Bayesian Nonparametric Autoregressive Models via Latent Variable Representation

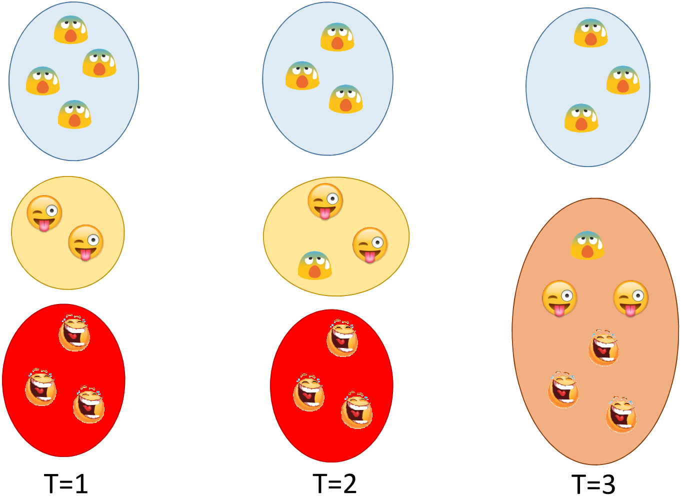
Maria De Iorio

Yale-NUS College
Dept of Statistical Science, University College London

Collaborators: Lifeng Ye (UCL, London, UK)
Alessandra Guglielmi (Polimi, Milano, Italy)
Stefano Favaro (Universitá degli Studi di Torino, Italy)

29th August 2018
Bayesian Computations for High-Dimensional Statistical Models

Motivation



Want

- flexible model for time-evolving distribution
- data driven clustering
- allow for covariates
- prediction
- feasible posterior inference
- framework for automatic information sharing across time
- possible easily generalizable tools for efficient sharing of information across data dimension (e.g. space)

Nonparametric Mixture

At time t, data $y_{1t}, \ldots, y_{nt}$

$$y_{i\mathbf{t}} \overset{iid}{\sim} f_{\mathbf{t}}(y) = \int K(y \mid \theta) G_{\mathbf{t}}(\,\mathrm{d}\theta)$$

where $G_{\mathbf{t}}$ is the mixing distribution at time t.

Assign flexible prior to G, e.g. DP, PT, Pytman-Yor, ...

$\Rightarrow$ Bayesian Nonparametric Mixture Models

Dirichlet Process (DP)

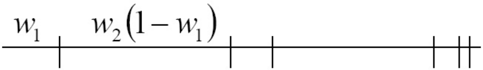
Probability model on distributions $G \sim DP(\alpha, G_0)$, with measure $G_0 = \mathbb{E}(G)$ and precision parameter α.

G is a.s. discrete

Sethuraman's stick breaking representation

$$
\begin{aligned}
G &= \sum_{h=1}^{\infty} w_h \delta_{\theta_h} \\
\xi_h &\sim \mathrm{Beta}(1, \alpha) \\
w_h &= \xi_h \prod_{i=1}^{h-1} (1 - \xi_i),
\end{aligned}
$$

scaled Beta distribution



$$
\theta_h \stackrel{iid}{\sim} G_0, \quad h = 1, 2, \ldots
$$

where $\delta(x)$ denotes a point mass at x, ψ_h are weights of point masses at locations θ_h.

Dirichlet Process Mixtures (DPM)

In many data analysis applications the discreteness is inappropriate.

<u>To remove discreteness:</u> convolution with a continuous kernel

$$\begin{aligned} f(y) &= \int p(y \mid \theta) \mathrm{d}G(\theta) \\ G &\sim \mathrm{DP}(\alpha, G_0) \end{aligned}$$

Dirichlet Process Mixtures (DPM)

...or with latent variables θ_i

$$\begin{aligned} G &\sim \mathrm{DP}(\alpha, G_0) \\ \theta_i &\sim G \\ f(y) &= p(y \mid \theta_i) \end{aligned}$$

Nice feature: Mixture is discrete with probability one, and with small α, there can be high probabilities of a finite mixture.

Often $p(y \mid \theta) = N(\beta, \sigma^2) \longrightarrow f(y) = \sum_{h=1}^{\infty} w_h N(\beta_h, \sigma^2)$

Comment

Under $G : p(\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i'}) > 0$

- Observations share the **same** $\boldsymbol{\theta}$ $\Rightarrow$ belong to the **same cluster**

$\Rightarrow$ DPM induces a **random partition** of the observations $\{1, \dots, n\}$

Note: in the previous model the clustering of observations depends only on the the distribution of y.

Models for collection of distributions

Observations are associated to different temporal coordinates

$$y_t \overset{ind}{\sim} f_t(y) = \int K(y \mid \theta) G_t(\, \mathrm{d}\theta)$$

Recent research focuses on developing models for a collection of random distributions

$$\{G_t; t \in \mathcal{T}\}$$

Interest: $\mathcal{T}$ discrete set.
Goal: induce dependence.

Reason: properties of the distributions f_t are thought to be similar in some way; e.g. similar means, similar tail behaviour, distance between them small.

Dependent Dirichlet Process

If $G \sim DP(G_0, \alpha)$, then using the constructive definition of the DP (Sethuraman 1994)

$$G_t = \sum_{k=1}^{\infty} w_{tk} \delta_{\theta_{tk}}$$

where $(\theta_{tk})_{k=1}^{\infty}$ are iid from some $G_0(\theta)$ and (w_{tk}) is a stick breaking process

$$\begin{aligned} w_{tk} &= z_{tk} \prod_{j<k} (1 - z_{tj}) \\ z_{tj} &\sim \text{Beta}(1, \alpha_t) \end{aligned}$$

Dependence has been introduced mostly in regression context, conditioning on level of covariates x.

Dependence Structure

Introduce dependence

- through the base distributions G_{0t} of conditionally independent nonparametric priors G_t $\Rightarrow$ Simple but restrictive
- dependence only in the atoms of the G_t $\Rightarrow$ efficient computations but not very flexible approach
- dependence structure in the weights $\Rightarrow$ complex and inefficient computational algorithms, limiting the applicability of the models
- alternative is to assume

$$G_t = \pi_t \widetilde{G} + (1 - \pi_t) G_t^{\star}$$

Dependence through Weights

- flexible strategy
- random prob measures share the same atoms
- under certain conditions we can approximate any density with any atoms
- varying the weights can provide prob measures very close (similar weights) or far apart

Temporal Dependence

- ✓ **Literature:** Griffin and Steel (2006), Caron, Davy and Doucet (2007), Rodriguez and ter Horst (2008), Rodriguez and Dunson (2009), Griffin and Steel (2009), Mena and Walker (2009), Nieto-Barajas et al. (2012), Di Lucca et al. (2013), Bassetti et al. (2014), Xiao et al. (2015), Gutirrez, Mena and Ruggiero (2016) ...
- ✓ There exist many related constructions for dependent distributions defined through Poisson-Dirichlet Process (e.g. Leisen and Lijoi (2011) and Zhu and Leisen (2015)) or correlated normalized completely random measures (e.g. Griffin et al. 2013, Lijoi et al. 2014)
- ✓ Related approach: **Covariate Dependent Random Partion Models**
- ✓ **Idea:** dynamic DP extension can be developed by introducing temporal dependence in the weights through a transformation of the Beta random variables and specifying common atoms across G_t.

Related Approaches

$$G_t = \sum_{k=1}^{\infty} w_{tk}\delta_{\theta_k}, \quad w_{tk} = \xi_{tk}\prod_{i=1}^{k-1}(1-\xi_{ti}), \quad \xi_{tk} \sim \text{Beta}(1,\alpha)$$

✗ **BAR Stick–Breaking Process** (Taddy 2010): defines evolution equation for w_t:

$$\begin{aligned} \xi_t &= (u_t v_t)\xi_{t-1} + (1-u_t) \sim Beta(1,\alpha) \\ u_t &\sim \text{Beta}(\alpha, 1-\rho), \qquad v_t \sim \text{Beta}(\rho, 1-\rho) \\ 0 &< \rho < 1, \qquad u_t \perp\!\!\!\perp v_t \end{aligned}$$

✗ DP marginal, $\text{corr}(\xi_t, \xi_{t-k}) = [\rho\alpha/(1+\alpha-\rho)]^k > 0$

✗ Prior simulations show that number of clusters and number of singletons at time $t = 1$ is different from other times.

✗ Very different clustering can corresponds to relatively similar predictive distributions.

✗ **Latent Gaussian time-series** (DeYoreo and Kottas 2018):

$$
\begin{aligned}
\xi_t &= \exp\left\{-\frac{\zeta^2+\eta_t^2}{2\alpha}\right\} \sim \mathrm{Beta}(\alpha,1) \\
\zeta &\sim \mathrm{N}(0,1) \\
\eta_t \mid \eta_{t-1},\phi &\sim \mathrm{N}(\phi\eta_{t-1}, 1-\phi^2), \qquad |\phi|<1 \\
w_{t1} &= 1-\xi_{t1}, \qquad w_{tk} = (1-\xi_{tk})\prod_{i=1}^{k-1}\xi_{ti}
\end{aligned}
$$

η_t AR(1) process.

✗ $\alpha \geq 1$ implies a 0.5 lower bound on $\mathrm{corr}(\xi_t, \xi_{t-k})$. As a result, same weights (For example, the first weight at time 1 and first weight at time 2) always have correlation above 0.5.

✗ η (AR component) is squared so that the correlation between different times is positive.

✗ These problems can be overcome by assuming time-varying locations.

Latent Autoregressive Process

Want: Autoregressive model for a collection of random function G_t.

Goal: to obtain flexible time-dependent clustering.

We borrows ideas from copula literature, in particular from Marginal Beta regression (Guolo & Varin 2014)

Recall Probability Integral Transformation:

If $\epsilon \sim \mathrm{N}(0, 1)$ and F is the cdf of a Beta distribution then

$$Y = F^{-1}(\Phi(\epsilon); a, b)$$

is marginally distributed as a Beta with parameters (a, b).

ARDP

✓ AR process:

$$\begin{aligned} \epsilon_1 &\sim \mathrm{N}(0,1) \\ \epsilon_t &= \psi\epsilon_{t-1} + \eta_t, \qquad t > 1 \\ \eta_t &\overset{iid}{\sim} \mathrm{N}(0, 1-\psi^2) \end{aligned}$$

✓ Stick-breaking construction

$$\begin{aligned} \{\epsilon_{tk}\} &\overset{iid}{\sim} \mathrm{AR}(1), \qquad k = 1, \ldots, \infty \\ \xi_{tk} &= F_t^{-1}(\Phi(\epsilon_{tk}); a_t, b_t) \\ w_{tk} &= \xi_{tk} \prod_{l<k} (1 - \xi_{tl}) \\ \theta_k &\overset{iid}{\sim} G_0 \\ G_t &= \sum_{k=1}^{\infty} w_{tk} \delta_{\theta_k} \end{aligned}$$

Comments

- Easy to generate $\{G_t, t = 1, \ldots\}$.
- The marginal distribution of ϵ_t is N(0,1) and therefore the marginal distribution of ξ_t is Beta with desired parameters.
- If $a_t = 1, b_t = \alpha$, then marginally each G_t is DP(α, G_0).
- The $\{\xi_t\}$ inherit the same Markov structure (AR(1)) of the $\{\epsilon_t\}$ process, and therefore also the $\{G_t, t = 1, 2, \ldots\}$ is AR(1).
- Easy to derive the $k-$step predictive densities G_{t+k} as it is easy to derive the weights.

Evolution of the weights through time

We can derive:

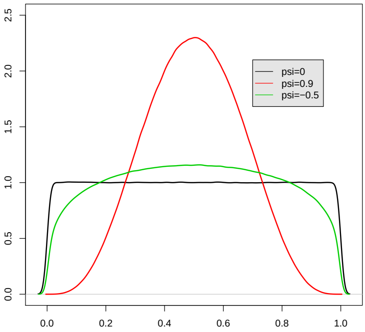
- ✓ The conditional distribution of $\xi_t \mid \xi_{t-1}$:

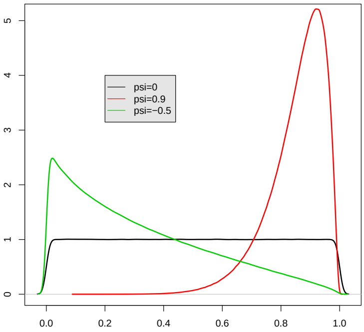
$$\begin{aligned} \mathcal{L}(\xi_t \mid \xi_{t-1}) &\stackrel{d}{=} 1-(1-\Phi(Z))^{1/\alpha} \\ Z &\sim \mathcal{N}\left(\psi\Phi^{-1}\left(1-(1-\xi_{t-1})^{\alpha}\right), 1-\psi^2\right). \end{aligned}$$

- ✓ The conditional law of ξ_t, given ϵ_{t-1}:

$$\begin{aligned} \mathcal{L}\left\{\xi_t \mid \epsilon_{t-1}\right\} &= \mathcal{L}\left\{F_t^{-1}(\Phi(\epsilon_t); a, b) \mid \epsilon_{t-1}\right\} \\ \epsilon_t \mid \epsilon_{t-1} &\sim \mathrm{N}(\psi\epsilon_{t-1}, 1-\psi^2) \\ \mathbb{P}(\xi_t \leq x \mid \epsilon_{t-1}) &= \Phi\left(\frac{\Phi^{-1}(1-(1-x)^{\alpha})-\psi\epsilon_{t-1}}{\sqrt{1-\psi^2}}\right) \end{aligned}$$

$\mathcal{L}(\xi_2 \mid \xi_1)$





$\xi_1 = 0.5$ $\xi_1 = 0.9$

$\alpha = 1$. Ability of accommodate a variety of distributional shapes over the unit interval.

Posterior Inference

- ✗ MCMC based on truncation of DP to L component.
- ✗ Blocked Gibbs sampler (Ishwaran and James 2001) is extended to include a Particle MCMC update (Andrieu et al. 2010) for the weights of the DP.
- ✗ Let $\mathbf{s}_t$ and $\mathbf{w}_t$ be the allocation and the weight vector at time t, respectively. We need to sample from

$$p(\mathbf{w}_1, \ldots, \mathbf{w}_T \mid \mathbf{s}_1 \ldots, \mathbf{s}_T, \psi)$$

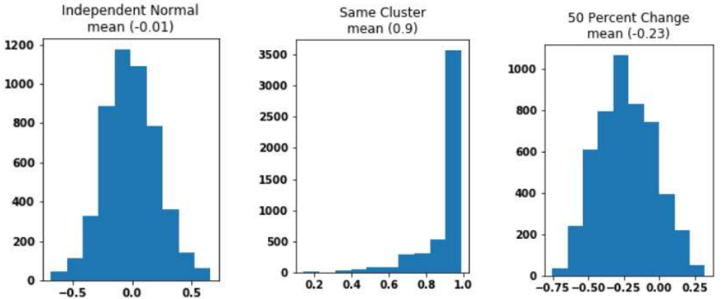
- ✗ Standard algorithm such as the FFBS are not applicable since 1-step predictive distributions, $p_\psi(\mathbf{w}_t \mid \mathbf{w}_{t-1})$, cannot be derived analytically.
- ✗ Employ PMCMC , using the prior to generate samples ϵ_t. We exploit

$$\mathcal{L}_\psi(\mathbf{w}_t \mid \mathbf{w}_{t-1}) = \mathcal{L}_\psi(\epsilon_t \mid \epsilon_{t-1})$$

SMC approximations:

$$\widehat{p}_\psi(\mathbf{w}_1, \ldots, \mathbf{w}_T \mid \mathbf{s}_1 \ldots, \mathbf{s}_T) = \sum_{r=1}^{R} \omega_T^r \delta_{\epsilon_{1:T}}$$

Simulations



- $T = 4$, $n = 100$, independent time points, $y_{it} \sim$ N.
- $T = 4$, $n = 100$. At $t = 1$ data generated from 2 clusters of equal size. For $t = 2, 3, 4$, individuals remain in the same cluster and the distribution of cluster will remain same over time.
- $T = 4$, $n = 100$. At $t = 1$ data generated from 2 clusters of equal size. For $t = 2, 3, 4$, individuals remain in the same cluster with probability 50% as $t = i - 1$, and with probability 50% , people will switch to the other cluster. Note that the distribution of each cluster will change over time.

Disease Mapping

- ✓ Disease incidence or mortality data are typically available as rates or counts for specified regions, collected over time.
- ✓ **Data:** breast cancer incidence data of 100 MSA (Metropolitan Statistical Area) from *United States Cancer Statistics: 1999–2014 Incidence, WONDER Online Database.*
- ✓ We use the data from year 2004 to year 2014.
- ✓ Population data of year 2000 and year 2010 is obtained from U.S. Census Bureau, Population Division. The population data of the remaining years are estimated by linear regression.
- ✓ **Primary goal of the analysis:**
  - ➜ identification of spatial and spatio-temporal patterns of disease (*disease mapping*)
  - ➜ spatial smoothing and temporal prediction (forecasting) of disease risk.

Space-Time Clustering

- Y_{it} = breast cancer incidence counts (number of cases), in region MSA_i at time t
- N_{it} = the number of individuals at risk
- R_{it} = disease rate
- Model:

$$\begin{aligned}
Y_{kt} \mid N_{it}, R_{it} &\sim \mathrm{Poisson}(N_{it} R_{it}), \qquad i = 1, \ldots, 100; \quad t = 1, \ldots 11 \\
\ln(R_{it}) &= \mu_{it} + \phi_i \\
\mu_{it} \mid G_t &\sim G_t \\
\{G_t, t \geq 1\} &\sim \mathrm{ARDP}(1) \\
G_0 &= \mathrm{N}(0, 10) \\
\phi \mid \mathbf{C}, \tau^2, \rho &\sim \mathrm{N}\left(\mathbf{0}, \tau^2 \left[\rho \mathbf{C} + (1-\rho)\mathbf{I}_n\right]^{-1}\right)
\end{aligned}$$

Spatial Component

$$\phi \mid \mathbf{C}, \tau^2, \rho \sim \mathrm{N}\left(\mathbf{0}, \tau^2 \left[\rho \mathbf{C} + (1-\rho)\mathbf{I}_n\right]^{-1}\right)$$

- variance parameter τ^2 controls the amount of variation between the random effect.
- the weight parameter ρ controls the strength of the spatial correlation between the random effects
- the elements of $\mathbf{C}$ are equal to

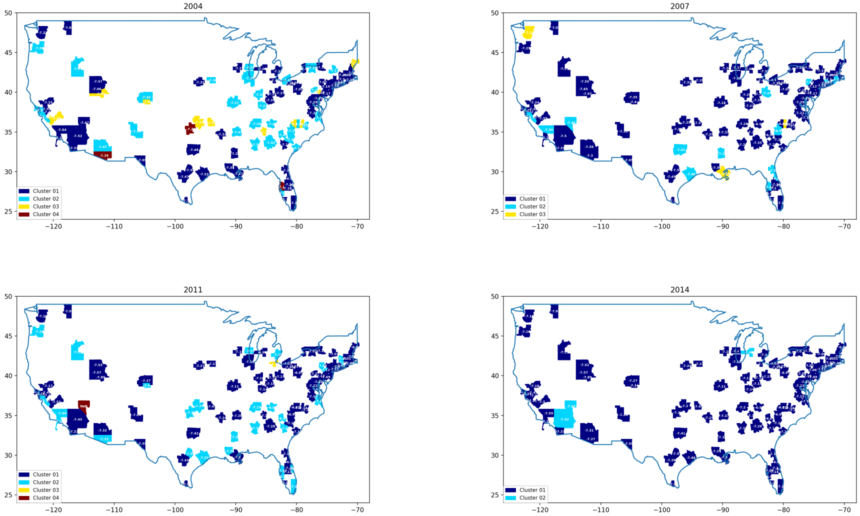
$$c_{jk} = \begin{cases} n_k, & \text{if } j = k \\ -1, & \text{if } j \sim k \\ 0, & \text{otherwise.} \end{cases}$$

where $j \sim k$ denotes area (j, k) are neighbours and n_k denotes the number of neighbours of area (k).

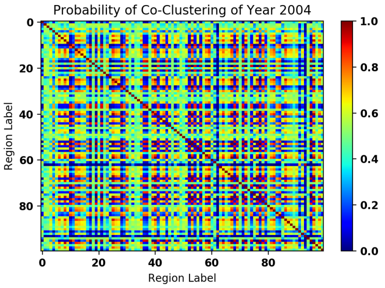
Prior specification:

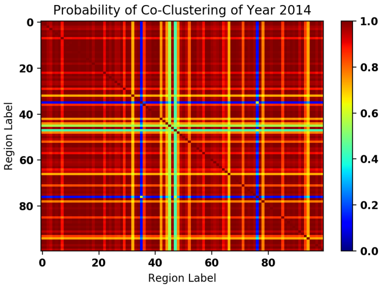
$$\tau^2 \sim \mathtt{U}(0, 10), \qquad \rho \sim \mathtt{discrete\ uniform}(0, 0.05, 0.10, 0.15, ..., 0.95)$$

Posterior Inference on Clustering

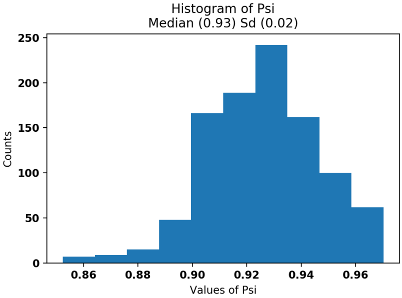


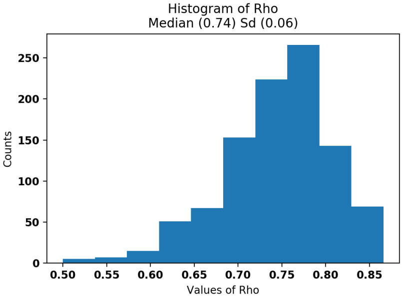
Co-Clustering Probability





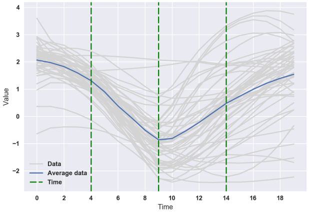
Posterior Inference on Correlations





Dose-escalation Study

- Data: wbc over time for $n = 52$ patients receiving high doses of cancer chemotherapy
- CTX: anticancer agent, known to lower a person's wbc
- GM-CSF: drugs given to mitigate some of the side-effects of chemotherapy



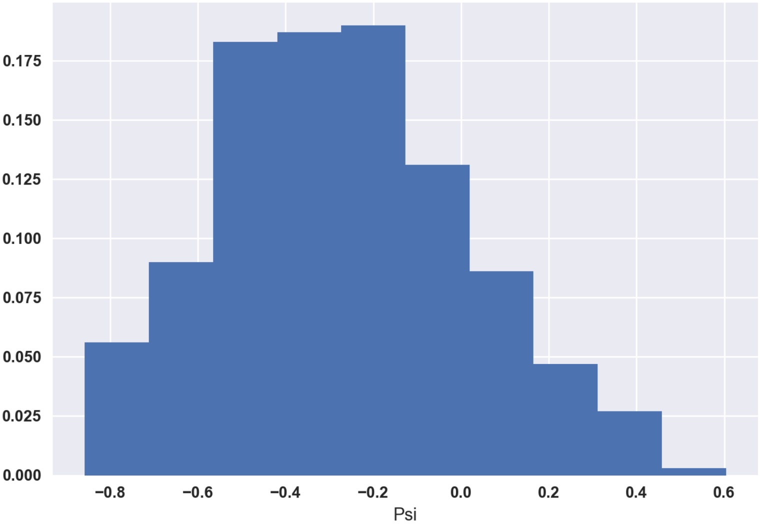
WBC profiles over time:

- initial baseline
- sudden decline when chemotherapy starts
- slow S-shaped recovery back to $\approx$ baseline after end of treatment
- interest in understanding the effect of dose on wbc in order to protect patients against severe toxicity

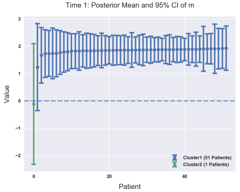
Model

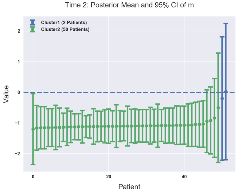
$$
\begin{aligned}
\log(Y_{it}) \mid \mu_{it}, \tau_{it} &\sim \mathrm{N}(\mu_{it}, (\lambda\tau_{it})^{-1}) \\
\mu_{it} &= m_{it} + \mathbf{x}_{it}\boldsymbol{\beta}_t \\
m_{it}, \tau_{it} \mid G_t &\sim G_t \\
\{G_t, t \geq 1\} &\sim \mathrm{ARDP}(1) \\
G_0 &\sim \mathrm{NormalGamma}(\mu_0, \lambda, \alpha, \gamma)
\end{aligned}
$$

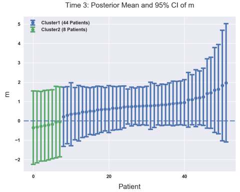
Posterior Inference on ψ



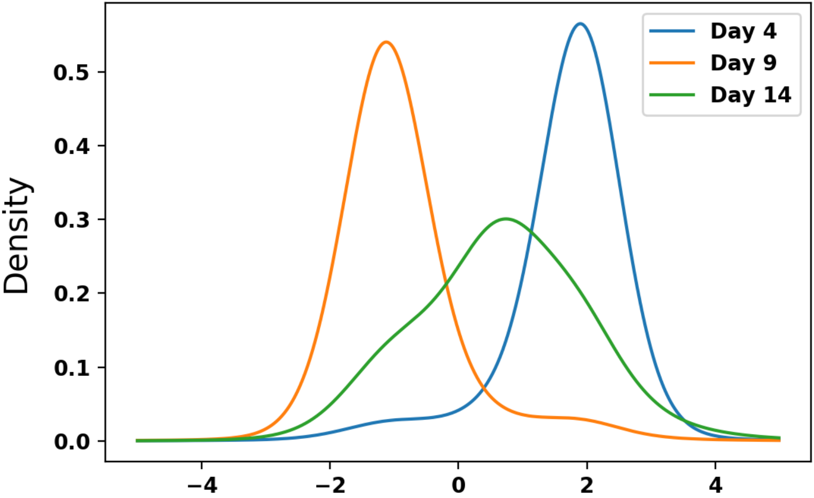
Posterior Inference on Clustering







Posterior Density Estimation



Conclusions

- ✓ Dependent process for time-evolving distributions based on the DP — could be generalised to GDP
- ✓ Introduce temporal dependence through latent stochastic process – Normal variables.
- ✓ Advantage – more general process on ϵ, e.g ARMA
- ✓ flexible – can accommodate different correlation pattern
- ✓ dependent clustering
- ✓ borrowing of information across time-periods
- ✓ general applicability
- ✓ allows for prediction of G_{t+1}
- ✓ posterior computations through PMCMC