STA 642: Homework # 4 Exercises

1. **Course mini-project thinking, proposal and development**. Continue progressing on this. Email &/or talk to MW and TA Jiurui as convenient (some of you are already clearly moving along well here ...). In the near future, you will be required to submit and be assessed on a 2 page outline of your project.

2. Review the DLM specification and forward filtering code for univariate DLMs in the course Matlab code using the Sales:Index time series example. There are a details of model and prior specification that we will move into next week, and it is critical that you are on top of the basic model structuring and sequential analysis technicalities prior to these next steps. One key aspect of this is that model specification of evolution variance matrices $\mathbf{W}_t$ uses discount factors based on the component discount specification of P&W Section 4.3.6, p131, and as in course slides that we will review next week. Read and digest that material in P&W.

   The code creates an example DLM for the Sales:Index analysis– a DLM with a trend component, a dynamic regression component, and a (Fourier) seasonal component. It then performs and summarizes forward filtering. Explore the model analysis as specified for the Sales:Index data as exemplified in class. Rerun the example and make sure you are really on top of the summaries of analysis produced. And, play with changes to the model specification and explore how summary "results" change. We will revisit this next week, and then in the next homework, with deeper investigation of issues of model choice and comparison.

3. Derive/reproduce the Bayesian filtering theory that is at the core of the Kalman filtering (+variance learning) equations: P&W Section 4.3.1 and 4.3.2, and as reviewed in class and summarized in class slides. This an extension (to dynamic contexts, and with the state evolution prior to posterior update) of multivariate normal/inverse-gamma Bayesian theory; "simply"... but key and critical. Working through the details will help you fix ideas (as well as become comfortable with the notation). Working through some of the code will help too.

4. Consider a DLM with time t observation variance v_t known. The evolution model setup is summarised by the time $t-1$ posterior $(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \sim N(\mathbf{m}_{t-1}, \mathbf{C}_{t-1})$ that evolves through the state equation $\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t$ where $\boldsymbol{\omega}_t \sim N(\mathbf{0}, \mathbf{W}_t)$ and $\boldsymbol{\theta}_{t-1} \perp\!\!\!\perp \boldsymbol{\omega}_t$.

   Suppose now the special case in which:

   - the model has a random walk state evolution, i.e., $\mathbf{G}_t = \mathbf{I}$ for all t, and
   - the evolution variance matrix at time t is defined by the specific formula $\mathbf{W}_t = \epsilon\mathbf{C}_{t-1}$ where $\epsilon = (1-\delta)/\delta$ for some *discount factor* $\delta \in (0, 1)$.

   (a) Show how the "Kalman filter" update equations for prior:posterior analysis at time t simplify in this special case.

   Here $\mathbf{a}_t = \mathbf{m}_{t-1}$ and $\mathbf{R}_t = \mathbf{C}_{t-1} + \mathbf{W}_t$ reduces to $\mathbf{R}_t = \mathbf{C}_{t-1}/\delta$. The usual update to $(\mathbf{m}_t, \mathbf{C}_t)$ then applies with these specific, simplified values.

   (b) Comment on the simplified structure and how it depends on the chosen/specified discount factor δ.

In the evolution step, overall uncertainty about the state vector is increased by $100\epsilon\%$ in the change from $\mathbf{C}_{t-1}$ to $\mathbf{C}_{t-1}/\delta$. A value of δ near 1 implies a small increase, consistent with the view that the state vector is changing only slightly over time; in contrast, a lower discount factor implies a larger increase consistent with more substantial change in the state vector elements. Note also that evolution step maintains correlations among elements of the state vector. Finally, the same discount rate applies to all elements of the state vector– in some contexts it may be desirable to model different rates of change of sub-vectors.

(c) Comment on the computational implications of this simplified structure.

Only that there is no linear algebra required in the evolution step.

Additional comments: The usual filtering equations (Kalman filter equations) can, of course, be written in the alternative "normal prior-posterior update" form using precision matrices; that is, for all $t > 0$,

$$\mathbf{m}_t = \mathbf{C}_t(\mathbf{R}_t^{-1}\mathbf{m}_{t-1} + v_t^{-1}\mathbf{F}_t y_t) \quad \text{and} \quad \mathbf{C}_t^{-1} = \mathbf{R}_t^{-1} + v_t^{-1}\mathbf{F}_t\mathbf{F}_t'.$$

This is general, for any DLM. The practical computational benefits of the Kalman filter form are obvious as they avoid matrix inversions. The alternative forms do provide insights into the role of discounting in this special class of models when $\mathbf{G}_t = \mathbf{I}$ for all t, and using the single discount factor δ. These special cases are, simply, dynamic regression models. In these cases, it follows that

$$\mathbf{C}_t^{-1}\mathbf{m}_t = \delta\mathbf{C}_{t-1}^{-1} + v_t^{-1}\mathbf{F}_t y_t \quad \text{and} \quad \mathbf{C}_t^{-1} = \delta\mathbf{C}_{t-1}^{-1} + v_t^{-1}\mathbf{F}_t\mathbf{F}_t'.$$

On recursing backwards over time, these give

$$\mathbf{C}_t^{-1}\mathbf{m}_t = \delta^t\mathbf{C}_0^{-1}\mathbf{m}_0 + \sum_{r=1:t}\delta^{t-r}v_r^{-1}\mathbf{F}_r y_r$$

and

$$\mathbf{C}_t^{-1} = \delta^t\mathbf{C}_0^{-1} + \sum_{r=1:t}\delta^{t-r}v_r^{-1}\mathbf{F}_r\mathbf{F}_r'.$$

The equations explicitly show the role of δ as an exponential decay factor on historical information: in inference on $\boldsymbol{\theta}_t$, the observation pair $y_r, \mathbf{F}_r$ at a previous time r is weighted by δ^{t-r} in computing $\mathbf{m}_t, \mathbf{C}_t$, and this weight decays as $t-r$ increases. As a point estimate of $\boldsymbol{\theta}_t$, the vector $\mathbf{m}_t$ is an exponentially (matrix) weighted regression estimate, more heavily weighting recent data than data at older time points.

5. This exercise concerns key components of theory for a main part of analysis of DLMs that we will move into soon– backward simulation of "time trajectories" of latent state-vectors in DLMs. This question essentially works through to prove the key results in P&W Section 4.3.5. This will help ensure understanding of the concepts, the role of the Markovian structure of

a DLM in retrospective analysis, and facility in manipulation of some of the main aspects of theory relevant to inference in DLMs.

Consider a DLM in which, for all time t, we have known observation variance v_t. The resulting forward filtering analysis is then based on the simple normal theory and resulting Kalman filtering-based equations. Then, given $\mathcal{D}_{t-1}$, the two consecutive state vectors $\boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_{t-1}$ are related linearly with Gaussian error, and so the two state vectors have a joint normal distribution $p(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1})$ in the implied $2p-$dimensions. We already know the mean vectors and variance matrices of the two $p-$dimensional margins, i.e.,

$$\begin{aligned} E(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) = \mathbf{m}_{t-1}, \quad E(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) = \mathbf{a}_t, \\ V(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) = \mathbf{C}_{t-1}, \quad V(\boldsymbol{\theta}_t|\mathcal{D}_{t-1}) = \mathbf{R}_t. \end{aligned}$$

So all we need to find to have all the parameters is the $p \times p$ covariance matrix between the two state vectors.

(a) Show that $C(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) = \mathbf{G}_t\mathbf{C}_{t-1}$, and hence that $C(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t|\mathcal{D}_{t-1}) = \mathbf{C}_{t-1}\mathbf{G}_t'$.

Using the state evolution equation it is trivial that

$$\begin{aligned} C(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) &= C(\mathbf{G}_t\boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \\ &= C(\mathbf{G}_t\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \; + \; C(\boldsymbol{\omega}_t, \boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \\ &= \mathbf{G}_t C(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) \; + \mathbf{0} \quad (\text{using } \boldsymbol{\theta}_{t-1} \perp\!\!\!\perp \boldsymbol{\omega}_t) \\ &= \mathbf{G}_t V(\boldsymbol{\theta}_{t-1}|\mathcal{D}_{t-1}) = \mathbf{G}_t\mathbf{C}_{t-1}. \end{aligned}$$

The transposed matrix immediately gives $C(\boldsymbol{\theta}_{t-1}, \boldsymbol{\theta}_t|\mathcal{D}_{t-1}) = \mathbf{C}_{t-1}\mathbf{G}_t'$.

(b) Deduce that the $p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1})$ is normal and prove that the mean vector and variance matrix are as given in P&W eqns. (4.12,13).

We have established that, conditional on $\mathcal{D}_{t-1}$,

$$\begin{pmatrix} \boldsymbol{\theta}_{t-1} \\ \boldsymbol{\theta}_t \end{pmatrix} \sim N\left(\begin{pmatrix} \mathbf{m}_{t-1} \\ \mathbf{a}_t \end{pmatrix}, \begin{pmatrix} \mathbf{C}_{t-1} & \mathbf{C}_{t-1}\mathbf{G}_t' \\ \mathbf{G}_t\mathbf{C}_{t-1} & \mathbf{R}_t \end{pmatrix} \right).$$

Then standard multivariate normal theory gives the required conditional as normal with homoskedastic linear regression, viz

$$\begin{aligned} E(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}) &= \mathbf{m}_{t-1} + \mathbf{B}_{t-1}(\boldsymbol{\theta}_t - \mathbf{a}_t) \\ V(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}) &= \mathbf{C}_{t-1} - \mathbf{B}_{t-1}\mathbf{R}_t\mathbf{B}_{t-1}' \end{aligned}$$

where $\mathbf{B}_{t-1} = \mathbf{C}_{t-1}\mathbf{G}_t'\mathbf{R}_t^{-1}$ (the latter is just the matrix of regression coefficients given by the covariance matrix $\mathbf{C}_{t-1}\mathbf{G}_t'$ “divided by” the variance matrix $\mathbf{R}_t$ of the conditioning information.)

(c) Suppose you are standing at a future time $n \geq t$ and so have available all the data up to that time, which we can write as $\mathcal{D}_n = \{\mathcal{D}_{t-1}, y_{t:n}\}$. What is the distribution $p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_n)$?

This *"look back one step"* conditional posterior is a central ingredient of retrospective analysis of time series using DLMs; see P&W Section 4.3.5.

In words: the model is 1st-order Markov in the states, so learning $\boldsymbol{\theta}_t$ renders $\boldsymbol{\theta}_{t-1}$ conditionally independent of all future states $\boldsymbol{\theta}_{t+k}$ for $k \geq 1$. Since each future observation y_{t+k} (for $k \geq 0$) is conditionally independent of $\boldsymbol{\theta}_{t-1}$ given its time $t+k$ state vector $\boldsymbol{\theta}_{t+k}$, it follows that the conditioning on $\boldsymbol{\theta}_t$ means future y values are also conditionally independent of $\boldsymbol{\theta}_{t-1}$. Hence

$$p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_n) \equiv p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}), \qquad \text{for all } n \geq t.$$

That is, simply the time $t-1$ conditional just derived above.

In pictures: the directed graphical model representing the dependence structure of the defined DLM is, as we know, that of the HMM

$$\begin{tikzcd}
 & y_{t-1} & y_t & y_{t+1} & & y_n \\
{} \arrow[r] & \boldsymbol{\theta}_{t-1} \arrow[u] \arrow[r] & \boldsymbol{\theta}_t \arrow[u] \arrow[r] & \boldsymbol{\theta}_{t+1} \arrow[u] \arrow[r] & \cdots & \boldsymbol{\theta}_n \arrow[u]
\end{tikzcd} .$$

The implied undirected graphical model of *the full joint distribution of all state vectors and observations* is then simply

$$\begin{tikzcd}[arrows=-]
 & y_{t-1} & y_t & y_{t+1} & & y_n \\
{} \arrow[r] & \boldsymbol{\theta}_{t-1} \arrow[u] \arrow[r] & \boldsymbol{\theta}_t \arrow[u] \arrow[r] & \boldsymbol{\theta}_{t+1} \arrow[u] \arrow[r] & \cdots & \boldsymbol{\theta}_n \arrow[u]
\end{tikzcd}$$

where we have dropped the arrowheads (and "moralised" the graph by marry any unmarried parents– there are none to marry). We know how to "read" this graphical model: conditioning on any node in the graph between two (or more) other nodes simply deletes the edges from the node conditioned upon. So conditioning on $\boldsymbol{\theta}_t$ cuts the graph into pieces, with $\boldsymbol{\theta}_{t-1}$ having its edges to the past and to $\boldsymbol{\theta}_t$ but being separated from all future states and observations. Separation (no edges) is precisely conditional independence. This is a formal proof that $p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_n) \equiv p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1})$ for all $n \geq t$.

In mathematics: You can do it via Bayes theorem too, but it is trivial and more relevant to do it in words or pictures, as that is where the intuition is. The mathematics simply bears out the picture:

$$p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_n) \propto p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}) p(y_{t:n}|\boldsymbol{\theta}_{t-1:t}, \mathcal{D}_{t-1})$$

by Bayes theorem. Then, conditional on $\boldsymbol{\theta}_t$ we know that the future data $y_{t:n}$ are conditionally independent of $\boldsymbol{\theta}_{t-1}$; knowing $\boldsymbol{\theta}_t$ breaks the dependence. So the conditional likelihood function here does not depend on $\boldsymbol{\theta}_{t-1}$ at all, and so $p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_n) \equiv p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1})$.

(d) Briefly comment on the role of this theory in quantifying the retrospective distribution for a full trajectory of states $p(\boldsymbol{\theta}_{1:n}|\mathcal{D}_n)$ at any chosen time point n.

FFBS– forward filtering to compute and save sufficient summaries at each t, then perfom backward or, better, retrospective simulation by (i) sampling $p(\boldsymbol{\theta}_n|\mathcal{D}_n)$, and then (ii) cascading back over times $t = n : -1 : 2$ to sample $p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1})$ with the just-simulated value of $\boldsymbol{\theta}_t$ in the conditioning. As in slides and P&W.

(e) Consider now a specific class of DLMs as in Exercise 4 above; that is, the special cases in which:

- the model has a random walk state evolution, i.e., $\mathbf{G}_t = \mathbf{I}$ for all t, and
- the evolution variance matrix at time t is defined by the specific formula $\mathbf{W}_t = \epsilon \mathbf{C}_{t-1}$ where $\epsilon = (1-\delta)/\delta$ for some *discount factor* $\delta \in (0, 1)$.

Show how the above results relevant to retrospective analysis change and simplify in these special cases, discussing both the role of δ as well as computational considerations.

Here $\mathbf{a}_t = \mathbf{m}_{t-1}$ and $\mathbf{R}_t = \mathbf{C}_{t-1}/\delta$ so that $\mathbf{B}_{t-1} = \delta\mathbf{I}$. As a result, the normal "one-step back" conditional normal distribution at time $t-1$ has moments

$$\begin{aligned} E(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}) &= \mathbf{m}_{t-1} + \delta(\boldsymbol{\theta}_t - \mathbf{m}_{t-1}) = (1-\delta)\mathbf{m}_{t-1} + \delta\boldsymbol{\theta}_t \\ V(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_{t-1}) &= (1-\delta)\mathbf{C}_{t-1}. \end{aligned}$$

Note that the retrospective conditional has a mean vector that is a convex linear combination of the estimate $\mathbf{m}_{t-1}$ made at the time, and the just-sampled value of the next state $\boldsymbol{\theta}_t$; the relative weights naturally reflect the thinking underlying levels of variation in the state evolution as defined by the specified value of δ.

Clearly these simplified formulæ lead to significant savings in both storage over times $1:n$ and then in linear algebra to computer– and simulate– retrospectively at each time.

6. Read and digest the web-linked supplement on harmonic (Fourier) analysis, taken from the W&H (yellow book) support text. The first sections are basic harmonic/Fourier analysis material that underlie practicable models for time-varying seasonal structure in time series broadly. Our initial examples have already entered into this class of models, and we will continue in the coming classes.