کاربرد روش‌های خودگردان‌ساز در استنباط بیزی

مریم خان‌احمدی
دانشکده علوم ریاضی، دانشگاه شاهرود

چکیده

خودگردان‌ساز (بوت‌استرپ) از روش‌های باز نمونه‌گیری است که امروزه کاربرد گسترده‌ای در علم آمار دارد. این روش انواع مختلفی دارد. در این مقاله روش خودگردان‌ساز بیزی، که خود شامل روش‌های ناپارامتری و پارامتری است مورد بررسی قرار گرفته، و روش به‌دست آوردن آن‌ها در مثال‌های جداگانه ارائه شده است.

واژه‌های کلیدی: خودگردان‌ساز بیزی، استنباط ناپارامتری.

۱ تاریخچه

خودگردان‌ساز[۱] رده بزرگی از روش‌هاست، که در آن‌ها بازنمونه‌گیری[۲] از مجموعه داده‌های اولیه انجام شده و به همین دلیل روش‌های باز نمونه‌گیری نامیده می‌شوند. روش خودگردان‌ساز برای اولین بار توسط افرون[۳] [۵] در مقاله‌ای با عنوان «خودگردان‌ساز: نگاهی دیگر به روش جک نایف» معرفی و از آن برای برآورد اریبی، واریانس و توزیع نمونه‌ای آماره‌ها استفاده شد. خودگردان‌ساز معرفی شده توسط افرون خودگردان‌ساز ناپارامتری بود.

بعد از معرفی روش خودگردان توسط افرون، انواع مختلف خودگردان‌ساز پارامتری، خودگردان‌ساز هموار، خودگردان‌ساز دوگانه و خودگردان‌ساز بیزی معرفی شدند. خودگردان‌ساز بیزی برای اولین بار در سال ۱۹۸۱ توسط روبین [۴] [۱۵] معرفی شد و از آن برای برآورد توزیع پسین پارامتر مورد نظر استفاده کرد. لو [۵] [۱۰، ۱۱] این روش را به جوامع متناهی و به داده‌های سانسور شده تعمیم داد. خودگردان‌ساز بیزی و خودگردان‌ساز وزنی در مقاله‌های متعددی توسط: افرون [۵]، لو [۹]،

[۱] Bootstrap
[۲] Resamlping
[۳] Efron
[۴] Rubin
[۵] Lo

ونگ [۶] [۱۶]، میسون و نیوتن [۷] [۱۲]، پیریستگارد و ولنر [۸] [۱۴]، گاسپارینی [۹] [۷] و جیمز [۱۰] [۸] مورد مطالعه قرار گرفت. بنکز [۱۱] [۳] در مطالعات شبیه سازی از خودگردان‌ساز بیزی و نوع هموار آن استفاده کرد. خودگردان‌ساز بیزی ناپارامتری توسط روبین [۱۵]، اوون [۱۲] [۱۳] مورد بررسی قرار گرفت.

۲ مقدمه

فرض کنید $\mathbf{X} = (X_1, \ldots, X_n)$ یک نمونه تصادفی از توزیع نامعلوم F با پارامتر θ و $x = (x_1, \ldots, x_n)$ یافته آن باشد. آماره $\hat{\theta} = \hat{\theta}(X)$ را به عنوان برآوردگر پارامتر θ در نظر می‌گیریم.

هدف برآورد توزیع نمونه‌ای آماره $\hat{\theta} = \hat{\theta}(X)$ می‌باشد. چون معمولاً از $\hat{\theta}$ برای ساختن فواصل اطمینان برای پارامتر نامعلوم θ (θ وابسته به F است) استفاده می‌کنیم به جای $\hat{\theta}$ یک ریشه آن یعنی $R_F = R(X_1, \ldots, X_n, F)$ را در نظر می‌گیریم. ریشه R_F ممکن است به صورت $\hat{\theta} - \theta$ باشد. پس نیاز داریم که توزیع نمونه‌ای $R_F = R(X, F) = \hat{\theta} - \theta$ را به‌دست آوریم، که این کار با استفاده از روش خودگردان‌ساز ناپارامتری امکان پذیر است. با محاسبه آماره خودگردان‌ساز $\hat{\theta}^* = \hat{\theta}(X^*)$ می‌توان از توزیع $R(X^*, F_n) = \hat{\theta}^* - \hat{\theta}$ به عنوان تقریب توزیع نمونه‌ای $R(X, F)$ استفاده کرد.

خودگردان‌ساز بیزی

خودگردان‌ساز بیزی [۱۳] (BB) مشابه خودگردان است، با این تفاوت که در روش خودگردان‌ساز توزیع نمونه‌ای آماره شبیه سازی می شود، درصورتی که در روش BB توزیع پسین پارامتر شبیه سازی می‌شود.

در استنباط بیزی فرض می‌شود θ متغیر تصادفی است که دارای یک توزیع پیشین است، و $\mathbf{X} = (X_1, \ldots, X_n)$ یک نمونه تصادفی از F (توزیع شرطی X به شرط θ) است. استنباط بیزی بر اساس توزیع پسین $R(\mathbf{X}, F)$ به شرط X است. اگر F متعلق به یک خانواده پارامتری با پارامترθ باشد، توزیع پسین $R(\mathbf{X}, F)$ در صورتی قابل محاسبه است که توزیع پیشین معلوم باشد.

۳ خودگردان‌ساز بیزی ناپارامتری

در این بخش ابتدا الگوریتم محاسبه خودگردان‌ساز بیزی ناپارامتری را ارائه می‌کنیم.[۱]

روبین [۱۵] یک روش خودگردان‌ساز بیزی ناپارامتری را در حالتی که توزیع پیشین نامعلوم است، ارایه کرد. هدف در این روش، پیدا کردن توزیع پسین $R(\mathbf{X}, F)$ است.

ابتدا $x_n, \ldots, x_1$ به عنوان یک نمونه تصادفی مستقل و هم توزیع از متغیرهای تصادفی $X_n, \ldots, X_1$ با تابع توزیع F و تابع توزیع تجربی F_n را در نظر بگیرید. فرض کنید θ برآورد مبتنی بر $x_n, \ldots, x_1$ است. می‌دانیم که خودگردان‌ساز ناپارامتری می‌تواند برای تقریب توزیع $\hat{\theta}$ به کار برده شود. هر تکرار خودگردان‌ساز بیزی به صورت یک احتمال پسین برای x_i است.

[۶] Weng
[۷] Mason and Newton
[۸] Preastgaard and Wellner
[۹] Gasparini
[۱۰] James
[۱۱] Banks
[۱۲] Owen
[۱۳] Bayesian Bootstrap

لو [۹] ساختار کلی روش BB را به صورت زیر در نظر گرفت.
$(n-۱)$ متغیر تصادفی در فاصله [۰،۱] در نظر بگیرید آماره‌های ترتیبی $u_{(n-۱)},\dots,u_{(۱)}$ به طوری که $u_{(۰)}=۰, u_{(n)}=۱$ را تشکیل می‌دهیم (چرنیک [۴]) و کمیت‌های

$$g_i = u_{(i)} - u_{(i-۱)}, \quad i = ۱,\dots,n,$$

را محاسبه می‌کنیم. g_iها اختلاف بین آماره‌های مرتب یکنواخت به صورت بردار $g=(g_۱,\dots,g_n)$ هستند که به عنوان بردار احتمالات برای مقادیر $x_n,\dots,x_۱$ در هر تکرار BB در نظر می‌گیریم. برای مثال n مشاهده با نمونه‌گیری با جایگذاری از $x_n,\dots,x_۱$ انتخاب شده اند اما به جای اینکه هر x_i دقیقا با احتمال $\frac{۱}{n}$ انتخاب شوند، از احتمال g_i انتخاب می‌شوند یعنی $x_۱$ از $g_۱$، $x_۲$ از $g_۲$ و ... . بنابراین با تکرار خودگردان‌ساز، یک مجموعه جدید از $n-۱$ عدد تصادفی یکنواخت از g_i ها داریم. اگر D_n تابع توزیع تصادفی با وزن‌های g_i برای هر x_i باشد، آنگاه با محاسبه $\bar{\theta}=\theta(D_n)$ می‌توان از توزیع شرطی $R(X,D_n)=\hat{\theta}-\bar{\theta}$ (به شرط X) به عنوان تقریب توزیع پسین $R(\mathbf{X},F)$ استفاده کرد.
به عنوان مثال فرض کنید θ، میانگین X است، می‌خواهیم توزیع پسین θ را به‌دست آوریم. در روش خودگردان‌ساز بیزی، بردار $g=(g_۱,\dots,g_n)$ را محاسبه می‌کنیم. با فرض اینکه $\theta=\mu$، هر تکرار BB یک میانگین به صورت $\bar{\mu}=\sum_{i=۱}^{n} g_i x_i$ تولید می‌کند. $\bar{\mu}=\sum_{i=۱}^{n} g_i x_i$ روی تمام تکرارهای BB توزیع بوت استرپی میانگین X است که به عنوان تقریب توزیع پسین μ می‌تواند استفاده شود.
حال از روش مونت‌کارلو برای تقریب توزیعBB، $\bar{\mu}$ استفاده می‌کنیم. مرحله بالا را به‌طور مستقل B بار تکرار می‌کنیم، در نهایت $\bar{\mu}^{B},\dots,\bar{\mu}^{۱}$ می‌تواند به عنوان تقریب توزیع خودگردان‌ساز بیزی $\bar{\mu}$ و در نتیجه تقریب توزیع پسین μ استفاده شود. (ایران‌پناه [۱])
از نظر کاربردی دو روش خودگردان‌ساز و BB تنها در تخصیص احتمالات به هر x_i با هم اختلاف دارند. افرون [۵] ثابت کرد که اگر

$$n_i = \text{فراوانی } x_i \text{ ها در نمونه خودگردان ساز},$$

آنگاه بردار $n=(n_۱,\dots,n_n)$ دارای توزیع چند جمله ای است که احتمال $\frac{۱}{n}$ را به هر جمله نسبت می‌دهد، و اگر

$$f_i = \text{فراوانی نسبی } x_i \text{ ها در نمونه خودگردان ساز},$$

آن گاه بردار $\mathbf{f}=(f_۱,\dots,f_n)$ دارای توزیع چند جمله‌ای $(n,\frac{۱}{n},\dots,\frac{۱}{n})$ می‌باشد که

$$\begin{aligned} E(f_i) &= \frac{۱}{n}, \\ Var(f_i) &= \frac{n-۱}{n^{۳}}, \\ Cov(f_i,f_j) &= -\frac{۱}{n^{۳}}, \quad i \neq j, \\ Corr(f_i,f_j) &= -\frac{۱}{n-۱}, \quad i \neq j. \end{aligned}$$

در روش BB، $g_i = u_{(i)} - u_{(i-۱)}$ دارای توزیع $Beta(۱,n-۱)$ می‌باشد (بهبودیان [۲])، آنگاه بردار $g=(g_۱,\dots,g_n)$ دارای توزیع دیریکله $(۱,\dots,۱)$ است.

(ویلکس [۱۷])

$$
\begin{aligned}
E(g_i) =& \frac{۱}{n},\\
Var(g_i) =& \frac{n-۱}{n^۲(n+۱)},\\
Cov(g_i,g_j) =& -\frac{۱}{n^۲(n+۱)}, \quad i \neq j,\\
Corr(g_i,g_j) =& -\frac{۱}{n-۱}, \quad i \neq j.
\end{aligned}
$$

اکنون به بررسی مثالی از روبین [۱۵] می‌پردازیم.

مثال ۱.۳. *فرض می‌کنیم X متغیری دو حالتی است (۰ یا ۱)، که پارامتر مورد بررسی (θ)، بردار احتمال $X = ۱$ است، همچنین $n_۱$ را تعداد $x_i = ۱$ و $n - n_۱$ را تعداد $x_i = ۰$ در نظر می‌گیریم.*
یک تکرار BB ، ابتدا n احتمال را از n (g_i) تولید شده از $n - ۱$ متغیر تصادفی از توزیع یکنواخت در بازه ی $(۰, ۱)$ $(U(۰, ۱))$ به‌دست آورده و سپس این احتمالات را به مقادیر مشاهده شده X نسبت می‌دهد، بنابراین یک تکرار BB، $n_۱$ احتمال را به $x_i = ۱$ $(P_۱ = \sum_{i=۱}^{n} g_i I(X_i = ۱))$ و $n - n_۱$ احتمال باقیمانده را به $x_i = ۰$ نسبت می‌دهد، در نتیجه مقدار BB پارامتر مورد بررسی، $P_۱$ است. تکرار بعدی BB، به همین روش اما با انتخاب $n - ۱$ متغیر جدید از توزیع یکنواخت در بازه ی $(۰, ۱)$ $(U(۰, ۱))$ و محاسبه مقدار جدید $P_۱$ به‌دست می‌آید، با ادامه تکرارها، توزیع بوت‌استرپی بیزی θ به‌دست می‌آید.
به دلیل این‌که g_iها دارای توزیع $Beta(۱, n-۱)$ می‌باشند، در نتیجه $P_۱$ که مجموع $n_۱$، g_i است دارای توزیع $Beta(n_۱, n - n_۱)$ است. بنابراین توزیع BB پارامتر مورد بررسی، $Beta(n_۱, n-n_۱)$ می‌باشد. (روبین [۱۵]) یادآوری می‌کنیم که $Beta(n_۱, n - n_۱)$ توزیع پسین θ در این نمونه است.

۴ خودگردان‌ساز بیز پارامتری

در حالت پارامتری ارتباط بین خودگردان‌ساز بیزی و غیر بیزی شفاف‌تر است.
فرآیند خودگردان‌ساز بیزی پارامتری را براساس یک مساله تک پارامتری ساده مطرح می‌کنیم.

مثال ۱.۴. *جدول ۱ شامل نمرات ۲۲ دانش‌آموز در دو درس بردار و مکانیک است.*
$\hat{\theta} = ۰/۴۹۸$، برآورد ضریب همبستگی نمونه‌ای این دو درس است. می‌خواهیم توزیع پسین پارامتر $\theta_۰ = corr(mech, vec)$ را محاسبه کنیم.

فرض می‌کنیم که نمرات هر دانش‌آموز به صورت

$$\mathbf{Y_i} = (y_{i۱}, y_{i۲}) \sim N_۲(\mu, \Sigma) \quad ; i = ۱, ۲, \ldots, ۲۲,$$

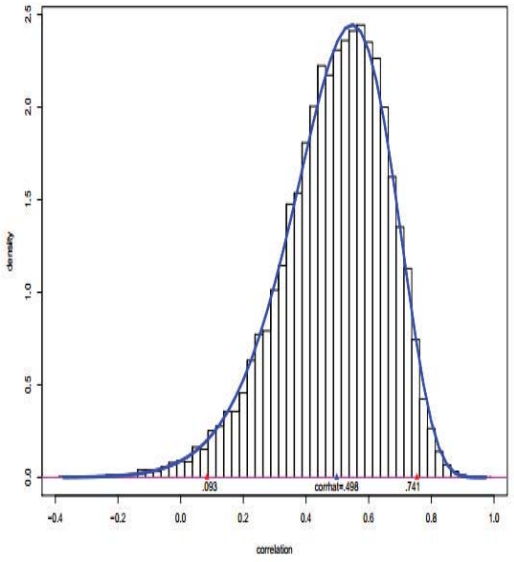
*که $(\hat{\mu}, \hat{\Sigma})$ برآورد درستنمایی ماکسیمم ،**(MLE)** (μ, Σ) هستند. بنابراین از توزیع $N_۲(\hat{\mu}, \hat{\Sigma})$ تحت خودگردان‌ساز پارامتری نمونه‌های تکراری تولید می‌کنیم و برای هر نمونه y^*، پاسخ خودگردان‌ساز $\hat{\theta}^*$، را محاسبه می‌کنیم.*
شکل ۱ بافت نگار ۱۰۰۰۰ تکرار خودگردان‌ساز برای ضریب همبستگی نمرات دانش‌آموزان را نشان می‌دهد که با تابع توزیع تجربی فیشر $f_\theta(\hat{\theta})$ مقایسه شده است.

$$
\begin{aligned}
f_\theta(\hat{\theta}) =& \\
& \frac{(n-۲)(۱-\theta^۲)^{(n-۲)/۲}(۱-\hat{\theta}^۲)^{(n-۴)/۲}}{\pi} \\
& \times \int_۰^\infty \frac{\mathrm{d}\omega}{(cosh\omega - \theta\hat{\theta})^{n-۱}},
\end{aligned}
$$

جدول ۱: نمرات ۲۲ دانش آموز در دو درس بردار و مکانیک

	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱
mech	۷	۴۴	۴۹	۵۹	۳۴	۴۶	۰	۳۲	۴۹	۵۲	۴۴
vec	۵۱	۶۹	۴۱	۷۰	۴۲	۴۰	۴۰	۴۵	۵۷	۶۴	۶۱
	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸	۱۹	۲۰	۲۱	۲۲
mech	۳۶	۴۲	۵	۲۲	۱۸	۴۱	۴۸	۳۱	۴۲	۴۶	۶۳
vec	۵۹	۶۰	۳۰	۵۸	۵۱	۶۳	۳۸	۴۲	۶۹	۴۹	۶۳

برآورد MLE، θ مقدار ۰/۴۹۸ $\hat{\theta} =$ است.



شکل ۱: مقایسه هیستوگرام با تابع چگالی فیشر

اکنون فرض کنید یک چگالی پیشین $\pi(\theta)$ برای پارامتر θ داریم و می‌خواهیم چگالی پسین $\pi(\theta|\hat{\theta})$ را محاسبه کنیم برای هر زیر مجموعه مانند A از فضای پارامتر $\Theta = [-1, 1]$ مطابق قضیه بیز داریم:

$$P(\theta \in A|\hat{\theta}) = \frac{\int_A \pi(\theta) f_\theta(\hat{\theta}) \mathrm{d}\theta}{\int_\Theta \pi(\theta) f_\theta(\hat{\theta}) \mathrm{d}\theta}.$$

عامل تبدیل $R(\theta)$ که نسبت تابع درستنمایی به چگالی خودگردان‌ساز است را به صورت $R(\theta) = \frac{f_\theta(\hat{\theta})}{f_{\hat{\theta}}(\theta)}$ تعریف می‌کنیم. در اینجا مقدار ۰/۴۹۸ $\hat{\theta} =$ است. تابع احتمال را به صورت

$$P(\theta \in A|\hat{\theta}) = \frac{\int_A \pi(\theta) R(\theta) f_{\hat{\theta}}(\theta) \mathrm{d}\theta}{\int_\Theta \pi(\theta) R(\theta) f_{\hat{\theta}}(\theta) \mathrm{d}\theta}, \qquad (۱)$$

بازنویسی می‌کنیم. به‌طور کلی اگر $t(\theta)$ تابعی از θ باشد، امید ریاضی پسین θ به صورت

$$E[t(\theta)|\hat{\theta}] = \frac{\int_\Theta t(\theta) \pi(\theta) R(\theta) f_{\hat{\theta}}(\theta) \mathrm{d}\theta}{\int_\Theta \pi(\theta) R(\theta) f_{\hat{\theta}}(\theta) \mathrm{d}\theta}, \qquad (۲)$$

محاسبه می‌شود.

انتگرال‌های روابط (۱) و (۲) برحسب تابع چگالی خودگردان‌ساز پارامتری $f_{\hat{\theta}}(.)$ هستند. از آن‌جایی که $\theta_B, \ldots, \theta_1$ نمونه تصادفی از $f_{\hat{\theta}}(.)$ هستند. انتگرال‌های بالا می‌توانند با میانگین نمونه‌ای برآورد شوند

$$\hat{E}[t(\theta)|\hat{\theta}] = \frac{\sum_{i=1}^{B} t_i \pi_i R_i}{\sum_{i=1}^{B} \pi_i R_i},$$

که $t_i = t(\theta_i)$، $\pi_i = \pi(\theta_i)$ و $R_i = R(\theta_i)$ هستند. توسط قانون اعداد بزرگ ثابت می‌شود که $\hat{E}[t(\theta)|\hat{\theta}]$ تقریبی از $E[t(\theta)|\theta]$ است، وقتی‌که $B \to \infty$.

در شکل ۲ نمودار پررنگ چگالی پسین $\pi(\hat{\theta}|\theta)$ برای ضریب همبستگی نمرات دانش‌آموزان را نشان می‌دهد.

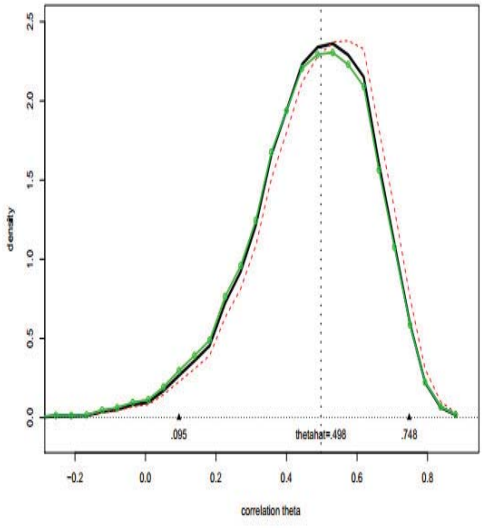
نمودار خط‌چین چگالی خودگردان‌ساز خام (بدون وزن)، و نمودار مهره دار چگالی خودگردان‌ساز وزن‌دار (BC_a) را نشان می‌دهد، که تقریبا به $\pi(\hat{\theta}|\theta)$ شبیه است.

در توزیع خودگردان‌ساز خام، θ_i در هر تکرار خودگردان‌ساز وزن $\frac{1}{B}$ را می‌گیرد. بنابراین $w_i = \pi_i R_i$ را تعریف می‌کنیم و به این ترتیب برآورد توزیع پسین به صورت

$$P(\theta \in A|\hat{\theta}) = \frac{\sum_{\theta_i \in A} w_i}{\sum_{i=1}^{B} w_i}, \qquad (۳)$$

به‌دست می‌آید.

وزن‌های (BC_a) به صورت $w_i^{BC_a} = \frac{\varphi(z_{\theta_i}/(1+az_{\theta_i})-z_0)}{(1+az_{\theta_i})^2\varphi(z_{\theta_i}+z_0)}$ که $[z_{\theta_i} = \Phi^{-1}\hat{G}(\theta_i) - z_0]$ تعریف می‌شوند. توزیع پسین به صورت $\pi^{BC_a}(\hat{\theta}|\theta)$ با همان فرمول (۳) برآورد می‌شود.



شکل ۲: مقایسه چگالی پسین وزن دار

۵ بحث و نتیجه‌گیری

به عقیده روبین خودگردان‌ساز و خودگردان‌ساز بیزی شبیه به هم هستند و ویژگی‌های مشترکی دارند. او معتقد است که محدودیت‌های خودگردان‌ساز بیزی ممکن است ناشی از محدودیت‌های خودگردان‌ساز ناپارامتری باشد. همچنین توصیه می‌کند که اگر چه خودگردان‌ساز بیزی در بعضی از مسایل مفید است اما چون توزیع پیشین (به دلیل ماهیت g_i ها) محدود می‌شود، نباید به عنوان ابزار اصلی استنباط به‌کار رود. همان‌طور که خودگردان‌ساز ناپارامتری نمی تواند به عنوان ابزار اساسی استنباط به کار رود. خودگردان‌ساز پارامتری زمانی‌که تابع چگالی پارامتر مورد نظر فرم پیچیده‌ای دارد، می تواند به عنوان روشی برای برآورد توزیع پسین پارامتر مورد نظر به‌کار رود.

مراجع

[۱] ایران‌پناه، ن. (۱۳۷۷). الگوریتم بوت استرپ بیزی. اندیشه آماری، سال سوم، شماره دوم، ۶۴-۶۹ .

[۲] بهبودیان، ج. (۱۳۸۳). آمار ناپارامتری. انتشارات دانشگاه شیراز. چاپ چهارم.

[3] Banks, D. L. (1988). Histospline smoothing the Bayesian bootstrap. Biometrika, 75(4), 673-684.

[4] Chernick, M. R. (2007). Bootstrap Methods: A Guide for Practitioners

and Researchers. 2nd ed, Wiley, New York.

[5] Efron, B. (1979). Bootstrap methods: another look at the jackknife. Annals of Statistics, 7, 1-26.

[6] Efron, B. (2011). The bootstrap and Markov chain Monte Carlo. Journal of Biopharmaceutical Statistics, 21(6), 1052-1062.

[7] Gasparini, M. (1995). Exact multivariate Bayesian bootstrap distribution of moments. Annals of Statistics, 23, 762-768.

[8] James, L. F. (1997). A study of a class of weighted bootstrap for censored data. Annals of Statistics, 25, 1595-1621.

[9] Lo, A. Y. (1987). A large-sample study of the Bayesian bootstrap. Annals of Statistics, 15, 360-375.

[10] Lo, A. Y. (1988). A Bayesian bootstrap for a finite population. Annals of Statistics, 16, 1684-1695.

[11] Lo, A. Y. (1993). A Bayesian bootstrap for censored data. Annals of Statistics, 21, 100-123.

[12] Mason, D. M. and Newton, M. A. (1992). A rank statistics opproach to the consistency of a general bootstrap. Annals of Statistics, 20, 1611-1624.

[13] Owen, A. B. (1990). Empirical likelihood ratio confidence regions. Annals of Statistics, 18, 90-120.

[14] Praestgaard, D. and Wellner, J. A. (1993). Exchangeably weighted bootstrap of the general empirical process. Annals of Probability, 21, 2053-2086.

[15] Rubin, D. B. (1981). The Bayesian bootstrap. Annals of Statistics, 9, 130-134.

[16] Weng, C. S. (1989). On a second-order asymptotic property of the Bayesian bootstrap mean. Annals of Statistics, 17, 705-710.

[17] Wilks, S. S. (1962). Mathematical Statistics, Wiley, New York.