Posterior consisteny of logistic random effect models

Yongdai Kim

Seoul National University, Korea

jointly with

Dohyun Kim

2007.08.05

1. Introduction

- In longitudinal study or cluster randomization designs, observations in each subject (or cluster) are likely to be dependent.
- An example is a genetic epidemiology where observations on members of one family are correlated.
- A statistic challenge is to assess the effect of covariates to a response variable with adjusting intra-correlation.
- Two popular used models are *Generlized estimating equations* (GEE) and *Generalizaed linear mixed models* (GLMM) which incorporate *random effects*.

Notations

- y_{ij} : binary response for jth ($j = 1, \ldots, n_i$) observation on ith ($i = 1, \ldots, n$) subject (or cluster).
- $\mathbf{x}_{ij}$: $p \times 1$ vector of covariates
- $\boldsymbol{\beta}$: $p \times 1$ vector of regression coefficients
- $\mathbf{z}_{ij}$: $q \times 1$ vector of covariates with random coefficients
- $\mathbf{b}_i$: $q \times 1$ vector of random effect $\sim F$
- $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})'$
- $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})'$.

Logistic-random effects model

- Conditional on $\mathbf{b}_i, i = 1, \ldots, m$, y_{ij} are independent with

$$\Pr(y_{ij} = 1|\mathbf{b}_i) = \frac{\exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{b}_i)}{1 + \exp(\mathbf{x}_{ij}^{'}\boldsymbol{\beta} + \mathbf{z}_{ij}^{'}\mathbf{b}_i)};$$

- $\mathbf{b}_1, \ldots, \mathbf{b}_m \overset{i.i.d.}{\sim} F$ for some distribution function F.
- In practice, a multivariate normal distribution with mean 0 and variance-covariance matrix Σ is widely used for F. However, when the true distribution of random effects is not Gaussian, estimators of the regression coefficients can be inconsistent (e.g. Neuhaus et.al.(1992)).
- So, we need to consider a nonparametric random effect distribution.

Objective of the talk

- Give sufficient conditions for the posterior consistency of the regression coefficients $\boldsymbol{\beta}$ as well as the random effect distribution F when we use
  - parametric prior on $\boldsymbol{\beta}$
  - nonparametric prior (e.g. Dirichlet process) on F
- Explain how they are used in the proof.
- Discuss limitations of the sufficient conditions and future works.

2. Literature Review

Logistic Random effects model

- Walker and Mallick (1997) : Polya Tree prior on F, GLMM and Frailty model, MCMC
- Mukhopadhyay and Gelfand (1997) : Dirichlet process prior on F, GLMM, MCMC
- And more ....

Posterior consistency

- General theorem
  - Schwartz (1965)
  - Barron, Schervish and Wasserman (1999)
  - Ghosal, Ghosh and Ramamoorthi (1999)
  - Walker (2004)

- Regression problem
  - Ameou-Atisso, Ghosal, Ghosh and Ramamoorthi (2003): Semiparametric Regression
    * $y = \mathbf{x}^{'}\boldsymbol{\beta} + \epsilon, \;\; \epsilon \sim F.$
    * Prior: $\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}), F \sim$ Polya Trees
  - Choudhuri, Ghosal and Roy (2004): Binary data with nonparametric link function
    *
    $$\Pr(Y = 1|\mathbf{x}) = \Phi(\mathbf{x}^{'}\boldsymbol{\beta})$$
    * $\boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}),$ logit $\Phi(z) \sim$ GP .
  - Choi and Schervish (2007): Nonparametric regression
    *
    $$y = f(\mathbf{x}) + \epsilon, \;\; \epsilon \sim N(0, \sigma^2).$$
    * Prior: $f \sim$ GP $, \sigma^2 \sim \pi(\sigma^2).$

3. Posterior consistency : General theory

Definition

The sequence of posteriors $\{\Pi(\cdot|Y^n)\}$ is said to be consistent at θ_0 if for every neighborhood W of θ_0,

$$\Pi(W^c|Y^n) \overset{n\to\infty}{\to} 0 \;\; a.s.\; [P_{\theta_0}],$$

where P_{θ_0} is a probability measure induced by the parameter value θ_0 and $Y^n \equiv (Y_1, \cdots, Y_n)$

An Extension of Schwartz Theorem

- **Existence of tests**
  - Let $W \subset \Omega$ be a given neighborhood of θ_0.
  - Suppose W^c can be decomposed to finitely many subsets $W_1^c, \ldots, W_k^c$. That is

$$W^c = \bigcup_{j=1}^{k} W_j^c.$$

  - For testing $H_0 : \theta = \theta_0$ *V.S.* $H_1 : \theta \in W_j^c$, we can construct a sequence of (exponentially separating) test functions $\Phi_n^j(Y_1, \cdots, Y_n)$ such that for some constants $C_1^j, C_2^j, C^j > 0$,

$$E_{P_{\theta_0}}(\Phi_n^j) \leq C_1^j e^{-nC^j}, \inf_{\theta \in W_j^c} E_{P_\theta}(\Phi_n^j) \geq 1 - C_2^j e^{-nC^j}.$$

- **Prior positivity**

  - Let Π be a prior on Ω. For all $\delta > 0$

    $$\Pi(\theta : K(P_{\theta_0}, P_\theta) < \delta) > 0$$

    where $K(P_{\theta_0}, P_\theta)$ is the K-L divergence.

- **Theorem:** If the above two conditions hold, then

$$\Pi(W^c|Y^n) \to 0 \;\; a.s.\; [P_{\theta_0}].$$

- **Remark.** Walker (2004) extended the theorem when k is countably infinite (i.e. $W^c = \cup_{j=1}^{\infty} W_j^c$) provided $\sum_{j=1}^{\infty} \sqrt{\pi(W_j^c)} < \infty$.

4. Posterior consistency of logistic random effect model

Notations

- $\Omega = \mathbb{R}^p \times \mathcal{F}$: Parameter space for $(\boldsymbol{\beta}, F)$
- μ : Prior on $\boldsymbol{\beta}$, Π_1 : Prior on F
- $\Pi = \mu \times \Pi_1$: Prior on $(\boldsymbol{\beta}, F)$
- $\boldsymbol{\beta}^*, F^*$: the true value

Basic assumption

- We only consider random intercept models. That is, $\mathbf{z}_{ij} = 1$.
- We only consider random-X design, which can be easily extended to fixed-X design.
- Since F is fully nonparametric, we don't need the intercept term. For identifiability, we assume $\Pr(a^{'}\mathbf{x} \neq 1) > 0$ for all $a \in R^p$.

Sufficient Conditions for the posterior consistency

1. $n_i \geq 1$ (For simplicity, we let $n_i = 1$ and drop the subscript j from the notations. That is, $(y_1, \mathbf{x}_1), \ldots, (y_n, \mathbf{x}_n)$ are i.i.d. copies of $(y, \mathbf{x})$.)

2. F has a support inside $[-M, M]$ for some $M > 0$.

3. Let $\mathbf{x} = (x_1, \ldots, x_p)$. Then

   (a) $\Pr(x_1 \in (a, b)) > 0$ for all $-\infty < a < b < \infty$.

   (b) Let $\mathbf{x}_{-1} = (x_2, \ldots, x_p)$. Then, there exists $\delta > 0$ and $\gamma > 0$ such that for any $\theta \in R^{p-1}$ with $\|\theta\|_2 = 1$

   $$\Pr(\theta' \mathbf{x}_{-1} < -\delta | x_1) > \gamma \text{ and } \Pr(\theta' \mathbf{x}_{-1} > \delta | x_1) > \gamma.$$

4. $\beta_1^* \neq 0$.

- **Main theorem**: Suppose the sufficient conditions holds. If the prior positivity condition also holds, then the posterior of $\boldsymbol{\beta}$ and F are consistent where the weak topology is used for F.
- **Remark.** For Dirichlet processes or Polya's trees, it can be easily shown that the prior positivity condition holds, and hence the posterior is consistent whenever the sufficient conditions hold.

Some explanation of the sufficient conditions

1. $n_i \geq 1$: it means that the model is identifiable even when we have only one observation in each subject. (nice but may make the other conditions too restrictive!)
2. The compact support of F
   - It is needed for some technical reasons.
   - It may not be a serious restriction since the intercept term can to be assumed be neither too small nor too large.
3. $\Pr(x_1 \in (a, b)) > 0$ for all $-\infty < a < b < \infty$.
   - It means that the support of x_1 is R.
   - It is necessary for the identifiability of F as well as β_1
   - It is a real limitation. We will be back on this issue later.
4. Conditions on $\mathbf{x}_{-1}$

- Simply speaking, it holds if
  - the all covariates in $\mathbf{x}_{-1}$ are centered
  - $\mathbf{x}$ are linearly independent with positive probability
- it can be satisfied easily in practice.

5. $\beta_1^* \neq 0$
   - It is required for identifiability of F.

Sufficient conditions for the consistency of MLE (Butler and Luois, 1997)

- $Pr(x_1 \in (a, b)) > 0$ for all $-\infty < a < b < \infty$.
- $\mathbf{x}$ is linearly independent (with positive probability).
- $n_i \geq p + q - 1$ (in the worst case).
- $\beta_1^* \neq 0$.

Remark

- It still require the support of x_1 covers R, which is for the identifiability of F.
- The number of samples in each subject depends on the dimension of covariates, which is not good in practice.
- F has non-compact support.
- Consider multivariate random effects models.

4. Proof (Sketch)

- For simplicity, we only consider $p = 2$ (i.e. $\mathbf{x}_{-1} = x_2$)

Review of the three steps of the proof of posterior consistency

1. Partition of $W^c = \bigcup_{j=1}^{k} W_j^c$
2. Constructing (exponentially separating) tests for $H_1 : \theta \in W_j^c$ for $j = 1, \ldots, k$.
3. Prior positivity

Partition of W^c

- $W = \{(\boldsymbol{\beta}, F) : |\beta_1 - \beta_1^*| \leq \epsilon_1, |\beta_2 - \beta_2^*| \leq \epsilon_2, d_w(F, F_0) \leq \epsilon_F\}$ where d_w is the metric from the weak topology.
- We decompose W^c by the union of
  1. $W_1^c = \{(\boldsymbol{\beta}, F) : |\beta_1 - \beta_1^*| > \epsilon_1\}$
  2. $W_2^c = \{(\boldsymbol{\beta}, F) : |\beta_1 - \beta_1^*| \leq \epsilon_1, |\beta_2 - \beta_2^*| > \epsilon_2\}$
  3. $W_3^c = \{(\boldsymbol{\beta}, F) : |\beta_1 - \beta_1^*| \leq \epsilon_1, |\beta_2 - \beta_2^*| \leq \epsilon_2, d_w(F, F_0) > \epsilon_F\}$.

Constructing a test for $H_1 : W_1^c$

- Further decomposition of W_1^c into
  - $W_{11}^c = \{(\boldsymbol{\beta}, F) : \beta_1 - \beta_1^* > \epsilon_1, \beta_2 \geq \beta_2^*\}$
  - $W_{12}^c = \{(\boldsymbol{\beta}, F) : \beta_1 - \beta_1^* > \epsilon_1, \beta_2 < \beta_2^*\}$
  - $W_{13}^c = \{(\boldsymbol{\beta}, F) : \beta_1 - \beta_1^* < -\epsilon_1, \beta_2 \geq \beta_2^*\}$
  - $W_{14}^c = \{(\boldsymbol{\beta}, F) : \beta_1 - \beta_1^* < -\epsilon_1, \beta_2 < \beta_2^*\}$
- Consider a test statistics

$$\Phi_n((Y_1, \mathbf{x}_1), \ldots, (Y_n, \mathbf{x}_n)) = \sum_{i=1}^{n} \Phi(Y_i, \mathbf{x}_i)/n$$

  where $\Phi(Y, \mathbf{x}) = Y I(\mathbf{x} \in A)$ (or $\Phi(Y, \mathbf{x}) = (1 - Y) I(\mathbf{x} \in A)$) for some $A \in R^2$.

- By Hoeffiding's lemma, for exponential separability, it suffices to show that

$$\inf_{(\boldsymbol{\beta},F)\in W_{ij}^c} \mathrm{E}_{\boldsymbol{\beta},F}(\Phi(Y,\mathbf{x})) - \mathrm{E}_{\boldsymbol{\beta}^*,F^*}(\Phi(Y,\mathbf{x})) > \gamma > 0. \qquad (1)$$

- From now on, what I am going to do is to find an appropriate set A for a given W_{ij}^c, which makes Φ satisfies (1).

- Note that

$$\mathrm{E}_{\boldsymbol{\beta},F}(\Phi(Y,\mathbf{x})|\mathbf{x}) = Pr_{\boldsymbol{\beta},F}(Y=1|\mathbf{x})I(\mathbf{x}\in A)$$

- For W_{11}^c,
  - It is easy to see that for all $x_1 > 0, x_2 > 0$,

$$Pr_{\beta_1,\beta_2,F}(Y = 1|\mathbf{x}) \geq Pr_{\beta_1^*,\beta_2^*,F}(Y = 1|\mathbf{x}).$$

  - Moreover, since F has compact support, we can show that there exists an open interval of positive length (a, b) such that there exists $\delta > 0$ such that for all $x_1 \in (a, b)$ and $x_2 > 0$,

$$Pr_{\beta_1,\beta_2,F}(Y = 1|\mathbf{x}) > Pr_{\beta_1^*,\beta_2^*,F^*}(Y = 1|\mathbf{x}) + \delta.$$

  - Let $A = \{\mathbf{x} : x_1 \in (a, b), x_2 > 0\}$. Since $P(A) > 0$ because the support of x_1 is R, the (1) is satisfied.
  - We can construct exponentially separating test functions for $W_{1j}^c, j = 2, 3, 4$ similarly.

Constructing a test for $H_1 : W_2^c$

- Further decomposition of W_2^c to
  - $W_{21}^c = \{(\beta, F) : |\beta_1 - \beta_1^*| \leq \epsilon_1, \beta_2 - \beta_2^* > \epsilon_2, F(0) > F^*(0)\}$
  - $W_{22}^c = \{(\beta, F) : |\beta_1 - \beta_1^*| \leq \epsilon_1, \beta_2 - \beta_2^* > \epsilon_2, F(0) < F^*(0)\}$
  - $W_{23}^c = \{(\beta, F) : |\beta_1 - \beta_1^*| \leq \epsilon_1, \beta_2 - \beta_2^* < -\epsilon_2, F(0) > F^*(0)\}$
  - $W_{24}^c = \{(\beta, F) : |\beta_1 - \beta_1^*| \leq \epsilon_1, \beta_2 - \beta_2^* < -\epsilon_2, F(0) < F^*(0)\}$
- For W_{21}^c
  - Choose $A = \{\mathbf{x} : x_2 \in (a, b) \subset (\delta, \infty), \beta_1^* x_1 + \beta_2^* x_2 \approx 0\}$ and $P(A) > 0$.

– Then for all $\mathbf{x} \in A$,

$$\begin{aligned} & P_{\beta_1,\beta_2,F}(Y=1|\mathbf{x}) - P_{\beta_1^*,\beta_2^*,F^*}(Y=1|\mathbf{x}) \\ \approx\ & P_{\beta_1^*,\beta_2,F}(Y=1|\mathbf{x}) - P_{\beta_1^*,\beta_2^*,F^*}(Y=1|\mathbf{x}) \\ \geq\ & P_{0,\epsilon_2,F}(Y=1|\mathbf{x}) - P_{0,0,F^*}(Y=1|\mathbf{x}) \ \ (\because \beta_1^* x_1 + \beta_2^* x_2 \approx 0) \\ >\ & P_{0,0,F}(Y=1|\mathbf{x}) - P_{0,0,F^*}(Y=1|\mathbf{x}) \\ =\ & F(0) - F^*(0) \geq 0. \end{aligned}$$

– Again, we use the Hoeffiding's lemma to construct a desirable test.

- Construction of tests for the other W_{2j}^c can be done similarly.

Constructing a test for $H_1 : W_3^c$

- Recall $W_3^c = \{(\boldsymbol{\beta}, F) : \beta_1 \approx \beta_1^*, \beta_2 \approx \beta_2^*, d_w(F, F^*) > \epsilon_F\}$.
- **Proposition:** There exists a **finitely many** real numbers $\xi_1, \ldots, \xi_m$ and positive number $\gamma > 0$ such that

$$W_3^c \subset \bigcup_{l=1}^{m} W_{3l}^c$$

where

$$W_{3l}^c = \{(\boldsymbol{\beta}, F) : \beta_1 \approx \beta_1^*, \beta_2 \approx \beta_2^*, |H_F(\xi_l) - H_{F^*}(\xi_l)| > \gamma\}$$

where

$$H_F(z) = \int \frac{\exp(z+s)}{1+\exp(z+s)} dF(s) (= Pr_F(Y = 1 | \mathbf{x}^{'} \beta^* = z)).$$

- Proof of Proposition
  - Due to the identifiability of location-mixture problem (Teicher, 1961), $d_w(F, F^*) > \epsilon_F$ implies that there exists $\gamma_0 > 0$ such that

$$\sup_{z \in R} |H_F(z) - H_{F^*}(z)| > \gamma_0.$$

  - Since $\{H_F(\cdot) : F\}$ is tight ($\because F$ has a compact support), there is a compact set C such that

$$\sup_{z \in C} |H_F(z) - H_{F^*}(z)| > \gamma_0.$$

  - Since $H_F(z)$ is sufficiently smooth ($\because$ the logistic function is smooth), we can prove there exists $\gamma > 0$ and $\{\xi_1, \ldots, \xi_m\}$ such that

$$\sup_{z \in \{\xi_1, \ldots, \xi_m\}} |H_F(z) - H_{F^*}(z)| > \gamma.$$

- Constructing a test for W_{3l}^c.
  - Decompose W_{3l}^c into two parts
    - $W_{3l+}^c = \{(\boldsymbol{\beta}, F) : \beta_1 \approx \beta_1^*, \beta_2 \approx \beta_2^*, H_F(\xi_l) - H_{F^*}(\xi_l) > \gamma\}$.
    - $W_{3l-}^c = \{(\boldsymbol{\beta}, F) : \beta_1 \approx \beta_1^*, \beta_2 \approx \beta_2^*, H_F(\xi_l) - H_{F^*}(\xi_l) < -\gamma\}$.
  - Construct a test for W_{3l+}^c
    - Let $A = \{\mathbf{x} : \mathbf{x}^{'}\boldsymbol{\beta}^* \approx \xi_l\}$ and $P(A) > 0$.
    - Then for all $x \in A$,

$$\begin{aligned} & P_{\beta_1,\beta_2,F}(Y = 1|\mathbf{x}) - P_{\beta_1^*,\beta_2^*,F^*}(Y = 1|\mathbf{x}) \\ \approx \quad & P_{\beta_1^*,\beta_2^*,F}(Y = 1|\mathbf{x}) - P_{\beta_1^*,\beta_2^*,F^*}(Y = 1|\mathbf{x}) \\ \approx \quad & H_F(\xi_l) - H_{F^*}(\xi_l) > \gamma > 0 \end{aligned}$$

    - Again, we use the Hoeffiding's lemma to construct a desirable test.
  - For W_{3l-}^c, we can construct a test function similarly.

Prior positivity

- Suppose $F \sim DP(\alpha)$ with a base measure $\alpha(\cdot)$.
- Assume $F^* << \alpha$ and $\mathrm{E}(\mathbf{x}) < \infty$.
- Then, we can easily show the prior positivity.
- Almost the same condition can be applied for the Polya tree prior, too.

5. Concluding Remark (future works)

About the sufficient conditions

- The most serious restriction among the given sufficient conditions is that the support of x_1 is R.
- This condition was crucially used for
  1. constructing a test for β_1
  2. constructing a test for F.
- Basically, the condition is strongly related to the fact that $H_F(z) = H_G(z)$ for all $z \in R$ implies $F = G$ (Identifiability of location-mixture problem, Teicher, 1961, Ann. Math. Statist.).

- Conjectures (and future works) to make the support of x_1 being compact
  - Reduce the space of $\mathcal{F}$ such that $H_F(z) = H_G(z)$ for z on a some compact set implies $F = G$.
  - A possible try(?): Mixture of parametric model and nonparametric model, say

$$F = \pi_1 N(0, \sigma^2) + \pi_2 G.$$

  - Also, restricting $n_i \geq 2$ would be helpful for β_1.

General multivariate random effects (e.g. multi-level model)

- Conjecture: The minimum of n_i should be related to the dimension of the random effect.
- The rank of $(\mathbf{x}, \mathbf{z})$ would play an important role.

General link function other than logistic

- Any link function which is symmetric at 0 and has nice tail properties (eg. exponential decreasing) would be OK.

Consistency of MLE

- We believe that the MLE is also consistent under our sufficient conditions.
- Our sufficient conditions, though it only considers a random intercept model, has some merits over those of the Butler and Louis.
- That is, techniques for posterior consistency can be used as an alternative method of proving the consistency of MLE.

References

- Andrew R. Barron, Mark. J. Schervish and Larry Wasserman (1999), The consistency of posterior distributions in nonparametric problems. Ann. Statist., 27 : 536-561
- Steven M. Butler, Thomas A. Louis(1997), Consistency of maximum likelihood estimators in general random effects models for binary data, The Annals of Statistics, 25.1.pp 351-377.
- N. Choudhuri, S. Ghosal, and A. Roy (2004) Bayesian estimation of the spectral density. J. Amer. Statist. Assoc., 99. pp 1050-1059,
- Thomas S. Ferguson(1973), A Bayesian Analysis of Some Nonparametric Problems, The Annals of Statistics, 1, pp 209-230
- Subhashis Ghosal, Jayanta. K. Ghosh and R. V. Ramamoorthi (1999) Posterior consistency of Dirichlet mixtures in density estimation, Ann. Statist., 27(1) : 143-158
- Amewou-Atisso Messan, Subhashis Ghosal, Jayanta. K. Ghosh and R. V. Ramamoorthi(2003), Posterior consistency for semiparametric regression problems, Bernoulli, vol 9, No 2, pp. 291-312
- S. Mukhopadhyay, A. E. Gelfand(1997), Dirichlet process mixed generalized linear models, Journal of the American Statistical Association, Vol 92, No 438.

pp 633-639

- J. M. Neuhaus, W. W. Hauck and J. D. Kalbfleisch(1992), The Effects of Mixture Distribution Misspecification when Fitting Mixed-Effects Logistic Models, Biometrika, 79.4. pp. 755-762.
- Lorraine Schwartz (1965), On Bayes procedures. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 4, pp 10-26
- Henry Teicher(1961), Identifiability of mixtures, Ann. Math. Statist. 32. pp 244-248.
- Stephen G. Walker, Bani K. Mallick(1997), Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparametric Mixing, Journal of the Royal Statistical Society. Series B, 59, 4, pp. 845-860.
- Stephen G. Walker(2004), New approaches to bayesian consistency. Ann. Statist., 32 pp 2028-2043