$ denotes the (annualized) percentage change, and the subscript $i = 1, 2, 3$ refers to the series: (1) personal consumption, (2) employment, and (3) industrial production. The index $t$ denotes the observations and runs from $t = 1$, Q4 of 1972 (dated 1 January 1973), to $T = 203$, Q2 of 2023 (dated 1 July 2023). The *transformed* data can be found in the files `data.xls`, `data.csv`, and `data.mat`, where the last file can be loaded directly into Matlab.

All estimations should be based on the sample from $t = 1$ to $T = 189$, Q4 of 2019 (dated 1 January 2020). That is, you should exclude the last 14 observations from the data file for the estimation and keep them separate for forecasting purposes.

## QUESTION 1 (Markov switching models)

**1.a)** Consider the series $y_{3,t}$, industrial production. Use maximum likelihood (ML) to estimate a Markov switching model with switching mean and variance, that is

$$y_{3,t} \sim \mathcal{N}(\mu_{S_t}, \sigma^2_{S_t}),$$

where $S_t$ is a random variable with $\Pr[S_t = 1|S_{t-1} = 1] = p_{11}$ and $\Pr[S_t = 2|S_{t-1} = 2] = p_{22}$. We use the convention that $S_t = 2$ denotes the high-volatility state. You are asked to investigate three possible initializations for the Hamilton filter. Initialize your model by imposing that the filter starts in

- state 1, that is, set $\Pr(S_0 = 1|\mathcal{I}_0) = 1$
- state 2, that is, set $\Pr(S_0 = 2|\mathcal{I}_0) = 1$
- the long-term mean, that is, set $\Pr(S_0 = 1|\mathcal{I}_0) = (1-p_{22})/(2-p_{11}-p_{22})$ and $P(S_0 = 2|\mathcal{I}_0) = (1 - p_{11})/(2 - p_{11} - p_{22})$.

As your starting parameter values for the estimation, take $p_{11} = p_{22} = 0.8$, set $\mu_1$ and $\mu_2$ equal to the sample mean, and take $\sigma_1$ and $\sigma_2$ equal to 0.5 and 1.5 times the sample standard deviation, respectively.

What are the numerical values of the estimated coefficients for these three initializations? Save the estimation results with corresponding log-likelihood values and the log-term means. Do not use a burn-in period in the estimation.

**1.b)** Construct the $h$-periods ahead out-of-sample forecasts for $y_{3,t}$, the corresponding forecast errors and MSFE with $h = 1, 2. \dots, 14$ using the initialization in state 1.

**1.c)** Construct the one-period ahead out-of-sample forecasts for $y_{3,t}$, the corresponding forecast errors and MSFE using the initialization in state 1 by iteratively expanding the sample (without re-estimating the parameter vector).

**1.d)** Use the EM algorithm for Markov-switching models to estimate the six parameters $p_{11}, p_{22}, \mu_1, \mu_2, \sigma_1, \sigma_2$ along with the starting points $\Pr(S_0 = 1) = \rho_1$ and $\Pr(S_0 = 2) = \rho_2$. Since $\rho_1 + \rho_2 = 1$, there are seven parameters to estimate. As your initial parameter values, take $p_{11} = p_{22} = 0.8$, set $\mu_1$ and $\mu_2$ equal to the sample mean, and take $\sigma_1$ and $\sigma_2$ equal to 0.5 and 1.5 times the sample standard deviation, respectively. For your first EM run, set $\hat{\boldsymbol{\xi}}_{0|0} = (\rho_1, \rho_2)'$ with $\rho_1 = \rho_2 = 0.5$. Each EM step should update seven parameters. Note the estimated parameters after 1000 iterations.

## QUESTION 2 (State space models)

This question considers the same data as the previous question. However, for the purpose of this question, you should de-mean the estimation sample of the three data series $y_{1,t}$, $y_{2,t}$ and $y_{3,t}$. Note that the mean should be calculated based on $T = 189$ (i.e. based on the sample before 1 January 2020) and not the full sample. From this point onwards, we use the symbol $y_{i,t}$ to denote the de-meaned series $y_{i,t}$ for $i = 1, 2, 3$.

**2.a)** Estimate a simple 'AR(1) plus noise' model,

$$\begin{aligned} y_t &= \mu_t + \varepsilon_t \\ \mu_{t+1} &= \phi \mu_t + \eta_{t+1} \end{aligned}$$

for each of the three univariate series $y_{1,t}$, $y_{2,t}$, and $y_{3,t}$, for $t = 1, 2, \ldots, T$, using maximum likelihood (ML). Use the Kalman filter with a diffuse initialization, that is, take $\widehat{\xi}_{0|0} = 0$ and $P_{0|0} = 10^6$ for each series. Compute the log likelihood using the prediction-error decomposition and do not use a burn-in period (that is, be sure to sum over all observations when computing the log likelihood). Save the numerical values of the estimated parameters, the log likelihood and the steady state value $\bar{P}$.

**2.b)** Forecast the last 14 observations out-of-sample and calculate the forecast error and MSFE for each series. Do not forget to add means back when making forecasts out-of-sample.

**2.c)** We suspect that the three series are driven by the same underlying factor, which is the state of the economy. For this reason, we consider the following factor model in state space form:

$$\begin{pmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \end{pmatrix} = \begin{pmatrix} h_1 & h_2 & h_3 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}' \begin{pmatrix} \xi_{0,t} \\ \xi_{1,t} \\ \xi_{2,t} \\ \xi_{3,t} \end{pmatrix},$$

with

$$\begin{pmatrix} \xi_{0,t+1} \\ \xi_{1,t+1} \\ \xi_{2,t+1} \\ \xi_{3,t+1} \end{pmatrix} = \begin{pmatrix} f_0 & 0 & 0 & 0 \\ 0 & f_1 & 0 & 0 \\ 0 & 0 & f_2 & 0 \\ 0 & 0 & 0 & f_3 \end{pmatrix} \begin{pmatrix} \xi_{0,t} \\ \xi_{1,t} \\ \xi_{2,t} \\ \xi_{3,t} \end{pmatrix} + \mathbf{v}_t,$$

and where $\mathbf{v}_t$ is independently normally distributed for all time $t$ with mean zero and covariance matrix

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & q_1 & 0 & 0 \\ 0 & 0 & q_2 & 0 \\ 0 & 0 & 0 & q_3 \end{pmatrix}.$$

Here, $\xi_{0,t}$ can be viewed as the state of the economy, the common component, or the factor that affects all observations $y_{i,t}$, for $i = 1, 2, 3$, via the factor loadings $h_1, h_2, h_3$. The three states $\xi_{i,t}$ for $i = 1, 2, 3$ represent idiosyncratic components that cannot be explained by the common component. For identification purposes, we have normalised the variance of the innovation of the common component to 1.

Adjust the Kalman filter demo code to accommodate the increased dimensionality of the system and estimate ten unknown parameters by ML. Use a diffuse initialisation in the Kalman filter: set $\xi_{i,1|0} = 0$ for each $i = 0, 1, 2, 3$ and $\mathbf{P}_{1|0} = 10^6 \times \mathbf{I}_4$, where $\mathbf{I}_4$ denotes the $4 \times 4$ identity matrix. Do not use a burn-in period when computing the log likelihood (that is, be sure to compute the log likelihood by summing over all observations). Save the estimation results and the log likelihood value.

*Hint 1:* As your starting point for the optimisation, take

$$(f_0, f_1, f_2, f_3, h_1, h_2, h_3, q_1, q_2, q_3) = (1, 0, 1, 1, 1, 1, 7, 10, 1, 1)$$

and try different optimisation algorithms. For example, in Matlab you can use the optimisers `fminunc`, `fmincon` or `fminsearch`.

*Hint 2:* The multivariate pdf functions might be slow and problematic. In the log likelihood, use your own expression for the multivariate normal pdf. Using Matlab syntax, the multivariate normal distribution with parameters `mu` and `Sigma` evaluated at `y` is given by
`1/sqrt(det(2*pi*Sigma))*exp(-1/2*(y-mu)'*inv(Sigma)*(y-mu))`

*Hint 3:* Also, the inverse functions in most of the languages can be slow or inaccurate. Sometimes a better alternative is to use (with some extra caution) the pseudo-inverse function, which is a bit more stable. In Matlab and Julia a better alternative is to use the so-called backslash operator `A\ B` rather than `inv(A)*B` for two matrices `A` and `B` of appropriate size. In this case, the multivariate normal distribution above reads
`1/sqrt(det(2*pi*Sigma))*exp(-1/2*(y-mu)'*((Sigma)\(y-mu)))`

**2.d)** As before, exclude the last 14 observations for each series. Consider jointly the three series $y_{1,t}$, $y_{2,t}$, and $y_{3,t}$, which are collected in the vector $\mathbf{y}_t = (y_{1,t}, y_{2,t}, y_{3,t})'$. Use the expectation maximisation (EM) algorithm to estimate the parameters of the unrestricted model

$$\mathbf{y}_t = \boldsymbol{\xi}_t + \mathbf{w}_t, \quad \boldsymbol{\xi}_{t+1} = \mathbf{F}\boldsymbol{\xi}_t + \mathbf{v}_{t+1},$$

where $\mathbf{w}_t$ and $\mathbf{v}_t$ are normally distributed with mean zero and covariance matrices $\mathbf{R}$ and $\mathbf{Q}$, respectively. Use the diffuse initialisation for $\boldsymbol{\xi}_0$ in each EM step similar to the previous question. Note your parameter estimates and log-likelihood values after 20 and 1000 iterations.

As your starting values for the EM algorithm, take $\mathbf{F} = 0.95 \times \mathbf{I}_3$, $\mathbf{R} = 0.2 \times \mathbb{V}[\Delta\mathbf{y}_t]$, and $\mathbf{Q} = 0.5 \times \mathbb{V}[\Delta\mathbf{y}_t]$, where $\Delta\mathbf{y}_t$ denotes the first difference of $\mathbf{y}_t$ and $\mathbb{V}[\Delta\mathbf{y}_t]$ is the sample covariance matrix of $\Delta\mathbf{y}_t$.

*Hint 1:* Due to bounded machine precision, the filtered/smoothed covariance matrix `P(:,:,t)` may become (slightly) asymmetric if the Kalman filter/smoother is computed many times. To prevent this from happening, you may calculate `P(:,:,t)` as usual and then set

```
P(:,:,t)=(P(:,:,t)+P(:,:,t)')/2
```

thereby entering a hard constraint that enforces symmetry.

*Hint 2:* Similarly, the covariance matrices in the M step of the EM-algorithm may become (slightly) asymmetric if the EM step is computed many times. To prevent this from happening, you may compute `Q` and `R` as usual, and then set `Q=(Q+Q')/2` and `R=(R+R')/2` to enforce symmetry before proceeding with another EM step.