Discussion on Default Priors and Robust Estimation for GLMs (Abel Rodríguez)

Alice Kirichenko, University of Warwick

8 September 2022

Context (LPEPs)

- **Idea:** Train the prior on imaginary data (sample size $n^\star$)
- Use a power trick on the likelihood:

$$\pi_k^{\text{PEP}}(\theta_k \mid M_k) = \int \frac{\tilde{f}_k(y^\star \mid \theta_k, M=k, \delta)\pi_k^N(\theta_k \mid M=k)}{\int \tilde{f}_k(y^\star \mid \theta_k, M=k, \delta)\pi_k^N(\theta_k \mid M=k)d\theta_k} m^\star(y^\star) p(\delta) dy^\star d\delta,$$

where $\tilde{f}_k(y^\star \mid \theta_k, M=k, \delta) = \frac{p_k^{1/\delta}(y^\star \mid \theta_k, M=k)}{\int p_k^{1/\delta}(y^\star \mid \theta_k, M=k)d\theta_k}$ is the normalized power likelihood.

- Generally difficult to work with for GLMs

Context (LPEPs)

- **Idea:** Train the prior on imaginary data (sample size $n^\star$)
- Use a power trick on the likelihood:

$$\pi_k^{\text{PEP}}(\theta_k \mid M_k) = \int \frac{\tilde{f}_k(y^\star \mid \theta_k, M = k, \delta)\pi_k^N(\theta_k \mid M = k)}{\int \tilde{f}_k(y^\star \mid \theta_k, M = k, \delta)\pi_k^N(\theta_k \mid M = k) d\theta_k} m^\star(y^\star) p(\delta) dy^\star d\delta,$$

where $\tilde{f}_k(y^\star \mid \theta_k, M = k, \delta) = \frac{p_k^{1/\delta}(y^\star \mid \theta_k, M=k)}{\int p_k^{1/\delta}(y^\star \mid \theta_k, M=k) d\theta_k}$ is the normalized power likelihood.

- Generally difficult to work with for GLMs
- **Solution:**
    - Use un-normalized likelihoods *(Fouskakis et al. (2018))*
    - Use Laplace approximations *(Porwal and Rodriguez (2021))*

Summary (LPEPs)

- The predictive distribution $m^\star(y^\star)$ does not depend on δ

Summary (LPEPs)

- The predictive distribution $m^{\star}(y^{\star})$ does not depend on δ
- Three choices for δ are suggested:
  - Use $\delta = n = n^{\star}$
  - Use hyper-g/n prior
  - Use a robust prior

Summary (LPEPs)

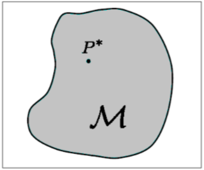
- The predictive distribution $m^\star(y^\star)$ does not depend on δ
- Three choices for δ are suggested:
  - Use $\delta = n = n^\star$
  - Use hyper-g/n prior
  - Use a robust prior
- Model selection consistency is guaranteed under some regularity conditions

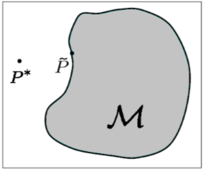
Summary (LPEPs)

- The predictive distribution $m^\star(y^\star)$ does not depend on δ
- Three choices for δ are suggested:
  - Use $\delta = n = n^\star$
  - Use hyper-g/n prior
  - Use a robust prior
- Model selection consistency is guaranteed under some regularity conditions
- Computationally tractable, can be incorporated in standard MCMC
- Good empirical performance

Bayesian GLMs under misspecification I

- Even when the model is wrong, we would still like our Bayesian methods to perform well (find the "closest" approximation to the truth):





- Standard Bayes does not always work:
  Kleijn and van der Vaart (2012); Müller (2013); Holmes and Walker (2017), etc.

Bayesian GLMs under misspecification II

- One can consider generalised posteriors as a remedy:

$$\pi(\theta \mid y) \propto p(y \mid \theta)^{\eta} \pi(\theta)$$

Bayesian GLMs under misspecification II

- One can consider generalised posteriors as a remedy:

$$\pi(\theta \mid y) \propto p(y \mid \theta)^{\eta} \pi(\theta)$$

- *de Heide et al. (2020)* shows that in the context GLMs, we have consistency and (almost) optimal convergence rates as long as
  - prior has continuous strictly positive density
  - $\eta < 1$ is sufficiently small
  - regularity conditions are satisfied

Bayesian GLMs under misspecification II

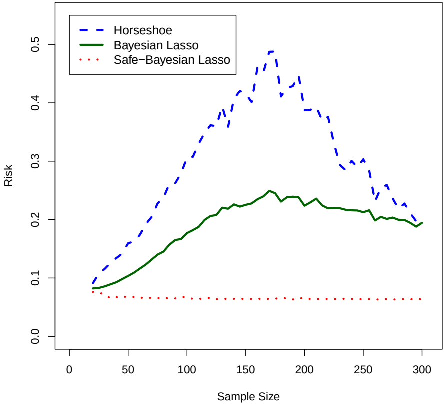
- One can consider generalised posteriors as a remedy:

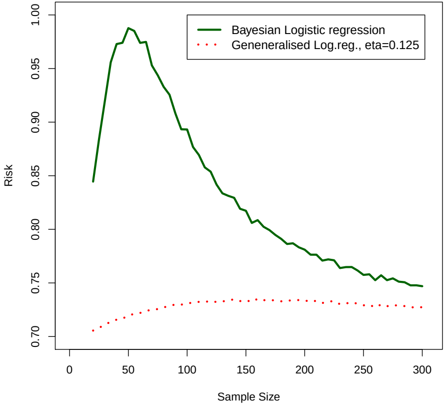
$$\pi(\theta \mid y) \propto p(y \mid \theta)^{\eta} \pi(\theta)$$

- *de Heide et al. (2020)* shows that in the context GLMs, we have consistency and (almost) optimal convergence rates as long as
  - prior has continuous strictly positive density
  - $\eta < 1$ is sufficiently small
  - regularity conditions are satisfied
- Selecting appropriate learning rate η is hard: *Grünwald and van Ommen (2017); Holmes and Walker (2017); Lyddon et al. (2019); Syring and Martin (2019).*

Bayesian GLMs under misspecification III

Good empirical performance (with the right η):





Questions

- Is there a way to address potential misspecification?
- Poor performance as the number of non-zero coefficients in the true model increases?
- What is needed to generalise to the case when p grows with n? Is there any hope for $p = n$ or $p > n$?

  Catalytic prior distributions *(Huang, Stein, Rubin, and Kou (2020))?*

Summary (Factor models)

For binary data $y_{ij} \in \{0, 1\}$

$$y_{ij} \sim \text{Ber}(\theta_{ij}), \quad \theta_{ij} = G_j(\mu_j + \alpha_j^T \beta_i),$$

where $\alpha_j, \beta_i \in \mathbb{R}^d$,$1 \leq i \leq I, 1 \leq j \leq J$, $\alpha_j, \beta_i \in \mathbb{R}^d$, and $d \ll J$.

Summary (Factor models)

For binary data $y_{ij} \in \{0, 1\}$

$$y_{ij} \sim \text{Ber}(\theta_{ij}), \quad \theta_{ij} = G_j(\mu_j + \alpha_j^T \beta_i),$$

where $\alpha_j, \beta_i \in \mathbb{R}^d$,$1 \leq i \leq I, 1 \leq j \leq J$, $\alpha_j, \beta_i \in \mathbb{R}^d$, and $d \ll J$.

Challenges:

- Selecting the correct dimension d

Summary (Factor models)

For binary data $y_{ij} \in \{0, 1\}$

$$y_{ij} \sim \text{Ber}(\theta_{ij}), \quad \theta_{ij} = G_j(\mu_j + \alpha_j^T \beta_i),$$

where $\alpha_j, \beta_i \in \mathbb{R}^d$,$1 \leq i \leq I, 1 \leq j \leq J$, $\alpha_j, \beta_i \in \mathbb{R}^d$, and $d \ll J$.

Challenges:

- Selecting the correct dimension d
- Choice between parametric vs nonparametric priors on the latent traits

Summary (Factor models)

For binary data $y_{ij} \in \{0, 1\}$

$$y_{ij} \sim \text{Ber}(\theta_{ij}), \quad \theta_{ij} = G_j(\mu_j + \alpha_j^T \beta_i),$$

where $\alpha_j, \beta_i \in \mathbb{R}^d$,$1 \leq i \leq I, 1 \leq j \leq J$, $\alpha_j, \beta_i \in \mathbb{R}^d$, and $d \ll J$.

Challenges:

- Selecting the correct dimension d
- Choice between parametric vs nonparametric priors on the latent traits
- Potential benefit from using geometry other than Euclidean

Summary (Factor models)

For binary data $y_{ij} \in \{0, 1\}$

$$y_{ij} \sim \text{Ber}(\theta_{ij}), \quad \theta_{ij} = G_j(\mu_j + \alpha_j^T \beta_i),$$

where $\alpha_j, \beta_i \in \mathbb{R}^d$,$1 \leq i \leq I, 1 \leq j \leq J$, $\alpha_j, \beta_i \in \mathbb{R}^d$, and $d \ll J$.

Challenges:

- Selecting the correct dimension d
- Choice between parametric vs nonparametric priors on the latent traits
- Potential benefit from using geometry other than Euclidean
  - Utility functions that use geodesic distances
  - Focus on spherical models: coming up with a right prior

Yu and Rodriguez (2020). A Bayesian Approach to Spherical Factor Analysis for Binary Data. arXiv preprint arXiv:2008.05109.

Bayesian analysis on manifolds

- Thomas Bayes' walk on manifolds *(Castillo, Kerkyacharian, and Picard (2014))*
- Bayesian manifold regression *(Yang and Dunson (2016))*
- Density estimation and modeling on symmetric spaces *(Li, Lu, Chevallier, and Dunson (2020))*
- Poisson process intensity estimation on manifolds *(Giordano, Kirichenko, and Rousseau (2022+))*

Bayesian analysis on manifolds

- Thomas Bayes' walk on manifolds *(Castillo, Kerkyacharian, and Picard (2014))*
- Bayesian manifold regression *(Yang and Dunson (2016))*
- Density estimation and modeling on symmetric spaces *(Li, Lu, Chevallier, and Dunson (2020))*
- Poisson process intensity estimation on manifolds *(Giordano, Kirichenko, and Rousseau (2022+))*

Remarks:

- Focus on the regression/density estimation
- Mostly Gaussian process based or piece-wise constant
- Mostly assume manifold is known

Questions II

- Can we benefit from the existing literature on priors on manifolds?
- Any guidance to choosing the geometry/manifold family? Are nested manifolds preferable?

Bibliography

[1] Fouskakis, Ntzoufras, and Perrakis (2018). Power-expected-posterior priors for generalized linear models. Bayesian Analysis, 13(3), 721-748.

[2] Porwal and Rodriguez (2021). Laplace Power-expected-posterior priors for generalized linear models with applications to logistic regression. arXiv preprint arXiv:2112.02524.

[3] Heide, Kirichenko, Grunwald, Mehta (2020). Safe-Bayesian generalized linear regression. AISTATS 2020.

[4] Huang, Stein, Rubin, and Kou (2020). Catalytic prior distributions with application to generalized linear models. Proceedings of the National Academy of Sciences, 117(22), 12004-12010.

[5] Castillo, Kerkyacharian, and Picard (2014). Thomas Bayes' walk on manifolds. Probability Theory and Related Fields, 158(3), 665-710.

[6] Yang and Dunson (2016). Bayesian manifold regression. The Annals of Statistics, 44(2), 876-905.

[7] Li, Lu, Chevallier, and Dunson (2020). Density estimation and modeling on symmetric spaces. arXiv preprint arXiv:2009.01983.

[8] Giordano, Kirichenko, and Rousseau (2022). Poisson process intensity estimation on manifolds. *In preparation.*