EST-46115: Modelación Bayesiana

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022 — Inferencia Aproximada.
Objetivo: En esta sección del curso estudiaremos métodos aproximados de inferencia. En particular nos concentraremos en estudiar inferencia variacional. Lo cual implica utilizar una familia de distribuciones dónde buscaremos un sustituto para nuestra distribución posterior. `Stan` utiliza un método de esta colección de aproximaciones.
Lectura recomendada: Una explicación de la aproximación Laplace la puedes encontrar en 2.4.4 de [5] y 13.3 de [2]. Es natural encontrar exposición de inferencia variacional en textos de ML como [6] o [1]. Aunque, también se puede consultar en la sección 13.7 de [2].

1. INTRODUCCIÓN

En inferencia Bayesiana definimos un modelo conjunto para `observaciones` y las `configuraciones` del proceso generador de datos. Esto nos permite utilizar el teorema de Bayes para actualizar nuestro conocimiento sobre los parámetros que no conocemos por medio de

$$\pi(\theta|y) \propto \pi(y|\theta)\,\pi(\theta)\,. \tag{1}$$

El procedimiento de inferencia dentro de este marco es sencillo y prácticamente directo pues se traduce en `reportar la distribución posterior` de las configuraciones en luz de las observaciones que tengamos.

A lo largo de este curso hemos establecido que de alguna u otra forma lo que necesitamos es reportar valores esperados (que se traduce en poder resolver integrales) utilizando dicho estado de conocimiento actualizado.

En esta sección estudiaremos mecanismos para utilizar aproximaciones al proceso de inferencia basado en el lado derecho de Eq. (1).

2. MUESTREO Y APROXIMACIONES

Hasta ahora lo que hemos visto son métodos de `aproximación de integrales`. En particular utilizando el `método Monte Carlo`. Hemos discutido que este es un método de estimación insesgado del cual se pueden esperar algunas propiedades bondadosas en el largo plazo.

Con los métodos de `simulación Markoviana` esperamos poder eliminar los problemas de complejidad computacional que usualmente se encuentran en aplicaciones. Por ejemplo, la incapacidad de utilizar generadores de números aleatorios para cualquiera que sea la distribución dada.

Los métodos Markovianos generan muestras, que esperamos sean, ligeramente correlacionadas y cuya distribución corresponda a la distribución que nos interesa.

Resta estudiar qué alternativas tenemos cuando no podemos esperar a que termine de correr nuestro muestreador.

2.1. Aproximación por curvatura

Estudiaremos alternativas para poder aproximar el problema de inferencia Bayesiana. La aproximación ya no es sobre la resolución de una integral. Ahora, vamos a aproximar la distribución misma utilizando ciertas nociones de optimización.

En clase hemos discutido que con un conjunto suficientemente grande de datos la distribución posterior se *parece* a una Gaussiana. Parece natural poder, entonces, construir una aproximación con estas características. Es decir,

$$\pi(\theta|y) \approx \text{Normal}\left(\theta|\hat{\theta}, \Sigma_{\hat{\theta}}\right) , \tag{2}$$

donde

$$\hat{\theta} = \text{moda}(\theta|y), \qquad \Sigma_{\hat{\theta}} = \left[-\nabla^2_{\theta} \log \pi(\hat{\theta}|y)\right]^{-1} . \tag{3}$$

La aproximación utiliza información de primero y segundo orden de la distribución posterior. Es decir

$$\hat{\theta} = \arg\max_{\theta} \pi(\theta|y) , \tag{4}$$

y Σ_θ nos da la información de la curvatura. Esta `aproximación cuadrática` se denomina `aproximación de Laplace`.

> Para que la aproximación de Laplace tenga sentido para un problema con parámetros restringidos, es usual transformar los parámetros a una escala sin restricciones. Por ejemplo, utilizando una transformación logarítmica o *logit* y recordando incorporar el término multiplicativo de la Jacobiana de dicha transformación.

La aproximación de Laplace nos permite sustituir un modelo de probabilidad por otro. Aunque en teoría podemos determinar la `calidad de la aproximación` –por medio de la `expansión de Taylor`– en la práctica puede resultar infactible dar una estimación de este error.

El problema de la aproximación de Laplace es que utiliza información local y puede fallar en capturar propiedades globales importantes de la distribución objetivo.

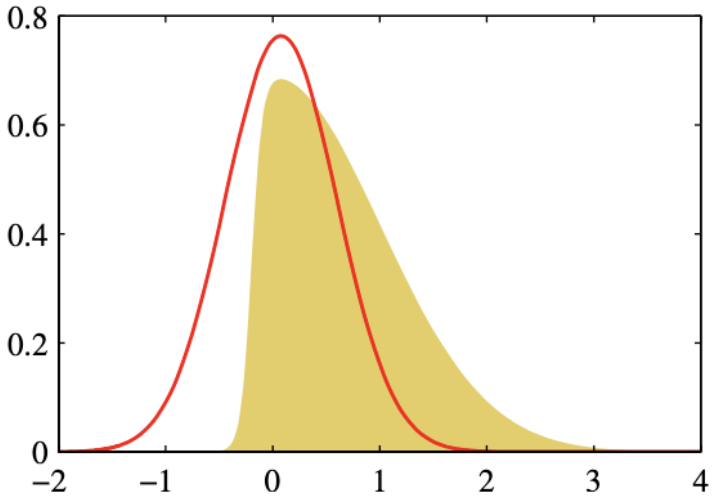


FIGURA 1. *Imagen tomada de [1]. Aproximación de Laplace en rojo. Distribución objetivo sombreada.*

2.2. Aproximación por optimización

Una alternativa es definir una familia de posibles distribuciones $\mathcal{Q}$ y encontrar dentro de esta familia de distribuciones la que `mejor se parezca` a nuestra distribución objetivo $\pi(\theta|y)$.

Lo importante es poder definir la noción de encontrar al mejor candidato dentro de $\mathcal{Q}$.

2.3. La solución

En inferencia aproximada buscamos sustituir nuestra distribución objetivo con aquella que resuelva el problema

$$\min_{q \in \mathcal{Q}} \mathsf{KL}\left(q(\theta) \middle\| \pi(\theta|y) \right), \tag{5}$$

donde la familia de distribuciones $\mathcal{Q}$ define la calidad/complejidad de nuestra aproximación.

El problema es que no podemos calcular la divergencia de KL, pues necesitamos calcular la constante de normalización en $\pi(\theta|y)$. Así que lo que hacemos es re-expresar

$$\mathsf{KL}(q(\theta) \| \pi(\theta|y)) = \mathbb{E}[\log q(\theta)] - \mathbb{E}[\log \pi(\theta, y)] + \log \pi(y)\,. \tag{6}$$

El lado derecho en el primer término define

$$\mathsf{ELBO}(q) := \mathbb{E}[\log \pi(\theta, y)] - \mathbb{E}[\log q(\theta)]\,. \tag{7}$$

la cual se denomina la cota inferior de evidencia (*evidence lower bound*, `ELBO`).

La cual podemos usar para re-expresar

$$\log \pi(y) = \mathsf{KL}(q(\theta) \| \pi(\theta|y)) + \mathsf{ELBO}(q)\,, \tag{8}$$

de donde podemos ver que lo que podemos buscar es `maximizar` el `ELBO` en lugar de `minimizar` la divergencia `KL`.

Nota que también podemos expresar

$$\mathsf{ELBO}(q) = \mathbb{E}[\log \pi(y|\theta)] - \mathsf{KL}(q(\theta) \| \pi(\theta))\,. \tag{9}$$

Lo cual nos dice que la distribución $q \in \mathcal{Q}$ que encontraremos será aquella que busque configuraciones afines al proceso generador de datos y que sea cercana a la distribución inicial.

En Fig. 2 se muestra la solución encontrada minimizando el criterio de `ELBO`.

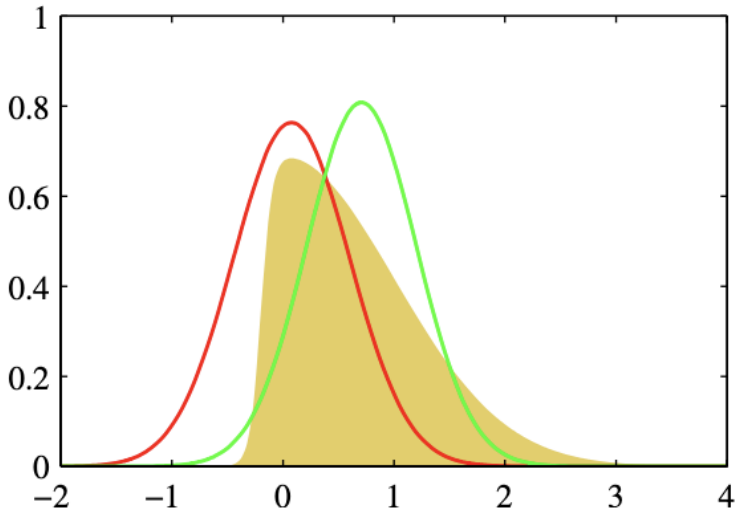


FIGURA 2. *Imagen tomada de [1]. Solución de `ELBO` se muestra en verde. Aproximación de Laplace en rojo.*

2.4. Dirección de KL

Hemos tomado la solución de $\mathsf{KL}(q\|\pi)$ por cuestiones numéricas y también discutimos que la solución tiene la interpretación de ser una aproximación de la posterior (justo lo que nos interesa).

Por ejemplo, en Fig. 3, bajo una familia de Gaussianas independientes para $\mathcal{Q}$ la solución de $\mathsf{KL}(q\|\pi)$, además, se puede ver como una distribución que se concentra en las zonas de `alta probabilidad`. Mientras que la solución de $\mathsf{KL}(\pi\|q)$ se concentra en zonas de `alta densidad`. Lo que nos habla que la formulación correcta se fijará en las propiedades que nos interesen.

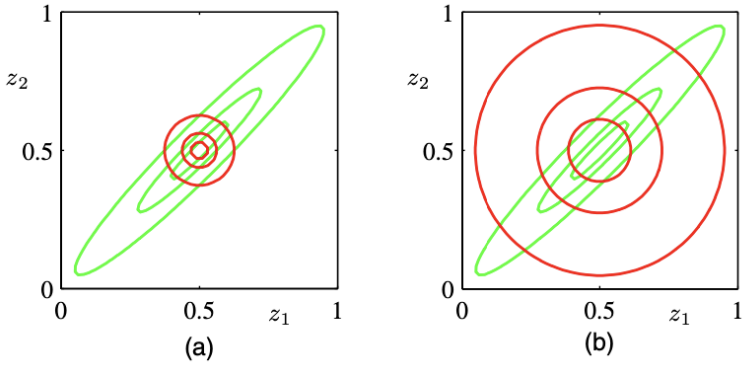


FIGURA 3. *Imagen tomada de [1]. En (a) se muestra* $\mathsf{KL}(q\|\pi)$ *y en (b) se muestra* $\mathsf{KL}(\pi\|q)$ *donde* $q \in \mathcal{Q}$ *y* π *es la distribución objetivo.*

El mismo efecto se muestra en Fig. 4 donde dependiendo de la formulación se pueden recuperar ciertas propiedades de la distribución objetivo.

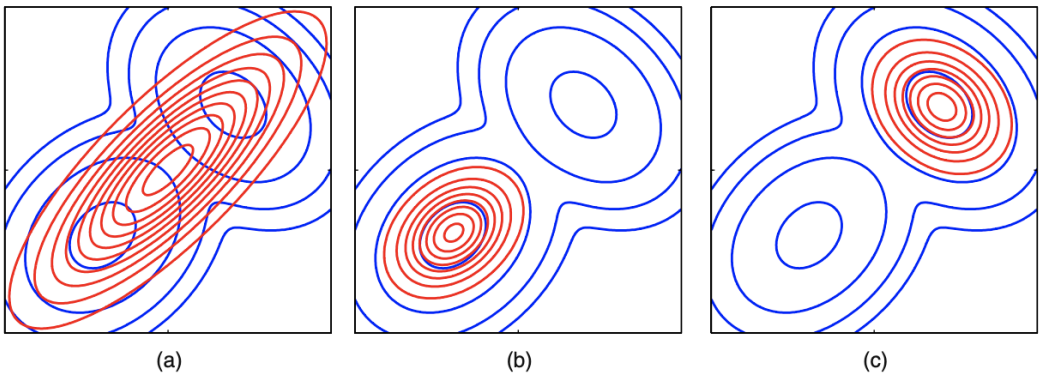


FIGURA 4. *Imagen tomada de [1]. Modelo objetivo basado en una mezcla de Gaussianas (azul). En (a) se muestra la aproximación que minimiza* $\mathsf{KL}(\pi\|q)$*. En (b) y (c) se muestran mínimos globales que corresponden a* $\mathsf{KL}(q\|\pi)$*.*

2.5. Conclusiones

La familia de distribuciones $\mathcal{Q}$ define la calidad de aproximación. Por simplicidad se utiliza una distribución Gaussiana con componentes independientes en el espacio de parámetros transformados (`solución de campo medio`, *mean field*).

Dado que la solución de

$$\min_{q\in\mathcal{Q}} \mathsf{KL}(q\|\pi)\,, \tag{10}$$

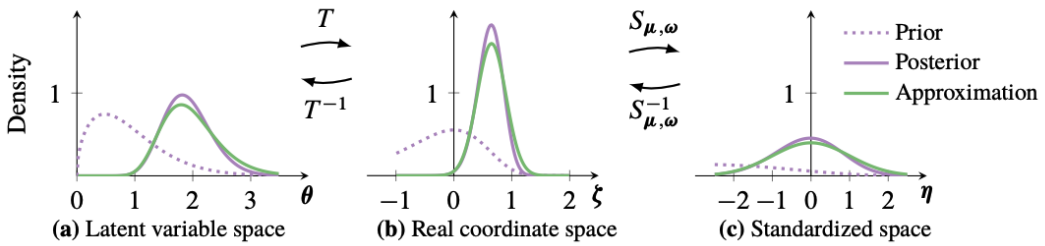


FIGURA 5. *Imagen tomada de [4].*

necesita resolverse en un espacio de funciones de probabilidad se utilizan herramientas de `cálculo de variaciones`. Por lo tanto, utilizar estas soluciones para resolver un problema de inferencia se llama `inferencia variacional` o `bayes variacional` ([1]).

Sin embargo, es usual considerar `familias parametrizadas` y buscar

$$\min_{\omega \in \Omega} \mathsf{KL}\left(q_\omega(\theta) \middle\| \pi(\theta|y) \right), \tag{11}$$

donde la búsqueda se realiza mediante $\omega \in \Omega$. Por ejemplo, el vector de medias y matriz de varianzas-covarianzas de las distribuciones Gaussianas.

3. INFERENCIA APROXIMADA

`Stan` en particular ofrece una solución basada en [4]. En el cual se formula el problema en términos de:

- Diferenciación automática.
- Una familia $\mathcal{Q}$ de distribuciones que operan bajo un espacio sin restricciones.
- La familia $\mathcal{Q}$ es la familia de distribuciones Gaussianas con componentes independientes (la matriz de varianzas es una matriz diagonal).
- Se puede utilizar un modelo con matriz de varianzas completas: `method = "fullrank"`.

4. EJEMPLO NUMÉRICO

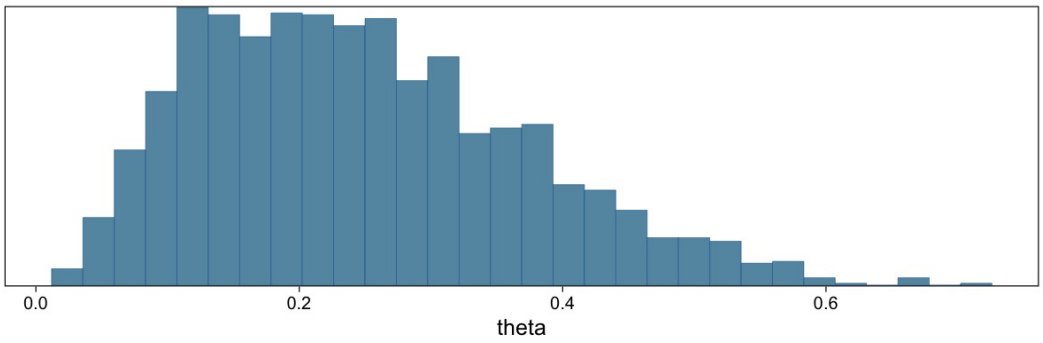
Tomado de la documentación de `Stan`

```
data {
  int<lower=0> N;
  array[N] int<lower=0,upper=1> y;
}

parameters {
  real<lower=0, upper=1> theta;
}

model {
  theta ~ beta(1, 1);
  y ~ bernoulli(theta);
}
```

```
stan_data ← list(N = 10, y = c(0,1,0,0,0,0,0,0,0,1))
posterior ← modelo$sample(stan_data, seed = 123, chains = 2, refresh = 1000)
```

```
posterior$summary() ▷ as.data.frame()
```

```
  variable   mean median   sd  mad     q5   q95 rhat ess_bulk ess_tail
1     lp__ -7.30  -7.03 0.72 0.38 -8.820 -6.75    1      902     1006
2    theta  0.25   0.23 0.12 0.13  0.079  0.47    1      762      712
```

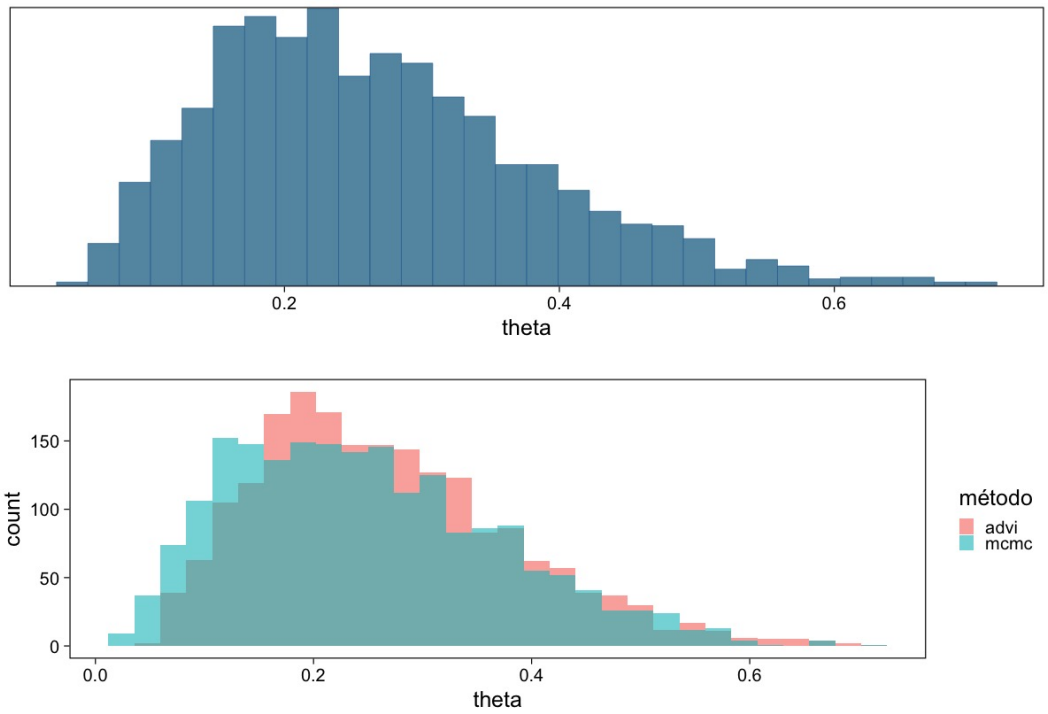



```
posterior.advi ← modelo$variational(stan_data, seed = 123,
                                    output_samples = 2000)
```

```
------------------------------------------------------------
EXPERIMENTAL ALGORITHM:
  This procedure has not been thoroughly tested and may be unstable
  or buggy. The interface is subject to change.
------------------------------------------------------------
Gradient evaluation took 6e-06 seconds
1000 transitions using 10 leapfrog steps per transition would take 0.06
    seconds.
Adjust your expectations accordingly!
Begin eta adaptation.
Iteration:   1 / 250 [  0%]  (Adaptation)
Iteration:  50 / 250 [ 20%]  (Adaptation)
Iteration: 100 / 250 [ 40%]  (Adaptation)
Iteration: 150 / 250 [ 60%]  (Adaptation)
Iteration: 200 / 250 [ 80%]  (Adaptation)
Success! Found best value [eta = 1] earlier than expected.
Begin stochastic gradient ascent.
  iter             ELBO   delta_ELBO_mean   delta_ELBO_med   notes
   100           -6.262             1.000            1.000
   200           -6.263             0.500            1.000
   300           -6.307             0.336            0.007   MEDIAN ELBO
       CONVERGED
Drawing a sample of size 2000 from the approximate posterior...
COMPLETED.
Finished in  0.1 seconds.
```

```
posterior.advi$summary() ▷
as.data.frame()
```

```
      variable   mean median   sd  mad     q5      q95
1           lp__ -7.18  -6.95 0.62 0.27 -8.36 -6.7508
2 lp_approx__ -0.52  -0.23 0.71 0.31 -2.06 -0.0029
3        theta  0.27   0.25 0.12 0.11  0.11  0.4799
```



5. SOLUCIÓN DE CAMPO MEDIO

El supuesto de factorización

$$q(\theta) = \prod_j q_j(\theta_j)\,, \tag{12}$$

es una estrategia bastante útil en física (*mean field theory*) y modelación de sistemas expertos (*message passing*, [3]).

La ventaja de esta factorización nos permite encontrar soluciones de forma cerrada donde podemos considerar para cada j

$$\log q_j^*(\theta_j) = \mathbb{E}_{q,i\neq j}\left(\log \pi(y,\theta)\right) + \text{const}\,, \tag{13}$$

donde usualmente la solución tiene un expresión analítica para miembros de la `familia exponencial`.

En la práctica, se desarrolla el modelo y se desarrollan las ecuaciones para resolver el problema de optimización variacional.

> ¿Cuánto tiempo tardas en escribir y resolver el problema variacional para una aplicación? El mismo tiempo que tarda en converger tu MCMC. —Frase popular en: Conferencia en inferencia aproximada.

6. CONCLUSIONES

- Inferencia variacional tiene poco de incorporarse a lenguajes de programación probabilística.
- Es una alternativa viable para poner en producción modelos bayesianos (por ejemplo, la sesión de conferencia de Smartly.io).
- Tema activo de investigación que logra conjuntar el estado del arte en ML con formulación probabilista.

> Variational inference is an extremely flexible and powerful approximation method. Its downside is that constructing the bound and update equations can be tedious. For a quick test, variational inference is often not a good idea. But for a deployed product, ita can be the most powerful tool in the box. —Philip Hennig, Probabilistic ML course, Tübingen U.

REFERENCIAS

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, New York, 2006. ISBN 978-0-387-31073-2. 1, 2, 3, 4, 5

[2] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*, volume 2. CRC press Boca Raton, FL, 2014. 1

[3] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2009. ISBN 978-0-262-01319-2. 7

[4] A. Kucukelbir, R. Ranganath, A. Gelman, and D. Blei. Automatic Variational Inference in Stan. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 5

[5] R. McElreath. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Texts in Statistical Science. Taylor and Francis, CRC Press, Boca Raton, Second edition, 2020. ISBN 978-0-367-13991-9. 1

[6] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, MA, 2012. ISBN 978-0-262-01802-9. 1