Computational Statistics II

Unit D.2: Approximate methods for probit and logit models

Tommaso Rigon

University of Milano-Bicocca

Ph.D. in Economics and Statistics

Unit D.2

Main concepts

- Laplace approximation for the logit model;
- Variational Bayes for logit models, Jaakkola and Jordan (2000) lower bound;
- Examples and comparisons on the Pima Indian dataset.

- Associated **R** code: `https://tommasorigon.github.io/CompStat/exe/un_D2.html`
- Additional **R** code (VB tutorial): `https://github.com/tommasorigon/logisticVB`

Main references

- Chopin, N. and Ridgway, J. (2017). Leave Pima indians alone: binary regression as a benchmark for Bayesian computation. *Statistical Science*, **32**(1), 64–87.
- Durante, D. and Rigon, T. (2019). Conditionally conjugate mean-field variational Bayes for logistic models. *Statistical Science*, **34**(3), 472–485.
- Jaakkola, T. S., and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, **10**(1), 25–37.

The logit model (recap)

- In this unit we will focus exclusively on the **logit model**, although similar strategies (Laplace, VB and EP) can be applied in the probit case as well.

- Let us recall once again that $\mathbf{y} = (y_1, \dots, y_n)^\intercal$ is a vector of the observed **binary responses**.

- Let $\mathbf{X}$ be the corresponding **design matrix** whose generic row is $\mathbf{x}_i = (1, x_{i2}, \dots, x_{ip})^\intercal$, for $i = 1, \dots, n$.

- In this unit we consider a logistic model such that

$$(y_i \mid \pi_i) \overset{\text{ind}}{\sim} \text{Bern}(\pi_i), \qquad \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}, \qquad \eta_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- As before, we assume a Gaussian prior $\pi(\beta) = \text{N}_p(\beta \mid \boldsymbol{b}, \boldsymbol{B})$.

EM algorithm and Laplace approximation

- The Laplace approximation relies on the MAP estimate $\hat{\beta}_{\text{MAP}}$ and on the **negative Hessian** matrix $\hat{\boldsymbol{M}}$, which in the logistic model case is

$$\hat{\boldsymbol{M}} = \boldsymbol{X}^\intercal \hat{\boldsymbol{H}} \boldsymbol{X} + \boldsymbol{B}^{-1},$$

where the vector $\hat{\boldsymbol{H}} = \text{diag}\{\hat{\pi}_1(1-\hat{\pi}_1), \dots, \hat{\pi}_n(1-\hat{\pi}_n)\}$ is evaluated at the MAP.

- We consider here an **EM algorithm** for finding $\hat{\beta}_{\text{MAP}}$ using the Pólya-gamma data augmentation, extending the approach we have described in **unit C.2** for the MLE.

- **Exercise**. Prove that the EM algorithm for logistic regression leads to the following iterative scheme:

$$\beta^{(r+1)} = (\boldsymbol{X}^\intercal \hat{\boldsymbol{Z}}^{(r)} \boldsymbol{X} + \boldsymbol{B}^{-1})^{-1}\{\boldsymbol{X}^\intercal(\boldsymbol{y} - 1/2) + \boldsymbol{B}^{-1}\boldsymbol{b}\},$$

where $\hat{\boldsymbol{Z}}^{(r)} = \text{diag}(\hat{z}_1^{(r)}, \dots, \hat{z}_n^{(r)})$, having defined

$$\hat{z}_i^{(r)} = \frac{\tanh(\boldsymbol{x}_i^\intercal \beta^{(r)}/2)}{2\boldsymbol{x}_i^\intercal \beta^{(r)}}, \qquad i = 1, \dots, n.$$

- The **Laplace approximation** then is $q(\beta) = \text{N}_p(\beta \mid \hat{\beta}_{\text{MAP}}, \hat{\boldsymbol{M}}^{-1})$.

Laplace approximation: implementation in R

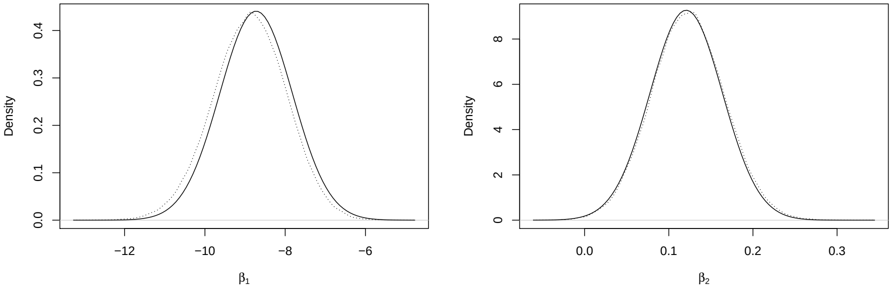
```r
logit_Laplace <- function(y, X, B, b, tol = 1e-16, maxiter = 10000) {
  # Initialization
  P <- solve(B) # Prior precision matrix
  Pb <- P %*% b # Term appearing in the EM algorithm
  logpost <- numeric(maxiter)
  Xy <- crossprod(X, y - 0.5)
  beta <- solve(crossprod(X / 4, X) + P, Xy + Pb)
  eta <- c(X %*% beta)
  w <- tanh(eta / 2) / (2 * eta); w[is.nan(w)] <- 0.25
  logpost[1] <- sum(y * eta - log(1 + exp(eta))) - 0.5 * t(beta) %*% P %*% beta

  # Iterative procedure
  for (t in 2:maxiter) {
    beta <- solve(qr(crossprod(X * w, X) + P), Xy + Pb)
    eta <- c(X %*% beta)
    w <- tanh(eta / 2) / (2 * eta); w[is.nan(w)] <- 0.25
    logpost[t] <- sum(y * eta - log(1 + exp(eta))) - 0.5 * t(beta) %*% P %*% beta

    if (logpost[t] - logpost[t - 1] < tol) { # Have we reached convergence?
      prob <- plogis(eta)
      return(list(
        mu = c(beta), Sigma = solve(crossprod(X * prob * (1 - prob), X) + P),
        Convergence = cbind(Iteration = (1:t) - 1, logpost = logpost[1:t])
      ))
    }
  }
  stop("The algorithm has not reached convergence")
}
```

Laplace approximation: results

- Using again the Pima indian dataset, we compare the performance of the Laplace approximation with the smoothed density obtained via MCMC (gold standard).

- Obtaining the Laplace approximation took **0.119 seconds**.

- In the picture are shown the marginal densities of β_1 and β_2 using MCMC (**dotted lines**) and the Laplace approximation (**solid lines**).



Variational Bayes

- The logistic regression case has been often presented as an example in which mean-field variational Bayes can not be applied; see for example Section 10.5 of Bishop (2006).
- The main "variational" alternative for a couple of decades was the Jaakkola and Jordan (2000) **lower bound**, which leads to a Gaussian approximation for logistic models.
- The JJ lower bound was introduced and motivated solely by convexity arguments.
- **Remark**. The JJ lower bound approach actually coincides with a genuine mean-field approximation based on the Pólya-gamma data augmentation. It is not a local method.

Main references

- Durante, D. and Rigon, T. (2019). Conditionally conjugate mean-field variational Bayes for logistic models. *Statistical Science*, **34**(3), 472–485.
- Jaakkola, T. S., and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, **10**(1), 25–37.

- Let $\boldsymbol{z} = (z_1, \dots, z_n)^\intercal$ be a vector of latent iid random variables following a $\text{PG}(1, 0)$.

- Then, recall that the **Pólya-gamma augmented likelihood** for a logistic model is

$$\pi(\boldsymbol{y}, \boldsymbol{z} \mid \beta) = \prod_{i=1}^{n} \frac{1}{2} \pi(z_i \mid 1, 0) \exp\{(y_i - 1/2)\boldsymbol{x}_i^\intercal \beta - z_i(\boldsymbol{x}_i^\intercal \beta)^2/2\},$$

  as described in **unit C.2**.

- We employ **mean-field** approximation, forcing the independence between and $\boldsymbol{z}$ and β, namely

$$q(\beta, \boldsymbol{z}) = q(\beta) q(\boldsymbol{z}).$$

- This means we can use the CAVI algorithm discussed in **unit D.1**.

The CAVI algorithm for logistic models

- The CAVI algorithm iterates between two simple steps.

- **Update $q(\beta)$**. The locally optimal variational distribution for $q(\beta)$ is

$$
\begin{aligned}
q(\beta) &\propto \exp\left[\mathbb{E}_q\{\pi(\boldsymbol{y}, \boldsymbol{z} \mid \beta) + \log \pi(\beta)\}\right] \\
&\propto \pi(\beta) \exp\left\{\sum_{i=1}^n (y_i - 1/2)\boldsymbol{x}_i^\intercal \beta - \frac{1}{2}\mathbb{E}_q(z_i)(\boldsymbol{x}_i^\intercal \beta)^2\right\}.
\end{aligned}
$$

Re-arranging the above equation, we obtain that $q(\beta) = \mathrm{N}_p(\beta \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$
\boldsymbol{\mu} = \boldsymbol{\Sigma}\{\boldsymbol{X}^\intercal(\boldsymbol{y} - 1/2) + \boldsymbol{B}^{-1}\boldsymbol{b}\}, \quad \boldsymbol{\Sigma} = (\boldsymbol{X}^\intercal \mathbb{E}_q(\boldsymbol{Z})\boldsymbol{X} + \boldsymbol{B}^{-1})^{-1},
$$

where $\boldsymbol{Z} = \mathrm{diag}(z_1, \dots, z_n)$ and its expectation is taken with respect to $q(\boldsymbol{z})$.

- Hence, the optimal variational distribution for β is **Gaussian**. This is an implication of the mean-field structure and not an assumption.

The CAVI algorithm for logistic models

- The second CAVI step involves the variational distribution $q(\boldsymbol{z})$.

- **Update $q(\boldsymbol{z})$**. The locally optimal variational distribution for $q(\boldsymbol{z})$ is

$$
\begin{aligned}
q(\boldsymbol{z}) &\propto \exp\left[\mathbb{E}_q\{\pi(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{\beta})\}\right] \\
&\propto \prod_{i=1}^n p(z_i \mid 1, 0) \exp\left\{-\frac{z_i}{2}\mathbb{E}_q(\eta_i^2)\right\}.
\end{aligned}
$$

  Re-arranging the above equation, we obtain that the following structure

$$
q(\boldsymbol{z}) = \prod_{i=1}^n \mathrm{PG}\left\{z_i \mid 1, \mathbb{E}_q(\eta_i^2)\right\}.
$$

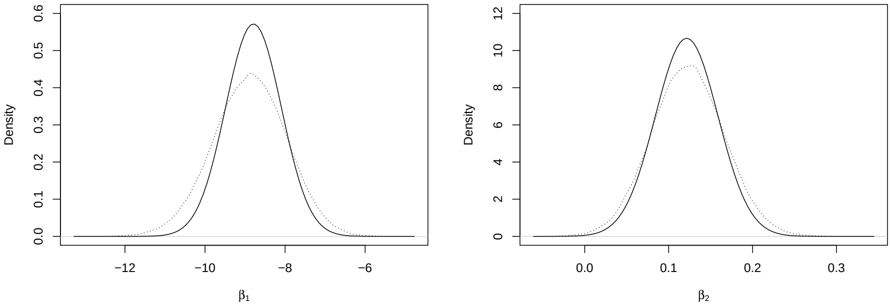
- Hence, the optimal variational distribution for $\boldsymbol{z}$ are **independent** Pólya-gamma distributions. As before, this is an implication and not an assumption.

Variational Bayes: implementation in **R**

```r
logit_CAVI <- function(y, X, B, b, tol = 1e-16, maxiter = 10000) {
  lowerbound <- numeric(maxiter)
  p <- ncol(X); n <- nrow(X)
  P <- solve(B); Pb <- c(P %*% b); Pdet <- ldet(P)

  # Initialization
  # ...
  # [Code omission, refer to the online Markdown D.2 file]

  # Iterative procedure
  for (t in 2:maxiter) {
    P_vb <- crossprod(X * omega, X) + P; Sigma_vb <- solve(P_vb)
    mu_vb <- Sigma_vb %*% (crossprod(X, y - 0.5) + Pb)

    # Update of xi
    eta <- c(X %*% mu_vb)
    xi <- sqrt(eta^2 + rowSums(X %*% Sigma_vb * X))
    omega <- tanh(xi / 2) / (2 * xi); omega[is.nan(omega)] <- 0.25

    lowerbound[t] <- 0.5 * p + 0.5 * ldet(Sigma_vb) + 0.5 * Pdet - 0.5 * t(mu_vb - b) %*% P %*% (mu_vb - b) +
        sum((y - 0.5) * eta + log(plogis(xi)) - 0.5 * xi) - 0.5 * sum(diag(P %*% Sigma_vb))

    if (abs(lowerbound[t] - lowerbound[t - 1]) < tol) {
      return(list(mu = c(mu_vb), Sigma = matrix(Sigma_vb, p, p)))
    }
  }
  stop("The algorithm has not reached convergence")
}
```

Variational approximation: results

- Obtaining the variational Bayes approximation took **0.082 seconds**.
- In the picture are shown the marginal densities of β_1 and β_2 using MCMC (**dotted lines**) and the variational approximation (**solid lines**).
- The variational approximation is clearly problematic. The **variance** is **much smaller** than that of the true posterior. The **posterior means** look approximately **correct**.



Hybrid Laplace (a possibly bad proposal)

- It is generally agreed that VB approximations leads to sensible **point estimates**, but they fail at quantifying the associated uncertainty.

- Some proposals to correct this distortion have been made (Giordano, Broderick and Jordan, 2017), but this is not straightforward to apply.

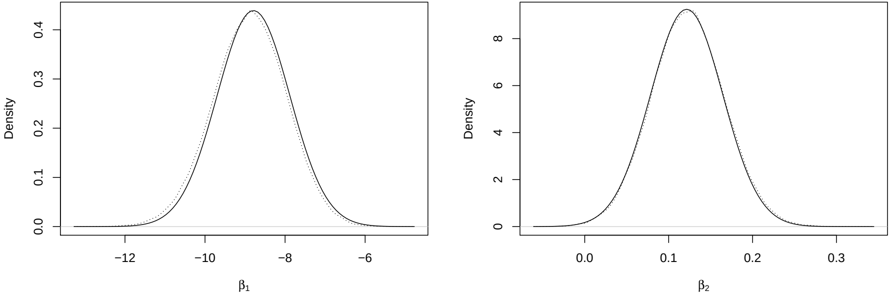
- In the logistic regression case, it seems that there is an easy fix. We could plug-in the VB estimates into the inverse Fisher information matrix.

```
logit_HL <- function(y, X, B, b, tol = 1e-16, maxiter = 10000) {
  fit_HL <- logit_CAVI(y, X, B, b, tol, maxiter)
  prob <- c(plogis(X %*% fit_CAVI$mu))
  fit_HL$Sigma <- solve(crossprod(X * prob * (1 - prob), X) + solve(B))
  fit_HL
}
```

- **(Difficult exercise)**. Can you prove that this procedure leads to an "optimal" Gaussian approximation, in some sense? Is it better than the usual mean-field VB? Is it better than the Laplace approximation?

Hybrid Laplace: results

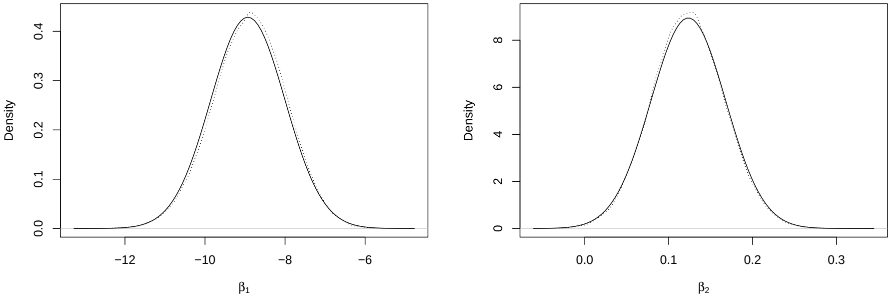
- Obtaining the hybrid Laplace requires almost the same time of the VB.
- In the picture are shown the marginal densities of β_1 and β_2 using MCMC (**dotted lines**) and the hybrid Laplace approximation (**solid lines**).
- The hybrid Laplace approximation is a sensible improvement over the VB. It is also an mild improvement over the Laplace approximation, as we will later clarify.



Expectation propagation: results

- Obtaining the EP approximation required **0.011 seconds** using the `EPGLM` package.

- In the picture are shown the marginal densities of β_1 and β_2 using MCMC (**dotted lines**) and the EP approximation (**solid lines**).



Final comparisons

- We conclude our discussion by comparing the various approximations with the "optimal" Gaussian distribution based on moment matching.
- In this experiment, the moments are obtained via MCMC.
- We consider the Kullback-Leibler divergence and the Wasserstein distance, which are both available in closed form in the Gaussian-Gaussian case.
- The hybrid Laplace and the EP perform best in this example.

Method	Kullback-Leibler	Wasserstein distance
Laplace approximation	0.029	0.027
Variational Bayes	0.275	0.065
Expectation Propagation	0.032	0.006
Hybrid Laplace	0.011	0.010