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

Mixture of Negative Binomial

Moments of MNB Distribution

Maximum Likelihood Estimation in MNB

Asymptotic Behavior

Estimating the Mixing Measure
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

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

where $\lambda$ 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 *exchangeability*, George and Bowman (1995) proposed the *full likelihood procedure* for analyzing correlated binary data.

Introduction

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

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

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

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

- By the celebrated de Finetti representation theorem,

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

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

- Immediately it follows

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

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

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

\( \{ \lambda_k : k = 0, 1, 2, \ldots \} \) \( \lambda_0 = 1 \) is complete monotone:

\[ (-1)^k \Delta^l \lambda_k \geq 0, \quad l = 0, 1, 2, \ldots \]

where \( \Delta \) 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 \( \{a_1, a_2, \ldots \} \).
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

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

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

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

- Then the variance of $Y$ is simply

$$\text{Var}(Y) = E(Y^2) - (E(Y))^2 = r(r + 1)\lambda_{-2} - r\lambda_{-1} - (r\lambda_{-1})^2.$$
Moments

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

\[
\mathbb{E}(Y) = r \lambda_{-1} = \frac{r}{p},
\]

\[
\text{Var}(Y) = r(r + 1) \lambda_{-2} - r \lambda_{-1} - (r \lambda_{-1})^2 = r(1 - p)/p^2.
\]
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

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

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

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

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

\[
\mathbb{E}(Y) = r\lambda_{-1} = \frac{4r}{3} \quad \text{and}
\]

\[
\text{Var}(Y) = r(r + 1)\lambda_{-2} - r\lambda_{-1} - (r\lambda_{-1})^2 = \frac{2r^2}{9} + \frac{2r}{3}.
\]
Maximum Likelihood Estimation

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

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

where $\lambda_y = (\lambda_r, \ldots, \lambda_y)$. Note that the number of parameters varies with observation $y$.

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

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

where $Y_n^* = \max(Y_1, \ldots, Y_n)$. Assume for now $r$ is known.
The MLE $\hat{\lambda}$ of $\lambda$ is the maximizer of $l_n(\lambda Y_n^*)$ subject to
\[ \lambda_r \leq 1, \quad (-1)^l \Delta^l \lambda_i \geq 0, \quad i \geq r, l \geq 0. \]

Let $\pi_y = P(Y = y) = \binom{y-1}{r-1} p_y$ 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 $c_{t,i} = \binom{t-r}{i-r} / \binom{i-1}{r-1}$. 

Efficient Inference In The Mixture of Negative Binomial Distributions
Maximum Likelihood Estimation

In terms of \( p = \{p_k : k = r, r+1, \ldots\} \), we write \( l_n(\lambda Y_n^*) \) as

\[
\ell_n(p_{Y_n^*}) = \frac{1}{n} \sum_{i=1}^{n} \log p_{Y_i} + C_n,
\]

The MLE \( \hat{p} \) of \( p \) 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

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

where \( A_y = \sum_{i=1}^{n} 1[Y_i = y] \).
Thus, the MLE $\hat{\lambda}$ of $\lambda$ can be obtained as

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

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* \( \lambda_t \): \( \mathbb{E}(\hat{\lambda}_t) = \lambda_t \).

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

\[
\sigma^2_t = n \text{Var}(\hat{\lambda}_t) = \frac{1}{n} \sum_{i=r}^{t} c_{ti}^2 \text{Var}(A_i) = \sum_{i=r}^{t} c_{ti}^2 \pi_i - \lambda_t^2.
\]

- The asymptotic covariance is

\[
C_{st} \equiv n \text{Cov}(\hat{\lambda}_s, \hat{\lambda}_t) = \sum_{i=r}^{s \wedge t} c_{si} c_{ti} \pi_i - \lambda_s \lambda_t, \ s, t = r, r + 1, \ldots.
\]

where \( \sum'_{i \neq j} \) denotes \( \sum_{i=r}^{s} \sum_{j \neq i, j=1}^{t} \) and \( s \wedge t = \min(s, t) \).
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 $\lambda_d = (\lambda_{t_1}, \cdots, \lambda_{t_d})^\top$ and $\hat{\lambda}_{nd} = (\hat{\lambda}_{d_1}, \cdots, \hat{\lambda}_{d_d})^\top$. Denote $\Sigma_d$ the $d \times d$ matrix with the $(i, j)$th entry $C_{t_it_j}$ when $t_i \neq t_j$ and the $(i, i)$ entry $\sigma_{t_i}^2$.

An application of the usual multivariate central limit theorem yields the asymptotic normality.

**Theorem***

$$\sqrt{n}(\hat{\lambda}_d - \lambda_d) \Rightarrow \mathcal{N}(0, \Sigma_d), \quad n \to \infty.$$
We now study the asymptotic efficiency of the stochastic process $\hat{\lambda} = \{\hat{\lambda}_k : k = r, r + 1, \ldots \}$. 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**

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

Efficient Inference In The Mixture of Negative Binomial Distributions
Sketches of Proof:

- Recall that a sequence of random elements $Y_n$ with values in a metric space converges in distribution to a random element $Y$ if

$$\mathbb{E} f(Y_n) \to \mathbb{E} f(Y), \quad n \to \infty$$

for every bounded, continuous $f$ from the metric space to reals $\mathbb{R}$.

- Let $S$ be a nonempty set and $\ell^\infty(S)$ be a set of bounded functions on $S$. Let $\mathcal{P}$ be a collection 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 \( X_n = \{X_{n,k} : k = r, r + 1, \ldots \} \) be the stochastic process given by

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

  Let \( X \) be the Gaussian process with marginal zero mean and the marginal covariance by \( C_{st}, \sigma^2_t \).

- Define \( \Pi_m \) the coordinate projection given by

  \( \Pi_m Y = (Y_k : k = r, r + 1, \ldots, r + m - 1) \) for a stochastic sequence \( Y = (Y_k : k = r, r + 1, \ldots) \).

- By Theorem*, the \( m \)-dimensional vector \( X_n \circ \Pi_m \) converges in distribution to \( X \circ \Pi_m \) for every positive integer \( m \).
Asymptotic behavior of the Stochastic Process

Sketches of Proof:

➤ Suffices to show

\[ \mathbb{E}f(X_n) \to \mathbb{E}f(X), \quad n \to \infty, \]

for every bounded and Lipschitz continuous function \( f \).

➤ Fix integer \( m \). Then

\[
|\mathbb{E}f(X_n) - \mathbb{E}f(X)| \leq |\mathbb{E}f(X_n) - \mathbb{E}f(X_n \circ \Pi_m)| \\
+ |\mathbb{E}f(X_n \circ \Pi_m) - \mathbb{E}f(X \circ \Pi_m)| + |\mathbb{E}f(X \circ \Pi_m) - \mathbb{E}f(X)|.
\]

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

Peng Efficient Inference In The Mixture of Negative Binomial Distributions
Asymptotic behavior of the Stochastic Process

Sketches of Proof:

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

   $$ |\mathbb{E}f(X_n) - \mathbb{E}f(X_n \circ \Pi_m)| \leq L\epsilon + LP(||X_n - X_n \circ T_m|| \leq \epsilon) $$

   $$ \leq L\epsilon + LP(m \leq Y^*_n, ||X_n - X_n \circ T_m|| \leq \epsilon) + LP(m > Y^*_n) $$

   $$ \leq L\epsilon + LP(m > Y^*_n) \rightarrow L\epsilon, $$

   by first fix $m$ and let $n \rightarrow \infty$ and then $m \rightarrow \infty$ and noting $Y^*_n \rightarrow \infty$ a.s. Here $L$ is the Lipschitz constant. Because $\epsilon$ is arbitrary, the desired result follows. \qed
Mixture of Geometric Distribution

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

\[
P(Z = z) = \sum_{k=0}^{z-1} (-1)^k \binom{z-1}{k} \lambda_{1+k}, \quad z = 1, 2, \ldots \quad (1)
\]

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

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

Because \( \hat{\lambda}_i \) is an unbiased estimator of \( \lambda_i \), we readily have

\[
\mathbb{E}\hat{F}_n(\theta) = F_n(\theta),
\]

where

\[
F_n(\theta) = \sum_{s \leq \lfloor n\theta \rfloor} \binom{n}{s} (-1)^{n-s} \Delta^{n-s} \lambda_s, \quad \theta \in [0, 1].
\]

It is well known that

\[
F_n(\theta) \to F(\theta)
\]

for every \( \theta \) in the set \( C(F) \) of continuity points of \( F \), see Feller(page 227, 1971).
Consistency

Accordingly,

**Theorem**

*At every continuous point \( \theta \) in \( C(F) \),*

\[
\hat{F}_n(\theta) \rightarrow F(\theta), \quad \text{a.s.}
\]

**For** \( \theta \in [0, 1] \), let \( V_n(\theta) = n \text{Var}(\hat{F}_n(\theta) - F_n(\theta)) \). Then

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

where

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

- By CLT,

**Theorem**

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

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

- Tough job 1: $\lim_{n \to \infty} V_n(\theta) = ?$
- 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