STA 642: Midterm 50% take-home

Do not derive standard and known results; in proving any results requested, be clear in statements about the results, but do not re-derive everything from scratch.

Print your name, sign and date below, and email this signed pdf honour form along with your exam solutions.

I confirm that this Midterm submission is my work alone. I have not copied nor adapted the work of others. I have not sought or provided help or advice on the exam; I have not discussed the exam with anyone at all during my work to complete and submit it.

Name: Ziyang Ding

Signature: Ziyang Ding

Date: 03/16/2020

1. Look at the detrended O2 isotope ratio time series that relates to the (topical) issues of longer-term cycles in climate and, perhaps, "ice-age" cycles in climatic indicators. As we have briefly discussed, this area has a long history related to the(so-called) Milankovitch cycles in the orbital dynamics of the Earth; e.g. https://en.wikipedia.org/wiki/Milankovitch_cycles. See also, for example, West (1996), *Bayesian time series: Models and computations for the analysis of time series in the physical sciences*, in *Maximum Entropy and Bayesian Methods 15*, Eds: K. Hanson & R. Silver, Kluwer, 1996, pp23–34.

   Reverse the time series so that left-to-right is now moving forward in time, with the end of the time series being "currently". Interest here will, in part, be on forecasting over the next many thousands of years. Some Matlab code to read in the data, reverse the time to be "forward", and plot etc., is linked here at the course Schedule page. Then, the course TVAR code for class examples can be used and slightly modified to address the following questions about analysis of this series.

   (a) Fit a TVAR(18) model to this forward-time oxygen isotope series. Use TVAR state discount factor $\delta = 0.99$ and volatility discount $\beta = 0.975$. Use the prior specifications:
   m0=zeros(p,1); m0(1)=1; n0=5; s0=0.02; C0=eye(p)/2

   Look at the "plug-in" estimated decomposition of the time series and the corresponding "plug-in" estimates over time of the wavelengths of the 4 dominant quasi-periodic components in the series. Discuss and summarise what these visuals suggest.

   (b) Use TVAR forecasting code tvarforecast.m to generate a reasonable number (several 000s) of synthetic futures over the next 240,000years. (That is, the next $k = 80$ time points, since our data are at 3kyear intervals).

   In general terms, do you think this is a "good" model for the series? From visual displays of predictions via synthetic futures compared to the real data, what kinds of (mismatch?) features stand-out? Include relevant supporting figures.

   (c) Peaks and troughs in the series relate to the earth orbital dynamics which link (through this noisy data) to quasi-periodicities in features of global climate. So predicting peaks and troughs is of interest. On the basis of your Monte Carlo predictions, make inferences on the following:

      i. The time (from now=T) to the next 240,000year *maximum* level in (detrended) O2, i.e., the time at which, over the coming 240,000years, the maximum level is predicted to be achieved– be careful to communicate full uncertainties with your forecast.
      ii. What do you predict for the O2 value that will be achieved at that first maximum?
      iii. As above, but now on the next 240,000year *minimum*.

2. Full FFBS in the TVAR(18) analysis of the detrended Oxygen isotope series of Exercise 1.

   Run FFBS (use the `tvarFFBS.m` function) to generate a reasonably large MC sample from the full posterior over trajectories of TVAR states and volatilities.

   (a) For each of these MC posterior samples, use the TVAR decomposition to compute the implied samples of the wavelengths and moduli of the 2 dominant (longest-wavelength) quasi-cyclical components in the series. Produce some relevant graphical summaries of some of the MC sampled trajectories and moduli over time. Comment on what is graphed in the context of retrospective analysis of this series and the use of TVAR DLMs, and why they are interesting in this applied study.

   (b) Climatological and geological questions have been generated over potential changes in climate forcing mechanisms around 1million, i.e., 1,000kyears, ago. These have been raised partly in response to the view that the 2 dominant earth-orbital dynamic mechanisms– of periods around 110kyears and 40kyears, respectively– seem to "look different" since that time that prior to that time.

   Use the MC samples of trajectories of moduli of the two dominant components to explore this via relevant posterior inferences. In particular:

   i. Compute, plot and discuss some sampled trajectories of $r_{t,1}/r_{t,2}$ over time where $r_{t,j}$ is the modulus of component j at time t.
   ii. Compute the MC estimate of $Pr(r_{t,1} > r_{t,2}|\mathcal{D}_T)$ for each time t and plot over time.

   On the basis of these summaries of your analysis, comment on the question of change a million years ago.

3. The beta-gamma discount random walk model that we have used normal precisions (inverse variances) can be used in other contexts where a positive model parameter is "locally constant" but varies over time. One example is time series of counts at relatively low levels, including zero values, where Poisson/gamma models are very relevant[1].

   Use/state theoretical results already defined in the normal DLM SV model– do not prove them from scratch. Refer to material in class slides and, particularly, HW#5 on the beta/gamma model, some of which recapitulates Problem 4 of P&W Chapter 4, Section 4.6, but with some differences in notation. That earlier context relates to the stochastic precision $\phi_t = v_t^{-1}$ in the normal DLM; the change here is that ϕ_t is a Poisson mean process. You have already worked through the relevant forward filtering, backwards smoothing and sampling results in HW#5 for the in the normal SV model. The application here to the Poisson time series model and essentials of analysis are the same with relevant notational changes.

   Suppose you have a time series problem with *count* data, $x_t = 0, 1, 2, \ldots$, over equally-spaced time t, and model the data as conditionally Poisson, $x_t|\phi_t \sim Po(\phi_t)$.

   At a time $t-1$, assume that the current posterior on the mean (a.k.a. level) of the Poisson time series is $\phi_{t-1}|\mathcal{D}_{t-1} \sim Ga(a_{t-1}, a_{t-1}/m_{t-1})$ for some shape parameter $a_{t-1} > 0$ and mean $E(\phi_{t-1}|\mathcal{D}_{t-1}) = m_{t-1} > 0$.

   (a) The Poisson mean evolves over $(t-1, t)$ to $\phi_t = \phi_{t-1}\eta_t/\beta$ where η_t is independent of ϕ_{t-1} with $\eta_t \sim Be(\beta a_{t-1}, (1-\beta)a_{t-1})$ for some discount factor $\beta \in (0, 1)$.
   What is the implied time t prior $p(\phi_t|\mathcal{D}_{t-1})$?

   (b) What is the 1-step ahead forecast mean $E(x_t|\mathcal{D}_{t-1})$?

   (c) What is the 1-step ahead predictive density $p(x_t|\mathcal{D}_{t-1})$?

   (d) Show that the posterior for $\phi_t|\mathcal{D}_t$ is gamma, $Ga(a_t, a_t/m_t)$ and give expressions for a_t, m_t. Show that the posterior mean $m_t \equiv E(\phi_t|\mathcal{D}_t)$ can be written as $m_t = r_t m_{t-1}$ where $r_t = (\beta a_{t-1} + x_t)/(\beta a_{t-1} + m_{t-1})$. Comment on how m_t depends on x_t, and on the effect of large values of a_{t-1}.

   (e) Assuming a time series observed over times $1 : T$, look back to time $t-1 < T$. Discuss how you can simulate the time $t-1$ retrospective distribution for $\phi_{t-1}|\phi_t, \mathcal{D}_T$ conditional on any value of ϕ_t. Comment on relevant theory and conditional independence structure.

   (f) How do you simulate from the full posterior of the trajectory $\phi_{1:T} = \{\phi_1, \ldots, \phi_T\}$ given the full observed series $\mathcal{D}_T$?

[1] Recall notation and structure of Poisson/gamma distributions:

- $x|\phi \sim Po(\phi)$, Poisson with mean $E(x|\phi) = \phi$ and p.d.f. $p(x|\phi) = \phi^x \exp(-\phi)/x!$ on $x = 0, 1, \ldots$.
- Gamma prior $\phi \sim Ga(a, b)$ with p.d.f. $p(\phi) = \phi^{a-1}\exp(-b\phi)b^a/\Gamma(a)$ on $\phi > 0$. Often parametrized with $b = a/m$ where $m = E(\phi)$.
- Implied marginal predictive distribution for x is negative binomial ("fatter tailed" than Poisson) $x \sim NB(a, p)$ with probability $p = b/(1+b) \equiv a/(a+m)$ and p.d.f. $p(x) = p^a(1-p)^x\Gamma(a+x)/\{\Gamma(a)\Gamma(x+1)\}$. In Matlab, the p.d.f. is evaluated via `nbinpdf(x,a,p)`, for example.

Aside just FYI: Applications involving discrete/count time series and this basic model setup include spatial time series of count data in demography and epidemiology, sales/consumer demand forecasting, and IT applications including dynamic network studies such as in the 2017 JASA paper Scalable Bayesian modeling, monitoring, and analysis of dynamic network flow data, by Xi Chen *et al*.

4. The time series intrusionevents.txt (in the course Data folder) represents a time series of monthly attempted intrusion events in a secure IT system over 52 biweekly periods. Assume that the Poisson/gamma discount model of Exercise 3 is adopted with the initial prior $\phi_0|\mathcal{D}_0 \sim Ga(a_0, a_0/m_0)$ with $a_0 = 25, m_0 = 14.5$.

   (a) Explore choice of discount factor β. Consider some values on the range $0.8{:}(0.01){:}0.99$. Choose a value you regard as relevant for this data set, and justify the choice.

   (b) At the chosen value of β, implement FFBS to generate Monte Carlo samples of the full trajectory $\phi_{1:T}$ over time up to the end $T = 52$. Use this to generate a large MC sample size of trajectories. Summarize inference in relevant graphical display(s).

   (c) The security protocols of the IT system were substantially modified in time period 44. Management is now interested in a formal statistical assessment of if– and, if so, how– that has impacted the level of intrusion events. For example, did the underlying level of intrusion events change as a result of those modifications? And, if so, how?

   Discuss and provide responses to these general questions, including (i) some comments on the x_t data series itself, and then (ii) formal inference on ϕ_{52}/ϕ_{44}.

   (d) Management is also interested in reliable predictions of the *total* number of intrusion events in the next k time periods. Describe how you would formally address the question.