Efficient Inference In The Mixture of Negative Binomial Distributions

Hanxiang Peng
Joint work with Gibson Rayner

Department of Mathematics
University of Mississippi

September 15, 2008

Outline

Introduction

- Negative Binomial. Consider a sequence of IID binary trials. Let Y be the number of trials required to get the first r successes. Then

$$\mathbb{P}(Y = y) = \binom{y-1}{r-1}\lambda^r(1-\lambda)^{y-r}, \quad y = r, r+1, \ldots$$

  where λ is the probability of success.
- A substitute for Poisson distribution because the mean and variance are not equal:

$$\mathbb{E}(Y) = r/\lambda \quad \text{and} \quad \mathbb{V}ar(Y) = r(1-\lambda)/\lambda^2.$$

- The NB does not assume a fixed sample size, so it provides an alternative sequential approach in modelling binary responses.

Introduction

- Independence assumption is not appropriate in many areas, e.g. in developmental toxicity study. Offspring from the same litter are correlated and may respond *more similarly* to a stimulus than fetuses from different litters.
- Relaxing *independence* to *exchangebility*, George and Bowman (1995) proposed the *full likelihood procedure* for analyzing correlated binary data.
- Under *exchangebility*, Rayner and Peng (2006), Wang and Peng (2006) proposed *mixture of negative binomial* to study correlated binary data.

Introduction

- A sequence $X_1, X_2, \ldots$ is *exchangeable* if for any finite subset $X_{i_1}, \cdots, X_{i_n}$,

$$\mathbb{P}(X_{\pi_1} = x_1, \cdots, X_{\pi_n} = x_n) = \mathbb{P}(X_{i_1} = x_1, \cdots, X_{i_n} = x_n),$$

where $\pi_1 \ldots \pi_n$ is a permutation of $i_1 \ldots i_n$ and $x_i = 0, 1, \forall i$.

- Rayner, Peng and Wang (2006) derived that the probability that the first r successes is realized in y trials is given by

$$\mathbb{P}(Y = y) = \binom{y-1}{r-1} \sum_{k=0}^{y-r} (-1)^k \binom{y-r}{k} \lambda_{r+k}, \; y = r, r+1, \ldots$$

Introduction

- By the celebrated de Finetti representation theorem,

$$\lambda_k = \int_0^1 u^k \, dQ(u), \quad k = 0, 1, \cdots ,$$

where Q is the probability measure on $[0, 1]$ uniquely determined by the infinite exchangeable sequence.

- Immediately it follows

$$\mathbb{P}(Y = y) = \int_0^1 \binom{y-1}{r-1} u^r (1-u)^{y-r} \, dQ(u), \quad y = r, r+1, \cdots .$$

- Written $Y \sim \mathrm{MNB}(\boldsymbol{\lambda}, r)$ with $\boldsymbol{\lambda} = (\lambda_r, \lambda_{r+2}, \ldots)$ where

$$\lambda_k = \mathbb{P}(X_1 = 1, \cdots, X_k = 1), \quad k = 1, 2, \ldots.$$

- The case $r = 1$ is the *mixture of geometric distributions (MG)*.
- Interestingly, MNB is equivalent to a "parametric distribution" with countably infinitely many parameters. i.e., MNB has infinitely many parameters.
- In this talk, we are interested in the *efficient estimation* of the infinitely many parameters.
- The efficiency criterion is that of *least dispersed regular estimates* based on the convolution theorems, see e.g. Schick (1986) or van der Vaart (1998).
- In this talk, we also shall give an MLE of the mixing measure Q.
- Estimating mixing measure, e.g., van der Geer (1996 (J. Nonparametric Statist.), 2003(Compu. Statist. & Data Analy.)), Genovese and Wasserman (2000, Ann. Statist.)

- $\{\lambda_k : k = 0, 1, 2, ...\}$ $(\lambda_0 = 1)$ is *complete monotone*:

$$(-1)^k \Delta^l \lambda_k \geq 0, \quad l = 0, 1, 2, ...$$

where Δ is the difference operator:

$$\Delta a_i = a_{i+1} - a_i, \quad \Delta^2 a_i = \Delta(\Delta a_i) = a_{i+2} - 2a_{i+1} + a_i,$$

for a sequence $\{a_1, a_2, ...\}$.

- Using de Finetti representation, the moment generating function of Y is

$$M_Y(t) = e^{tr} \int_0^1 u^r [1 - (1-u)e^t]^{-r} Q(u),$$

in some neighborhood of the origin.

- We formally define

$$\lambda_{-k} = \int_0^1 \frac{dQ(u)}{u^k}, \quad k = 1, 2, \ldots.$$

Moments

Theorem

- *If $\lambda_{-1} < \infty$, then the mean of Y exists and is given by*

$$\mathbb{E}(Y) = M'_Y(0) = r\lambda_{-1}.$$

If $\lambda_{-2} < \infty$, then the second moment of Y exists and is given by

$$\mathbb{E}(Y^2) = M''_Y(0) = r(r+1)\lambda_{-2} - r\lambda_{-1}.$$

- *Then the variance of Y is simply*

$$\mathbb{V}ar(Y) = \mathbb{E}(Y^2) - \left(\mathbb{E}(Y)\right)^2 = r(r+1)\lambda_{-2} - r\lambda_{-1} - (r\lambda_{-1})^2.$$

Moments

- If the mixing measure Q is a point mass concentrated on $p \in (0,1)$, then the resulting distribution is the negative binomial $\mathrm{NB}(p,r)$. Indeed,

$$\lambda_k = \int_0^1 u^k \, dQ(u) = p^k, \quad \lambda_{-k} = \int_0^1 \frac{dQ(u)}{u^k} = \frac{1}{p^k},$$

  Hence, all moments exists.
- In particular, we recover the mean and variance of $Y \sim \mathrm{NB}(p,r)$,

$$\mathbb{E}(Y) = r\lambda_{-1} = r/p,$$

$$\mathbb{V}ar(Y) = r(r+1)\lambda_{-2} - r\lambda_{-1} - (r\lambda_{-1})^2 = r(1-p)/p^2.$$

Moments

Suppose Q has a density q w.r.t. the Lesbegue measure.

1 If $q(u) = 1$, then

$$\lambda_k = \int_0^1 u^k(1)\ du = \frac{1}{k+1} \quad k = 0, 1, 2, \cdots .$$

In this case, for all $k = 1, 2, \cdots$, $\lambda_{-k} = \int_0^1 \frac{dQ(u)}{u^k} = \int_0^1 \frac{1}{u^k}\ du$, does not exist; therefore, none of the moments of the MNB exist either.

Moments

2 Now suppose $q(u) = 2u$. Then we have

$$\lambda_k = \int_0^1 u^k(2u)\ du = \frac{2}{k+2} \quad k = 0, 1, 2, \cdots .$$

In this case, $\lambda_{-1} = \int_0^1 \frac{2u}{u}\ du = 2$. Consequently, the mean of Y is given by $\mathbb{E}(Y) = r\lambda_{-1} = 2r$. However, the variance and higher moments still do not exist.

Moments

3 Finally, suppose $q(u) = 4u^3$. Then we have

$$\lambda_k = \int_0^1 u^k(4u^3)\ du = \frac{4}{k+4} \quad k = 0, 1, 2, \cdots .$$

In this case,

$$\lambda_{-1} = \int_0^1 \frac{4u^3}{u}\ du = \frac{4}{3}, \quad \lambda_{-2} = \int_0^1 \frac{4u^3}{u^2}\ du = 4\int_0^1 u\ du = 2.$$

Consequently, the mean and variance of Y are given by
$\mathbb{E}(Y) = r\lambda_{-1} = \frac{4r}{3}$ and
$\mathbb{V}ar(Y) = r(r+1)\lambda_{-2} - r\lambda_{-1} - (r\lambda_{-1})^2 = \frac{2r^2}{9} + \frac{2r}{3}$.

Maximum Likelihood Estimation

- Let $Y \sim \mathrm{MNB}(\boldsymbol{\lambda}, r)$. Then for $y = r, r+1, \ldots,$

$$f(y; \boldsymbol{\lambda}_y, r) = \mathbb{P}(Y = y) = \binom{y-1}{r-1} \sum_{k=0}^{y-r} (-1)^k \binom{y-r}{k} \lambda_{r+k},$$

where $\boldsymbol{\lambda}_y = (\lambda_r, \cdots, \lambda_y)$. Note that the number of parameters varies with observation y.

- For $Y_1, Y_2, \cdots, Y_n$ i.i.d. copies of Y, the average of the log-likelihood function is

$$l_n(\boldsymbol{\lambda}_{Y_n^*}) = \frac{1}{n} \sum_{i=1}^{n} \log f(Y_i; \boldsymbol{\lambda}_{Y_i}, r)$$

where $Y_n^* = \max(Y_1, \cdots, Y_n)$. Assume for now r is known.

Maximum Likelihood Estimation

- The MLE $\hat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$ is the maximizer of $l_n(\boldsymbol{\lambda}_{Y_n^*})$ subject to

$$\lambda_r \leq 1, \quad (-1)^l \Delta^l \lambda_i \geq 0, \quad i \geq r, l \geq 0.$$

- Let $\pi_y = \mathbb{P}(Y = y) = \binom{y-1}{r-1} p_y$ where

$$p_y = \sum_{k=0}^{y-r} (-1)^k \binom{y-r}{k} \lambda_{r+k}, \quad y = r, r+1, \ldots$$

Reversing these equations yields

$$\lambda_t = \sum_{i=0}^{t-r} (-1)^i \binom{t-r}{i} p_{r+i} = \sum_{i=0}^{t-r} (-1)^i c_{t,i+r} \pi_{r+i}, \quad t = r, r+1, \ldots$$

where $c_{t,i} = \binom{t-r}{i-r} / \binom{i-1}{r-1}$.

Maximum Likelihood Estimation

- In terms of $\mathbf{p} = \{p_k : k = r, r+1, ...\}$, we write $l_n(\boldsymbol{\lambda}_{Y_n^*})$ as

$$\ell_n(\mathbf{p}_{Y_n^*}) = \frac{1}{n}\sum_{i=1}^{n} \log p_{Y_i} + C_n,$$

- The MLE $\hat{\mathbf{p}}$ of $\mathbf{p}$ is the maximizer of the above subject to

$$p_y \geq 0, y \geq r, \quad \sum_{y=r}^{\infty} \binom{y-1}{r-1} p_y = 1.$$

- By the Lagrange multipliers, the MLE can be found as

$$\hat{p}_y = A_y / \binom{y-1}{r-1} n, \quad y = r, r{+}1, \cdots, Y_n^*; \quad \hat{p}_y = 0, \quad y > Y_n^*,$$

where $A_y = \sum_{i=1}^{n} \mathbf{1}[Y_i = y]$.

Moments

- Thus, the MLE $\hat{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$ can be obtained as

$$\hat{\lambda}_t = \sum_{i=0}^{t-r} (-1)^i \binom{t-r}{i} \hat{p}_{r+i}, \quad t = r, r+1, \cdots, Y_n^*.$$

and $\hat{\lambda}_t = 0, t > Y_n^*$.

Unbiasedness

- Easily verified

Theorem

For every $t = r, r+1, \ldots$, $\hat{\lambda}_t$ is unbiased est. of λ_t: $\mathbb{E}(\hat{\lambda}_t) = \lambda_t$.

- The asymptotic variance of $\hat{\lambda}_t$ for $t = r, r+1, \ldots$,

$$\sigma_t^2 = n\mathbb{V}ar(\hat{\lambda}_t) = \frac{1}{n}\sum_{i=r}^{t} c_{ti}^2 \mathbb{V}ar(A_i) = \sum_{i=r}^{t} c_{ti}^2 \pi_i - \lambda_t^2.$$

- The asymptotic covariance is

$$C_{st} \equiv n\mathbb{C}ov(\hat{\lambda}_s, \hat{\lambda}_t) = \sum_{i=r}^{s\wedge t} c_{si} c_{ti} \pi_i - \lambda_s \lambda_t, \;\; s, t = r, r+1, \ldots.$$

where $\sum'_{i\neq j}$ denotes $\sum_{i=r}^{s} \sum_{j\neq i, j=1}^{t}$ and $s \wedge t = \min(s, t)$.

Asymptotic Normality

- To stress the dependence of $\hat{\lambda}_t$ on the n observations $Y_1, \cdots, Y_n$, we write $\hat{\lambda}_t = \hat{\lambda}_{nt}$. For d positive integers $t_k \geq r$ where $k = 1, \cdots, d$, let $\boldsymbol{\lambda}_d = (\lambda_{t_1}, \cdots, \lambda_{t_d})^\top$ and $\hat{\boldsymbol{\lambda}}_{nd} = (\hat{\lambda}_{d_1}, \cdots, \hat{\lambda}_{t_d})^\top$. Denote Σ_d the $d\mathbf{x}d$ matrix with the (i,j)th entry $C_{t_i t_j}$ when $t_i \neq t_j$ and the (i,i) entry $\sigma^2_{t_i}$.
- An application of the usual multivariate central limit theorem yields the asymptotic normality.

Theorem*

$$\sqrt{n}(\hat{\boldsymbol{\lambda}}_d - \boldsymbol{\lambda}_d) \Longrightarrow \mathcal{N}(0, \Sigma_d), \quad n \to \infty.$$

Asymptotic behavior of the Stochastic Process

- We now study the asymptotic efficiency of the stochastic process $\hat{\boldsymbol{\lambda}} = \left\{ \hat{\lambda}_k : k = r, r+1, \ldots \right\}$. The following theorem states that we can estimate almost the parameters asymptotically.

Theorem

If $0 < \lambda_1 < 1$ *then* $\mathbb{P}(\lim_{n\to\infty} Y_n^* = \infty) = 1$.

- By asymptotic theory of semiparametric models (e.g. Bickel, Klassen, Ritov and Wellner (1991), or van der Vaart(1998)), we can show

Theorem

$\hat{\boldsymbol{\lambda}}$ *is an efficient estimate of* $\boldsymbol{\lambda}$.

Asymptotic behavior of the Stochastic Process

Sketches of Proof:

- Recall that a sequence of random elements $\mathbf{Y}_n$ with values in a metric space *converges in distribution* to a random element $\mathbf{Y}$ if
$$\mathbb{E}f(\mathbf{Y}_n) \to \mathbb{E}f(\mathbf{Y}), n \to \infty$$
for every bounded, continuous f from the metric space to reals $\mathcal{R}$.
- Let S be a nonempty set and $\ell^\infty(S)$ be a set of bounded functions on S. Let $\mathcal{P}$ be a collecdtion of probability measures.

Asymptotic behavior of the Stochastic Process

Sketches of Proof: Theorem 25.48, van der Vaart(1998).

Theorem

(Efficiency in $\ell^\infty(S)$) Suppose $\psi : \mathcal{P} \mapsto \ell^\infty(S)$ is differentiable at P, and suppose that $T_n(s)$ is asymptotically efficient at P for estimating $\psi(P)(s)$, for every $s \in S$. Then T_n is asymptotically efficient at P provided that the sequence $\sqrt{n}(T_n - \psi(P))$ converges under P in distribution to a tight limit in $\ell^\infty(S)$.

Asymptotic behavior of the Stochastic Process

Sketches of Proof:

- Let $\mathbf{X}_n = \{X_{n,k} : k = r, r+1, \cdots\}$ be the stochastic process given by

$$X_{n,k} = n^{-1/2} \sum_{i=1}^{n} (\mathbf{1}[Y_i = k] - \pi_k), \quad k = r, r+1, \cdots$$

  Let $\mathbf{X}$ be the Gaussian process with marginal zero mean and the marginal covariance by C_{st}, σ_t^2.
- Define Π_m the coordinate projection given by $\Pi_m \mathbf{Y} = (Y_k : k = r, r+1, \cdots, r+m-1)$ for a stochastic sequence $\mathbf{Y} = (Y_k : k = r, r+1, \cdots)$.
- By Theorem*, the m-dimensional vector $\mathbf{X}_n \circ \Pi_m$ converges in distribution to $\mathbf{X} \circ \Pi_m$ for every positive integer m.

Asymptotic behavior of the Stochastic Process

Sketches of Proof:

- Suffices to show

$$\mathbb{E}f(\mathbf{X}_n) \to \mathbb{E}f(\mathbf{X}), \quad n \to \infty,$$

for every bounded and Lipschitz continuous function f.

- Fix integer m. Then

$$\begin{aligned} |\mathbb{E}f(\mathbf{X}_n) - \mathbb{E}f(\mathbf{X})| &\leq |\mathbb{E}f(\mathbf{X}_n) - \mathbb{E}f(\mathbf{X}_n \circ \Pi_m)| \\ &+ |\mathbb{E}f(\mathbf{X}_n \circ \Pi_m) - \mathbb{E}f(\mathbf{X} \circ \Pi_m)| + |\mathbb{E}f(\mathbf{X} \circ \Pi_m) - \mathbb{E}f(\mathbf{X})|. \end{aligned}$$

Now the last term goes to zero as m tends to infinity by the Lipschitz continuity of f and the boundedness of Gaussian process $\mathbf{X}$. The second term goes to zero by the Portmanteau theorem and Theorem*.

Asymptotic behavior of the Stochastic Process

Sketches of Proof:

- Fix $\epsilon > 0$. For the first term, we have, with L a Lipschitz constant,

$$
\begin{aligned}
|\mathbb{E}f(\mathbf{X}_n) - \mathbb{E}f(\mathbf{X}_n \circ \Pi_m)| &\leq L\epsilon + L\mathcal{P}(\|\mathbf{X}_n - \mathbf{X}_n \circ T_m\| \leq \epsilon) \\
&\leq L\epsilon + L\mathcal{P}(m \leq Y_n^*, \|\mathbf{X}_n - \mathbf{X}_n \circ T_m\| \leq \epsilon) + L\mathcal{P}(m > Y_n^*) \\
&\leq L\epsilon + L\mathbf{1}[m < \infty] + L\mathcal{P}(m > Y_n^*) \to L\epsilon,
\end{aligned}
$$

by first fix m and let $n \to \infty$ and then $m \to \infty$ and noting $Y_n^* \to \infty$ a.s. Here L is the Lipschitz constant. Because ϵ is arbitrary, the desired result follows. $\square$

Mixture of Geometric Distribution

- When $r = 1$, we have the *mixture of geometric distribution* $\mathrm{MGB}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_k : k = 1, 2, ...)$. The probability is

$$\mathcal{P}(Z = z) = \sum_{k=0}^{z-1} (-1)^k \binom{z-1}{k} \lambda_{1+k}, \quad z = 1, 2, ... \tag{1}$$

- Denote F the distribution function under the mixing measure Q. Based on the estimates $\hat{\lambda}_1, \hat{\lambda}_2, ...$, we propose to estimate $F(\theta)$ by

$$\hat{F}_n(\theta) = \sum_{1 \le s \le [n\theta]} \binom{n}{s} (-1)^{n-s} \Delta^{n-s} \hat{\lambda}_s, \quad \theta \in [0, 1]. \tag{2}$$

Unbiasedness

- Because $\hat{\lambda}_i$ is an unbiased estimator of λ_i, we readily have $\mathbb{E}\hat{F}_n(\theta) = F_n(\theta)$, where

$$F_n(\theta) = \sum_{s \le [n\theta]} \binom{n}{s} (-1)^{n-s} \Delta^{n-s} \lambda_s, \quad \theta \in [0,1]. \tag{3}$$

- It is well known that

$$F_n(\theta) \to F(\theta) \tag{4}$$

for every θ in the set $C(F)$ of continuity points of F, see Feller(page 227, 1971).

Consistency

- Accordingly,

Theorem

At every continuous point θ in $C(F)$,

$$\hat{F}_n(\theta) \to F(\theta), \quad a.s.$$

- For $\theta \in [0,1]$, let $V_n(\theta) = n\mathbb{V}ar(\hat{F}_n(\theta) - F_n(\theta))$. Then

$$V_n(\theta) = A_n(\theta) - F_n^2(\theta).$$

where

$$A_n(\theta) = \sum_{i=[n\bar{\theta}]+1}^{n} \pi_i \left(\sum_{s=n-i+1}^{[n\theta]} (-1)^s \binom{n}{s} \binom{s-1}{n-i} \right)^2.$$

Asymptotic Normality

- By CLT,

Theorem

For every $\theta \in [0,1]$, $\hat{F}_n(\theta)$ is asymptotically normal:

$$V_n(\theta)^{-1/2}\sqrt{n}(\hat{F}_n(\theta) - F_n(\theta)) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1).$$

- Tough job 1: $\lim_{n\to\inf} V_n(\theta) = ?$
- Tough job 2: Convergence Rate of the MLE of the mixing measure: $\hat{\lambda}$ determines an estimate $\hat{Q}$ of Q. How to construct $\hat{Q}$? How fast does $\hat{Q}$ converges to Q? In terms of Hellinger distance:

$$h^2(P,Q) = (1/2)\int(\sqrt{dP} - \sqrt{dQ})^2.$$

THANK YOU