Advances on Discrete Spike-and-Slab Priors for Variable Selection

Marina Vannucci

Department of Statistics
Rice University
Houston, TX
USA

ISBA 2022, Montreal, CANADA

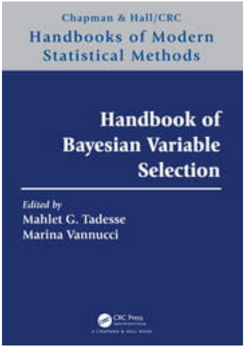
Outline of the Talk

- Handbook of Bayesian Variable Selection
- Variable selection via *spike-and-slab* priors
  - Discrete vs continuous constructions
  - Bayesian hidden Markov models with variable selection for seizure risk assessment

Handbook of Bayesian Variable Selection

Edited by Mahlet G. Tadesse and Marina Vannucci
Published December 20, 2021 by Chapman and Hall/CRC

- Comprehensive review of theoretical, methodological and computational aspects of BVS
- Divided into four parts: *Spike-and-Slab Priors*; *Continuous Shrinkage Priors*; *Extensions to various Modeling* (causal inference, state-space models, edge selection in graphical models); *Other Approaches to BVS* (Bayes factors, decision trees, partition models)
- Contributions by experts in the field



Chapman & Hall/CRC
Handbooks of Modern Statistical Methods

Handbook of Bayesian Variable Selection

Edited by
Mahlet G. Tadesse
Marina Vannucci

CRC Press

Available on Taylor & Francis eBooks
Buy it on Amazon

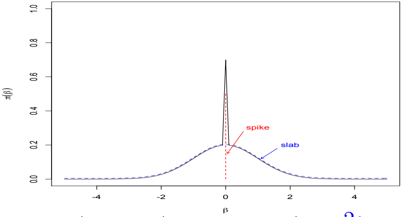
Spike-and-slab Variable Selection Priors

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I)$$

Introduce latent indicators $\gamma = (\gamma_1, \ldots, \gamma_p)'$

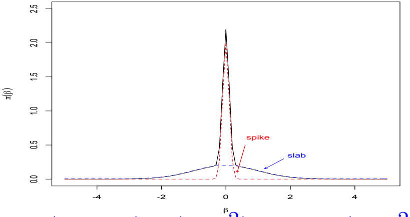
$$\begin{cases} \gamma_j = 1 & \text{if variable } j \text{ included in model} \\ \gamma_j = 0 & \text{otherwise} \end{cases}$$

Discrete *Spike-and-Slab*



$$\beta_j \sim (1-\gamma_j)\delta_0 + \gamma_j N(0, \sigma_\beta^2),$$

Continuous *Spike-and-Slab*



$$\beta_j \sim (1-\gamma_j) N(0, \sigma_0^2) + \gamma_j N(0, \sigma_1^2)$$

Notes on misconceptions

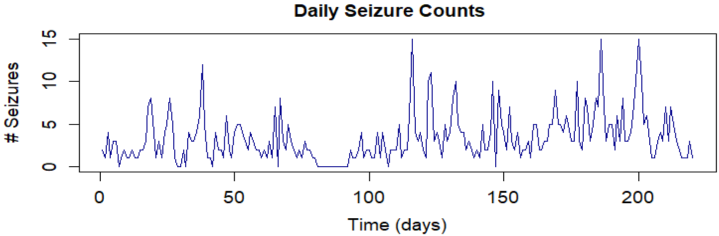
- Computational aspects:
    - With conjugate priors we can marginalize the $\beta's$ out
    - Incorrect to believe that discrete SS need *reversible jump* in non-conjugate (or non-Gaussian) settings
    - Originally used in George & McCulloch (1993, 1997)
    - Gottardo & Raftery (2008 JCGS) formulate reversible jump as a mixture of singular distributions.
    - Sample (β, γ) jointly, as in Savitsky et al. (2011, Stat Science) with standard Metropolis.
    - Can handle non-conjugate and non-Gaussian settings (via DA)
    - Handbook ch.1 (Vannucci) and ch.5 (Griffin & Steel)

- Theoretical aspects:
  - Continuous SS priors are more amenable to theoretical developments - Handbook ch.3 (Narisetty) & ch.4 (Bai et al.)
  - Results are now also available for the discrete SS in terms of optimal support recovery, posterior contraction rate and consistent variable selection (Castillo *et al.* 2015, AoS)
  - Handbook ch. 2 (Zhou & Pati)
- Applications:
  - Both extend to various modeling settings– Handbook Part III - ch.9-14
  - Both scalable (EM; VB) - Handbook ch. 1,2 and 4
  - Continuous SS priors have two variance parameters to tune

Next: Application of discrete SS to HMM for count data

Motivating application: Assessing Seizure Risk

- 60 million people (1% of the population) have epilepsy
- Seizures unpredictable and severely affect patients' quality of life.
- Electronic dairies: $Y_{it} \equiv$ daily seizure counts; $X_{it} \equiv$ time-varying covariates, $i = 1, \ldots, N,\ t = 1, \cdots, T_i$



- Main goals:
  - Estimate underlying seizure risk at subject level.
  - Identify risk factors contributing to seizure risk.

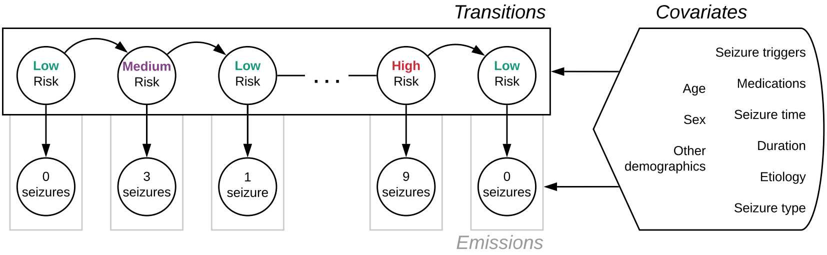
Existing approaches

- Currently, clinical decision-making heavily depends on raw seizure counts and decisions about treatment primarily on increase/decrease of seizure frequency after intervention.

- Recent notion that seizures are a stochastic realization of periods of heightened seizure risk (Goldenholz et al., 2018) and that raw seizure counts are only a surrogate measure of a patient's true seizure risk.

- In Chiang et al. (2018, *Epilepsia Open*) we developed a hidden Markov model (HMM) to provide a probabilistic estimation of discrete seizure risk (assumed Poisson observations; monthly granularity). Validated against specialized epilepsy clinician experts (Chiang et al., 2020 *Epilepsia*).

A Bayesian HMM for Assessing Seizure Risk



- Negative binomial emissions allow overdispersion and daily granularity.
- Incorporate covariates and identify risk factors contributing to seizure risk.
- Bayesian framework.
- Spike-and-slab for variable selection.

Hidden Markov model

Given data, $Y_{it} \equiv$ number of seizure of patient i at time t.

Let ξ_{it} be the latent risk state of patient i at time t

- **Transitions:** Multinomial logit-link

$$Pr(\xi_{it} \mid \xi_{i,(t-1)}, \dots, \xi_{i1}) = Pr(\xi_{it} \mid \xi_{i,(t-1)}) \quad (Markovian)$$

$$Pr(\xi_{it} = k | \xi_{i,t-1} = k') = \frac{exp(\boldsymbol{X}_{i,t-1}^T \boldsymbol{\beta}_{k'k})}{1 + \sum_{l=1}^{K-1} exp(\boldsymbol{X}_{i,t-1}^T \boldsymbol{\beta}_{k'l})}$$

- VS on $\boldsymbol{\beta}_{k'}$ determines covariates associated with worsening or improvement of seizure risk.
- Closed-form updates for $\boldsymbol{\beta}_k$ via Polya-Gamma data augmentation (Polson et al. (2013)).

- **Emissions:** zero-inflated Negative binomial distribution

$$[Y_{it} \mid \xi_{it} = k] \sim ZINB(r_k, \psi_{itk}, p_k)$$

Negative binomial allows for overdispersion ($\sigma^2 > \mu$)

Zero-inflated NB tailored towards data with excess zeros

Mixture of a NB and a point mass at zero:

$$[Y_{it} \mid r, \psi, p] \ \sim \ p \cdot 1_{\{Y_{it}=0\}} + (1-p) \cdot NB(r, \psi)$$

- Reparametrize the NB with dispersion r_k and subject- and state-dependent success probability ψ_{itk}

$$\psi_{\mathrm{itk}} = \frac{exp(\boldsymbol{X}_{it}^T \boldsymbol{\rho}_k)}{1 + exp(\boldsymbol{X}_{it}^T \boldsymbol{\rho}_k)}$$

  with ρ_k a state-dependent vector of regression coefficients.

- Closed-form updates for state-dependent ρ_k via Polya-Gamma data augmentation.

- Mean parameters can be recovered as $\mu_{itk} = \frac{\psi_{itk} r_k}{1-\psi_{itk}}$.

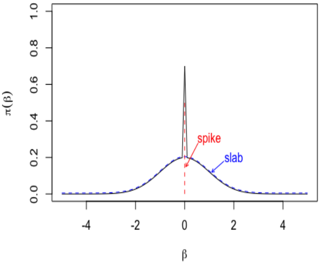
- VS on $\boldsymbol{\rho}_k$ determines covariates associated with increases or decreases in seizure frequency, conditional on the latent risk state.

Variable selection priors

- Spike-and-slab variable selection priors on regression coefficients (George & McCulloch (1993,1997); Brown *et al.* (1998 & 2002))

$$[\beta_{j,k'k} \mid -] \sim \gamma_{j,k'k} N(\mu_\beta, \sigma^2_\beta) + (1 - \gamma_{j,k'k})\delta_0(\beta_{j,k'k}),$$
$$[\rho_{jk} \mid -] \sim \delta_{jk} N(\mu_\rho, \sigma^2_\rho) + (1 - \delta_{jk})\delta_0(\rho_{jk}),$$

$\boldsymbol{\gamma}$, $\boldsymbol{\delta}$: inclusion indicators
(Bernoulli prior)
(1 for important covariate, 0 if not)



MCMC Algorithm

Gibbs sampler:

1. Joint update of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$ via stochastic search with add/delete/swap combined with Pólya-Gamma data augmentation.
2. Update hidden states $\boldsymbol{\xi}$ via Forward-Backward algorithm.
3. Joint update of $(\boldsymbol{\rho}, \boldsymbol{\delta})$, similarly to the update of $(\boldsymbol{\beta}, \boldsymbol{\gamma})$.
4. Update overdispersion $\mathbf{r}$ via data augmentation.
5. Update zero-inflation $\mathbf{p}$ from the full conditionals.
6. Update other auxiliary variables.

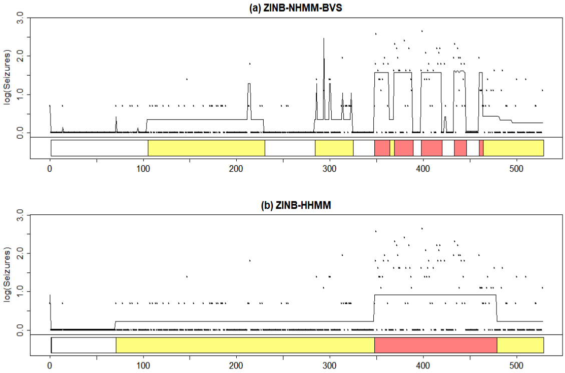
Case Study on Dravet Syndrome

Daily seizure counts from SeizureTracker - electronic seizure diary with over 2 million seizures logged by >30,000 patients since 2006.

- $n = 133$ patients with Dravet syndrome, with ages between 2 months and 47 years.
- $34,431$ generalized tonic-clonic seizures (GTCs) recorded by these patients between 2007-2020, spanning over $141,499$ person-days.
- $p = 37$ covariates including 23 classes of medications, 10 common seizure triggers, and 4 other patient characteristics.
- Prior specification as in simulations
- Optimal number of states, K, chosen based on deviance information criterion (DIC) over a grid of possible values
- $K = 3$ (low, moderate and high risk)

Clinical findings

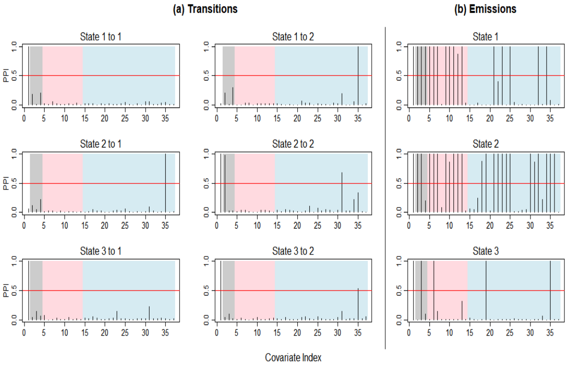
Model produces subject-specific estimated states ξ_{it} and subject- and state-specific estimates of expected number of seizures μ_{itk}.



Accounting for external modulatory factors improves accuracy of the estimates for seizure risk states.

Clinical findings

Thresholding marginal PPIs of $\boldsymbol{\beta}_{k'}$, by covariate, and $\boldsymbol{\rho}_k$, by state and covariate, at 0.5 identifies drivers of risk cycles



Clinical variables that effect number of seizures at time t, given the current state, vs more long-term effect on transitioning at time $t + 1$.

Transitions (partial results)

Transition	Covariate	Post. mean (SD)	MPPI	95% CI
2 → 2	Age	2.99 (0.59)	0.98	(1.81, 4.16)
2 → 2	Zonisamide	5.47 (0.94)	0.68	(3.29, 7.06)
1 → 2	Cannabidiol	3.15 (0.49)	1.00	(2.33, 4.12)
2 → 1	Cannabidiol	5.50 (0.85)	1.00	(3.54, 6.62)
3 → 2	Cannabidiol	-2.36 (0.70)	0.53	(-3.90, -1.13)

Emissions (state 2 - partial results)

	Covariate	Post. mean (SE)	PPI	95% CI
	Age	-1.17 (0.09)	1.00	(-1.35, -0.99)
	Gender	-0.15 (0.03)	1.00	(-0.22, -0.09)
	Bad mood	1.43 (0.13)	1.00	(1.17, 1.68)
	Change in medications	1.84 (0.05)	1.00	(1.75, 1.93)
	Triple or potassium bromide	-1.99 (0.49)	1.00	(-3.02, -1.12)
	Verapamil	-1.21 (0.34)	1.00	(-1.92, -0.60)

Cannabidiol associated with **greater** likelihood of remaining in states 1 & 2 than transitioning to state 3.

Patient age and treatment with zonisamide **increase** chance of remaining in state 2

Bad mood, sudden changes in medications, illness, and tiredness were strongly associated with a **greater** expected n. of seizures (Haut et al, 2007).

Triple or potassium bromide and verapamil associated with **reducing** expect n. seizures (Yoshitomi, 2019).

- Wang, Chiang, Haneef, Rao, Moss and Vannucci (2022, *Annals Applied Stats*)

- RNS Data from surgically implanted devices (Chiang et al. 2021, *Brain stimulation*)

Summary and Conclusions

- *Spike-and-slab* priors for variable selection are well suited for applications.
- Flexible structure for the incorporation of external information
- Methodologies can be extended beyond Gaussian data (e.g., count data).
- Computational schemes can embed data augmentation schemes for efficient posterior sampling.
- Improved performance over competitive penalized approaches.

Acknowledgments

- **Emily Wang**, Data Scientist at CommonSpirit Health
  Ph.D. Thesis title: "Bayesian State-Space Models with Variable Selection for Neural Count Data" (NLM Training Program in Biomedical Informatics)
- Robert Moss, collaborator at SeizureTracker
- **Sharon Chiang** and Vikram Rao, UCSF
- Zulfi Haneef, Baylor College of Medicine
- Stephen Cleboski, NeuroPace, Inc.

Gulf Coast Consortia
QUANTITATIVE BIOMEDICAL SCIENCES

Seizure Tracker™
be aware. track it.

NEUROPACE

UCSF
University of California
San Francisco

Baylor
College of
Medicine