A Short Course on Bayesian Nonparametrics

Lecture 5 - Nonparametric priors on collections of Distributions

Abel Rodriguez - UC, Santa cruz

Universidade Federal Do Rio de Janeiro
March, 2011

Modeling collections of distributions

- Models for collections of measures:

$$G_{\mathcal{S}} = \{G_s : s \in \mathcal{S} \subset \mathbb{R}^d\}$$

- We already discussed some examples (when we discussed unrestricted exchangeability).
- We now interested in general mechanisms that allow us to deal with distribution that vary with the level of some covariate:
  - Covariates.
  - Spatial models.
  - Temporal models.

Modeling collections of distributions

- Dependent stick-breaking processes:
  - ANOVA DDPs (De Iorio et al., 2004).
  - Spatial DDPs (Gelfand et al., 2005).
  - DLM DDPs (Rodriguez & ter Horst, 2008).
  - Kernel stick-breaking processes (Dunson & Park, 2008)
  - Probit stick-breaking processes (Rodriguez & Dunson, 2011).
  - Order dependent DPs (Griffin & Steel, 2006).

  There is a strong link with mixture of experts/latent class models.
- Linear combinations of DPs.
  - Density regression (Dunson et al., 2007).

Dependent Dirichlet processes

- Start with the constructive definition of the DP,

$$G(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\vartheta_k}$$

$\vartheta_k \sim_{idd} H$ and $\omega_k = z_k \prod_{l<k}\{1 - z_l\}$ and $z_l \sim_{iid}$ beta$(1, \alpha)$.

- Now replace the iid realizations from probability distributions by iid realizations from stochastic processes, so that

$$G_{\mathcal{S}}(\cdot) = \sum_{k=1}^{\infty} \omega_{\mathcal{S},k} \delta_{\vartheta_{\mathcal{S},k}}$$

$\vartheta_{\mathcal{S},k} \sim_{idd} H_{\mathcal{S}}$ and $\omega_{\mathcal{S},k} = z_{\mathcal{S},k} \prod_{l<k}\{1 - z_{\mathcal{S},l}\}$ and $z_{\mathcal{S},k} \sim_{iid}$ come from a stochastic process with beta$(1, \alpha(s))$.

- The marginals $G_s(\cdot)$ are Dirichlet processes for every $s \in \mathcal{S}$.

Dependent Dirichlet processes

- The HDP and the NDP are examples of DDP priors.
  - For the HDP, remember the representation,

$$G_i(\cdot) = \sum_{k=1}^{\infty} \varpi_{i,k} \delta_{\vartheta_k} \qquad \varpi_i \sim \mathsf{DP}(\alpha, \gamma) \qquad \gamma \sim \mathsf{SB}(\beta)$$

    (dependence in weights, but not in atoms). Well, not quite a DDP: The marginals are not really DPs.
  - For the NDP, note that

$$G_i(\cdot) = \sum_{k=1}^{\infty} \varpi_{i,k} \delta_{\vartheta_{l,k}}$$

    where $(\{\varpi_{l,k}\}, \{\vartheta_{l,k}\})$ come from a Pólya urn (dependence on both atoms and weights).

"Single-p" models

- Simplest construction: Replace the atoms with stochastic processes, but leave the weights constant.
- Computational simplicity: The single-p DDP is just a DP mixture of stochastic processes

$$G_{\mathcal{S}}(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\vartheta_{\mathcal{S},k}}$$

  so all computational tools we discussed for the DP are applicable.
- Drawbacks:
  - You might need replicates at each level s in order to estimate the process.
  - Cannot produce independent collections of distributions.

ANOVA DDP

- Same setting as an ANOVA model: $y_{i,j}$ is the value of a *continuous* outcome associated with the j-th replicate at covariate level i. For a one-way ANOVA it is typical to take

$$y_{i,j} = \mu + \alpha_i + \epsilon_{i,j} \qquad\qquad \epsilon_{i,j} \sim \mathrm{N}(0, \sigma^2)$$

  for $j = 1, \ldots, J_i$ and $i = 1, \ldots, I$. This implies that outcomes are normally distributed for each level of the covariate.

- The model can written as

$$y_{i,j} = d'_{i,j}\theta + \epsilon_{i,j} \qquad\qquad \epsilon_{i,j} \sim \mathrm{N}(0, \sigma^2).$$

  where $\theta = (\mu, \alpha_1, \ldots, \alpha_I)'$, and $d_{i,j,k} = 1$ for $k = 1$ and $k = i + 1$ and zero otherwise (i.e., $d_{i,j} = (1, 0, \ldots, 1, \ldots, 0)'$).

ANOVA DDP

- In a Bayesian framework, the model needs a prior for θ (usually a normal prior for conjugacy).

$$y_{i,j} \sim \mathrm{N}\left(d_{i,j}'\theta, \sigma^2\right) \qquad \theta \sim \mathrm{N}(\theta_0, \Omega)$$

- Generalize the model to a DP mixture

$$y_{i,j} \sim \mathrm{N}\left(d_{i,j}'\theta_{i,j}, \sigma^2\right) \quad \theta_{i,j} \sim \tilde{G} \quad \tilde{G} \sim \mathrm{DP}\{\alpha, \mathrm{N}(\theta_0, \Omega)\}$$

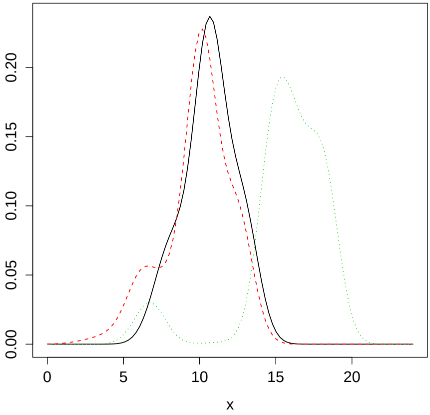
- Observations from one of an infinite number of ANOVAs.
- Equivalent to writing (after carefully picking θ_0 and Ω)

$$y_{i,j} \sim \mathrm{N}\left(\theta_{i,j}^*, \sigma^2\right) \qquad \theta_{i,j}^* \sim G_i \qquad G_i = \sum_{k=1}^{K} \omega_k \delta_{\vartheta_{i,k}^*}$$

where $\vartheta_{i,k}^* = m_k + A_{i,k}$, $m_k \sim_{iid} \mathrm{N}(\varphi_m, \tau_m^2)$, $A_{1,k} = 0$, and $A_{i,k} \sim_{iid} \mathrm{N}(\varphi_{A_i}, \tau_{A_i}^2)$.

Samples from the ANOVA DDP prior

Realizations of the ANOVA-DDP with $I = 3$.



ANOVA DDP

- Easy to generalize to multivariate continuous outcomes, or k-way ANOVAs.
- Not so easy to generalize to other types of outcomes (like mixtures of GLMs) $\Rightarrow$ Main difficulty is that computational simplicity is lost.
- The default contrasts used by De Iorio et al (2004) do not imply balanced priors for all groups $\Rightarrow$ Not clear how much of a problem that is.
- Could potentially add variable selection priors in the baseline measure (nobody has done it, that I know of, except maybe Dunson & Yi 2009 in a slightly different context).

Sampling for ANOVA DDP

- Because of conjugacy, exploit collapsed Gibbs samplers.
- Introduce component indicators $\{\xi_{i,j}\}$, where $\theta_{i,j} = \vartheta_{\xi_{i,j}}$.
- Conditionally on the component indicators, we can sample ϑ_k from the posterior of a one-way ANOVA model based on observations $\{y_{i,j} : \xi_{i,j} = k\}$.

$$\theta_k | \cdots \sim \mathsf{N}\left(\left\{ \Omega^{-1} + \sum_{\{(i,j):\xi_{i,j}=k\}} \frac{d_i d_i'}{\sigma^2} \right\}^{-1} \left\{ \Omega^{-1}\theta_0 + \sum_{\{(i,j):\xi_{i,j}=k\}} \frac{d_i y_i}{\sigma^2} \right\} \right.$$
$$\left. \left\{ \Omega^{-1} + \sum_{\{(i,j):\xi_{i,j}=k\}} \frac{d_i d_i'}{\sigma^2} \right\}^{-1} \right)$$

- The indicators $\{\xi_{i,j}\}$ can be sampled using the Pólya urn.
- Hyperparameters can be sampled conditionally on $\{\vartheta_k\}$.

Sampling for ANOVA DDP

- If you just want to apply the default ANOVA DDP (rather than include as part of a hierarchical model), you can do it through `DPpackage`.
- Look at the function `LDDPdensity` $\Rightarrow$ It implements a slight generalization of the model (which also includes the variance of the observations in the mixture).

Spatial DDPs

- Consider now creating a prior for an uncountable collection of random variables.

$$G_{\mathcal{S}}(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\vartheta_{\mathcal{S},k}}$$

where $\vartheta_{\mathcal{S},k} \sim_{iid} \mathrm{GP}(\mu(s), \gamma(s, s'))$.

- Hence for any finite set of locations $s_1, \ldots, s_n$ we have $G_{s_i} = \sum_{k=1}^{\infty} \omega_k \delta_{\vartheta_{s_i,k}}$ and

$$\begin{pmatrix} \vartheta_{s_1,k} \\ \vdots \\ \vartheta_{s_n,k} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu(s_1) \\ \vdots \\ \mu(s_n) \end{pmatrix}, \begin{pmatrix} \gamma(s_1, s_1) & \cdots & \gamma(s_1, s_n) \\ \vdots & \ddots & \vdots \\ \gamma(s_n, s_1) & \cdots & \gamma(s_n, s_n) \end{pmatrix} \right)$$

Spatial DDPs

- Assume that we observe T realizations of the process at locations $s_1, \dots, s_n$

$$y_t = (y_t(s_1), \dots, y_t(s_n))' \qquad t = 1, \dots, T$$

- One way to use the spatial DDP is to build a hierarchy where $\theta_t = (\theta_{s_1,t}, \dots, \theta_{s_n,t})'$, so that

$$y_t \sim \mathrm{N}(\theta_t, \sigma^2 I) \qquad\qquad \theta_t \sim \sum_{k=1}^{\infty} \omega_k \delta_{\vartheta_k}$$

  and $\vartheta_k = (\vartheta_{s_1,k}, \dots, \vartheta_{s_n,k})'$.
- Since the locations are the same for each T, this is merely a DP mixture of multivariate Gaussians.
- One surface for each $y_t \Rightarrow$ Global surface selection.

Spatial DDPs

- In a slightly different version of the model let

$$y_t(s_i) \sim \mathrm{N}(\theta_t(s_i), \sigma^2) \qquad \theta_t(s_i) \sim G_{s_i} \qquad G_{s_i} = \sum_{k=1}^{\infty} \omega_k \delta_{\vartheta_{s_i,k}}$$

with

$$\begin{pmatrix} \vartheta_{s_1,k} \\ \vdots \\ \vartheta_{s_n,k} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu(s_1) \\ \vdots \\ \mu(s_n) \end{pmatrix}, \begin{pmatrix} \gamma(s_1,s_1) & \cdots & \gamma(s_1,s_n) \\ \vdots & \ddots & \vdots \\ \gamma(s_n,s_1) & \cdots & \gamma(s_n,s_n) \end{pmatrix} \right)$$

- One whole surface for each $y_j(s_i) \Rightarrow$ Local surface selection.
- The first approach makes sense if each vector of replicates $y_t = (y_t(s_1), \ldots, y_t(s_n))'$ are separately exchangeable, the second if they are unrestrictedly exchangeable.

Spatial DDPs

- Note that

$$\mathrm{E}(y(s)) = \mathrm{E}_H(\vartheta(s))$$

$$\mathrm{Cov}\{y(s), y(s')\} = \frac{1}{1+\alpha}\mathrm{Cov}_H(\vartheta(s), \vartheta(s'))$$

  Hence, if the baseline process is stationary, then the model is *a priori* a stationary.

- However, *a posteriori*, the model is non-stationary because

$$\mathrm{E}(y(s)|G_{\mathcal{S}}) = \sum_{k=1}^{\infty} \omega_k \theta_{s,k}$$

$$\mathrm{Cov}\{y(s), y(s')|G_{\mathcal{S}}\} = \left\{ \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \omega_k \omega_l \theta_{s,k} \theta_{s',l} \right\} - \left\{ \sum_{k=1}^{\infty} \omega_k \theta_{s,k} \right\} \left\{ \sum_{k=1}^{\infty} \omega_k \theta_{s',k} \right\}$$

Spatial DDPs

- Given the indicators, we have K conditionally independent GP regression models, which share the same prior mean and covariance functions. For example, in the global surface selection model:
  - The surfaces for each of the K GPs are sampled as

$$\vartheta_k | \ldots \sim \mathrm{N}\left(\left\{ \Gamma^{-1} + \frac{m_k}{\sigma^2} I \right\}^{-1} \left\{ \Gamma^{-1}\mu + \frac{1}{\sigma^2} \sum_{t:\xi_t = k} y_t \right\}, \left\{ \Gamma^{-1} + \frac{m_k}{\sigma^2} I \right\}^{-1} \right)$$

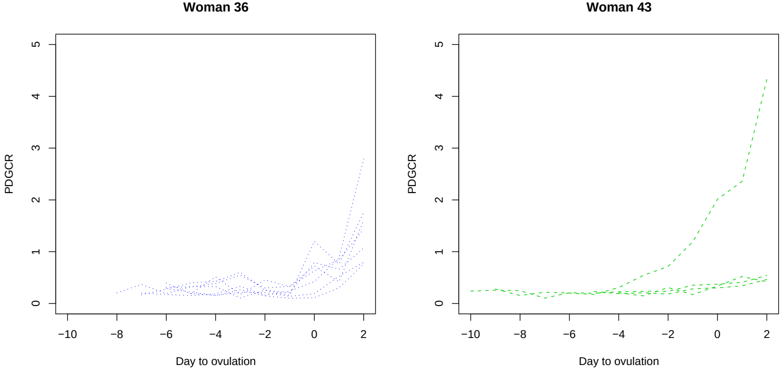
  - The parameter η of the baseline measure are sampled using standard MH steps.

$$p(\eta | \cdots) \propto \left\{ \prod_{k=1}^{K} \mathrm{N}(\vartheta_k | \mu_\eta, \Gamma_\eta) \right\} p(\eta)$$

- Indicators sampled using the Pólya urn.

Spatial DDPs: A non-spatial example

- Functional clustering in reproductive function studies: Let $y_j(s_i)$ be the level of progesterone during the s_i day of j-th menstrual cycle of a woman.



DLM-DDPs

- Let $G_t = \sum_{k=1}^{\infty} \omega_k \delta_{\vartheta_{t,k}}$ for $t = 1, \ldots, T$ where

$$\vartheta_{t,k} | \vartheta_{t-1,k} \sim \mathsf{N}\left(B_t \vartheta_{t-1,k}, W_t\right) \qquad \vartheta_{0,k} \sim \mathsf{N}(m_0, C_0)$$

- Natural setting $\Rightarrow$ Distribution evolving in time.
- To complete the model, set

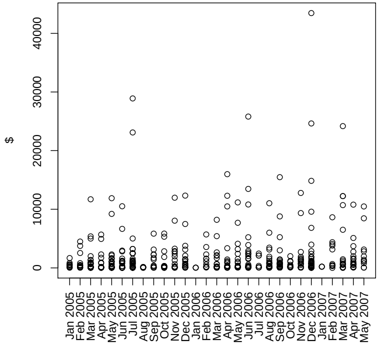
$$y_{t,i} \sim \mathsf{N}(A_{t,i}\theta_{t,i}, \sigma^2) \qquad \theta_{t,i} \sim G_t$$

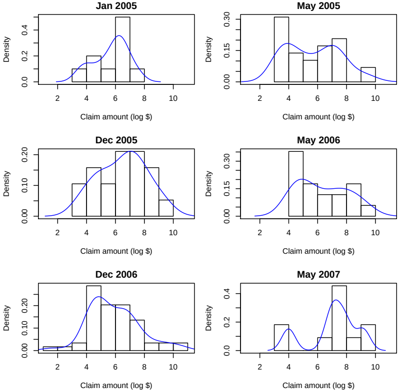
An application of the DLM-DDP

- We can exploit all we know about DLMs to build models that incorporate trends, periodicities, etc.
- We can slightly extend the model to include the variance σ^2 in the mixture $\Rightarrow$ Adaptive bandwith in space.
- We could also modify it to handle adaptive bandwidths in time.
- Filtered and smoothed density estimates can be obtained.
- Sampling follows along the same lines as other single-p models. Efficient sampling for the $\vartheta_{t,k}$s uses FFBS $\Rightarrow$ a slight adaptation is needed to account for missing data and more than one observation at each time point.

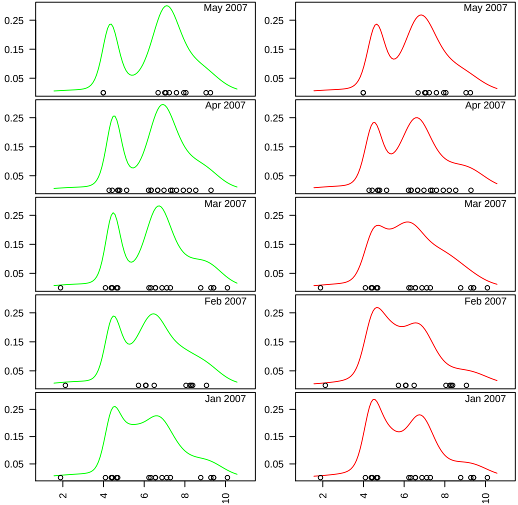
An application of the DLM-DDP

Modeling the value of reimbursement claims: $T = 29$ data points.





An example


May 2007
Apr 2007
Mar 2007
Feb 2007
Jan 2007
0.25
0.15
0.05
2
4
6
8
10

Single-p DDPs

- Note that all these examples of single-p DDPs focus on normals distributions $\Rightarrow$ This is because of computational tractability.
- The only other natural single-p model that nobody seems to have discussed is a mixture of CAR models $\Rightarrow$ Not clear what an application is.

Building dependence through the weights

- In many applications it might be more natural to use the same atoms for all distributions and build dependence through the atoms.
- Building dependence through the weights allows us to easily incorporate non-Gaussian kernels.
- In some cases, models that build dependence through the atoms seem to be identifiable even if only one observation is collected at value of s.
- However, building models for beta processes for which inference is simple is tricky!!

Probit stick-breaking processes (PSBP)

- First, the single-distribution case: Start with the stick-breaking construction,

$$G(\cdot) = \sum_{k=1}^{\infty} \omega_k \delta_{\vartheta_k} \qquad \vartheta_k \sim_{iid} H \qquad \omega_k = u_k \prod_{l<k} \{1 - u_l\}$$

- In the DP, we have $u_l \sim \text{beta}(1, \gamma)$. Instead, let

$$u_l = \Phi(\alpha_l) \qquad\qquad \alpha_l \sim \mathrm{N}(\mu, \sigma^2)$$

- If $\mu = 0$ and $\sigma = 1$ then $u_l \sim \text{Uni}[0, 1] \Rightarrow \text{DP}(\gamma = 1, H)$.

Properties of the PSBP

- Remember that for each measurable B, $G(B)$ is a random variable.

$$\mathrm{E}\{G(B)\} = H(B) \quad \mathrm{Var}\{G(B)\} = \frac{\beta_2}{2\beta_1 - \beta_2} H(B)\{1 - H(B)\}$$

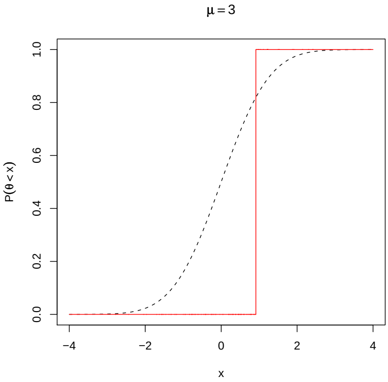
where

$$\beta_1 = \mathrm{E}(u_l) = \Phi(\mu/\sqrt{1+\sigma^2}) \quad \beta_2 = \mathrm{E}(u_l^2) = \Pr(T_1 > 0, T_2 > 0)$$

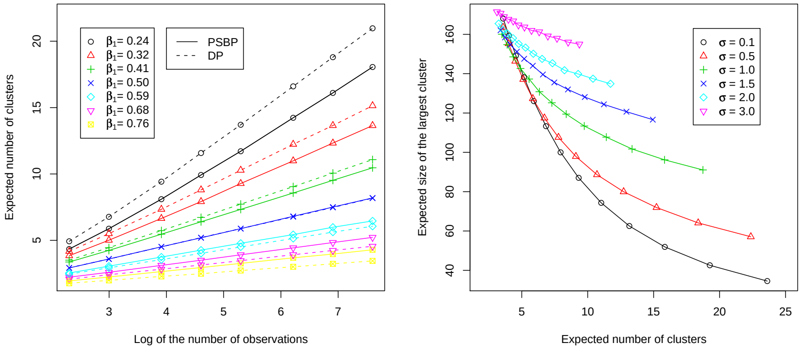
and

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma^2 + 1 & \sigma^2 \\ \sigma^2 & \sigma^2 + 1 \end{pmatrix} \right)$$

Properties of the PSBP



Properties of the PSBP



Properties of the PSBP

- As the truncated version of the DP, a truncated PSBP converges to the infinite version when $N \to \infty$.
- The PSBP implies a Pólya urn representation, but the expressions are very strightforward. For example, if $\theta_1, \theta_2 \sim G$ and $G \sim \mathrm{PSBP}(\mu, \sigma, H)$ then

$$\theta_1 \sim H \qquad \theta_2|\theta_1 \sim \frac{\beta_2}{2\beta_1 - \beta_2}\delta_{\theta_1} + \frac{2\beta_1 - 2\beta_2}{2\beta_1 - \beta_2}H$$

  (this is because $\Pr\{\theta_1 = \theta_2\} = \sum_{k=1}^{\infty} \omega_k^2$).
- General expressions can be obtained using results in Pitman (1995). This could be an interesting short project!!!!

Why probits?

- In principle, the probit prior on the stick-breaking ratios is almost as arbitrary as the betas associated with the DP.
- One advantage of probits is that we can simplify computation by introducing latent random variables.
- To be specific, consider the PSBP mixture

$$y_i|\theta_i \sim \psi(y_i|\theta_i) \quad \theta_i|G \sim G \quad G = \sum_{k=1}^{\infty} \left\{ u_k \left(\prod_{l<k} \{1 - u_l\} \right) \right\} \delta_{\vartheta_k}$$

Why probits?

- Remember the slice sampler $\Rightarrow$ Rewrite the model as $y_i | \{\xi_i\}, \{\vartheta_k\} \sim \psi(y_i | \vartheta_{\xi_i})$, with $z_{i,k} | \alpha_k \sim \mathrm{N}(\alpha_k, 1)$ and $\xi_i | \{z_{i,k}\}$ is deterministically given by $\xi_i = k$ iif $z_{i,l} < 0$ for $l < k$ and $z_{i,k} \geq 0$.
- If the $z_{i,k}$s and the ξ_is are integrated out, we recover the hierarchical formulation in the previous page. Close links to the continuation-ratio probit models of (Agresti, 1990; Chib and Hamilton, 2002).
- Conditionally on the $z_{i,k}$s, it is easy to sample the α_ks (under a normal prior for μ!!!).
- Conditionally on the α_ks and the ξ_is, the $z_{i,k}$s are just truncated normals.
- We can either implement the sampler using truncations or slice samplers.

Dependent PSBPs

- Models with constant atoms. For $s \in \mathcal{S}$ define

$$y_j(s)|\theta_j(s) \sim \psi(\cdot|\theta_j(s)) \quad \theta_j(s) \sim G_s \quad G_s(\cdot) = \sum_{k=1}^{\infty} \omega_k(s)\delta_{\vartheta_k}(\cdot)$$

where $w_k(s) = \Phi(\alpha_k(s)) \prod_{l<k}\{1 - \Phi(\alpha_l(s))\}$ and $\{\alpha_l(s) : s \in \mathcal{S}\}_{l=1}^{\infty}$ are stochastic processes with Gaussian margins.

- This implies

$$\begin{aligned} \mathrm{E}\{G_s(B)\} &= H(B) \\ \mathrm{Var}\{G_s(B)\} &= H(B)\{1 - H(B)\}\left\{\frac{\beta_2(s)}{2\beta_1(s) - \beta_2(s)}\right\} \end{aligned}$$

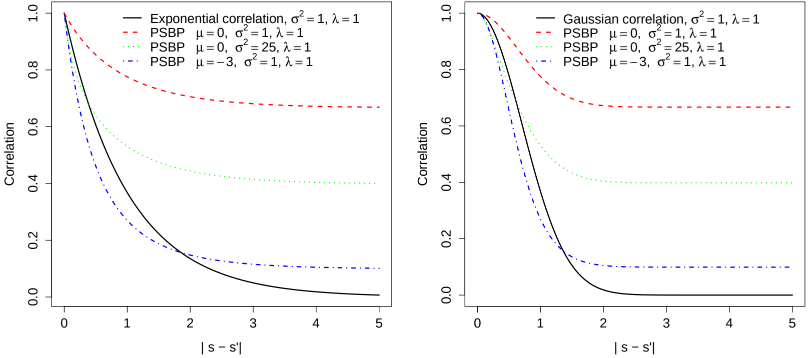
Properties of dependent PSBPs

- Covariance

$$\text{Cov}(G_s(B), G_{s'}(B)) = \frac{\beta_2(s,s')}{\beta_1(s) + \beta_1(s') - \beta_2(s,s')} H(B)\{1 - H(B)\}$$

- If the processes $\{\alpha_l(s) : s \in S\}_{l=1}^{\infty}$ are second-order stationary, then the same can be said for $G_s(B)$.
- As $s' \to s$ then $\text{Cov}(G_s(B), G_{s'}(B)) \to \text{Var}(G_s(B))$ and therefore $\text{Cor}(G_s(B), G_{s'}(B)) \to 1$.

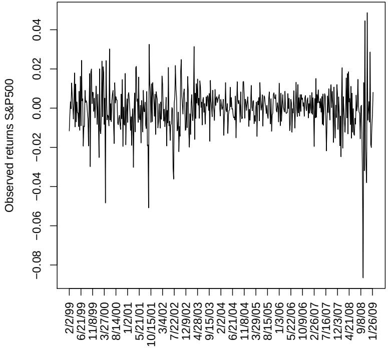
Properties of dependent PSBPs

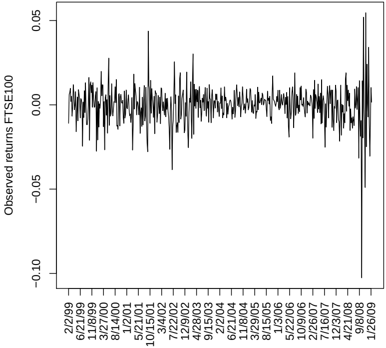


Properties of dependent PSBPs

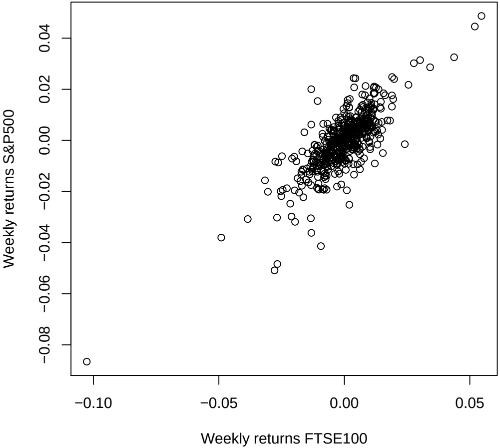
- We can augment the model as we did with the single-distribution case $\Rightarrow$ We can exploit normality to build very rich models for collections of distributions.
  - ANOVA PSBPs.
  - Spatial PSBPs.
  - DLM PSBPs.
  - Random effects PSBPs.
  - Factor models for distributions.
- Unlike the single-p models, we can implement these approaches with discrete data (e.g. binary and count data) very easily.

A stochastic volatility model based on the PSBP





A stochastic volatility model based on the PSBP



A stochastic volatility model based on the PSBP

- Let r_t^i be the return of index i at time t. The model is

$$\begin{pmatrix} r_t^1 \\ r_t^2 \end{pmatrix} \sim \mathsf{N}(\mu_t, \Sigma_t) \quad (\mu_t, \Sigma_t) \sim G_t \quad G_t = \sum_{k=1}^{K} \omega_{t,k} \delta_{(\mu_k^*, \Sigma_k^*)}$$

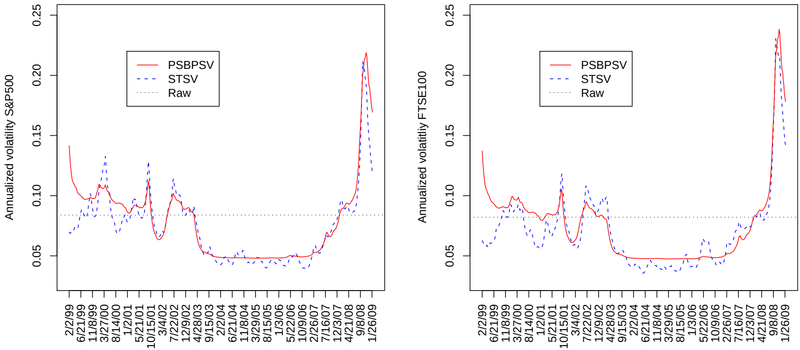
where

$$(\mu_k^*, \Sigma_k^*) \sim \mathsf{NIW}(\mu_0, \kappa_0, \nu_0, \Sigma_0)$$
$$\omega_{t,k} = \Phi(\alpha_{t,k}) \prod_{l<k} \{1 - \Phi(\alpha_{t,l})\}$$

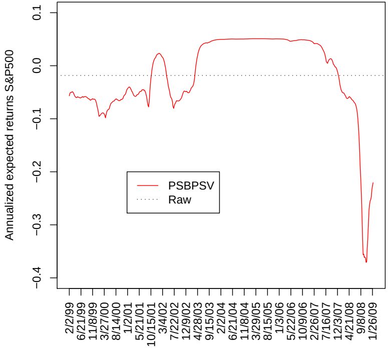
and $\alpha_{t,k} \sim \mathsf{N}(\alpha_{t-1,k}, \tau^2)$, $\alpha_{0,k} \sim \mathsf{N}(\mu, \sigma^2)$.

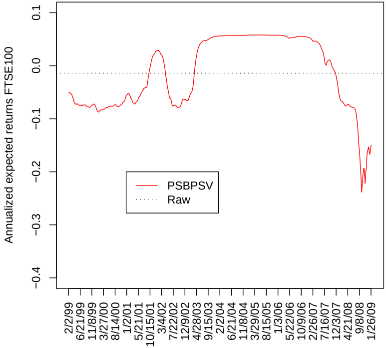
- Sampling using truncations.
- Missing values (market closed) can be imputed if we assume missingness at random.

A stochastic volatility model based on the PSBP

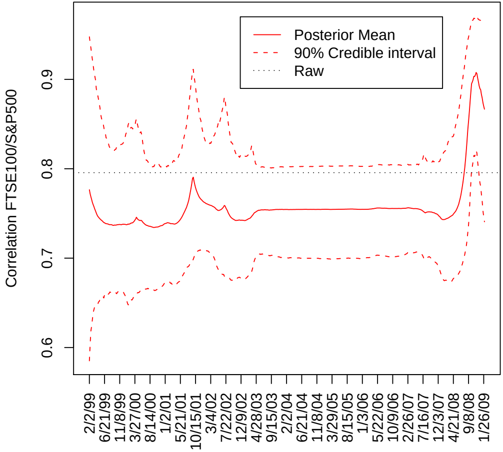


A stochastic volatility model based on the PSBP





A stochastic volatility model based on the PSBP



A possible project: nonparametric dynamic factor models

- Standard factor model:

$$y_i \sim \mathrm{N}(\Lambda\eta_i, \Psi) \qquad \eta_i \sim \mathrm{N}(0, I)$$

where Λ and Ψ are unknown.

- Extend the work of Dunson (200?) to a dynamic, nonparametric version

$$y_t \sim \mathrm{N}(\Lambda\eta_t, \Psi) \qquad \eta_t \sim G_t \qquad G_t = \sum_{k=1}^{K} \omega_{t,k}\delta_{\theta_k}$$

- The specific structure of the $\omega_{t,k}$s depends on the application, but might include periodic and/or AR components.

Some possible topics for the last lecture

- Alternative nonparametric priors for density estimation (Poisson-Dirichlet processes, normalized random meassures).
- Species sampling models.
- Nonparametric regression through density estimation.
- More on models for dependent collections of distributions.
- Contingency tables and generalized link functions.