Brazilian Journal of Probability and Statistics
2018, Vol. 32, No. 3, 583–596
https://doi.org/10.1214/17-BJPS355
© Brazilian Statistical Association, 2018

Sums of possibly associated multivariate indicator functions: The Conway–Maxwell-Multinomial distribution

Joseph B. Kadane[a] **and Zhi Wang**[b]

[a]*Carnegie Mellon University*
[b]*Columbia University*

Abstract. The Conway–Maxwell-Multinomial distribution is studied in this paper. Its properties are demonstrated, including sufficient statistics and conditions for the propriety of posterior distributions derived from it. An application is given using data from Mendel's ground-breaking genetic studies.

1 The Conway–Maxwell-Multinomial distribution

The Conway–Maxwell-Multinomial (COMM) Distribution has probability mass function (for fixed m)

$$P\{\boldsymbol{X}=\boldsymbol{k}|(\boldsymbol{p},\nu)\}=\binom{m}{\boldsymbol{k}}^{\nu}\prod_{i=1}^{r}p_i^{k_i}\Big/\sum_{\boldsymbol{j}\in D}\binom{m}{\boldsymbol{j}}^{\nu}\prod_{i=1}^{r}p_i^{j_i}, \qquad \boldsymbol{k}\epsilon D,$$

where

$$\boldsymbol{p}=(p_1,\dots,p_r), \qquad p_i>0, \qquad \sum_{i=1}^{r}p_i=1,$$

$$\boldsymbol{k}=(k_1,\dots,k_r), \qquad k_i\geq 0, \qquad \sum_{i=1}^{r}k_i=m, \qquad k_i\text{'s integers},$$

$$\binom{m}{\boldsymbol{k}}=\frac{m!}{k_1!k_2!\cdots k_r!}$$

and D is the set of vectors of integers $\boldsymbol{j}$ satisfying $j_i\geq 0$ and $\sum_{i=1}^{r}j_i=m$.

This distribution is a generalization of other distributions as follows:

$r=2$,	$\nu=1$	binomial,
$r>2$,	$\nu=1$	multinomial,
$r=2$,	$\nu\neq 1$	Conway–Maxwell Binomial.

Key words and phrases. Conjugate prior distributions, proper prior distributions, natural exponential family.

Received July 2016; accepted February 2017.

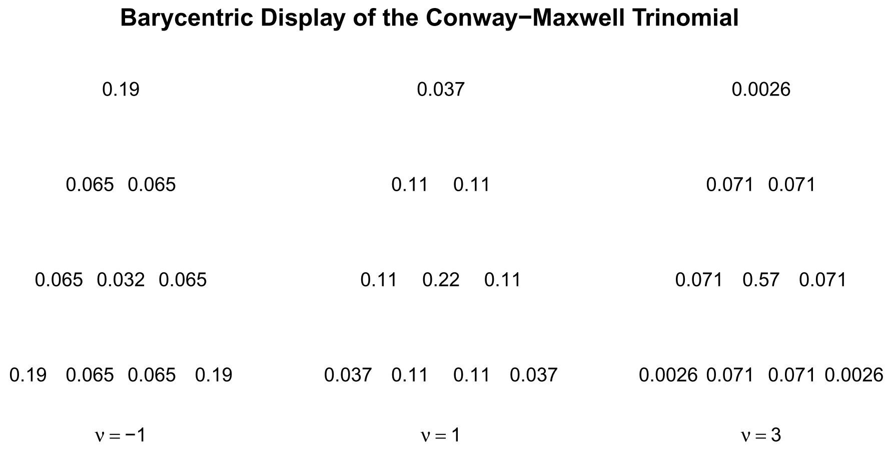

Figure 1

Compared to the multinomial distribution, the novelty is in the parameter ν. As illustrated in Figure 1 (for the special case $m = 3$, $p = (1/3, 1/3, 1/3)$, when $\nu > 1$, the center of the distribution is upweighted relative to the multinomial distribution, indicating negative association among the underlying multivariate indicator functions. Conversely, when $\nu < 1$, the tails of the distribution are upweighted relative to the multinomial distribution, indicating positive association.

The remainder of this paper is organized as follows: Section 2 gives an example of the use of the COMM distribution to a small part of the controversial Mendel data. Since the question there is whether Mendel's data are "too good to be true," *that is*, whether they fit the multinomial hypothesis too well, it seems ideal as an example. Section 3 gives several of the properties of the COMM distribution, results that are generalizations of the principal findings of Kadane (2016) for the Conway–Maxwell Binomial distributions. The proofs are in the Appendix.

2 An example: Mendel's data and Fisher's analysis

Gregor Mendel, an Augustinian monk published (1866) results of his experiments on garden peas, and proposed the model of genetics since known as Mendelian genetics. Fisher (1936), following Weldon (1902), did an analysis suggesting that Mendel's data were "too good to be true," that is, that they fit a binomial or multinomial distribution too well. This finding has led to intense speculation about how that might have occurred. However, since Mendel's papers were destroyed after his death, it is unlikely that additional evidence on this matter will now be discovered. Some recent work on the general matter can be found in Franklin et al. (2008) and Pires and Branco (2010).

Since Fisher is such a dominant figure in both statistics and genetics, it might be useful to give a fast gloss of his argument. Fisher uses chi-square measures of goodness of fit for Mendel's data, treating them as independent across different experiments. By adding the chi-squares and the degrees of freedom, he finds an extra-ordinary degree of coincidence between Mendel's theory and his data. The chi-square analysis of Fisher (started by E. Pearson about 1900) was (and still is) used as the basis of a test of significance. If the chi-square is too large, the null hypothesis is rejected. According to Fisher (1959), if the null hypothesis is rejected, "The force with which such a conclusion is supported is that of the simple disjunction: *Either* an exceptionally rare chance has occurred, *or* the theory of random distribution is not true" (p. 39). Fisher's theory does not permit one to say which of the two possibilities is the case, nor to give a probability for it. Furthermore, if significance is not achieved, nothing can be concluded. In order for the probability distribution that forms the basis of a chi-square test to be valid, the hypothesis to be tested must be declared before the data are examined.

Viewed in this light, there are several gaps between Fisher's calculations and his conclusion. Fisher is rejecting the multinomial null hypothesis if the chi-square is too small, which would be legitimate if the hypothesis test were declared before Weldon pointed the way, or if Fisher routinely used a two-tailed chi-square test. Neither is the case. And one still has Fisher's disjunction to contend with. Nonetheless, Fisher is a superb data-analyst, and we should not be interpreted as challenging his conclusion.

Our intent here is not to enter into these controversial waters, but simply to show the kind of contribution that the COMM distribution might make in this setting. As explained above, when $\nu = 1$ the usual multinomial distribution results. When $\nu < 1$, the tails of the distribution are upweighted relative to the center. Conversely, when $\nu > 1$ the center of the distribution is upweighted relative to the tails. Because ν is a continuous parameter ($-\infty \leq \nu \leq \infty$), the COMM distribution gracefully handles both positive and negative association.

In the case of the Mendel data, the notion that Mendel's data are "too good to be true" is translated into the hypothesis that $\nu > 1$. To illustrate how this works, we took a simple trinomial piece of Mendel's data, namely the first line of Fisher's Table 1 (1936). The data are 36, 60 and 28 for plants in a Bifactoral experiment, where the expected proportions are $(1/4, 1/2, 1/4)$. Along with the COMM likelihood, we must declare a prior. In this case it makes sense to have an opinionated prior on (p_1, p_2 and p_3), namely that we are sure the correct values are, respectively, $(1/4, 1/2, 1/4)$, as predicted by Mendelian genetics. For a prior on ν, we'll take ν to have a unit normal distribution centered at 1. In this way, the prior on ν does not influence the analysis toward or away from the space $\nu > 1$.

The posterior distribution is displayed in Figure 2. Approximately 80% of the probability is above $\nu = 1$, confirming that, in this small piece of Mendel's data, the data are likely to be more negatively associated than they would be were the distribution trinomial. As this example shows, a Bayesian analysis with a COMM likelihood can deliver the probability that Mendel's data are more negatively associated than would be expected under the trinomial hypothesis. Again, this should

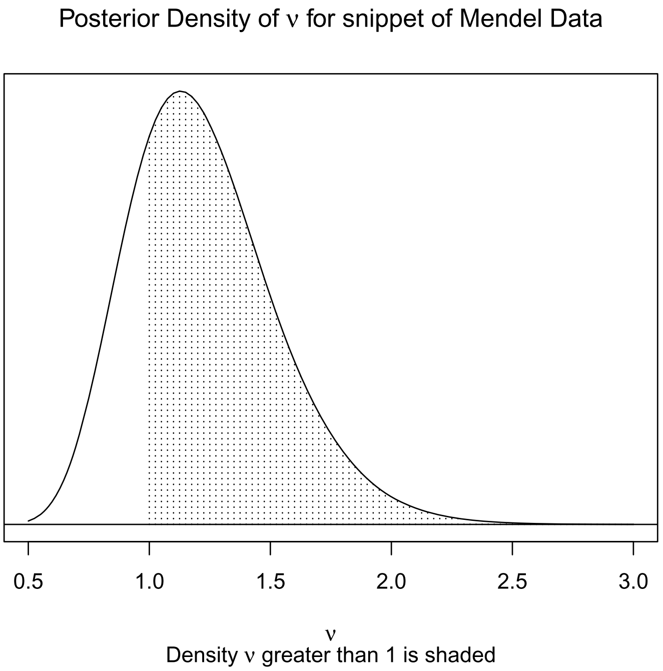

Figure 2

not be taken as a full analysis of Mendel's data, and says nothing about how the data came to be the way they are. A fuller analysis of the Mendel data would address the complete dataset.

3 Properties of the Conway–Maxwell Multinomial distribution

To begin, we should explain how the names Conway and Maxwell came to be associated with this distribution. They wrote a short paper (Conway and Maxwell (1962)) in which they discuss a generalization of the Poisson distribution having pmf proportional to

$$\lambda^x/(x!)^\nu,$$

where $x = 0, 1, 2, \ldots$, and λ and ν are parameters. When $\nu = 1$, the usual Poisson distribution results. This distribution turned out to be useful to model count data because it allows for heavier or lighter tails than the Poisson model (Boatwright, Borle and Kadane (2003), Borle et al. (2005)). Because Conway and Maxwell proposed it first, it became known as the Conway–Maxwell–Poisson distribution (Shmueli et al. (2004)).

It is a simple calculation to show that if X_1 has a Poisson distribution with parameter λ_1, and is independent of X_2, a Poisson random variable with parameter

λ_2, then the distribution of X_1 conditional on the event $X_1 + X_2 = m$ is Binomial with parameter $p = \lambda_1/(\lambda_1 + \lambda_2)$ and n. It is only a slightly more complicated calculation to show the multivariate generalization: If $X_1, X_2, \dots, X_r$ are independent Poisson random variables with means $\lambda_1, \lambda_2, \dots, \lambda_\nu$, then the distribution of $(X_1, X_2, \dots, X_r)$ conditional on the event that $\sum_{i=1}^r X_i = m$ is multinomial with parameter vector $\boldsymbol{p} = (\lambda_1, \lambda_2, \dots, \lambda_r)/\sum_{i=1}^r \lambda_i$ and m.

Shmueli et al. (2004) generalizes the relationship between the Poisson and binomial distributions to the Conway–Maxwell–Poisson as follows: Suppose X_1 has a Conway–Maxwell–Poisson (CMP) distribution with parameters (λ_1, ν), and X_2 is independently distributed CMP (λ_2, ν). Then X_1 conditional on the event $X_1 + X_2 = m$ has a Conway–Maxwell Binomial distribution with parameters $p = \lambda_1/(\lambda_1 + \lambda_2)$, ν and m. It is reasonable to hope, then, that the conditional distribution of r independent Conway–Maxwell–Poisson distributions with respective parameters $\lambda_1, \dots, \lambda_r$ and ν, would have a Conway–Maxwell Multinomial distribution. That this is the case is the content of Proposition 1:

Proposition 1. *Suppose $X_1, \dots, X_r$ are independently distributed with probability mass function Conway–Maxwell Poisson $X_i \sim \mathrm{CMP}(\lambda_i, \nu)$:*

$$P\{X_i = s_i \mid \lambda_i, \nu\} = \frac{\lambda_i^{s_i}}{(s_i!)^\nu Z(\lambda_i, \nu)},$$

where $Z(\lambda_i, \nu) = \sum_{j=0}^\infty \frac{\lambda_i^j}{(j!)^\nu}$.

Then $\boldsymbol{X} \mid \sum_{i=1}^r X_i = m$ has a COMM distribution with parameters $p_i = \lambda_i/\lambda$ and ν, where $\lambda = \sum_{i=1}^r \lambda_i$.

Proof. Let $S = \sum_{i=1}^r X_i$, $\boldsymbol{\lambda} = (\lambda, \dots, \lambda_r)$ and $G(\boldsymbol{p}, \nu) = \sum_{\boldsymbol{j} \in D} \binom{m}{\boldsymbol{j}}^\nu \prod_{i=1}^r p_i^{j_i}$.

Then

$$\begin{aligned} P\{S = m\} &= \sum_{\boldsymbol{j} \in D} \prod_{i=1}^r \frac{\lambda_i^{j_i}}{(j_i!)^\nu Z(\lambda_i, \nu)} \\ &= \frac{\lambda^m}{(m!)^\nu} \cdot \frac{1}{\prod_{i=1}^r Z(\lambda_i, \nu)} \cdot \sum_{\boldsymbol{j} \in D} \prod_{i=1}^r (\lambda_i/\lambda)^{j_i} \binom{m}{\boldsymbol{j}}^\nu \\ &= \frac{\lambda^m}{(m!)^\nu} \frac{1}{\prod_{i=1}^r Z(\lambda_i, \nu)} G(\boldsymbol{\lambda}/\lambda, \nu). \end{aligned}$$

Hence,

$$\begin{aligned} P\{\boldsymbol{X} = \boldsymbol{k} \mid S = m\} &= \prod_{i=1}^r \frac{\lambda_i^{k_i}}{(k_i!)^\nu Z(\lambda_i, \nu)} \Big/ \frac{\lambda^m}{(m!)^\nu} G(\boldsymbol{\lambda}/\lambda, \nu) \prod_{i=1}^r \frac{1}{Z(\lambda_i, \nu)} \\ &= \prod_{i=1}^r (\lambda_i/\lambda)^{k_i} \binom{m}{\boldsymbol{k}}^\nu \Big/ G(\boldsymbol{\lambda}/\lambda, \nu) \qquad \text{for } \boldsymbol{k} \in D \end{aligned}$$

which is the probability mass function of the COMM distribution with parameters $\boldsymbol{p} = \boldsymbol{\lambda}/\lambda$ and ν. □

Proposition 1 implies that there is a constructive way to generate Conway–Maxwell-Multinomial distributions for $\nu > 0$. However, the case for unrestricted ν is not so obvious. The next proposition shows a constructive way of generating (not-necessarily independent) multivariate discrete random variables that have an arbitrary sum.

Proposition 2. *Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be r-dimensional indicator random variables, i.e., each is an r-vector of 0's and 1's. Let $P\{\boldsymbol{S} = \boldsymbol{k}\} = p_{\boldsymbol{k}} \geq 0$, where $\sum_{\boldsymbol{k} \in D} p_{\boldsymbol{k}} = 1$. Then there exists a unique distribution on $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m$ such that $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_m$ are order m exchangeable and $\sum_{i=1}^{m} \boldsymbol{X}_i$ has the same distribution as does $\boldsymbol{S}$.*

Proof. For each $\boldsymbol{k} \in D$, there are $\binom{m}{\boldsymbol{k}}$ different arrangements of 1's and 0's such that each vector component i has k_i 1's and $(m - k_i)$ 0's. Let each such arrangement have probability $p_{\boldsymbol{k}}/\binom{m}{\boldsymbol{k}}$. Then $P(\sum_{i=1}^{r} \boldsymbol{X}_i = \boldsymbol{k}) = p_{\boldsymbol{k}}$ and the $\boldsymbol{X}_i$'s are exchangeable of order m. To show uniqueness, if the sum of the probabilities of the sequences with k_i 1's in the ith vector component for each i were not $p_{\boldsymbol{k}}$, the sum constraint would not be met. If they did not have equal probability, order m exchangeability would be violated. □

It is also useful to display the COMM distribution as a member of the exponential family, and in particular to show its sufficient statistics. This is done in Proposition 3.

Proposition 3. *The COMM distribution has the following sufficient statistics:*

$$S_0 = \sum_{j=1}^{n} \log[k_{ij}! \cdots k_{rj}!],$$

$$S_i = \sum_{j=1}^{n} k_{ij}, \qquad i = 1, \ldots, r-1,$$

where k_{ij} is the ith component of the jth sample. The COMM distribution is a member of the exponential family.

Proof.

$$p(\boldsymbol{k}_1, \ldots, \boldsymbol{k}_n \mid \boldsymbol{p}, \nu) = \prod_{j=1}^{n} \left[\binom{m}{\boldsymbol{k}_j}^{\nu} \prod_{i=1}^{r} p_i^{k_{ij}} / G(\boldsymbol{p}, \nu) \right]$$

$$\propto p_r^{nm} (m!)^{\nu n} \prod_{j=1}^{n} \left[\prod_{i=1}^{r-1} (p_i/p_r)^{k_{ij}} \right] \left[\prod_{i=1}^{r} k_{ij}! \right]^{-\nu}$$

$$\propto \exp\left(\sum_{i=1}^{r-1} \log(p_i/p_r) \sum_{j=1}^{n} k_{ij} - \nu \sum_{j=1}^{n} \log\left(\prod_{i=1}^{r} k_{ij}!\right)\right)$$
$$= e^{\sum_{i=1}^{r-1} \log(p_i/p_r) S_i - \nu S_0}.$$

This shows the sufficient statistics, and also shows that COMM is a member of the exponential family. □

Because the COMM distribution is in the exponential family, it has a conjugate distribution. However, when that distribution is proper is not so obvious, and is the content of Theorem 1 below.

Let $\boldsymbol{\psi} = (\log(p_1/p_r), \dots, \log(p_{r-1}/p_r))$ and $t(\boldsymbol{k}) = -\log(\prod_{i=1}^{r} k_i!)$, then the COMM probability mass function can be re-expressed as

$$P(\boldsymbol{X} = \boldsymbol{k}|\boldsymbol{\psi}, \nu) = e^{\boldsymbol{\psi}\cdot(k_1,k_2,\dots,k_{r-1}) + \nu t(\boldsymbol{k}) - M(\boldsymbol{\theta})},$$

where, $\boldsymbol{\theta} = (\boldsymbol{\psi}, \nu)$, and $M(\boldsymbol{\theta}) = \log(\sum_{\boldsymbol{j}\in D} e^{\boldsymbol{\psi}\cdot(j_1,j_2,\dots,j_{r-1}) + \nu t(\boldsymbol{j})})$.

Consider a conjugate family of the form

$$\pi(\boldsymbol{\theta}|\boldsymbol{a}, b, c) \propto e^{\boldsymbol{\psi}\cdot\boldsymbol{a} + b\nu - cM(\boldsymbol{\theta})}, \tag{1}$$

where $\boldsymbol{a} = (a_1, a_2, \dots, a_{r-1})$.

Then the updating of the hyper parameters using the sufficient statistics can be accomplished using

$$\begin{aligned} a_i' &= a_i + S_i, \qquad i = 1, \dots, r-1, \\ b' &= b - S_0, \\ c' &= c + n. \end{aligned}$$

Let a_r be a pseudo hyperparameter such that $\frac{a_r}{c} = m - \sum_{i=1}^{r-1} \frac{a_i}{c}$, then the conjugate prior π can be written in a symmetric form with respect to $(\boldsymbol{p}, \nu)$.

The Jacobian matrix of transformation is $J = \frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\rho}}$, $\boldsymbol{\rho} = (p_1, p_2, \dots, p_{r-1}, \nu)$, with determinant $|J| = (\prod_{i=1}^{r} p_i)^{-1}$, using the matrix determinant lemma. Therefore, we have

$$\begin{aligned} &h(\boldsymbol{p}, \nu|\boldsymbol{a}, a_r, b, c) \\ &\quad = \pi(\boldsymbol{\theta}|\boldsymbol{a}, b, c) \cdot \|J\| \\ &\quad \propto e^{\boldsymbol{\psi}\cdot\boldsymbol{a} + b\nu - cM(\boldsymbol{\theta})} \left(\prod_{i=1}^{r} p_i\right)^{-1} \\ &\quad = \left[\sum_{\boldsymbol{j}\in D} \frac{1}{(\prod_{i=1}^{r} j_i!)^{\nu}} p_1^{-\frac{a_1}{c}+j_1} \cdots p_{r-1}^{-\frac{a_{r-1}}{c}+j_{r-1}} p_r^{-\frac{a_r}{c}+j_r} e^{-\frac{b}{c}\nu}\right]^{-c} \\ &\qquad \times \left(\prod_{i=1}^{r} p_i\right)^{-1}. \end{aligned} \tag{2}$$

Note that when $\nu \neq 1$ the conjugate prior of this form depends on m, which can be regarded as the total number of trials. When $\nu = 1$, the COMM reduces to the multinomial distribution with conjugate prior the Dirichlet distribution, which does not depend on m.

Theorem 1. *The conjugate prior of COMM distribution in* (2) *is proper if and only if*

$$\text{(i)} \quad c > 0,$$

$$\text{(ii)} \quad a_i > 0, \qquad i = 1, 2, \ldots, r, \qquad \sum_{i=1}^{r} \frac{a_i}{c} = m,$$

$$\text{(iii)} \quad -\log(m!) < \frac{b}{c} < -\sum_{i=1}^{r} \left[\left(\frac{a_i}{c} - \left\lfloor \frac{a_i}{c} \right\rfloor \right) \log \left\lceil \frac{a_i}{c} \right\rceil + \log \left\lfloor \frac{a_i}{c} \right\rfloor ! \right].$$

The proof of Theorem 1 is in the Appendix.

Sometimes conjugate prior distributions are not the most convenient. For example, the snippet of Mendel data is treated with a prior that is not conjugate. Theorem 2 shows that a finite moment generating function suffices.

Theorem 2. *Let $g(\boldsymbol{\psi}, \nu)$ be a probability distribution that has a finite moment generating function. Then the prior proportional to*

$$\exp(\boldsymbol{\psi} \cdot a + \nu b - cM(\boldsymbol{\psi}, \nu)) g(\boldsymbol{\psi}, \nu)$$

is proper for all a, b, and c.

The proof of Theorem 2 is in the Appendix.

The generating functions are straightforward, and worth recording as well:

Let $\boldsymbol{t} = (t_1, \ldots, t_r)$. Then the probability generating function is

$$E(\boldsymbol{t}^{\boldsymbol{x}}) = \sum_{\boldsymbol{j} \in D} \boldsymbol{t}^{\boldsymbol{j}} p^{\boldsymbol{j}} \binom{m}{\boldsymbol{j}}^{\nu} \Big/ G(\boldsymbol{p}, \nu)$$
$$= G(\boldsymbol{t}\boldsymbol{p}, \nu)/G(\boldsymbol{p}, \nu).$$

Similarly, the moment generating and characteristic functions are, respectively,

$$G(e^{\boldsymbol{t}} \boldsymbol{p}, \nu)/G(\boldsymbol{p}, \nu)$$

and

$$G(e^{i\boldsymbol{t}} \boldsymbol{p}, \nu)/G(\boldsymbol{p}, \nu).$$

4 Conclusion

The results of Kadane (2016) extend to the multivariate case (except his Theorem 2). The hard work is in extending Theorem 1 to the multivariate case. The Conway–Maxwell Multinomial distribution can be part of a statistician's toolkit.

Appendix

Diaconis and Ylvisaker (1979) showed that for exponential family $\{P(\boldsymbol{\theta}) : dP(\boldsymbol{\theta}) = e^{\boldsymbol{x}\boldsymbol{\theta}-M(\boldsymbol{\theta})}\,d\mu(\boldsymbol{x}), \theta \in \Theta, \boldsymbol{x} \in \mathbb{R}^d\}$ where μ a is σ-finite measure on $\mathcal{B}(\mathbb{R}^d)$ and $\Theta = \{\theta : M(\theta) < \infty\}$, its conjugate prior family $\{\pi_{n_0,\boldsymbol{x}_0}(\boldsymbol{\theta}) : d\pi_{n_0,\boldsymbol{x}_0}(\boldsymbol{\theta}) = e^{n_0\boldsymbol{x}_0-n_0M(\boldsymbol{\theta})}\,d\boldsymbol{\theta}\}$ has the following property: if $\Theta = \mathbb{R}^d$

$$\pi_{n_0,\boldsymbol{x}_0}(\Theta) < \infty \quad \Leftrightarrow \quad \boldsymbol{x}_0 \in \mathcal{X}, \qquad n_0 > 0,$$

where $\mathcal{X}$ is the interior of the convex hull of μ's support. For our problem, $\mathcal{X}$ is the interior of the convex hull of $\tilde{A} = \{(k_1, k_2, \dots k_{r-1}, t(\boldsymbol{k}) : \boldsymbol{k} \in D\}$. In order to take advantage of the simplicity of symmetric forms, we first add the dimension k_r and consider $A = \{(\boldsymbol{k}, t(\boldsymbol{k})) : \boldsymbol{k} \in D\}$ in the lemmas to follow, and then relate A to $\tilde{A}$ in the proof of Theorem 1.

Here are some notations: $L = \{(z, h) : \sum_{i=1}^r z_i = m\}$ is a subspace of $\mathbb{R}^{r+1}$ and thus we have $A \subseteq \mathrm{Conv}(A) \subseteq L$. Let H denote the interior of the convex hull of A under the subspace topology, then its closure $\bar{H} = \mathrm{Conv}(A)$. Let $\Omega = \{\boldsymbol{x} \in \mathbb{N}^r : m - r + 1 \le \sum_{i=1}^r x_i \le m - 1\}$. If $\boldsymbol{x} \in \Omega$, then the following set $C_{\boldsymbol{x}}$ generated by $\boldsymbol{x}$ is not empty,

$$C_{\boldsymbol{x}} = \left\{\boldsymbol{k} : k_j = x_j \text{ or } x_j + 1, \sum_{j=1}^r k_j = m\right\}.$$

It is clear that $|C_{\boldsymbol{x}}| \ge \binom{r}{1} = r$ and $C_{\boldsymbol{x}} \subseteq D$.

The geometric intuition behind our proofs is that we first divide the "surface" of the convex hull into pieces of facets, then study the hyperplanes that contain these facets. The general steps of proof are:

1. Show that $\{\mathrm{Conv}(C_{\boldsymbol{x}}) : \boldsymbol{x} \in \Omega\}$ exhausts $\mathrm{Conv}(D)$
2. Show that for every $z \in \mathrm{Conv}(C_{\boldsymbol{x}})$, $(z, h) \in \bar{H}$ if and only if $a(z) \le h \le b_{\boldsymbol{x}}(z)$. The graphs of $a(z)$ and $b_{\boldsymbol{x}}(z)$ are, respectively, the bottom facet and a dominant facet indexed by $\boldsymbol{x}$.
3. Relate that above characterization of $\bar{H}$ to the necessary and sufficient conditions for the priors to be proper.

Lemma 1. $\{\mathrm{Conv}(C_{\boldsymbol{x}}) : \boldsymbol{x} \in \Omega\}$ *exhausts* $\mathrm{Conv}(D)$, *that is*, $\bigcup_{\boldsymbol{x}\in\Omega} \mathrm{Conv}(C_{\boldsymbol{x}}) = \mathrm{Conv}(D)$.

Proof. Suffices to show that $\mathrm{Conv}(D) \subseteq \bigcup_{\boldsymbol{x}\in\Omega} \mathrm{Conv}(C_{\boldsymbol{x}})$. If $\boldsymbol{z} \in \mathrm{Conv}(D)$, then $\sum_{i=1}^r z_i = m$. We need to find $\boldsymbol{x} \in \Omega$ such that $\boldsymbol{z} \in \mathrm{Conv}(C_{\boldsymbol{x}})$.

(1) $\sum_{j=1}^r \lfloor z_j \rfloor \leq m-1$.
Let $\boldsymbol{x} = (\lfloor z_1 \rfloor, \dots, \lfloor z_r \rfloor)$, then $\boldsymbol{x} \in \Omega$ as $m - 1 \geq \sum_{j=1}^r \lfloor z_j \rfloor > \sum_{j=1}^r (z_j - 1) = m - r$. We will show that $\mathrm{Conv}(C_{\boldsymbol{x}}) = \{\boldsymbol{z} | x_j \leq z_j \leq x_j + 1, \sum_{j=1}^r z_j = m\} := R_{\boldsymbol{x}}$.

For all $\boldsymbol{z}_1, \boldsymbol{z}_2 \in R_{\boldsymbol{x}}$ and $\lambda \in [0,1]$, we have $\lambda \boldsymbol{z}_1 + (1-\lambda)\boldsymbol{z}_2 \in R_{\boldsymbol{x}}$, therefore, $R_{\boldsymbol{x}}$ is a convex set. $C_{\boldsymbol{x}} \subseteq R_{\boldsymbol{x}} \Rightarrow \mathrm{Conv}(C_{\boldsymbol{x}}) \subseteq R_{\boldsymbol{x}}$.

Next, we prove that elements in $R_{\boldsymbol{x}} \setminus C_{\boldsymbol{x}}$ can not be vertices of $R_{\boldsymbol{x}}$. Assume $\tilde{\boldsymbol{z}} \in R_{\boldsymbol{x}} \setminus C_{\boldsymbol{x}}$ and is a vertex of $R_{\boldsymbol{x}}$. Without loss of generality, let $\tilde{z}_1, \tilde{z}_2$ be such that $\tilde{z}_1 \in (x_1, x_1 + 1)$, $\tilde{z}_2 \in (x_2, x_2 + 1)$. For $\epsilon > 0$ sufficiently small, both $\boldsymbol{z}_\epsilon^+ = (\tilde{z}_1 + \epsilon, \tilde{z}_2 - \epsilon, \dots, \tilde{z}_r)$ and $\boldsymbol{z}_\epsilon^- = (\tilde{z}_1 - \epsilon, \tilde{z}_2 + \epsilon, \dots, \tilde{z}_r)$ are elements of $R_{\boldsymbol{x}}$. Consequently, $\tilde{\boldsymbol{z}} = \frac{1}{2}(\boldsymbol{z}_\epsilon^+ + \boldsymbol{z}_\epsilon^-)$, which contradicts with the assumption that $\tilde{\boldsymbol{z}}$ is a vertex of $R_{\boldsymbol{x}}$. Hence, $R_{\boldsymbol{x}} \subseteq \mathrm{Conv}(C_{\boldsymbol{x}})$.

As a result, $\boldsymbol{z} \in R_{\boldsymbol{x}} = \mathrm{Conv}(C_{\boldsymbol{x}})$.

(2) $\sum_{j=1}^r \lfloor z_j \rfloor = m$.
We know $\sum_{j=1}^r z_j = m$, so $z_1, z_2, \dots, z_r$ must be integers. As $m \geq 1$, there exists $z_i \geq 1$ for some i. Let $\boldsymbol{x} = (z_1, \dots, z_{i-1}, z_i - 1, z_{i+1}, \dots, z_r) \in \Omega$, then $\boldsymbol{z} \in C_{\boldsymbol{x}} \subseteq \mathrm{Conv}(C_{\boldsymbol{x}})$. □

The following proposition is a step towards the proof of Lemma 1.

Proposition 4. *For every $\boldsymbol{x} \in \Omega$ and $\boldsymbol{z} \in D$,*

$$t(\boldsymbol{z}) \leq b_{\boldsymbol{x}}(\boldsymbol{z}) = t(\boldsymbol{x}) - \sum_{i=1}^r (z_i - x_i)\log(x_i + 1).$$

Equality holds if and only if $\boldsymbol{z} \in C_{\boldsymbol{x}}$.

Proof. Let $g(x) = -\log(\Gamma(x+1))$, then $-\log(x_i + 1) = g(x_i + 1) - g(x_i)$, $1 \leq i \leq r$. Also, $t(\boldsymbol{x}) = -\log(\prod_{i=1}^r \Gamma(x_i + 1)) = -\sum_{i=1}^r \log(\Gamma(x_i + 1)) = \sum_{i=1}^r g(x_i)$. Note that g is strictly concave, so $\forall \boldsymbol{x} \in \Omega, \boldsymbol{z} \in D$ and $i \in 1, \dots, r$

$$(z_i - x_i)\big(g(x_i + 1) - g(x_i)\big) \geq g(z_i) - g(x_i).$$

Equality holds if and only if $z_i = x_i$ or $x_i + 1$. Therefore,

$$\sum_{i=1}^r \big(g(x_i) + (z_i - x_i)\big(g(x_i + 1) - g(x_i)\big)\big) \geq \sum_{i=1}^r g(z_i).$$

Consequently,

$$t(\boldsymbol{x}) + \sum_{i=1}^r (z_i - x_i)\big(g(x_i + 1) - g(x_i)\big) \geq t(\boldsymbol{z}).$$

Equality holds if and only if

$$\begin{cases} z_i = x_i \quad \text{or} \quad z_i + 1, & i = 1, 2, \ldots, r, \\ \sum_{i=1}^{r} z_i = m \end{cases} \quad \Leftrightarrow \quad z \in C_{\boldsymbol{x}}. \qquad \square$$

Lemma 2. *For every $\boldsymbol{z} \in \operatorname{Conv}(C_{\boldsymbol{x}})$, $(\boldsymbol{z}, h) \in \bar{H}$ if and only if $-\log(m!) \le h \le b_{\boldsymbol{x}}(z)$.*

Proof. Assume $(\boldsymbol{q}_1, t(\boldsymbol{q}_1)), \ldots, (\boldsymbol{q}_{|A|}, t(\boldsymbol{q}_{|A|}))$ are the elements in A.

(i) $\forall \boldsymbol{k} \in D$, we have $\binom{m}{\boldsymbol{k}} \ge 1$. Hence, $t(\boldsymbol{k}) = \log\binom{m}{\boldsymbol{k}} - \log(m!) \ge -\log(m!)$. If $(\boldsymbol{z}, h) \in \bar{H} = \operatorname{Conv}(A)$, then there exists $\alpha_j \ge 0$, $\sum_{j=1}^{|A|} \alpha_j = 1$ such that $(z, h) = \sum_{j=1}^{|A|} \alpha_j (\boldsymbol{q}_j, t(\boldsymbol{q}_j))$. We have $h = \sum_{j=1}^{|A|} \alpha_j t(\boldsymbol{q}_j) \ge -\log(m!) \sum_{j=1}^{|A|} \alpha_j = -\log(m!)$. Again, by Proposition 4 and linearity of $b_{\boldsymbol{x}}(\boldsymbol{z})$, we get $b_{\boldsymbol{x}}(\boldsymbol{z}) = b_{\boldsymbol{x}}(\sum_{j=1}^{|A|} \alpha_j \boldsymbol{q}_j) = \sum_{j=1}^{|A|} \alpha_j b_{\boldsymbol{x}}(\boldsymbol{q}_j) \ge \sum_{j=1}^{|A|} \alpha_j t(\boldsymbol{q_j}) = h$.

(ii) Here we show the "only if" part of the proof. Suffices to show that $B = \{(\boldsymbol{z}, h) : \boldsymbol{z} \in \operatorname{Conv}(D), h = -\log(m!)\} \subseteq \bar{H}$ and $S_{\boldsymbol{x}} = \{(\boldsymbol{z}, h) : \boldsymbol{z} \in \operatorname{Conv}(C_{\boldsymbol{x}}), h = b_{\boldsymbol{x}}(\boldsymbol{z})\} \subseteq \bar{H}$. Let $\boldsymbol{e}_i$ denote the unit vector with a 1 in the ith coordinate and 0 elsewhere. Now consider $m\boldsymbol{e}_i \in D, 1 \le i \le r$. Note that $t(m\boldsymbol{e}_i) = -\log(m!)$ and $\operatorname{Conv}(m\boldsymbol{e}_i : 1 \le i \le r) = \operatorname{Conv}(D)$, thus we have $B \subseteq \bar{H}$. From Proposition 4, we know for every $\boldsymbol{z} \in C_{\boldsymbol{x}}, t(\boldsymbol{z}) = b_{\boldsymbol{x}}(\boldsymbol{z})$. Again, $\boldsymbol{z}$ can be written as $\sum_j \lambda_j \boldsymbol{k}_j$, where $\boldsymbol{k}_j \in C_{\boldsymbol{x}}$, $\lambda_j \ge 0$ and $\sum_j \lambda_j = 1$. Hence, $b_{\boldsymbol{x}}(\boldsymbol{z}) = b_{\boldsymbol{x}}(\sum_j \lambda_j \boldsymbol{k}_j) = \sum_j \lambda_j b_{\boldsymbol{x}}(\boldsymbol{k}_j) = \sum_j \lambda_j t(\boldsymbol{k}_j)$, which implies $S_{\boldsymbol{x}} \subseteq \operatorname{Conv}((\boldsymbol{k}_j, t(\boldsymbol{k}_j)) : \boldsymbol{k}_j \in C_{\boldsymbol{x}}) \subseteq \bar{H}$. □

Using these results, we now resume the proof of Theorem 1.

Proof of Theorem 1. Because there are only finite number of elements in D, the parameter space $\Theta = \{\boldsymbol{\theta} | M(\boldsymbol{\theta}) < \infty\} = \mathbb{R}^r$. Let $\tilde{H}$ denote the interior of the convex hull of $\tilde{A}$. The conjugate prior is proper, according to Diaconis and Ylivisacker's theorem, if and only if

$$\text{(i)} \quad c > 0, \qquad \text{(ii)} \quad \left(\frac{a_1}{c}, \frac{a_2}{c}, \ldots, \frac{a_{r-1}}{c}, \frac{b}{c}\right) \in \tilde{H}. \tag{A.1}$$

The following mapping $\boldsymbol{T}$ between $\tilde{A}$ and A is bijective,

$$\boldsymbol{T} : \tilde{A} \to A,$$

$$\tilde{\boldsymbol{q}}_i = (k_1^{(i)}, \ldots, k_{r-1}^{(i)}, t(\boldsymbol{k}^{(i)})) \mapsto \boldsymbol{q}_i = (\boldsymbol{k}^{(i)}, t(\boldsymbol{k}^{(i)})).$$

We can therefore convert the problem to a symmetric one by adding one more dimension. As $\boldsymbol{T}$ is bijective, we have $|\tilde{A}| = |A| := K$.

$$\begin{aligned}
&(\tau_1, \tau_2, \dots, \tau_r) \in \mathrm{Conv}(\tilde{A}) \\
\Leftrightarrow\quad & \exists \alpha_1, \alpha_2, \dots, \alpha_K \geq 0, \qquad \sum_{j=1}^{K} \alpha_j = 1 \\
& \text{s.t. } (\tau_1, \tau_2, \dots, \tau_r) = \sum_{j=1}^{K} \alpha_j \tilde{\boldsymbol{q}}_j \\
\Leftrightarrow\quad & \exists \alpha_1, \alpha_2, \dots, \alpha_K \geq 0, \qquad \sum_{j=1}^{K} \alpha_j = 1 \\
& \text{s.t. } \left(\tau_1, \tau_2, \dots, \tau_{r-1}, m - \sum_{j=1}^{r-1} \tau_j, \tau_r\right) = \sum_{j=1}^{K} \alpha_j \boldsymbol{q}_j \\
\Leftrightarrow\quad & \left(\tau_1, \tau_2, \dots, \tau_{r-1}, m - \sum_{j=1}^{r-1} \tau_j, \tau_r\right) \in \mathrm{Conv}(A).
\end{aligned}$$

Hence,

$$\begin{aligned}
&(\tau_1, \tau_2, \dots, \tau_r) \in \tilde{H} \\
\Leftrightarrow\quad & \exists \tilde{\varepsilon} > 0, \qquad \text{s.t. } B_{\tilde{\varepsilon}}(\tau_1, \tau_2, \dots, \tau_r) \subseteq \mathrm{Conv}(\tilde{A}) \\
\Leftrightarrow\quad & \exists \varepsilon > 0, \qquad \text{s.t. } L \cap B_{\varepsilon}\left(\tau_1, \dots, \tau_{r-1}, m - \sum_{j=1}^{r-1} \tau_j, \tau_r\right) \subseteq \mathrm{Conv}(A) \\
\Leftrightarrow\quad & \left(\tau_1, \dots, \tau_{r-1}, m - \sum_{j=1}^{r-1} \tau_j, \tau_r\right) \in H,
\end{aligned}$$

where $B_{\tilde{\varepsilon}}(\tau_1, \tau_2, \dots, \tau_r)$ is an open r dimensional sphere centered at $(\tau_1, \tau_2, \dots, \tau_r)$ with radius $\tilde{\varepsilon}$ and $B_{\varepsilon}(\tau_1, \dots, \tau_{r-1}, m - \sum_{j=1}^{r-1} \tau_j, \tau_r)$ is an open $r+1$ dimensional sphere centered at $(\tau_1, \dots, \tau_{r-1}, m - \sum_{j=1}^{r-1} \tau_j, \tau_r)$ with radius ε. As a result,

$$\left(\frac{a_1}{c}, \frac{a_2}{c}, \dots, \frac{a_{r-1}}{c}, \frac{b}{c}\right) \in \tilde{H} \quad \Leftrightarrow \quad \left(\frac{a_1}{c}, \frac{a_2}{c}, \dots, \frac{a_r}{c}, \frac{b}{c}\right) \in H. \tag{A.2}$$

Let $F(z) = \sum_{i=1}^{r} [(z_i - \lfloor z_i \rfloor) \log \lceil z_i \rceil + \log \lfloor z_i \rfloor !]$. We claim that

$$\bar{H} = \left\{(z, h) : -\log(m!) \leq h \leq F(z), z_j \geq 0, \sum_{j=1}^{r} z_j = m\right\}.$$

To show $\bar{H} \subseteq RHS$, assume $(z, h) \in \bar{H} = \mathrm{Conv}(A)$, then $z \in \mathrm{Conv}(D) = \{z : \sum_{j=1}^{r} z_j = m, z_j \geq 0\}$. According to Lemma 1, there exists $\boldsymbol{x} \in \Omega$ such that $z \in \mathrm{Conv}(C_{\boldsymbol{x}})$. Apply Lemma 2, we get $-\log(m!) \leq h \leq b_{\boldsymbol{x}}(z)$. Now, choose $\boldsymbol{x}$ as suggested by the proof of Lemma 1. If $\sum_{j=1}^{r} \lfloor z_j \rfloor \leq m-1$, let $\boldsymbol{x} = (\lfloor z_1 \rfloor, \ldots, \lfloor z_r \rfloor)$, then we have $b_{\boldsymbol{x}}(z) = F(z)$. Otherwise, $\sum_{j=1}^{r} \lfloor z_j \rfloor = m$, we have $z \in C_{\boldsymbol{x}}$ for every possible choice of $\boldsymbol{x}$, so $b_{\boldsymbol{x}}(z) = t(z) = F(z)$ holds.

Conversely, we show that $RHS \subseteq \bar{H}$. If $\sum_{j=1}^{r} z_j = m$ and $z_j \geq 0$ for $1 \leq j \leq r$, then $z \in \mathrm{Conv}(D)$ which implies $z \in \mathrm{Conv}(C_{\boldsymbol{x}})$ for some $\boldsymbol{x}$ by Lemma 1. We have seen that $F(z) = b_{\boldsymbol{x}}(z)$. Again, use Lemma 2, we get $(z, h) \in \bar{H}$.

Consequently,

$$H = \left\{ (z, h) \middle| -\log(m!) < h < F(z), z_j > 0, \sum_{j=1}^{r} z_j = m \right\}. \tag{A.3}$$

Now, let $z = (\frac{a_1}{c}, \ldots, \frac{a_r}{c})$, we have

$$F\left(\frac{a_1}{c}, \frac{a_2}{c}, \ldots, \frac{a_r}{c}\right) = -\sum_{i=1}^{r} \left[\left(\frac{a_i}{c} - \left\lfloor \frac{a_i}{c} \right\rfloor \right) \log \left\lceil \frac{a_i}{c} \right\rceil + \log \left\lfloor \frac{a_i}{c} \right\rfloor ! \right]. \tag{A.4}$$

By combining equation (A.1), (A.2), (A.3) and (A.4), we get Theorem 1. □

Proof of Theorem 2. Let $h_{\boldsymbol{j}}(\boldsymbol{\theta}) = \boldsymbol{\psi} \cdot (\boldsymbol{j}_1, \boldsymbol{j}_2, \ldots, \boldsymbol{j}_{r-1}) + \nu t(\boldsymbol{j})$, which is linear in $\boldsymbol{\theta}$. Let $R_{\boldsymbol{j}} = \{\boldsymbol{\theta} \mid \mathrm{sgn}(c) h_{\boldsymbol{j}}(\boldsymbol{\theta})$ is minimized over $\boldsymbol{j} \epsilon D\}$.

The $R_{\boldsymbol{j}}$'s are exhaustive and measurable. Let $D^* = \{\boldsymbol{j} \mid R_{\boldsymbol{j}}(\boldsymbol{\theta}) \neq \varnothing\}$. D^* is not empty.

Choose $\boldsymbol{j}^* \epsilon D^*$. For $\boldsymbol{\theta} \epsilon R_{\boldsymbol{j}^*}$,

$$\mathrm{sgn}(c) \sum_{\boldsymbol{j} \epsilon D} \exp(h_{\boldsymbol{j}}(\boldsymbol{\theta})) \geq \mathrm{sgn}(c) |D| \exp(h_{\boldsymbol{j}^*}(\boldsymbol{\theta})),$$

so

$$\mathrm{sgn}(c) M(\boldsymbol{\theta}) \geq \mathrm{sgn}(c) K h_{\boldsymbol{j}^*}(\boldsymbol{\theta}),$$

where $K = \log(|D|) > 0$.

Hence, $\int_{R_{\boldsymbol{j}^*}} \exp(\boldsymbol{\psi} \cdot \boldsymbol{a} + \nu b - cM(\boldsymbol{\theta})) g(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \int \exp(\boldsymbol{\psi} \cdot \boldsymbol{a} + \nu b - cK h_{\boldsymbol{j}^*}(\boldsymbol{\theta})) \times g(\boldsymbol{\theta}) \, d\boldsymbol{\theta} < \infty$ because $\boldsymbol{\psi} \cdot \boldsymbol{a} + \nu b - cK h_{\boldsymbol{j}^*}(\boldsymbol{\theta})$ is linear in $\boldsymbol{\theta}$ and g has a finite moment generating function. It follows that $\int \exp(\boldsymbol{\psi} \cdot \boldsymbol{a} + \nu b - cM(\boldsymbol{\theta})) g(\boldsymbol{\theta}) \, d\boldsymbol{\theta} \leq \sum_{\boldsymbol{j}^* \epsilon D^*} \int_{R_{\boldsymbol{j}^*}} \exp(\boldsymbol{\psi} \cdot \boldsymbol{a} + \nu b - cM(\boldsymbol{\theta})) g(\boldsymbol{\theta}) \, d\boldsymbol{\theta} < \infty$. □

Acknowledgment

We thank a very helpful referee for simplifying the proof of our theorem.

References

Boatwright, P., Borle, S. and Kadane, J. B. (2003). A model of the joint distribution of purchase quantity and timing. *J. Amer. Statist. Assoc.* **98**, 564–572. MR2011672

Borle, S., Boatwright, P., Kadane, J. B., Nunes, J. C. and Shmueli, G. (2005). The effect of product assortment changes on customer retention. *Mark. Sci.* **4**, 616–622.

Conway, R. and Maxwell, W. (1962). A queing model with state dependent service rates. *J. Ind. Eng.* **12**, 132–136.

Diaconis, P. and Ylvisaker, D. (1979). Conjugate priors for exponential families. *Ann. Statist.* **7**, 269–281. MR0520238

Fisher, R. A. (1936). Has Mendel's work been rediscovered? *Ann. of Sci.* **1**, 115–137.

Fisher, R. A. (1959). *Statistical Methods and Scientific Inference*, 2nd ed. Edinburgh: Oliver and Boyd.

Franklin, A., Edwards, A. W. F., Fairbanks, D. J., Hartl, D. L. and Seidenfeld, T. (2008). *Ending the Mendel–Fisher Controversy*. Pittsburgh: University of Pittsburgh Press.

Kadane, J. B. (2016). Sums of possibly associated Bernoulli variables: The Conway–Maxwell-Binomial distribution. *Bayesian Anal.* **11**, 403–420.

Pires, A. M. and Branco, J. A. (2010). A statistical model to explain the Mendel–Fisher controversy. *Statist. Sci.* **25**, 545–565. MR2807770

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S. and Boatwright, P. (2004). A useful distribution for fitting discrete data: Revival of the COM-Poisson. *J. R. Stat. Soc., Ser. C, Appl. Stat.* **54**, 127–142. MR2134602

Weldon, W. R. F. (1902). Mendel's law of alternative inference in peas. *Biometrika* **1**, 228–254.

Department of Statistics
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
USA
E-mail: kadane@stat.cmu.edu

Department of Statistics
Columbia University
New York, New York 10027
USA
E-mail: zw2393@columbia.edu