Supplemental Materials for "Game Changers: Detecting Shifts in Overdispersed Count Data"*

Matthew Blackwell†

August 14, 2017

1 Markov Chain Monte Carlo Estimation Strategy

1.1 Priors and hyperparameters

The complete model requires proper priors on all parameters following Frühwirth-Schnatter et al. (2009) and Fox et al. (2011). It is possible to fix the values of the hyperparameters, α, κ, and γ. Following, Fox et al. (2011), I put diffuse priors on these parameters to allow the data to partially determine their value. It is easier to work with transformations of these parameters, $(\alpha + \kappa)$, and $\theta = \kappa/(\alpha + \kappa)$. With these in hand, I use the following independent priors:

$$(\alpha + \kappa) \sim \text{Ga}(1, 0.1); \tag{1}$$

$$\gamma \sim \text{Ga}(1, 0.1); \tag{2}$$

$$\theta \sim \text{Beta}(100, 1); \tag{3}$$

$$\rho_k \propto \rho_k(\rho_k + 10)^{-4}; \tag{4}$$

$$\beta_k \sim \mathcal{N}(0, 25). \tag{5}$$

*To appear online.

†Assistant Professor, Department of Government, Harvard University, Institute for Quantitative Social Science, 1737 Cambridge Street, Cambridge, MA 02138 (email: mblackwell@gov.harvard.edu, web: mattblackwell.org)

1.2 Block sampling the latent regimes and HDP parameters

To draw the latent states and the HDP parameters, I use the blocked sampler from Fox et al. (2011). Sampling for the HDP parameters, $\boldsymbol{\delta}$, α, κ, and γ are complicated and require a heavy notational burden, so I refer the interested reader to Appendix D and Appendix E of Fox et al. (2011). To draw the latent states, Fox et al. (2011) rely on a forward-backward procedure similar to the algorithm of Chib (1998). Note that we can write the full conditional posterior of $\mathbf{s}$ as

$$p(s_T|s_{T-1}, \mathbf{y}, \Theta, \boldsymbol{\pi}) \times p(s_{T-1}|s_{T-1}\mathbf{y}, \Theta, \boldsymbol{\pi}) \times \cdots \times p(s_t|s_{t-1}, \mathbf{y}, \Theta, \boldsymbol{\pi}) \times \cdots \times p(s_1|\mathbf{y}, \Theta, \boldsymbol{\pi}), \tag{6}$$

where $\Theta = (\boldsymbol{\beta}, \boldsymbol{\rho}, \boldsymbol{\eta})$ is the collection of the model parameters. From this derivation, we can see that we can sample s_1 from its full posterior, then sample s_2 conditional on that value of s_1, and so on. To calculate the form of these distribution, however, requires the calculation of a series of "messages" passed from s_t to s_{t-1}. These messages are defined recursively as:

$$m_{t,t-1}(s_{t-1}) \propto \begin{cases} \sum_{s_t} p(s_t|\pi_{s_{t-1}}) p(y_t|\beta_{s_t}, \eta_t) p(\eta_t|\rho_{s_t}) m_{t+1,t}(s_t), & t \leq T; \\ 1, & t = T+1; \end{cases} \tag{7}$$

$$\propto p(y_{t:T}|s_{t-1}, \Theta, \boldsymbol{\pi}) \tag{8}$$

With these messages in hand, we can calculate distributions above as:

$$p(s_t|s_{t-1}, \mathbf{y}, \Theta, \boldsymbol{\pi}) \propto p(s_t|\boldsymbol{\pi}_{s_{t-1}}) p(y_t|\beta_{s_t}, \eta_t) p(\eta_t|\rho_{s_t}) m_{t+1,t}(s_t). \tag{9}$$

With these states in hand, it is straightforward to draw the transition probabilities as a function of the priors and the number of transitions observed in $\mathbf{s}$. That is, I draw

$$\boldsymbol{\pi}_j|\mathbf{s}, \alpha, \kappa, \boldsymbol{\delta} \sim \text{Dirichlet}(\alpha\delta_1 + n_{j1}, \ldots, \alpha\delta_j + \kappa + n_{jj}, \ldots, \alpha\delta_K + n_{jK})$$

for $j = 1, \ldots, K$. Here, n_{jk} is the number of times subsequences in $\mathbf{s}$ with $s_{t-1} = j$ and $s_t = k$.

1.3 Drawing the model parameters

Now that we have draws of the latent states, we need to take draws of the model parameters in each regime (β_k, ρ_k). The non-linear nature of the distributions involved eliminate the possibility of closed-form posterior distributions. This makes

the straightforward application of Gibbs sampling impossible. To avoid the inefficiencies of other MCMC approaches, I draw on the auxiliary mixture sampling approach of Frühwirth-Schnatter et al. (2009). This approach augments the data with a set of latent variables τ_{t1} and τ_{t2} which contain all the distributional information about the outcome y and whose distribution can be approximated by a mixture of Normals. With draws of $\tau_t = (\tau_{t1}, \tau_{t2})$ and mixture component indicators $r_t = (r_{t1}, r_{t2})$, we can turn this non-linear problem into a linear Gaussian regression problem. That is, conditional on τ_t, r_t, and η_t, posterior inference on the β_k is simply a Bayesian linear regression. I block sample the negative binomial parameters ρ_k and η_t, using slice sampling to draw ρ_k conditional on $\mathbf{y}$ and $\boldsymbol{\beta}$. With these in hand, ν_t is distributed Gamma with shape $\rho_{s_t} + y_t$ and scale $\rho_{s_t} + \exp(X_t \beta_{s_t})$.

Putting all of these steps together, we have the following draws for a single iteration of the MCMC algorithm:

1. Draw $\mathbf{s}|\mathbf{y}, \Theta, \boldsymbol{\pi}$ as described above.
2. Draw $(\boldsymbol{\rho}, \boldsymbol{\eta})|\mathbf{y}, \boldsymbol{\beta}, \mathbf{s}$:
    a) Draw $\rho_k|\mathbf{y}, \boldsymbol{\beta}$ unconditional on $\boldsymbol{\eta}$ using slice sampling (Neal, 2003).
    b) Draw $\eta_t|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\rho}, \mathbf{s} \sim \text{Gamma}\left(\rho_{s_t} + y_t, \rho_{s_t} + \exp(X_t \beta_{s_t})\right)$, for $t = 1, \ldots, T$.
3. Sample $\boldsymbol{\tau}, \mathbf{r}|\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\eta}, \boldsymbol{\rho}$ using the auxiliary mixture approach Frühwirth-Schnatter et al. (2009).
4. Draw from $\boldsymbol{\beta}|\boldsymbol{\tau}, \mathbf{r}, \boldsymbol{\eta}$ using the auxiliary mixture approach of Frühwirth-Schnatter et al. (2009).
5. Draw $\boldsymbol{\pi}_j|\mathbf{s}, \alpha, \kappa, \boldsymbol{\delta}$ from $\text{Dirichlet}(\alpha\delta_1 + n_{j1}, \ldots, \alpha\delta_j + \kappa + n_{jj}, \ldots, \alpha\delta_K + n_{jK})$, for $j = 1, \ldots, K$.
6. Draw β, $\alpha + \kappa$, θ and γ as in Fox et al. (2011).

One issue with this approach is that assessing convergence is a difficult process due to the number of parameters and the label-switching between draws of the sampler. It is possible to avoid these issues by developing a variational approximation approach to estimating such a model (Jordan et al., 1999). The benefit of this would be to side-step the issue of convergence since it is both guaranteed and is easy to assess. Furthermore, it would also alleviate the label-switching issue because it would find a single "canonical" labeling as the estimate. A downside to this

approach is that some common approaches to variational inference will underestimate the posterior variance relative to MCMC approaches (Grimmer, 2011). Despite this, developing a variational approximation to this model would be a valuable topic for future research.

1.4 A simulation study

To show how the present model compares to other approaches, I apply it to simulated datasets from three different data generating processes, one with no covariates (unconditional) and two conditional models, one with easily detectable changepoints (high power) and one with harder to detect changepoints (low power).[1] In the unconditional model, there are $T = 200$ observations with four regimes with 50 observations each. I simulated the data in each regime with a simple intercept, so that $\boldsymbol{\beta} = (6, 3, 6, 3)$ and with overdispersion parameters $\boldsymbol{\rho} = (1.5, 0.5, 3, 1.5)$. In the conditional models, there is one covariate distributed Unif$(0, 2)$ in the high-powered scenario and Unif$(0, 0.5)$ in the low-powered scenario, with three regimes with coefficients $\beta_1 = (1, 1)$, $\beta_2 = (1, -2)$, and $\beta_3 = (1, 2)$. Each regime lasts 50 observations and the overdispersion parameters are $\boldsymbol{\rho} = (1.5, 0.5, 3)$. The two conditional models differ in the variance of the covariate, with the higher variance covariate leading to smaller sampling variance for the coefficients in each regime. To investigate the properties of the various models, I draw 100 datasets from each DGP and apply a series of models to each draw from the DGPs:

- the sticky HDP-HMM with a negative binomial outcome distribution,
- a HDP-HSMM with a negative binomial outcome distribution (Johnson and Willsky, 2013),
- a series of fixed-number-of-regime models with negative binomial outcome distributions (varying from 0 to 7 changepoints),
- a left-to-right model (similar to that of Chib (1998) or Park (2010)) with no fixed number of changepoints,
- the sticky HDP-HMM with a Poisson outcome distribution.

The left-to-right model alters the Chib (1998) prior structure in two ways: (1) the prior distribution of the state for the first period is uniform over the possible states,

[1] Thanks to an anonymous reviewer who suggested these simulation settings.

and (2) there is no requirement that all of the states are visited so that the last state is drawn from its posterior rather than fixed to final state. Peluso, Chib, and Mira (2016) proposed a similar prior structure that would maintain the left-to-right nature of the Chib (1998) model without fixing the number of regimes a priori. The first four models compare various ways of allowing the number of changepoints to vary with ergodic (sticky HDP-HMM) and non-erogic (HSMM and the left-to-right) models. The last model fixes the sticky HDP-HMM structure and explores how misspecifying the outcome distribution affects the posterior distribution of the number of changepoints.

For each of these models, I ran the MCMC sampler with an upper bound of 10 states for 5,000 iterations, thinned by 5, with a burin period of 5,000 iterations. In each of the fixed-number-of-regimes models, I also calculated the marginal likelihood of the model using the approach of Chib (1995). This allows me to infer the probability distribution over the number of changepoints, assuming a uniform prior over the models. With these in hand, I created average posterior probabilities from these model across the 100 draws from each DGP. This allows us to see what the average posterior probability of a changepoint is for a given model and what the average posterior probability over the number of changepoints is. Ideally, we would want to see each method selecting the true number of changepoints with high average posterior probability and placing those changepoints near the true values (with some variance).

The average posterior probabilities for changepoint locations for the unconditional model and the high-power conditional model are presented in Figures 1 and 3. Both of these show a similar pattern: the negative binomial models all give results that are close to the true distribution of changepoints, whereas the Poisson sticky HDP-HMM gives a massive number of changeopints spread throughout the data. This result reveals one danger of the HDP-HMM approach: because the model is ergodic, the state variable can move freely between states a fairly fast rate. Usually the sticky version of the HDP-HMM can overcome this and produce more coherent clusters. Here, though, the model misspecification of the variance of the count data leads the model to quickly switch between the large counts and the small counts. Once properly specified with a negative binomial outcome model, the different ways of modeling the number of changepionts (sticky HDP-HMM, HDP-HSMM, marginal likelihood, and open-ended left-to-right) all give very similar answers, with the left-to-right approach having slightly higher variance.

The average posterior distributions over the number of changeopints is given in Figure 2 for the unconditional model and Figure 4 for the high-powered conditional model. These show that all of the models that use the negative binomial out-

come distribution place high posterior probability on the true number of changepoints, with slightly different variances for each method. These differences are to be expected because each method has a different prior structure for the number of changepoints and it is difficult (and sometimes impossible) to encode the exact same priors across models. In spit of this, the negative binomial methods all tend to recover roughly the correct number of changepoints, at least across draws from the DGP. The Poisson sticky HDP-HMM again places its posterior mass on a large number of changeopints, again due to the overdispersed nature of the data.

Note that these results do not mean that the Poisson model is not useful. In separate tests not reported here, the Poisson sticky HDP-HMM recovers correct inferences about the number of changepoints when the data is, in fact, distributed Poisson. The take-away from these results should be that the exact method for allowing arbitrary numbers of changepoints is relatively less important than the correct specification of the model within regimes. If the model is improperly specified, we may overestimate the number of changepoints.

Finally, the low-powered simulations in Figures 5 and 6 show how harder-to-detect changepoints affect these estimators. We can see that all of the methods have much higher average posterior variance over the number of changepoints, though the location of the changepoints appears to be correct when they are identified. One interesting feature of this simulation is that comparing marginal likelihoods gives very different answers than the rest of the models, placing high posterior mass on a model with 0 changepionts. There are a couple of reasons this might be occurring. First, there could be computational problems with the marginal likelihood calculations due to not being able to fully explore the posterior distribution. This appears somewhat plausible because the ML calculations in these models can be very sensitive to the point which the posterior ordinate is being calculated using the Chib (1995) approach. Second, the ML approach might be placing different implicit criteria on the inference via the uniform prior over the number of changepoints. A generalized comparison of model selection via ML and via various Bayesian nonparametric approaches is a good avenue for future research. Of course, these are just three different data generating processes and it may be the case that any of these methods may outperform another in a different scenario.

1.5 Consistency simulations

Miller and Harrison (2014) showed that Dirichlet process mixture models with a fixed concentration parameter and no hierarchical structure is inconsistent for the true number of regimes because those models tend to overestimate the number

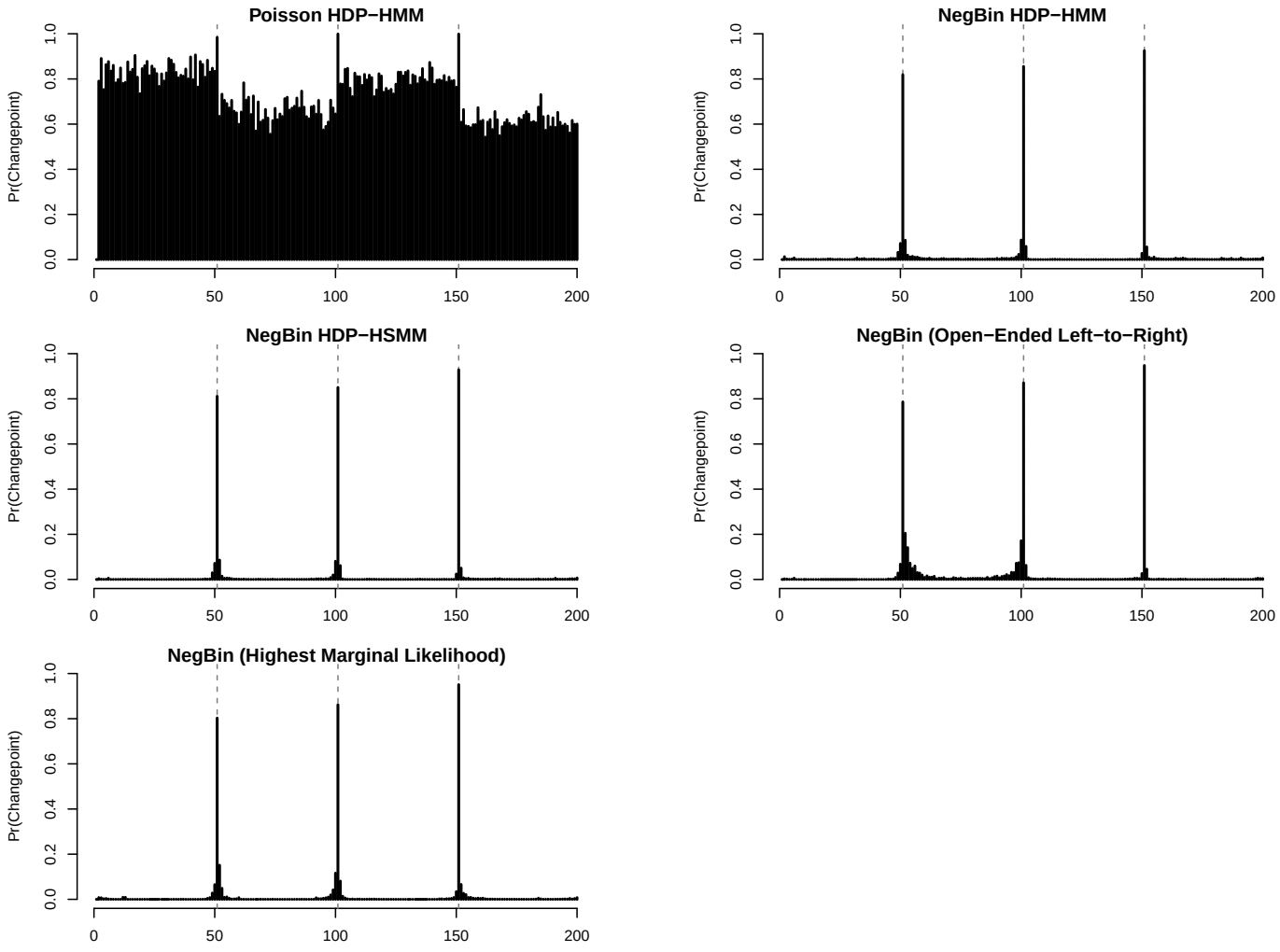


Figure 1: Average posterior probabilities of a changepoint at a given time period for the various changepoint models for the unconditional simulations with true changepoints at $t = 51, 101$ and 151, averaging over 100 draws from the DGP.

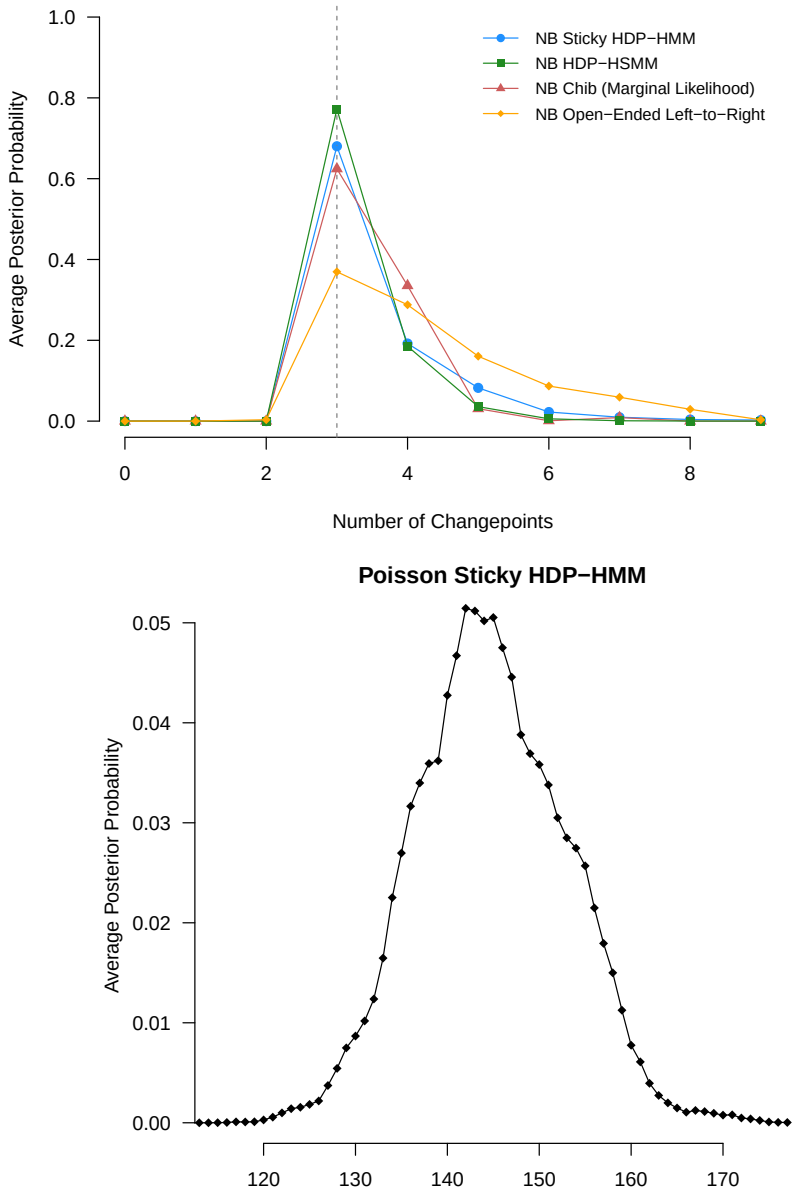


Figure 2: Average posterior probability distributions over the number of changepoints for the unconditional simulations for the methods with a negative binomial outcome distribution (top) and the sticky HDP-HMM with a Poisson outcome distribution. The true number of changepoints is 3.

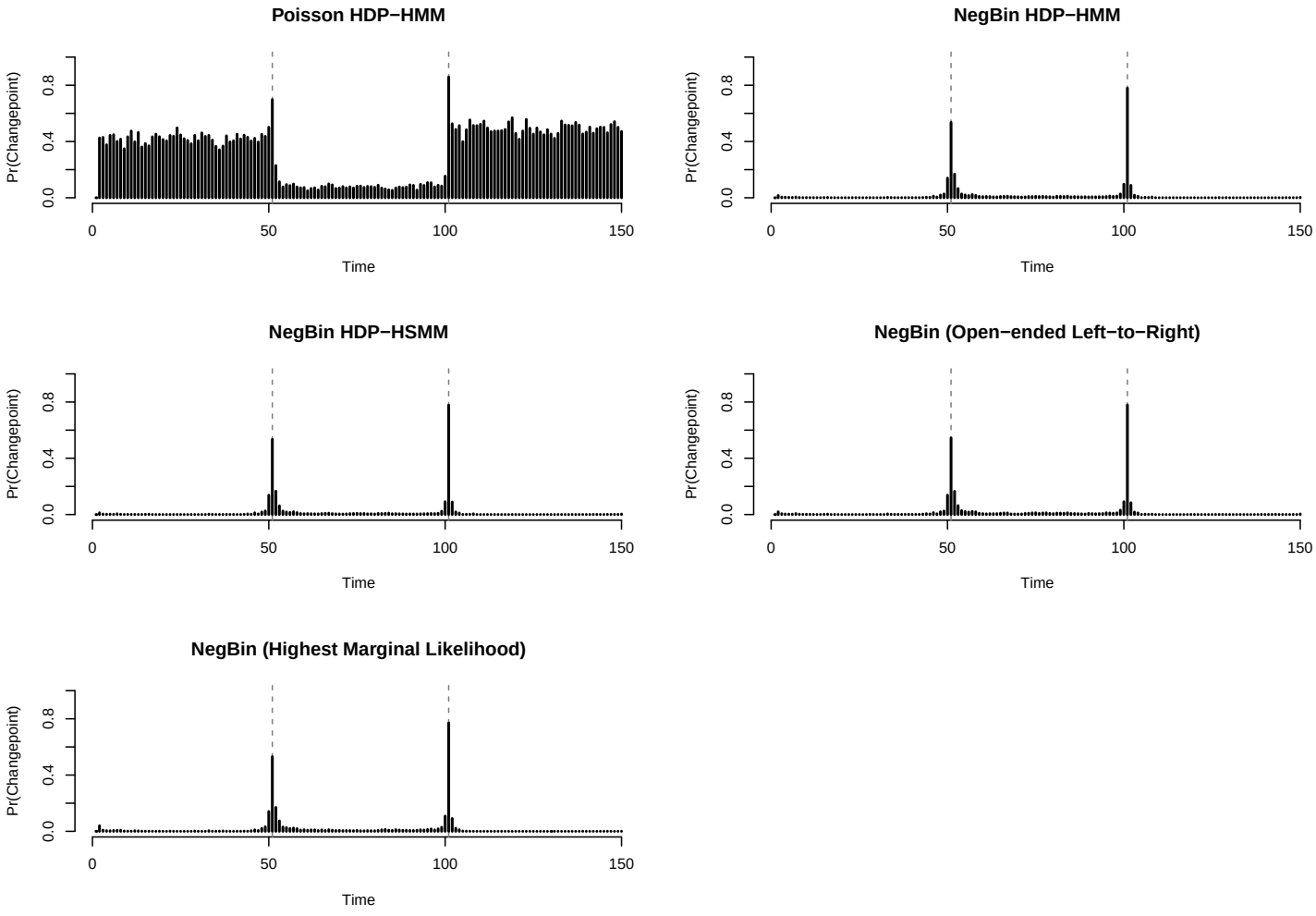


Figure 3: Average posterior probabilities of a changepoint at a given time period for the various changepoint models for the conditional simulations with true changepoints at $t = 51$ and 101, averaging over 100 draws from the DGP.

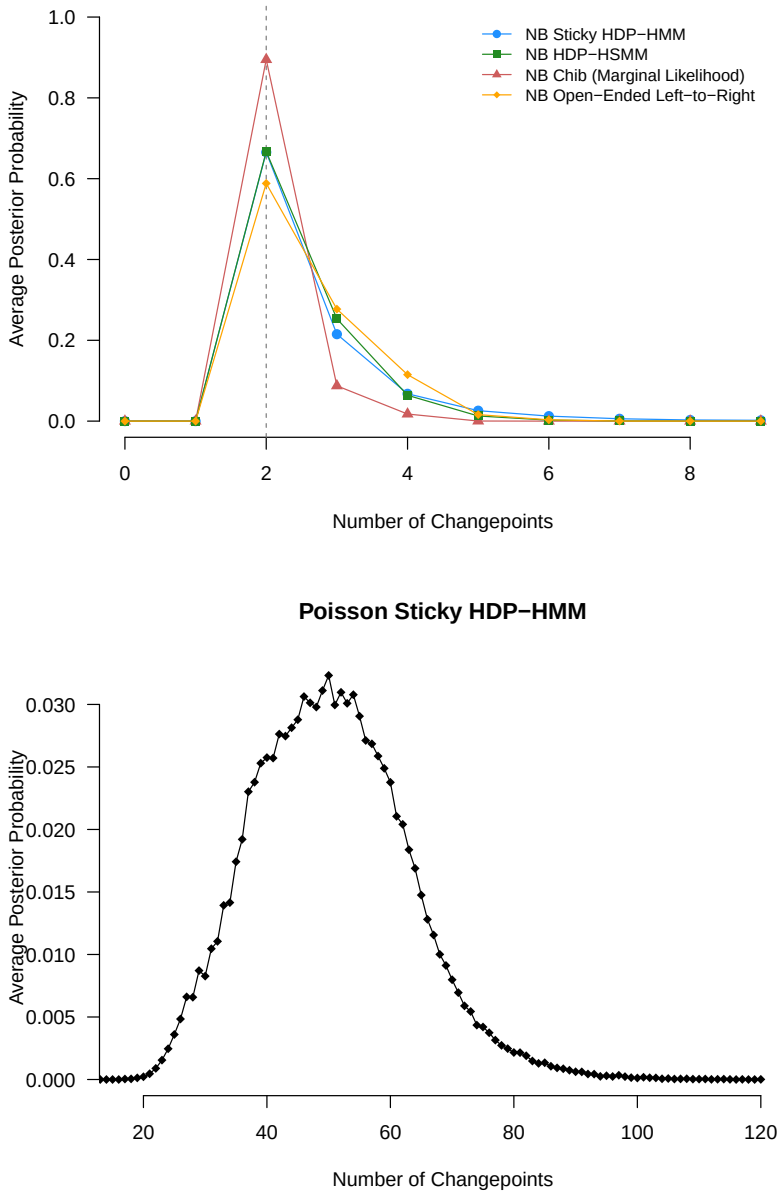


Figure 4: Average posterior probability distributions over the number of changepoints for the conditional simulations for the methods with a negative binomial outcome distribution (top) and the sticky HDP-HMM with a Poisson outcome distribution. The true number of changepoints is 2.

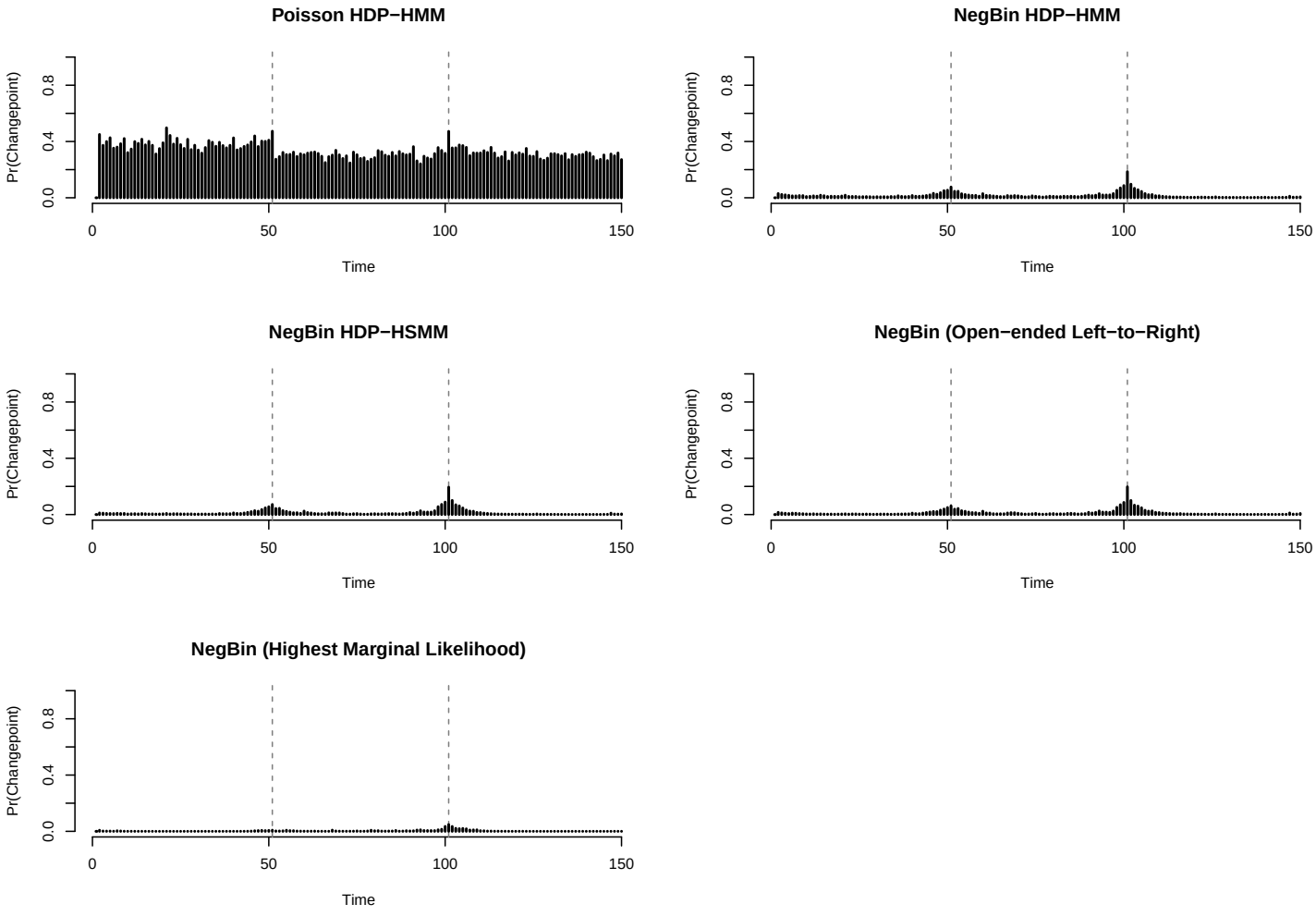


Figure 5: Average posterior probabilities of a changepoint at a given time period for the various changepoint models for the low-power conditional simulations with true changepoints at $t = 51$ and 101, averaging over 100 draws from the DGP.

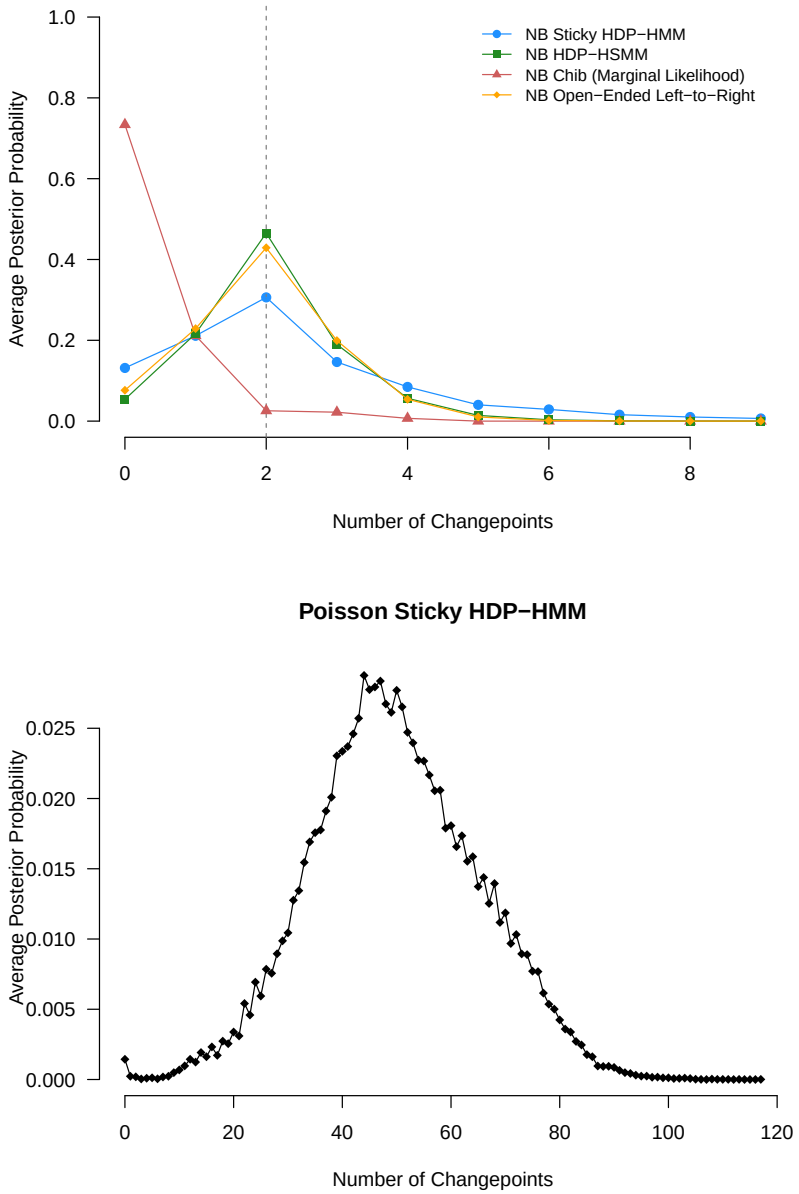
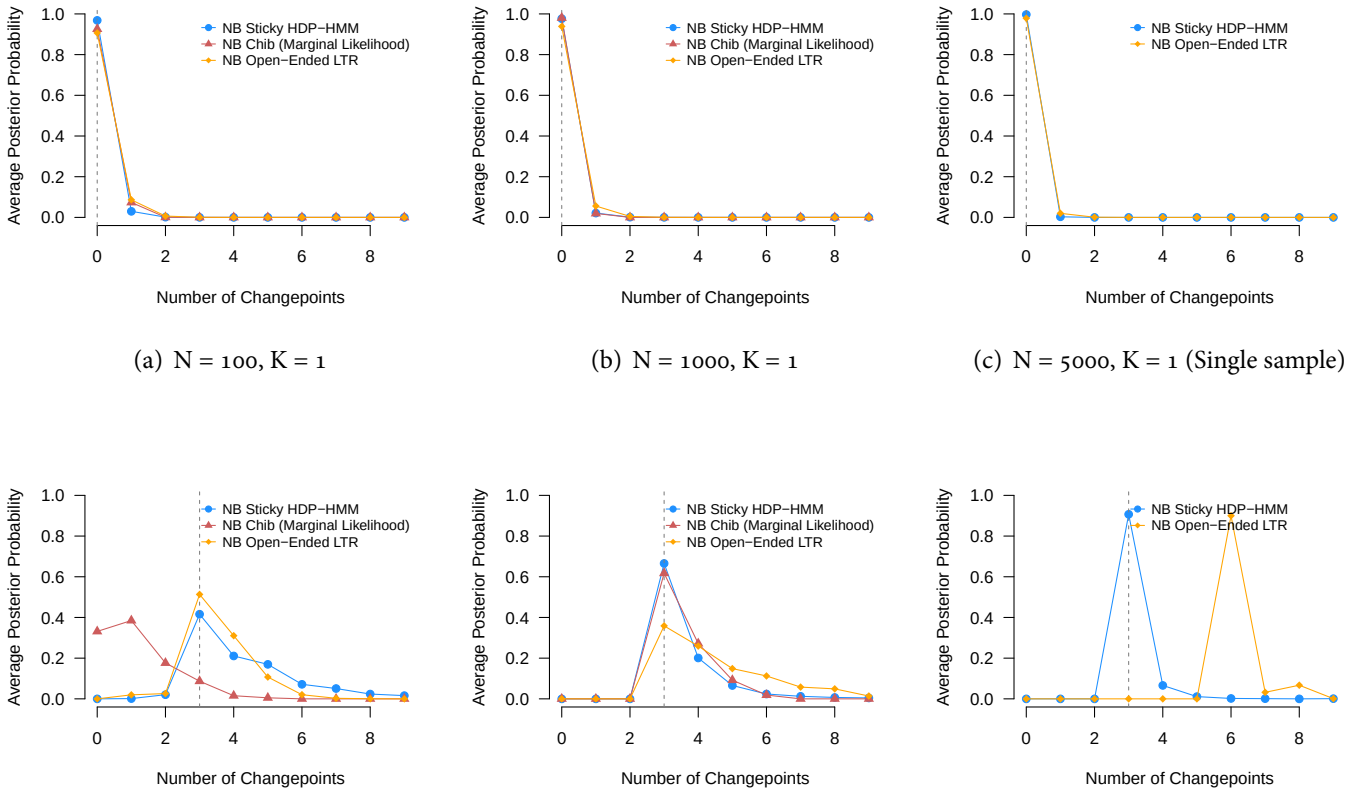


Figure 6: Average posterior probability distributions over the number of changepoints for the low-power conditional simulations for the methods with a negative binomial outcome distribution (top) and the sticky HDP-HMM with a Poisson outcome distribution. The true number of changepoints is 2.

of regimes. To investigate this, we take a simple simulation setup that compares the average posterior probability of draws from data with either no changepoints or three changepoints ($K = 4$) to see if the posteriors fail to concentrate at the true number as they do in the simulation studies of Miller and Harrison (2014). I ran the NB sticky HDP-HMM, the NB open-ended left-to-right model, and the separate NB models with fixed changepoints with marginal likelihood used to calculate posterior probability of the changepoints. The DGP with $K = 1$ that had posterior mean $e^3 \approx 20$ and overdispersion parameter $\rho = 0.5$ and the $K = 4$ model had regime parameters $\boldsymbol{\beta} = (3, 0.5, 4, 0.25)$ and $\boldsymbol{\rho} = (0.5, 1, 0.75, 2)$. For each DGP, I generated three different sample sizes of 100, 1000, and 5000, with the changepoints evenly spaced in the $K = 4$ DGP for each sample size. I drew 5000 MCMC samples after a burnin period of 5000 draws and I thinned the chain by 5. For the $N = 100$ and $N = 1000$ sample sizes, I took 100 draws from each of these DGPs and averaged the posterior probabilities of changepoints over the draws. For the $N = 5000$ sample size, I only ran the NB sticky HDP-HMM and the NB open-ended left-to-right model for a single draw of the above DGP due to the computational burden of the data at this size. Unfortunately, the HDP-HSMM is very computationally inefficient as the sample size grows unless one puts a cap on the maximum length of a regime, which would obviously induce the incorrect posterior on the number of regimes.

The results in Figure 7 show that, at least for the sticky HDP-HMM, posterior probability is overwhelmingly on finding the correct number of changepoints in these situations and there is relatively little change as the sample size increases. Thus, it appears as though the results Miller and Harrison (2014) do not necessarily apply to either of these model. One reason that consistency appears to hold in this case may be that, in the sticky HDP-HMM, the concentration parameters are being estimated from the data, rather than being set a priori as in the Miller and Harrison (2014) setting. Another explanation might be that there is some very small, though non-zero, posterior mass being left on situations with the incorrect number of changepoints. This would be consistent with the results of Miller and Harrison (2014) but indicate that the non-consistency results have less practical relevance in this setting. Of course, this is one DGP and there may be other cases where more serious issues do arise. We do see an inconsistency with the open-ended left-to-right model, which may be a model that is closer to the type that Miller and Harrison (2014) discuss. This would be a fruitful avenue for future research since the simulations here are somewhat limited by the $N = 5000$ case



(a) N = 100, K = 1 (b) N = 1000, K = 1 (c) N = 5000, K = 1 (Single sample)

(d) N = 100, K = 4 (e) N = 1000, K = 4 (f) N = 5000, K = 4 (Single sample)

Figure 7: Average posterior probability distributions over the number of changepoints sample sizes 100, 1000, and 5000 and with either 0 or 3 changepoints. For the $N = 100$ and $N = 1000$ sample sizes, posteriors are averaged across 100 draws from the DGP. For the $N = 5000$ sample size, the posteriors are for a single draw of the DGP.

being a single draw from the DGP.

References

Chib, Siddhartha (1995). "Marginal Likelihood from the Gibbs Output." *Journal of the American Statistical Association* 90.432, (cit. on pp. 5, 6).

— (1998). "Estimation and comparison of multiple change-point models." *Journal of Econometrics* 86.2, pp. 221–241 (cit. on pp. 2, 4, 5).

Fox, Emily B et al. (2011). "A sticky HDP-HMM with application to speaker diarization." *The Annals of Applied Statistics* 5.2A, pp. 1020–1056 (cit. on pp. 1–3).

Frühwirth-Schnatter, Sylvia et al. (2009). "Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data." *Statistics and Computing* 19.4, pp. 479–492 (cit. on pp. 1, 3).

Grimmer, Justin (2011). "An Introduction to Bayesian Inference via Variational Approximations." *Political Analysis* 19.1, pp. 32–47. eprint: http://pan.oxfordjournals.org/content/19/1/32.full.pdf+html (cit. on p. 4).

Johnson, Matthew J and Alan S Willsky (2013). "Bayesian Nonparametric Hidden Semi-Markov Models." *Journal of Machine Learning Research* 14.Feb, pp. 673–701 (cit. on p. 4).

Jordan, Michael et al. (1999). "An introduction to variational methods for graphical models." *Machine Learning* 37, pp. 183–233 (cit. on p. 3).

Miller, J W and M T Harrison (2014). "Inconsistency of Pitman-Yor process mixtures for the number of components." *Journal of Machine Learning Research* 15, pp. 3333–33370 (cit. on pp. 6, 13).

Neal, Radford M (2003). "Slice sampling." *The Annals of Statistics* 31.3, pp. 705–767 (cit. on p. 3).

Park, Jong Hee (2010). "Structural Change in U.S. Presidents' Use of Force." *American Journal of Political Science* 54.3, pp. 766–782 (cit. on p. 4).

Peluso, Stefano, Siddhartha Chib, and Antonietta Mira (2016). "Semiparametric Multivariate and Multiple Change-Point Modelling." *Working Paper* (cit. on p. 5).