STA 642: Spring 2020 – Homework #1 Exercises

Exercises run through basics of AR(1) models and some extensions, to ensure familiarity with key concepts and models. Models/contexts appearing in exercises will be recurrent: they are chosen as a mix of drill, extending class notes and discussion, anticipating material coming along, as building blocks for more elaborate models, and as vehicles for new concepts and methods.

Hand-in solutions to *all* questions, generating understanding and mastery of basic material (as well as good preparation for the mid-term exam), but understand that *not all will be assessed*.

Unless explicitly requested (some questions will explicitly request, most do not ..) do not work derivations of theoretical results from scratch if they relates to known theory. For example, (i) a linear combination of normal random quantities is normal: just quote and define the implied mean and variance, and (ii) a normal prior for a normal, linear regression parameter leads to a normal posterior with mean and variance given by standard formulæ– quote the formulæ, you do not need to reinvent the (quadratic form completion etc etc) wheels! If a question asks "what is the distribution of ... "and the result is a distribution in a known family (e.g., normal, T, etc) then simply state the result; you do not need to derive or write-out density functions.

Hand-in LaTeX drafted solutions to the assessed questions; include whatever numerical summaries and graphs you regard as relevant to support your stories and exploratory conclusions in data analyses, as well as detailed mathematical derivations/solutions.

1. Read and work through the supplementary notes as well as indicated sections of P&W, and explore course code examples. Beyond learning about AR models, this includes revision of basic Bayesian analysis in linear regression– Section 1.2 of Chapter 1 and reference Bayesian posterior distributions arising as laid out there; the Matlab code supports this, and the script for AR(1) model explorations shows use of this in fitting this little model to data (including the SOI time series). Makes sure you are completely on top of the notation– the basic Bayesian regression ideas were covered extensively in prerequisite linear models (and predictive modelling) courses, but notation is always a variable!

   Read ahead into course notes and slides for the coming week(s), and get intimate with relevant sections of the P&W text. Read, digest and anticipate.

2. Become proficient in Matlab. Rerun class examples using class code, explore code scripts and support functions, modify as you like, etc.

3. Consider the stationary model $x_t \leftarrow AR(1|(\phi, v))$ with $s = 1$. Simulate series of length $n = 100$ from the distribution of $x_{1:n}$ *conditional* on an pre-initial value x_0 for the two cases $x_0 = 0$ and $x_0 = 10$. Repeat this a few times with different values of ϕ (large, positive and negative in particular). Describe and interpret the resulting realizations, in comparison to realizations from the stationary joint distribution $p(x_{1:n})$.

   Sample paths starting at a value of x_0 that is not sampled, or distributed, as $N(0, s)$ do not represent samples from the full joint distribution. However, the short-memory of an AR(1) process means that the effects of such chosen initial values decay exponentially ($|\phi|^t$) so that, after some time steps, the realization of the process *converges*; eventually for t large enough,

$x_{t:(t+n)}$ does indeed represent a sample from $p(x_t, x_{t+1}, \ldots, x_{t+n})$. The effect of initial values chosen like this is longer for larger $|\phi|$.

Generating sample paths from $x_0 = 0$ leads to trajectories that are really almost immediately indistinguishable from those generated from a random initial value $x_0 \sim N(0, s)$ since this distribution is centred at 0. Starting at $x_0 = 10$ when $s = 1$, however, starts the process at what is a very unlikely value; the process then quickly moves towards 0, either decreasing (for cases with $\phi > 0$) or oscillating (when $\phi < 0$) and– once values move into the plausible range of the $N(0, s)$ stationary marginal distribution.

4. Figure 1 plots the monthly changes in the US S&P stock market index over 1965 to 2016. The course Schedule page has a little bit of Matlab to read in this (and related) data and create this plot (see link under this HW#1 on the web page). Consider an AR(1) model as a very

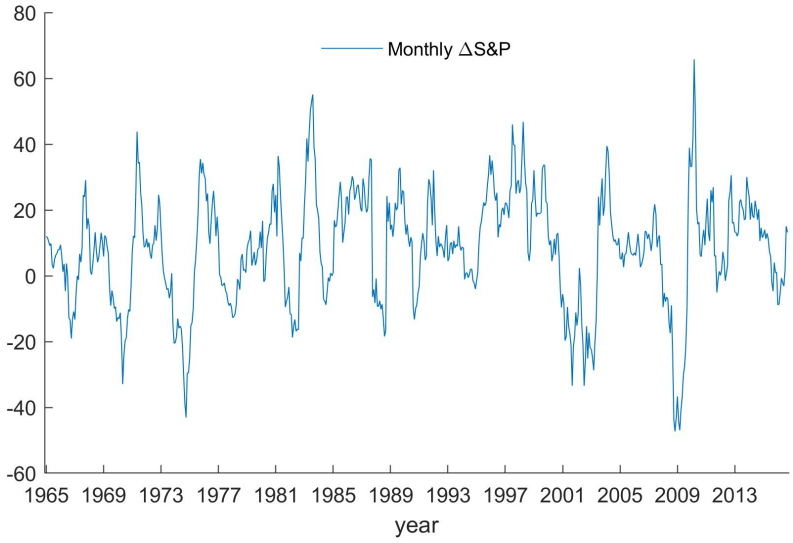


Figure 1: Monthly changes in the US S&P stock market index over 1965–2015

simple exploratory model– for understanding local dependencies but not for forecasting more than a month or two ahead. Then, we know there is a great deal of variation across the years in the market economy and that we might expect "change" that an AR(1) model does not capture. To get started on a first, very basic investigation of this, we can simply fit the AR(1) model to shorter sections of the data and examine the resulting inferences on parameters to see if they seem to vary across time. Do this as follows. The full series has $T = 621$ months of data; look at many separate time series by selecting a month m and taking some number k months either side; for example, you might take $k = 84$ and for any month m analyse the data over the "windowed period" from $m - k$ to $m + k$ inclusive. Repeat this for each month m running from $m = k + 1$ to $m = T - k$. These repeat analyses will define a "trajectory" of AR(1) analyses over time, one for each sub-series.

For each sub-series, subtract the sub-series mean and then compute the summaries of the reference posterior for an AR(1) model to just that $2k + 1$ time points– just treating each selected sub-series separately. Using the theoretical posterior T distribution for the ϕ parameter, compute and compare (graphically) the exact posterior 90% credible intervals

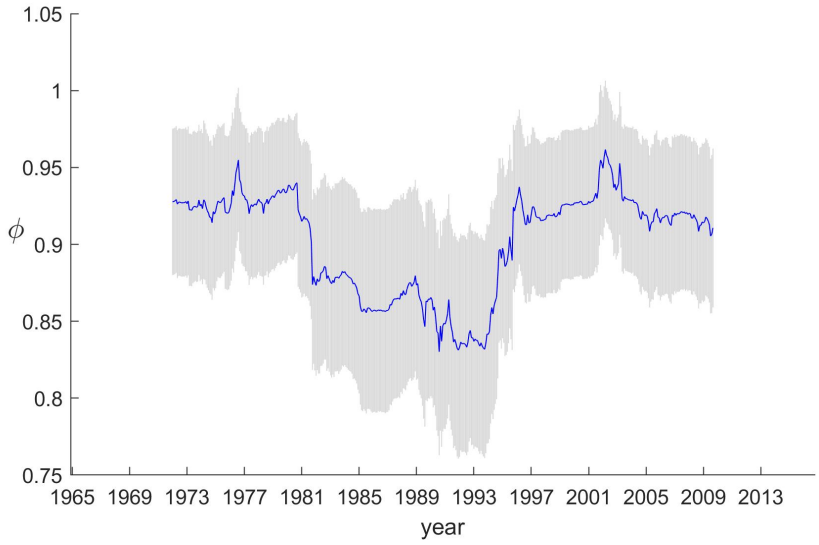


Figure 2: Trajectory of "local" estimates and 90% credible intervals for AR(1) parameter with moving-window subseries of ΔS&P data

(a) Comment on what you see in the plot and comparison, and what you might conclude in terms of changes over time.

See plotted trajectory of the posterior means of ϕ anchored at each window of data, along with corresponding 90% credible intervals, in Figure 2. The plot suggests high dependence with inferred values of ϕ in the 0.8-0.95 range. It also suggests somewhat lower values during the 1980s and early 1990s.

(b) Do you believe that short-term changes in S&P have shown real changes in month-month dependencies since 1965?

Although intervals overlap to a degree the global variation in seems clear, so implying that there is may well be real variation in the parameter (assuming the adequacy of the AR(1) model form of course). Even if the AR model *form* is appropriate for this data, the specific parameter values that are appropriate might vary; the suggestion– at this exploratory level– is that persistence of the dependence structure seemed to be weaker in the 1980s-1990s than earlier or in more recent years. Why ... ?

How to more formally assess this? And what about reanalysis with different k?

(c) How would you suggest also addressing the question of whether or not the underlying mean of the series is stable over time?

Fit a model with a mean μ to each subseries, and explore a similar analysis for μ.

Bonus credit for doing something for μ.

(d) What about the innovations variance?

Explore a similar analysis for v.

Bonus credit for doing something for v.

(e) What does this suggest for more general models that might do a better job of imitating this data?

Comment: Maybe an appropriate, more general model is some kind of *time-varying parameter* model, in which $x_t = \phi_t x_{t-1} + \nu_t$ where ν_t varies "slowly" according to some kind of second-stage model. This ad-hoc local estimation idea gives some suggestions and insights as to what kinds of second-stage models might make sense.

5. Sequential Bayesian learning and $1-$step ahead forecasting in AR(1) models. This question starts an investigation of the details of how posterior distributions for (ϕ, v) are sequentially updated as new data arises, forming a prelude to filtering in state-space models generally as well as for extension to time-varying parameter AR models.

Suppose $x_t \leftarrow AR(1|(\phi, v))$ with (ϕ, v) uncertain. At any time t write $\mathcal{D}_t$ for the past data and information, including all past observations. If no additional information arises over the time interval $(t-1, t]$, then $\mathcal{D}_t$ sequentially updates as the new observation is made via– simply– $\mathcal{D}_t = \{\mathcal{D}_{t-1}, x_t\}$.

Now suppose you are standing at the end of time interval $t-1$ so that you have current information set $\mathcal{D}_{t-1}$. Suppose the current posterior for $\theta = (\phi, v)$ based on this information has a conjugate normal-inverse gamma form written as

$$(1) \qquad (\phi|v, \mathcal{D}_{t-1}) \sim N(m_{t-1}, C_{t-1}(v/s_{t-1})),$$

$$(2) \qquad (v^{-1}|\mathcal{D}_{t-1}) \sim Ga(n_{t-1}/2, n_{t-1}s_{t-1}/2)$$

with known defining parameters. This would be the case, for example, of a reference posterior based on the first $t-1$ observations. Here m_{t-1} and $s_{t-1} > 0$ are natural point estimates of ϕ and v respectively, while $C_{t-1} > 0$ and $n_{t-1} > 0$ relate to uncertainty.

(a) What is the current marginal posterior for ϕ, namely $p(\phi|\mathcal{D}_{t-1})$?

The implied margin from the normal-inverse gamma, simply $T_{n_{t-1}}(\phi|m_{t-1}, C_{t-1})$.

(b) Show that, conditional on v and marginalizing over ϕ, the implied $1-$*step ahead forecast distribution* for x_t given v is

$$(x_t|v, \mathcal{D}_{t-1}) \sim N(f_t, q_t v/s_{t-1})$$

with $f_t = m_{t-1}x_{t-1}$ and $q_t = s_{t-1} + C_{t-1}x_{t-1}^2$.

x_t is a linear combination of the conditionally normal and independent quantities ϕ and ϵ_t, and so is normal.

The mean is $E(x_t|v, \mathcal{D}_{t-1}) = E(\phi x_t|v, \mathcal{D}_{t-1}) + E(\epsilon_t|v, \mathcal{D}_{t-1}) = m_{t-1}x_{t-1} + 0$. The variance is

$$V(x_t|v, \mathcal{D}_{t-1}) = V(\phi x_t|v, \mathcal{D}_{t-1}) + V(\epsilon_t|v, \mathcal{D}_{t-1}) = v\{C_{t-1}x_{t-1}^2/s_{t-1} + 1\} = vq_t/s_{t-1}$$

with q_t as stated.

(c) Now marginalize also over v to find the implied $1-$*step ahead forecast distribution* for x_t, namely $p(x_t|\mathcal{D}_{t-1})$, i.e., the distribution you will use in practice to predict x_t $1-$step ahead. What is this distribution?

Observe that $(x_t, v|\mathcal{D}_{t-1})$ have a normal-inverse gamma distribution whose margin for x_t is the required forecast distribution; immediately, this is $T_{n_{t-1}}(x_t|f_t, q_t)$.

(d) Now move to time t and observe the outcome x_t. Show that the *time t posterior* $p(\phi, v|\mathcal{D}_t)$ is also normal-inverse gamma, having the same form as in at time $t-1$ above but now with $t-1$ updated to t and updated defining parameters $\{m_t, C_t, n_t, s_t\}$ that can be written in the following forms:

- $m_t = m_{t-1} + A_t e_t$,
- $C_t = r_t(C_{t-1} - A_t^2 q_t)$,
- $n_t = n_{t-1} + 1$,
- $s_t = r_t s_{t-1}$ with $r_t = (n_{t-1} + e_t^2/q_t)/n_t$,

where

- $e_t = x_t - f_t$ is the realized $1-$*step ahead (point) forecast error*, and
- $A_t = C_{t-1}x_{t-1}/q_t$ is called the *adaptive coefficient*.

This is standard prior-to-posterior updating in normal models, with some algebra to rearrange the expressions for the four sufficient statistics (m_t, C_t, n_t, s_t) into the forms above.

For v and using $p(x_t|v, \mathcal{D}_{t-1}) = N(x_t|f_t, vq_t)$ from above, the marginal posterior is $p(v|\mathcal{D}_t) \propto IG(v|n_{t-1}/2, n_{t-1}s_{t-1}/2)N(x_t|f_t, vq_t)$ which is inverse gamma with updated parameters $n_t = n_{t-1} + 1$ and $n_t s_t = n_{t-1}s_{t-1} + e_t^2/q_t$ so that $s_t = (n_{t-1}s_{t-1} + e_t^2/q_t)/n_t$ as required.

Second, conditional on v, we have

$$p(\phi|v, \mathcal{D}_t) \propto N(\phi|m_{t-1}, vC_{t-1}/s_{t-1})N(x_t|x_{t-1}\phi, v)$$

which is normal with

$$E(\phi|v, \mathcal{D}_t) = m_{t-1} + vC_{t-1}x_{t-1}(x_t - m_{t-1}x_{t-1})/\{s_{t-1}(v + vC_{t-1}x_{t-1}^2/s_{t-1})\} = m_{t-1} + A_t e_t,$$

and variance

$$vC_{t-1}/s_{t-1} - v^2 C_{t-1}^2 x_{t-1}^2/\{s_{t-1}^2(v + vC_{t-1}x_{t-1}^2/s_{t-1})\} = v(C_{t-1} - A_t^2 q_t)/s_{t-1} = vC_t/s_t$$

where $C_t = r_t(C_{t-1} - A_t^2 q_t)$ and $r_t = s_t/s_{t-1}$ as defined. These give $N(\phi|m_t, vC_t)$ with (m_t, C_t) as in the above list.

(e) Comment on these expressions, giving particular attention to the following:

i. How (m_t, C_t) depend on the new data x_t relative to the prior values (m_{t-1}, C_{t-1}).

The update for m_t is a "predictor/corrector" form: The prior or "predicted" value for ϕ, namely m_{t-1}, is corrected by the weighted forecast error. A large forecast error implies a large correction, and vice-versa.

ii. The role of the adaptive coefficient in the update $(m_{t-1}, C_{t-1}) \longrightarrow (m_t, C_t)$.

If A_t is large in absolute value, the correction of m_{t-1} to m_t will be larger for any forecast error. This will happen when either (i) C_{t-1}/q_t is large– this occurs naturally enough when $p(\phi|\mathcal{D}_{t-1})$ is diffuse; and/or (ii) when $|x_{t-1}|$ is large– this is the usual regression/design effect, that seeing an "extreme" value of a predictor/covariate implies increased information/precision.

A larger $|A_t|$ also decreases the posterior variance more relative to the prior variance.

The sign of A_t is that of x_{t-1}. Hence the correction of m_{t-1} to m_t with a positive forecast error moves the estimate up if x_t is positive, otherwise corrects it down.

iii. The updates for the degrees of freedom n_t and point estimate s_t and how they depend on x_t.

The degrees of freedom n_t increases by one per observation, naturally– whatever the data x_t.

The posterior error variance estimate s_t is updated as a scalar multiple of the prior estimate s_{t-1}; the scaling factor r_t depends on the prior degrees of freedom n_{t-1} and the squared forecast error scaled by q_t. A larger forecast error leads to an inflation of the estimate of error variance, i.e., $r_t > 1$ when $e_t^2 > q_t$, while smaller errors shrink the estimate, i.e., $r_t < 1$ when $e_t^2 < q_t$.

(f) Consider an example in which the forecast error is very large relative to expectation, resulting in a value of e_t^2/q_t much greater than 1. Comment on how the posterior for (ϕ, v) responds.

The posterior mean m_t for ϕ will be substantially different to the prior mean m_{t-1} in such a case, while the posterior uncertainty reflected in C_t much larger than that of the prior. Correspondingly, the posterior will favour larger values of v.

STA 642: Fall 2018 – Homework #2 Exercises

1. Work through the AR(p) notes and especially the AR(2) examples as introduced in class. Among other things, this provides a foundation for a lot of basic/exploratory analysis of time series with AR models, as well as core theory and more building-blocks for general linear state space models. Read ahead into course notes and slides for the coming week(s), and get intimate with relevant sections of the P&W text.

   In particular, read and digest the material on AR(p) model order assessment using marginal likelihoods and information criteria (AIC, BIC) in P&W section 2.3.4. This relates to the question of selection of, and more broadly "inference on", the AR model order p, with more formal inferential/model-based ideas that complement the exploratory uses of ACF, PACF and other exploratory ideas in regression that you might use.

   The course code repository includes the Matlab function `arpcompare.m`. Open and review that function– it simply codes up the formal Bayesian marginal likelihood computation for model order p based on the conditional reference analysis, and also generates the AIC & BIC measures. It is used in examples in the class examples code for AR(p) models.

2. This excercise adds some theoretical structure to the stationary AR(1)+noise model (a.k.a. hidden Markov model). As part of this, the example here gets into ARMA models, and you may find the use of the backshift operator strategy useful (– though by no means necessary–) to explore parts of this question.

   In the stationary AR(1)+noise model– a first state space/hidden Markov model– we observe

   $$y_t = x_t + \nu_t \quad \text{where} \quad x_t \leftarrow AR(1|(\phi, v))$$

   with $\nu_t \sim N(0, w)$ and assuming $\nu_t \perp\!\!\!\perp \nu_r$ and $\nu_t \perp\!\!\!\perp \epsilon_r$ for all t, r. Clearly $y_t \sim N(0, q)$ with $q = s + w$ where $s = v/(1 - \phi^2)$.

   This exercise shows that y_t is *not* an AR(1) process; it is an ARMA(1,1) process as in the resulting AR-like form of the implied model for y_t, the driving innovations are correlated at $\text{lag}-1$.

   (a) Show that $y_t = \phi y_{t-1} + \eta_t$ where $\eta_t = \epsilon_t + \nu_t - \phi\nu_{t-1}$.

   This is most easily seen by applying the backshift operator $\phi(B) = 1 - \phi B$ to both sides of the identity $y_t = x_t + \nu_t$ to give $\phi(B)y_t = \phi(B)x_t + \phi(B)\nu_t$ or just $y_t - \phi y_{t-1} = x_t - \phi x_{t-1} + \nu_t - \phi\nu_{t-1}$ and the result follows since $x_t - \phi x_{t-1} = \epsilon_t$.

   The use of the backshift operator here is, of course, just a shorthand for working the algebra directly. That is– in longhand, by substitution– we have

   $$y_t = x_t + \nu_t = \phi x_{t-1} + \epsilon_t + \nu_t = \phi(y_{t-1} - \nu_{t-1}) + \epsilon_t + \nu_t$$

   and the result follows.

(b) Show that the lag-1 correlation in the η_t sequence is $-\phi w/(w(1+\phi^2)+v)$.

First, $V(\eta_t) = V(\epsilon_t + \nu_t - \phi\nu_{t-1}) = V(\epsilon_t) + V(\nu_t) + \phi^2 V(\nu_{t-1})$ by independence. This reduces to $V(\eta_t) = v + (1+\phi^2)w$. Second,

$$\begin{aligned} Cov(\eta_t, \eta_{t-1}) &= Cov(\epsilon_t + \nu_t - \phi\nu_{t-1}, \epsilon_{t-1} + \nu_{t-1} - \phi\nu_{t-2}) \\ &= Cov(-\phi\nu_{t-1}, \nu_{t-1}), \quad \text{by independence,} \\ &= -\phi V(\nu_{t-1}) \;=\; -\phi w. \end{aligned}$$

The stated correlation follows.

Additional comment: $Cor(\eta_t, \eta_{t-1})$ has the opposite sign to ϕ and is smaller in absolute value.

(c) Find an expression for the lag$-k$ autocorrelation of the y_t process in terms of k, ϕ and the signal:noise ratio s/q. Comment on this result. (We already worked through this in class; do it again!)

The covariance at lag$-k$ is

$$\begin{aligned} E(y_t y_{t-k}) &= E((x_t + \nu_t)(x_{t-k} + \nu_{t-k})) \\ &= E(x_t x_{t-k}) + E(x_t \nu_{t-k}) + E(x_{t-k}\nu_t) + E(\nu_t \nu_{t-k}) \\ &= \phi^k s + 0 + 0 + 0 \end{aligned}$$

so the correlation is $\phi^k s/q$. That is, relative to the lag$-k$ autocorrelation ϕ^k of the underlying signal process, that of the observed data process is directly reduced by the signal:noise ratio $s/q \equiv s/(s+w)$.

(d) Is y_t an AR(1) process? Is it Markov? Discuss and provide theoretical rationalisation.

Credit for any sensible comments or theory, including exploration of graphical model structure as detailed below. One obvious point is that the acf of y_t is *not* that of an AR(1) process (unless $w = 0$)– this shows that y_t is not AR(1), as already discussed in class.

To explore further, see the graphical modelling theory and development below in an appended Question 2. To add to that look in more detail at the structure of the innovations series η_t impacting the y_t process. From above, $\eta_t = \epsilon_t + \nu_t - \phi\nu_{t-1}$ and we already found that $Cov(\eta_t, \eta_{t-1})$ is non-zero. Now, for $k > 2$, the innovation term $\eta_{t-k} = \epsilon_{t-k} + \nu_{t-k} - \phi\nu_{t-k-1}$ has no terms in common with η_t, hence $Cov(\eta_t, \eta_{t-k}) = 0$ for $k > 1$. This means η_t must be a moving-average process of order 1, i.e., MA(1)– and any such process can be written as $\eta_t = \delta_t + \theta\delta_{t-1}$ for some i.i.d., zero mean normal series δ_t, where $|\theta| < 1$ and the lag-1 correlation satisfies $Cor(\eta_t, \eta_{t-1}) = \theta/(1+\theta^2)$. Here θ is known at the moving average coefficient (see Example 2.5, p 65 in P&W). Now, whatever the value of θ may be, this implies that $y_t = \phi y_{t-1} + \delta_t + \theta\delta_{t-1}$, which is an ARMA(1,1) process (Section 2.5 in P&W). This is not a Markov process. It is easiest to see this using the backshift operater B notation as follows. We have $(1-\phi B)y_t = (1-\theta B)\delta_t$ so that

$(1-\phi B)y_t = (1-\theta B)\delta_t$ or $(1-\theta B)^{-1}(1-\phi B)y_t = \delta_t$. Expanding the inverse operator term here gives

$$(1+\theta B+\theta^2 B^2+\theta^3 B^3+\cdots)(1-\phi B)y_t = \delta_t$$

or

$$\{1+(\theta-\phi)B+\theta(\theta-\phi)B^2+\theta^2(\theta-\phi)B^3+\cdots\}y_t = \delta_t$$

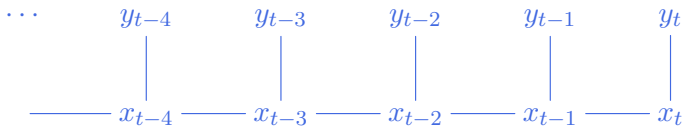
so that

$$y_t = \alpha_1 y_{t-1}+\alpha_2 y_{t-2}+\alpha_3 y_{t-3}\cdots+\delta_t$$

where $\alpha_j = \theta^{j-1}(\theta-\phi)$ for $j = 1, 2, \ldots$. This shows that y_t depends on the full past history of the series, i.e., is an infinite order AR representation.

Here is an extension of discussion of Question 2 to explore graphical model structure. The following is not expected or necessary as part of the solution to the exercise, but is really key as part of the bigger picture. And, by the way, could be presented as a solution to part of the question. The process y_t is not Markovian. While it may seem initially likely– since we are just adding noise to an AR(1) process– the lag-1 dependencies induced in the η_t innovations of the y_t process should raise a doubt. Intuitively, note the following. First, we have shown that η_t is correlated with η_{t-1}; second, η_{t-1} is directly dependent on ν_{t-2} which itself directly influences y_{t-2}; so this suggests y_t is related to y_{t-2} via a path $y_t \leftarrow \eta_t \leftarrow \eta_{t-1} \leftarrow y_{t-2}$ as well as directly via the AR(1)-induced path from $y_{t-1} \leftarrow y_{t-1} \leftarrow y_{t-2}$ (unless $w = 0$, of course). Let us exhibit and explore graphical models for this context.

(a) We know that the (undirected) graph of the full joint distribution $p(y_{1:t}, x_{0:t})$ has the form over $t-4:t$, of, simply,

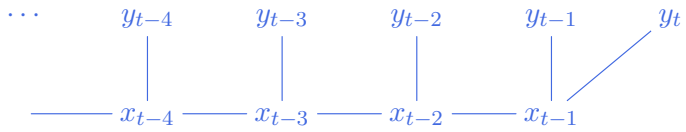


As we know, we "read-off" conditional dependencies based on edges in the graph. For example, we know/see that the conditional distribution for x_t given *all the other quantities up to and including time* t, depends only on its neighbours in the graph, namely x_{t-1}, y_t. Explicitly, x_t is conditionally independent of $x_{0:t-2}, y_{1:t-1}$ given x_{t-1}, y_t.

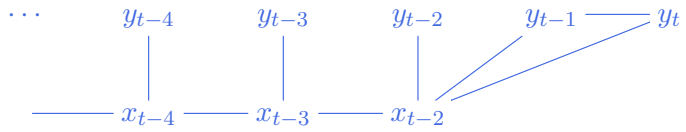
(b) We also know that *marginalization over a variable* just (i) deletes that variable from the graph, and (ii) adds edges between all variables/nodes having an edge to the variable removed. This second point is key: *when I leave, all my neighbours become neighbours (if they are not already)*: this is the graphical recognition that two nodes that may be conditionally independent in the full joint distribution become dependent under marginalization of a variable that "separates" them.

(c) Let us look at marginalization over x_t, to reduce to the marginal distribution $p(y_{1:t}, x_{0:t-1})$. The "join up all my neighbours when I leave" rule results in a new edge between x_{t-1}

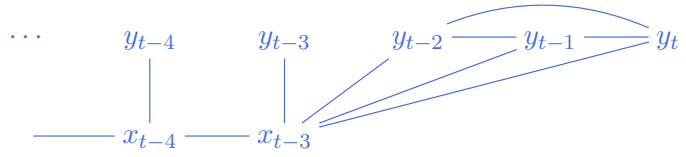
and y_t in the marginalized graph:



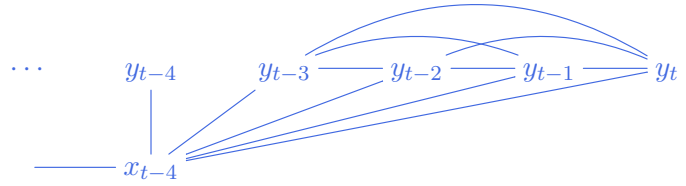
(d) Now marginalize over x_{t-1} to generate the resulting graph of the marginalized distribution $p(y_{1:t}, x_{0:t-2})$. Since x_{t-1} has neighbours x_{t-2}, y_{t-1}, y_t, these nodes all become neighbours in the marginalized graph:



(e) Next marginalize over x_{t-2} to generate the resulting graph of the marginalized distribution $p(y_{1:t}, x_{0:t-3})$. Since x_{t-2} has neighbours $x_{t-3}, y_{t-2}, y_{t-1}, y_t$, they all become neighbours in the marginalized graph:



(f) Take the next step: marginalize over x_{t-3} to define the resulting graph of the marginalized distribution $p(y_{1:t}, x_{0:t-4})$. Joining up all the neighbours of x_{t-3} gives:



(g) Imagine continuing the marginalization process to remove x_{t-4}, then sequentially back in time to remove all x_t down to and including x_0. This will yield the graph of $p(y_{1:t})$, the marginal distribution of the sequence of $y_{1:t}$ marginalized over the hidden signal $x_{0:t}$ altogether. In this distribution, we can then see which past does y_t depend on. That is, in $p(y_t|y_{1:t-1})$ which of the past y_s, $s = 1 : t-1$, matter. Do this by continuing the process above, to see that this induces edges between all pairs of y nodes, so that the resulting graph is a complete graph. Hence $p(y_t|y_{1:t-1})$ depends on *all past* $y_s, s = 1 : t-1$. So we have shown that adding independent noise to a first-order Markov process destroys the Markovian structure. Intuitively, we can see this by remembering that each y_s provides information on the full sequence of x_t values (recall smoothing when inferring the hidden signals). so marginalizing over the signal process links up all of the y_t. We have now shown this directly via the graphical models.

(h) For a series of length n, the variance matrix $V(x_{1:n}) = \mathbf{\Sigma}_n$ has inverse– the precision matrix of the AR(1) process $x_{1:n}$– given by

$$\mathbf{\Sigma}_n^{-1} = v^{-1} \begin{pmatrix} 1 & -\phi & 0 & \cdots & 0 & 0 \\ -\phi & 1+\phi^2 & -\phi & \cdots & 0 & 0 \\ 0 & -\phi & 1+\phi^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & -\phi & 0 \\ 0 & 0 & 0 & -\phi & 1+\phi^2 & -\phi \\ 0 & 0 & 0 & 0 & -\phi & 1 \end{pmatrix}.$$

We also know that the variance matrix of the $y_{1:n}$ process is $\mathbf{\Sigma}_n + w\mathbf{I}$ where $\mathbf{I}$ is the $n \times n$ identity. The above graphical explorations shows us that there are no zeros in the off-diagonal elements of the corresponding precision matrix $(\mathbf{\Sigma}_n + w\mathbf{I})^{-1}$ This is clear since $p(y_t|y_{1:n\setminus t})$ depends on *all* quantities in $y_{1:n\setminus t}$. A zero only occurs in row t, column s when $y_t \perp\!\!\!\perp y_s \,|\, y_{1:n\setminus(t,s)}$, and in this case there are no such occurrences. Incidentally, this shows how simply adding a constant to the diagonal of a matrix with a sparse inverse destroys that sparsity.

Bonus points for developing some of the graphical model aspects.

3. This question relates to an alternative state space representation of an AR(p) model; this is a special case of the state space representation of ARMA models (P&W, top of page 75; note that the AR(p) is the special case when $q = 0, m = p$ in the notation there).

Work this exercise explicitly in the case of $p = 2$; the linear algebra in this special case is easy and the special case illuminating of the more general AR(p) case. Provide solutions to the case of $p = 2$ for assessment. The general case is just a bit more linear algebra and will be given bonus credit, but is otherwise optional.

In the standard state space representation of the AR(p) model we have state vector $\mathbf{x}_t = (y_t, y_{t-1}, \ldots, y_{t-p+1})'$ and model equations $y_t = \mathbf{F}'\mathbf{x}_t$ and $\mathbf{x}_t = \mathbf{G}\mathbf{x}_{t-1} + \mathbf{F}\epsilon_t$ where

$$\mathbf{F} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad \mathbf{G} = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_{p-1} & \phi_p \\ 1 & 0 & \cdots & 0 & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

with AR parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)'$ and innovations $\epsilon_t \sim N(0, v)$.

Define the $p \times p$ symmetric matrix $\mathbf{A}$ by

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \phi_2 & \phi_3 & \phi_4 & \cdots & \phi_{p-1} & \phi_p \\ 0 & \phi_3 & \phi_4 & \phi_5 & \cdots & \phi_p & 0 \\ \vdots & \vdots & \vdots & & \iddots & \iddots & \vdots \\ \vdots & \vdots & \vdots & \iddots & \iddots & & \vdots \\ 0 & \phi_{p-1} & \phi_p & 0 & \cdots & \cdots & 0 \\ 0 & \phi_p & 0 & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

(a) Verify that the matrix product $\mathbf{AG}$ is given by

$$\mathbf{AG} = \begin{pmatrix} \phi_1 & \phi_2 & \phi_3 & \phi_4 & \cdots & \phi_{p-1} & \phi_p \\ \phi_2 & \phi_3 & \phi_4 & \phi_5 & \cdots & \phi_p & 0 \\ \phi_3 & \phi_4 & \phi_5 & \phi_5 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \iddots & \iddots & \vdots \\ \vdots & \vdots & \vdots & \iddots & \iddots & & \vdots \\ \phi_{p-1} & \phi_p & 0 & 0 & \cdots & \cdots & 0 \\ \phi_p & 0 & 0 & 0 & \cdots & \cdots & 0 \end{pmatrix}$$

noting that this is also symmetric.

Special case of $p = 2$: In this case

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & \phi_2 \end{pmatrix}$$

so that

$$\mathbf{AG} = \begin{pmatrix} 1 & 0 \\ 0 & \phi_2 \end{pmatrix} \begin{pmatrix} \phi_1 & \phi_2 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \phi_1 & \phi_2 \\ \phi_2 & 0 \end{pmatrix}$$

which has the stated– symmetric– form.

General case: This is easily verified simply by working through the matrix product entry-by-entry. And $\mathbf{AG}$ is evidently symmetric.

(b) Show or deduce that:

i. For a proper AR(p) model in which $\phi_p \neq 0$, then $|\mathbf{A}| \neq 0$ so that $\mathbf{A}$ is non-singular.

Special case of $p = 2$: In this case

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 0 & \phi_2 \end{pmatrix}.$$

Since $\phi_2 \neq 0$ then $|A| = \phi_2 \neq 0$ so $\mathbf{A}^{-1}$ exists.

General case: From the form of $\mathbf{A}$, is is clear that $|\mathbf{A}| = \phi_p^{p-1}$ which is non-zero hence $\mathbf{A}$ is non-singular.

ii. $\mathbf{AGA}^{-1} = \mathbf{G}'$ (*you can do this* without *trying to invert* $\mathbf{A}$).

Special case of $p = 2$*:* Here, directly,

$$\mathbf{A}^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \phi_2^{-1} \end{pmatrix}.$$

Then direct matrix multiplication gives the stated result. *Or*, follow the logic– below– for the case of general p that uses symmetry argument and does not need you to invert any matrices.

General case: Since $\mathbf{AG}$ is symmetric, then $\mathbf{AG} = (\mathbf{AG})' = \mathbf{G}'\mathbf{A}' = \mathbf{G}'\mathbf{A}$ as $\mathbf{A}$ is also symmetric. Thus, since $\mathbf{A}$ is non-singular, we deduce $\mathbf{AGA}^{-1} = \mathbf{G}'$.

iii. $\mathbf{AF} = \mathbf{F}$ and, as a result, $\mathbf{F}' = \mathbf{F}'\mathbf{A}^{-1}$.

This is immediate from the forms of $\mathbf{F}$ and $\mathbf{A}$.

(c) Hence show that an equivalent state space AR(p) form is given by $y_t = \mathbf{F}'\mathbf{z}_t$ and $\mathbf{z}_t = \mathbf{G}'\mathbf{z}_{t-1} + \mathbf{F}\epsilon_t$ based on a new $p \times 1$ state vector $\mathbf{z}_t = \mathbf{A}\mathbf{x}_t$ and where the state evolution matrix is $\mathbf{G}'$, i.e.,

$$\mathbf{G}' = \begin{pmatrix} \phi_1 & 1 & 0 & \cdots & 0 \\ \phi_2 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ \phi_{p-1} & 0 & 0 & \cdots & 1 \\ \phi_p & 0 & 0 & \cdots & 0 \end{pmatrix}$$

From the initial state space form, we see that

$$\mathbf{z}_t = \mathbf{A}\mathbf{x}_t = \mathbf{AG}\mathbf{x}_{t-1} + \mathbf{AF}\epsilon_t = \mathbf{AGA}^{-1}\mathbf{A}\mathbf{x}_{t-1} + \mathbf{AF}\epsilon_t = \mathbf{G}'\mathbf{z}_{t-1} + \mathbf{F}\epsilon_t$$

using the results above. Further, $y_t = \mathbf{F}'\mathbf{x}_t = \mathbf{F}'\mathbf{A}^{-1}\mathbf{z}_t = \mathbf{F}'\mathbf{z}_t$. The result follows.

(d) What is the interpretation of the elements of the transformed state vector $\mathbf{z}_t$?

Special case of $p = 2$*:* Here $\mathbf{z}_t' = (y_t, \phi_2 y_{t-1})$ so the first element is just as in the original model– the current value of the series– while the second element is the *contribution to the linear regression* for y_t from the lag-2 value.

General case: The lead element of the transformed state vector $\mathbf{z}_t$ is just y_t as in the original model. The later values are linear combinations of lagged y_{t-j} elements that contribute to the linear (auto-)regression for y_t. Not so interpretable, at all, as the original state space formulation, right? But the latter representation is nevertheless common in some areas of application.

One unstated point that we have reviewed in class: *Any* $p \times p$ matrix $\mathbf{A}$ such that $\mathbf{AGA}^{-1}$ has the same eigenvalues as $\mathbf{G}$ could be used to develop this analysis– transforming to a new state vector $\mathbf{z}_t = \mathbf{A}\mathbf{x}_t$ without changing the form of the model. In this example, $\mathbf{G}'$ is the result– and of course has the same eigenvalues as $\mathbf{G}$. Generally, it is wise to choose the "simplest" form and one that has easiest interpretation from such a class of "similar" models.

Bonus points for developing some of the theory in the general case.

STA 642: Fall 2018 – Homework #3 Exercises

1. Continue to work through the AR(p) notes and connect with relevant sections of the P&W text. Then, read ahead on material in the introductory sections of P&W Chapter 4 on general DLMs, and the support slides on the course web page. We will move there next class after spending more time on AR(p) models and review of this homework with its broader connections.
2. **Course mini-project proposal and development**. Re-read the expectations and process on the course assessment pdf linked to first class on the web site. See also the advance information on the *minproject-interim.pdf* file linked with this homework. You must be progressing on this. Before Fall break, you will be required to submit a 1-2 page outline of project ideas. Per the start of semester information and comments on this, that later mid-term checkpoint will be assessed (as part of the mid-term take-home exam) based on your identification of a topic, initial outline/sketch of the problem area, comments on data available/sources, and comments on initial goals for modelling, time series analysis and/or forecasting.

   Make sure to discuss with TA and/or myself– contact via email to discuss your initial thoughts, questions, and bounce-off project ideas. Nothing to hand-in here this week .. but, be sure to get focused on project planning!
3. The exploratory AR(p) Matlab code has, at the end, some basic exploration of quarterly US macro-economic data. See data in Figure 1. That example just reads in data, selects quarterly inflation rates and converts to *quarterly changes (i.e., difference values of quarterly actual inflation levels)*, and fits one or two AR(p) models. Earlier in the Matlab code there are examples of simulating the reference posterior in any AR(p) model, and generating synthetic futures representing the posterior distribution via Monte Carlo samples. The former can be used to explore (by Monte Carlo– as in the example using the SOI time series worked through in detail in that example code, and from class) the posterior for moduli and wavelengths of any quasi-periodic components suggested by the model. You may/should find this code useful in exploring this data further.

   Fit the reference analysis of an AR(8) model to this data, and address the following.

   (a) If the AR(8) is accepted as a good model for this data, do you think the data-model match supports stationarity? Give full numerical support for this based on the reference posterior.

   As in the example code, we can trivially generate a direct Monte Carlo sample from the posterior for the characteristic roots (a.k.a. eigenvalues) of the AR(8) model, simply by direct transformation of samples of the AR vector ϕ. Transforming to the roots and saving the $p = 8$ moduli $r_{1:8}$ gives us a sample from the posterior for these moduli, to be summarized. The proportion of samples with at least one r_j value lying outside $(-1, 1)$ can be trivially calculated– the Monte Carlo estimate of the posterior probability of non-stationarity.

   Figure 2 shows a Monte Carlo sample of posterior draws for the 8 moduli $r_{1:8}$, in terms of boxplots; it seems clear they all lie below $r = 1$, and it is trivially checked that in

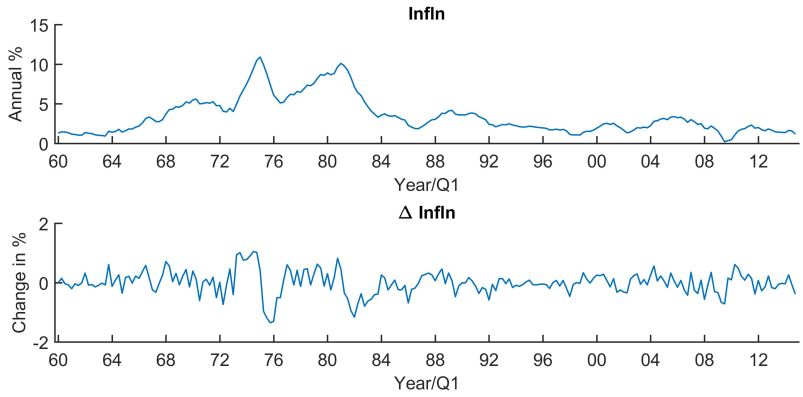


Figure 1: Quarterly US inflation and ΔInfln

fact they all do. Repeating this with larger samples maintains this, i.e., the posterior probability of non-stationarity is zero, up to Monte Carlo error.

Additional comments: The sampling of posterior for eigenvalues inherently suffers from a basic identification question. The simulation here orders Monte Carlo sampled values down by moduli, and saves the corresponding wavelengths (which are 0 or ∞ in cases of real eigenvalues). There are a number of details to then address in summarizing inferences. First, the boxplots of moduli reflect the ordering; it also shows that most of the samples give 4 pairs of complex conjugate eigenvalues, while a (very) few give 3 pairs plus 2 real eigenvalues of low moduli. Second, boxplots here just match wavelengths with moduli based on the initial ordering of the moduli. Then, to infer the ordered wavelengths reordering is needed. You can see that, in cases of two or more components with similar moduli, ordering by these will confound inference on wavelengths. This does not arise with this example, but does and may with others.

(b) Assuming that there is some indication of quasi-periodic behaviour under this posterior, summarise inferences on the *maximum wavelength (a.k.a. period)* of (quasi-)periodic components.

For each posterior sample of the roots computed above, we need to identify any complex conjugate roots and save the resulting values of the wavelength. It turns out that the posterior for this analysis very strongly supports at least one quasi-periodic component, so that repeat Monte Carlo draws almost surely include at least one sampled wavelength. Matlab code to do this simply adds this check on each sampled ϕ vector.

Figure 2 shows the histogram of the samples of the posterior for the maximum wavelength (and also the second largest). In this (and multiple repeat simulations) *all* Monte Carlo samples have at least one complex conjugate root, so that the histogram is of the

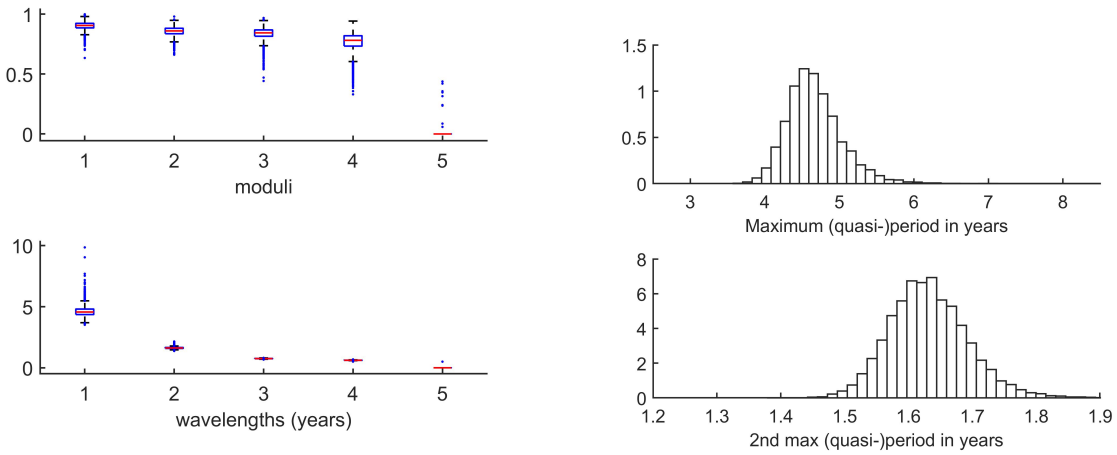


Figure 2: Boxplots of posteriors of moduli $r_{1:8}$ and histogram of the posterior for the largest two wavelengths of quasi-periodic components, in the reference AR(8) analysis of ΔInfln. Note that component are ordered down by moduli, and the wavelengths are then summarized by reordering.

full number simulated. This shows, by the way, the the posterior probability of at least one quasi-periodic component in the series (assuming we like the model) is basically one.

Then, some summaries of the posterior for the maximum wavelength (up to one decimal place) are as follows: mean and median are 4.6 years, interquartile range 4.3-4.8 years, 95% equal-tails interval 4.0-5.4 years.

(c) Explore and discuss aspects of inference on the implied decomposition of the series into underlying components implied by the eigenstructure of the AR model.

Posterior inferences for the sampled latent components are shown in Figure 3. The dominant component is a quasi-cyclical "business cycle" while the higher-order components refine that.

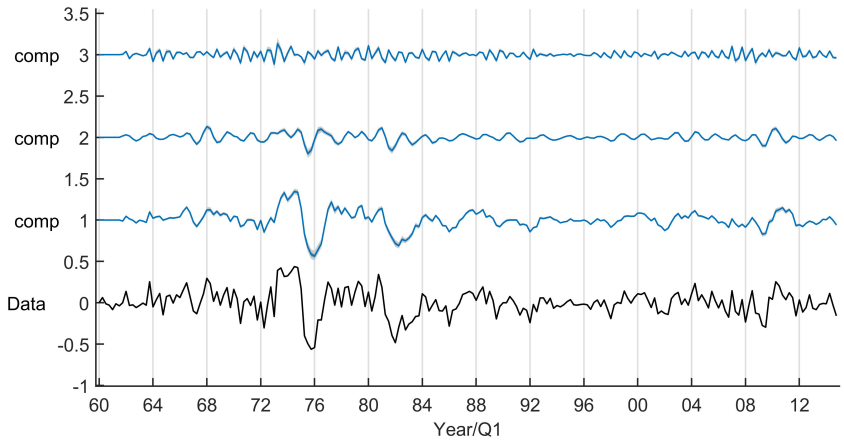


Figure 3: Posterior inferences on latent components in ΔInflnbased on AR(8) analysis.

(d) Produce and display graphical summaries– in terms of (Monte Carlo based) posterior medians, upper and lower quartiles, and 10% and 90% points of the predictive distributions of *actual inflation* over the 12 quarters following the end of the data series.

Each Monte Carlo sample of ϕ can be used to generate a path of future ΔInfln over the next 12 quarters, just recursively simulating the model into the future. Figure 4 shows the requested summaries, quarter-by-quarter, for ΔInfln. Now, to map to predictions of actual inflation we need to cumulatively sum over the coming quarters. Write y_t for actual % inflation in quarter t and x_t for ΔInfln, so that $x_t = y_t - y_{t-1}$ or $y_t = y_{t-1} + x_t$. Hence, from any "current" quarter $t = 0$, with a known value of current inflation y_0, we project over the next $h = 12$ quarters recursively; this implies the cumulative form $y_t = y_0 + \sum_{r=1:t} x_t$. From the above Monte Carlo prediction of ΔInfln with 20,000 samples, the posterior predictive samples of x_t over the next 12 time steps from the end of the data can be trivially mapped to actual inflation. Notice that the cumulation effect expands uncertainty as we go further ahead; see Figure 4.

As a bonus, note also the impact of the quasi-cyclicity in the forecasts, i.e., the Monte Carlo estimate of the forecast function.

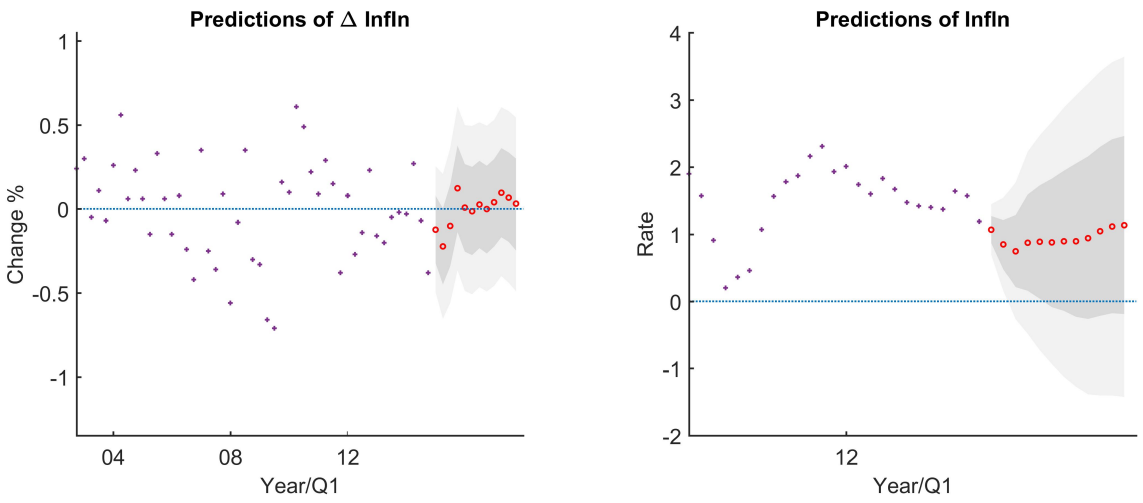


Figure 4: Synthetic futures of ΔInflnand actual Inflation.

(e) Assuming an AR(p) model is agreeable, do you think $p = 8$ makes sense for the differenced inflation series? Consider, for example, features of the fitted residuals from the model. In addition to exploratory and graphical analysis– and all you know about applied evaluation of linear regression model "fits"– the `arpcompare.m` function (noted and advised to review in Homework #2) may be of interest.

Some exploration of ACF and PACF summaries in Figure 5 indicate some quasi-periodic behaviour consistent with $p > 1$ for sure, while are strongly suggestive of $p > 5$, less strongly but evidently $p \approx 8$.

AIC, BIC and reference marginal likelihood for model order in Figure 5 also suggest $p = 8$ is interesting and relevant.

The qqplot and acf/pacf summaries of the fitted residuals from the AR(p) model, in Figure 6 indicate little evidence of non-normality which supports the normal innovation assumption. The residual correlations and autocorrelations suggest much of the dependence structure has been explained by the model, but there are perhaps some questions raised about the apparent "annual" pattern in PACF values, even though they are not obviously significant. The little bumps in both ACF and PACF at exactly 2 years might also raise a question– maybe a higher order model? or, maybe add one longer-term lag but not the intervening ones?

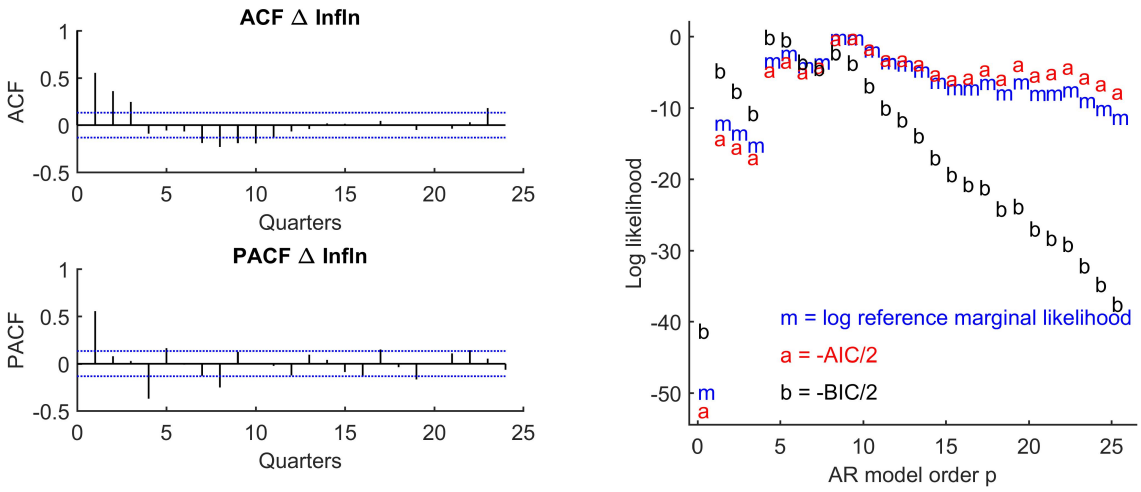


Figure 5: Sample correlations and model order assessments for Δlnfln.

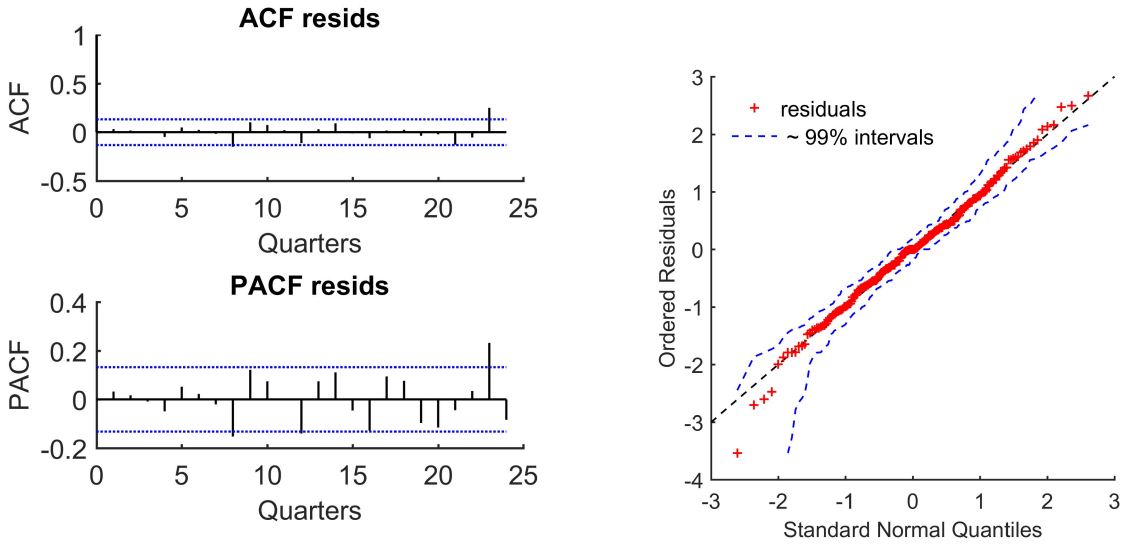


Figure 6: Sample correlations and normal qq plot for AR(8) residuals .

Bonus points for discussion of any additional analyses of other models with different values of p, comparisons, general discussion of model selection questions, etc. For example:

- Fit a higher-order model, such as with $p = 12$ or $p = 16$. Exploration of this can/may find some indication of longer wavelength quasi-periodic structure, but of course accompanied by a high level of uncertainty. Reflective and maybe rationalizing

some of why economists argue so much about what the business cycle is.

- Perhaps the data plots in Figure 1 suggest that there may be different levels of variation at different time periods, i.e., changes in volatility. How might this impact these results? How to explore?

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 $p(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_t, \mathcal{D}_n)$?

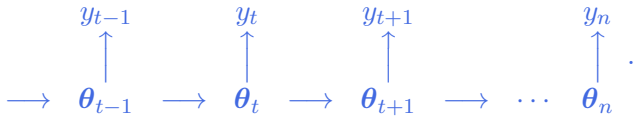
This *"look back one step"* conditional posterior is a central ingredient of retrospective analysis of time series using DLMs; see P&W Section 4.3.5.

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



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

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.

STA 642: Homework # 5 Exercises

1. **Course mini-project proposal and development**. Continue progressing on this. Talk/email to MW and TAs as convenient. The next homework will be based on (only) the summary proposal for your project.
2. **TVAR models.** Next week we will start looking at another class of DLMs– revisiting AR models and then their key practical and widely-used extensions to time-varying autogressive (TVAR) models. Read P&W Section 5.1 (skip over 5.1.3, or at least take it "very lightly" for now) and 5.2; and review the detailed course slides on TVAR models linked to the web page.

   There is code and examples on TVAR models in the course code base. Example code/scripts explore TVAR analyses of the quarterly changes in US inflation series and a subset of the EEG series from class discussion/slides.
3. **A start in stochastic volatility (SV) modelling.** Before we get into TVAR models, we will first look a bit more about the recurrent question of observational variances in DLMs that may/appear to change over time. We will focus in detail on the venerable and widely-used discount *stochastic volatility model* detailed in P&W, Section 4.3.7, and in (very) detailed course slides on this SV model linked to the web page. Read and explore that in preparation, and in helping with Question 4 in this homework.

   Some/most of you will likely want to integrate SV into projects later in semester, and the basic ideas and methodology in the univariate model underlie a main initial class of multivariate volatility models– i.e., time-varying variance matrices– in one of our first multivariate DLM settings coming along (P&W chapter 10). This basic model is also useful as a first model for time-varying rates in time series models for integer counts (e.g., for flows in networks ....); again, we will see examples later on.
4. The basic distribution theory in this question underlies the venerable and widely-used discount volatility model (P&W, Section 4.3.7). We will review the basic model, ideas and results in class; here, you will visit and work through some basic theory in advance to develop conceptual and technical understanding. We will also build on this later, for other kinds of models as noted above.

   The theory in this exercise concerns aspects of a bivariate distribution for two positive scalars ϕ_0 and ϕ_1. Mapping to the SV model in DLMs is made by noting the same setup arises there with, at any times $t-1, t$, the match $\phi_0 \leftarrow v_{t-1}^{-1}$ and $\phi_1 \leftarrow v_t^{-1}$, and then the bivariate distribution relates to $p(v_{t-1}, v_t|\mathcal{D}_t)$.

   Two positive scalar random quantities ϕ_0 and ϕ_1 have a joint distribution defined by:

   - the margin $p(\phi_0)$ given by $\phi_0 \sim Ga(a, b)$ for some scalars $a > 0, b > 0$; and
   - the conditional $p(\phi_1|\phi_0)$ that is implicitly defined by

   $$\phi_1 = \phi_0\eta/\beta, \quad \text{where} \quad \eta \sim Be(\beta a, (1-\beta)a) \quad \text{and} \quad \eta \perp\!\!\!\perp \phi_0,$$

   and where $\beta \in (0, 1)$ is a known, constant discount factor.

(a) What is $E(\phi_1|\phi_0)$?

Directly, $E(\phi_1|\phi_0) = \phi_0 E(\eta)/\beta = \phi_0$ since $E(\eta) = \beta a/[\beta a + (1-\beta)a] = \beta$.

(b) What are $E(\phi_0)$ and $E(\phi_1)$?

$E(\phi_0) = a/b$ and so $E(\phi_1) = E[E(\phi_1|\phi_0)] = E(\phi_0) = a/b$ as well.

(c) Starting with the joint density $p(\phi_0)p(\eta)$ (a product form since ϕ_0 and η are independent), make the bivariate transformation to (ϕ_0, ϕ_1) and show that

$$p(\phi_0, \phi_1) = c\, \mathrm{e}^{-b\phi_0}\, \phi_1^{\beta a - 1}\, (\phi_0 - \beta\phi_1)^{(1-\beta)a-1}\,, \quad \text{on } 0 < \phi_1 < \phi_0/\beta,$$

being zero otherwise. Here c is a normalizing constant that does not depend on the conditioning value of ϕ_0 (and we do not care about the value of c for the derivations here).

Since $\eta = \beta\phi_1/\phi_0$, the transformation from (ϕ_0, η) to (ϕ_0, ϕ_1) has reverse Jacobian β/ϕ_0. Further, we know $0 < \eta < 1$ so the transformed range is immediately $0 < \phi_1 < \phi_0/\beta$. Hence, substituting $\eta = \beta\phi_1/\phi_0$ in the $Be(\beta a, (1-\beta)a)$ p.d.f. and multiplying by the Jacobian– and ignoring some constants but being careful to account for all terms involving ϕ_0, ϕ_1– we have

$$p(\phi_1, \phi_0) \propto \frac{\beta}{\phi_0}\, \{\phi_0^{a-1}\mathrm{e}^{-b\phi_0}\} \left(\frac{\beta\phi_1}{\phi_0}\right)^{\beta a - 1} \left(1 - \frac{\beta\phi_1}{\phi_0}\right)^{(1-\beta)a-1}, \quad \text{on } 0 < \phi_1 < \phi_0/\beta,$$

which, after simplifying terms, yields

$$p(\phi_1, \phi_0) \propto \mathrm{e}^{-b\phi_0}\, \phi_1^{\beta a - 1}\, (\phi_0 - \beta\phi_1)^{(1-\beta)a-1}, \quad \text{on } 0 < \phi_1 < \phi_0/\beta,$$

as stated.

(d) Derive the p.d.f. $p(\phi_1)$ (up to a proportionality constant). Deduce that the marginal distribution of ϕ_1 is $\phi_1 \sim Ga(\beta a, \beta b)$.

Directly,

$$(1) \qquad p(\phi_1) \;=\; \int p(\phi_0, \phi_1) d\phi_0 \;\propto\; \phi_1^{\beta a - 1} \int_{\beta\phi_1}^{\infty} \mathrm{e}^{-b\phi_0} (\phi_0 - \beta\phi_1)^{(1-\beta)a-1} d\phi_0.$$

We can now use $\gamma = \phi_0 - \beta\phi_1$ to transform from ϕ_0 to γ in the integrand, giving

$$\begin{aligned} p(\phi_1) &\propto \phi_1^{\beta a - 1} \mathrm{e}^{-\beta b \phi_1} \int_0^{\infty} \mathrm{e}^{-b\gamma}\, \gamma^{(1-\beta)a-1} d\gamma \\ &\propto \phi_1^{\beta a - 1} \mathrm{e}^{-\beta b \phi_1} \end{aligned}$$

as the integral over γ is that of $\gamma \sim Ga((1-\beta)a, b)$ so depends only on a, b, β.

Hence the margin for ϕ_1 is $\phi_1 \sim Ga(\beta a, \beta b)$.

(e) Using the technical details of your derivations above (and without much more work), show that the reverse conditional $p(\phi_0|\phi_1)$ is implicitly defined by

$$\phi_0 = \beta\phi_1 + \gamma \quad \text{where} \quad \gamma \sim Ga((1-\beta)a, b) \quad \text{with } \gamma \perp\!\!\!\perp \phi_1.$$

This is implicit in the integrand of eqn. (1) above in the derivation of $p(\phi_1)$. Since the integrand is proportional to $p(\phi_1, \phi_0)$, then

$$p(\phi_0|\phi_1) \propto p(\phi_1, \phi_0) \propto (\phi_0 - \beta\phi_1)^{(1-\beta)a-1}\, \mathrm{e}^{-b\phi_0}, \quad \text{on } \phi_0 > \beta\phi_1.$$

Transform the random quantity ϕ_0 to $\gamma = \phi_0 - \beta\phi_1$; the Jacobian is 1 and we see that the resulting $p(\gamma|\phi_1)$ is that of $\gamma \sim Ga((1-\beta)a, b)$ independently of ϕ_1. The result follows.

5. P&W Chapter 4, Section 4.6: Problem 3.

Use the results derived in Question 4 above to answer this. (Do not redevelop technical results already shown.)

Matching notations $\phi_0 \leftarrow v_{t-1}^{-1}$, $\phi_1 \leftarrow v_t^{-1}$, $\eta \leftarrow \gamma_t$, and matching parameters $a \leftarrow n_{t-1}/2, b \leftarrow d_{t-1}/2 = n_{t-1}s_{t-1}/2$, means that the setup of this question is precisely that of the discount volatility model of eqn. (4.17) in P&W. Our result in part 4d above then proves the evolution theory of eqn. (4.18) in P&W.

6. P&W Chapter 4, Section 4.6: Problem 4.

Again, use the results derived in Question 4 above to answer this. (Do not redevelop technical results already shown.)

(a) With the matches of notation made in the previous question, and additionally matching $\gamma \leftarrow \nu_{t-1}^*$, the $1-$step back recursion in the discount volatility model follows from our result in part 4e above.

(b) Use the notation for precisions given by $\phi_t = v_t^{-1}$. The first-order Markov evolution of the v_t, hence the ϕ_t, means that, conditional on ϕ_t, ϕ_{t-1} is conditionally independent of all future ϕ_r as well as data at times t onward. The result follows immediately. (This is precisely the same use of conditional independence in first-order Markov models that we have already used in developing backward recursions for retrospective smoothing and sampling of states in DLMs.)

(c) Starting with $E(\phi_T|\mathcal{D}_T) = s_T^{-1}$, recurse back over time $t = T, T-1, \ldots,$ at each time evaluating expectations in the defining equation $\phi_{t-1} = \beta\phi_t + (1-\beta)\nu_{t-1}^*$. Immediately, this gives the backward smoothing recursion

$$E(\phi_{t-1}|\mathcal{D}_T) = E\{E(\phi_{t-1}|\phi_t, \mathcal{D}_{t-1})|\mathcal{D}_T\} = \beta E(\phi_t|\mathcal{D}_T) + (1-\beta)s_{t-1}^{-1}.$$

(d) Starting with a sample $v_T = \phi_T^{-1}$ from the inverse gamma posterior $p(v_T|\mathcal{D}_T)$, recurse back over time $t = T, T-1, \ldots,$ at each time $t-1$ sampling $v_{t-1} = \phi_{t-1}^{-1}$ via $v_{t-1}^{-1} = \beta v_t^{-1} + \nu_{t-1}^*$ where $\nu_{t-1}^* \sim Ga((1-\beta)n_{t-1}/2, n_{t-1}s_{t-1}/2)$ independently.