Construyendo modelos Bayesianos

Modelos Bayesianos con aplicaciones ecológicas

Dr. Cole Monnahan
University of Concepción, Chile

Enero, 2018

Recap

- La posterior es:
- $P(\theta|y)=cP(\theta)P(y|\theta)$
- Posterior = (constant)(prior)(likelihood)
- Se usa MCMC para hacer integración, con muestras correlacionadas
- Usa las muestras para aproximar las probabilidades
- Hay que chequear para señales de no hay convergencia

Construyendo modelos Bayesianos

- Gelman et al (2014) recomiende tres pasos básicos:

1. Hacer un modelo colectivo por todos los cantidades (datos y parámetros) del problema
2. Condicionar el modelo a los datos observados y estimar la probabilidad *a posteriori*
3. Evaluar el ajuste, realizar si necesario, y después hacer inferencia (calcular probabilidades).

Defina el proceso del los datos

- Especifica el proceso de los datos
  - Cual es la estructura del modelo?
  - Que tipo de dato puede ser observado?
  - Esos definen la verosimilitud
- E.g., crecimiento no puede ser negativo y las longitudes no pueden ser negativas
- Observaciones imperfectas puede ser bajas de la verdad pero nunca negativa.
- Se elige una curva VB con una verosimilitud log-normal

Elige la prior

- Identifica los parámetros desconocidos del proceso y especifica priors
- A veces hay información de estudios previos, o conocimiento del un experto
- A veces no es caro lo que debes usar
- La pregunta es: Que sabes del sistema antes de observar los datos

Consejo para elegir priors

- Palabras "*vague*", "*weakly informative*" etc. no son definidos
- A menudo es mejor cuando la escala de los parámetros es cerca de uno (*unit scale*)
  - E.g., estandardiza los predictores, divide por la escala
  - Así que las priors son mas fácil de entender
- Evita restricciones duras, a menos que hay una razón física, e.g., $\theta>0$ or $0<p<1$.
- E.g., si piensas $0< \theta <1$, usa N(.5,.5) en lugar de U(0,1)

https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Exploración de la priors

- Se recomienda plotear la prior vs la posterior después de ajustarlo
- Explorar varias priors asi puede Exploring different priors like this can help gauge the sensitivity

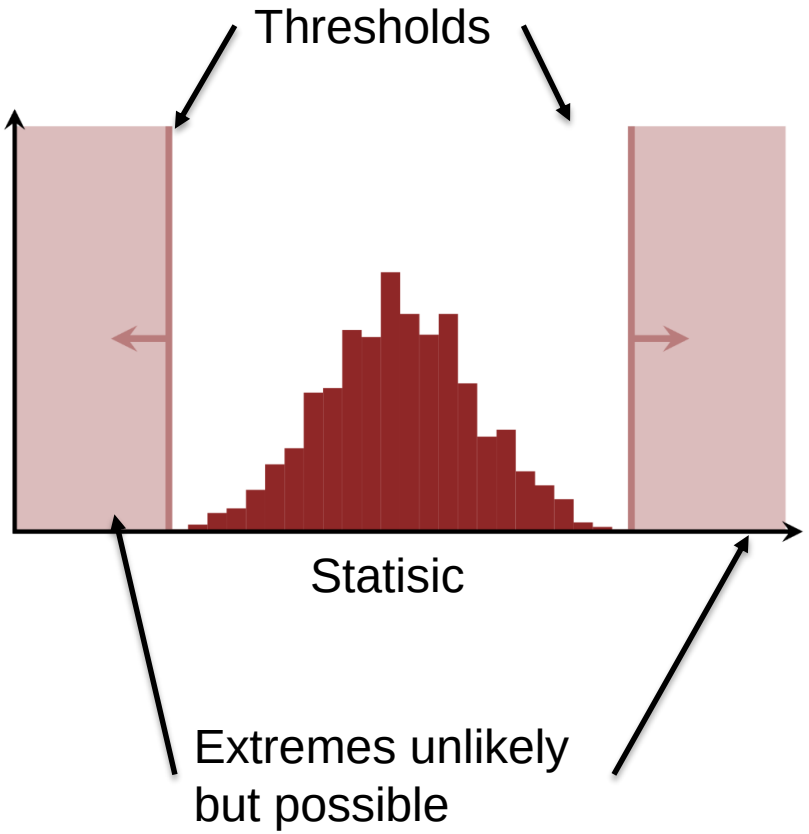
The prior predictive distribution

- Casi siempre se sabe algo, E.g.
  "we can be fairly sure that we won't observe any particularly healthy [birds] cruising near the speed of light"[1]
- Identifica una estadística importante (e.g., velocidad) con un umbral y verificar que:
  - Pocos valores son mas alla del umbral (*extreme is unreasonable but not impossible*)
  - Pero no muchos
  - Es antes de saber los datos

[1] https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html

The prior predictive distribution

- Corre el modelo con solo las priors
- En R o JAGS
- Plotea un histograma de la estadística
- Es realista?
- Si no, las priors no reflejan tu conocimiento



[1] https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html

Ejemplo

- Se marca 20 animales, y esperas que entre 10 y 20 vayan a sobrevivir (S) (i.e., binomial verosimilitud)
- La probabilidad de sobrevivencia =p=ilogit(theta) donde ilogit=1/(1+exp(-theta))
- Suponga que la prior es: theta~N(0,100)
- En R: Plotea la prior implicada de p, y la *prior predictive distribution* de S

Ejemplo

- Se marca 20 animales, y esperas que entre 10 y 20 vayan a sobrevivir (=S)
- Es una binomial verosimilitud
- La probabilidad de sobrevivencia =p=ilogit(theta) donde ilogit=1/(1+exp(-theta))
- Suponga que la prior es: theta~N(0,100)
- En R: Plotea la prior implicada de p, y la *prior predictive distribution* de S
- Por lo menos N(1, 0.5) es una mejor prior

Ejemplo

- Entonces introduce los datos al modelo y adjusta con JAGS
- Chequea por senals de no hay convergencia [Demo in R]

```
model {
   # Prior
   theta~dnorm(mu, 1/(sigma*sigma))
   logit(p) <- theta
   # likelihood
   y ~ dbin(p, N)
}
```

Ejercicio

- Actualiza el modelo JAGS (logistic2.jags) y código de R para usar datos de tres sitios independientes

```
dat2 <- list(y=c(15, 12, 11, 12, 4, 15, 17, 1
    2, 16, 14),
             N=rep(20, 10), R=10, mu=1,
sigma=0.5)
```

- [Pista: hay que usar *for loop*]
- Hace una comparación de la prior (PDF) vs posterior (histograma)

Posterior predictive distribution

- Recuerda que MCMC es un método para generar muestras de la posterior
- Pero MCMC convergencia no significa que el ajuste del modelo es bueno.
- Entonces, como se puede saber si es bueno? Que la estructura del modelo es suficiente compleja?
- Una manera es a *posterior predictive distribution*.

Posterior predictive distribution

- La idea es replicar el proceso que genera los datos, dado la posterior
- Se puede hacerlo en R o JAGS
- Después, compara los datos observados (reales) con los que **habría sido observado** (*the posterior predicted data*)
- Los reales deben parecer como los precedidos, y si no el modelo no es suficiente

Posterior predictive distribution

- Es decir que si tenemos una muestra θ^* después podemos simular otro dato y^*
- E.g., en nuestro caso y*=rbinom(1, 20, p*). Se produce un dato.
- Repite por todos las muestras de θ^* para formar una distribución (*posterior predictive*)
- Finalmente, compara eso (a veces visualmente) a los datos observados para encontrar patrones malos

Exercise

- Take the previous JAGS model and add a posterior predictive distribution, one for each replicate for each sample
- [Hint: ypred[i] ~ dbin(p,N[i]) will actually do random number generation like rbinom]
- Plot the distribution of site vs posterior predictive and then add the real data on top
- [Hint: use jittering for visual clarity]

References

- Hooten, M. B. and N. T. Hobbs (2015). "A guide to Bayesian model selection for ecologists." Ecological Monographs **85(1): 3-28.**
- Gelman, A., J. B. Carlin, et al. (2014). Bayesian data analysis, Taylor & Francis.