

## **Analisis Bayes Data DNA Flositometri**

**Mohd. Kidin Shahrin**

*Jabatan Statistik*

*Fakulti Sains Matematik*

*Universiti Kebangsaan Malaysia*

*43600 UKM Bangi, Selangor Darul Ehsan, Malaysia*

Received 18 October 1993

### **ABSTRAK**

Cadangan model baru menggunakan fungsi  $\iota$  yang telah dibina oleh Mohd. Kidin (1986, 1992) digunakan untuk menganggar perkadaran sel dalam fasa-fasa G1-, S- dan G2+M dari kitaran sel. Pendekatan Bayes dicadangkan menerusi taburan campuran tiga komponen. Kaedah ini terlebih dahulu diuji pada data simulasi sebelum diaplikasikan pada data nyata.

### **ABSTRACT**

A newly proposed model using  $\iota$  function developed by Mohd. Kidin (1986, 1992) was used to estimate cell proportions in the G1-, S- and G2+M – phases of the cell cycle. Bayesian approach is proposed via a mixture of three components distribution. The method of analysis is tested on simulated data before being applied to real data.

**Kata kunci:** analisis bayesian, data DNA, flositometri

### **PENGENALAN**

Perkadaran sel dalam fasa-fasa kitaran sel adalah sangat penting dalam kajian kesan penggunaan agen khusus fasa kitaran sel seperti dadah sitostatik. Kepentingan kuantiti tersebut semakin jelas terutamanya dalam kes di mana agen tersebut digunakan untuk menaikkan kesan pembunuhan sel tumor dalam terapi barah (Baisch dan Beck 1978). Disebabkan kepentingan yang menonjol seperti ini, maka terdapat banyak perkembangan dalam aspek pemodelan fungsi pendasar fasa S dan kaedah analisis data FCM (flositometrik) yang tujuan asasnya untuk menganggar perkadaran sel dalam fasa-fasa kitaran sel. Umpamanya Dean (1980) menggunakan model kuadratik dan kaedah kuasa dua terkecil dalam analisisnya, Bruni drk (1983) menggunakan model seragam dengan kaedah masalah songsang; dan Mohd. Kidin (1986) menggunakan model  $\iota$  yang telah dibina berasaskan fenomena fizik replikasi DNA. Dalam analisisnya, beliau menggunakan kaedah penganggaran kebolehjadian maksimum. Dalam kertas ini, beliau mencadangkan penggunaan kaedah Bayes sebagai teknik analisis pilihan.

**KAEDAH ANALISIS BAYES**

Katakan  $x_1, \dots, x_n$  sampel rawak saiz  $n$  dari taburan dengan ketumpatan  $h(\underline{x}; \underline{\theta})$ . Unsur penting dalam analisis Bayesian ialah teorem Bayes iaitu

$$h(\underline{\theta} | \underline{x}) = \frac{g(\underline{\theta}) h(\underline{x} | \underline{\theta})}{g(\underline{x})} \quad [1]$$

yang

$g(\underline{\theta})$  : ketumpatan tercantum prior bagi vektor parameter anu  $\underline{\theta}$ ;

$h(\underline{x} | \underline{\theta})$  : sebagai fungsi  $\underline{\theta}$  adalah fungsi kebolehdjian;

$h(\underline{\theta} | \underline{x})$  : ketumpatan posterior tercantum parameter anu;

$g(\underline{x})$  : diberikan sampel  $\underline{x}$ , adalah pemalar kenormalan.

Berasaskan kepada sampel yang diberi persamaan (1) boleh ditulis sebagai

$$h(\underline{\theta} | \underline{x}) \propto g(\underline{\theta}) \cdot h(\underline{\theta} | \underline{x}) \quad [2]$$

Dengan demikian, ketumpatan posterior tercantum  $h(\underline{\theta} | \underline{x})$  merangkumi maklumat parameter sebelum persampelan menerusi ketumpatan prior  $g(\underline{\theta})$ ; dan maklumat daripada sampel menerusi fungsi kebolehdjian  $h(\underline{x} | \underline{\theta})$ .

*Fungsi Kebolehdjian*

Katakan  $x$  ialah keamatan pendarflor bagi amaun DNA  $y$  dalam satu sel tunggal, yang mana  $x$  dianggap berkadaran terus kepada  $y$ . Bagi keseluruhan populasi sel tunggal, ketumpatan  $x$  diberikan oleh (lihat Mohd. Kidin 1986).

$$h(\underline{x}; \underline{\theta}) = \rho_1 h_1(x_1 | A) + \rho_3 \int_A^{2A} h_3(x | y) g(y) dy + \rho_2 h_2(x_2 | 2A) \quad [3]$$

yang

$\rho_1$  : Perkadaran sel dalam fasa G1;

$\rho_3$  : Perkadaran sel dalam fasa S;

$\rho_2$  : Perkadaran sel dalam fasa G2+M;

$x_1$  : Keamatan pendarflor bagi amaun DNA A; (sel berada dalam fasa G1);

$x_2$  : Keamatan pendarflor bagi amaun DNA 2A; (sel berada dalam fasa G2+M);

$h_i(\cdot | \cdot)$  : ( $i = 1, 2, 3$ ) Ketumpatan bersyarat ralat sukatan  $x$  bagi sel-sel dalam

fasa G1, G2+M dan S.

Fungsi  $g(y)$  ditakrifkan sebagai (Mohd. Kidin 1993)

$$g(y) = \frac{K}{[y-A(1-\alpha)]^{1.95} [2A(1+\alpha) - y]^{0.95}} \quad [4]$$

yang

- K : pemalar kenormalan;
- $\alpha$  : pemalar bentuk dalam selang  $[0, 1]$ ;
- A & 2A : seperti yang telah ditakrifkan.

Dalam amalinya, X yang disukat dalam sekala diskret dinamakan nombor alur (integer) seperti berikut:

*Sebarang sel yang mempunyai keamatan pendarflor  $x$  yang  $x \in [i - 1, i)$  ( $i = 1, 2, \dots, L$ ) maka  $x$  dikatakan mempunyai nilai  $i$ ,*

yang L adalah jumlah nombor alur dalam sistem. Dengan itu kebarangkalian sebarang sel mempunyai nombor alur  $i$  adalah

$$\rho_i = \int_{i-1}^i h_i(x;\theta) dx ; \quad i = 1, \dots, L ; \quad [5]$$

yang  $h(x;\theta)$  diberi dalam (3)

Sekarang katakan sebanyak  $n_i$  sel telah direkod pada nombor alur  $i$ , dan  $\sum n_i = N_s$ , saiz populasi sel; maka fungsi kebolehhjadian ditakrifkan sebagai

$$h(\underline{x} | \underline{\theta}) = \prod_{i=1}^L \{p_i\}^{n_i} \quad [6]$$

#### *Parameter dan Ketumpatan Prior*

Dalam masalah yang diberi, parameter anu yang hendak dianggarkan ialah

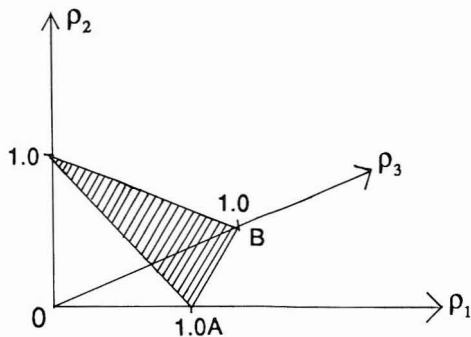
- $x_1$  dalam selang  $-\infty < x_1 < \infty$  ;  $\rho_1$  dalam selang  $0 < \rho_1 < 1$ ;
- $\rho_3$  dalam selang  $0 < \rho_3 < 1$  ;  $\gamma$  dalam selang  $0 \leq \gamma \leq \infty$  ;
- $\alpha$  dalam selang  $0 \leq \alpha \leq 1$  .

dengan kekangan  $\rho_1 + \rho_3 + \rho_2 = 1$ . Untuk tujuan menseseragamkan selang

parameter kepada  $(-\infty, \infty)$  maka dilakukan jelmaan berikut; dan seterusnya ditakrifkan parameter  $\underline{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$  sebagai

$$\begin{aligned} \theta_1 &= x_1; & \theta_2 &= \ln \frac{\rho_1}{1-\rho_1} \\ \theta_3 &= \ln \frac{\rho_3}{1-\rho_1-\rho_3} & \theta_4 &= \ln \gamma; \text{ dan} \\ \theta_5 &= \tan \{\pi (\alpha - 0.5)\}. \end{aligned}$$

Secara tabiinya  $x_1$ ;  $\gamma$  dan  $\alpha$  adalah tak bersandar antara satu sama lain dan dianggap seragam dalam selang masing-masing; maka dalam ruang asalnya, mereka mempunyai prior pemalar. Sementara  $\rho_1$ ,  $\rho_3$  dan  $1 - \rho_1 - \rho_3$  (atau  $\rho_2$ ) adalah bersandar dan dianggap tertabur seragam dalam ruang segitiga berlorek (seperti dalam *Rajah 1*)



*Rajah 1. Ruang taburan parameter  $\rho_1$ ,  $\rho_3$  dan  $1 - \rho_1 - \rho_3$*

Dengan memparameterkan dalam sebutan  $\rho_1$  dan  $\rho_3$ , maka prior mereka malar dalam ruang mendatar segitiga OAB. Oleh itu ketumpatan prior tercantum adalah

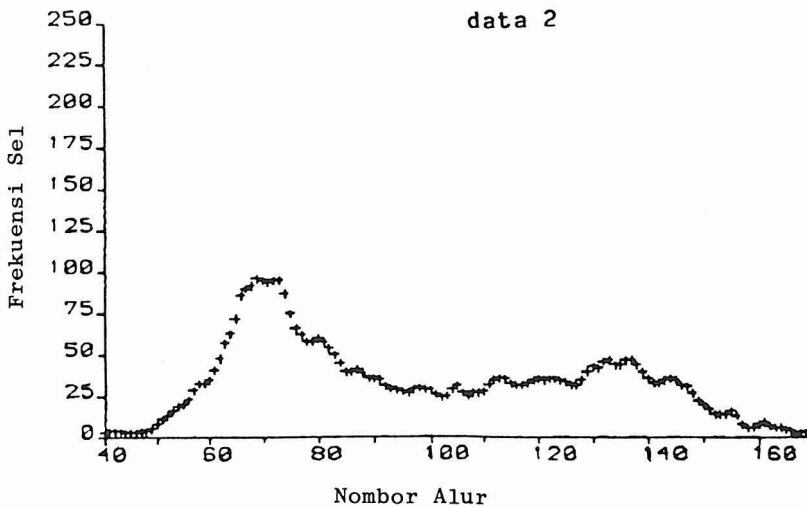
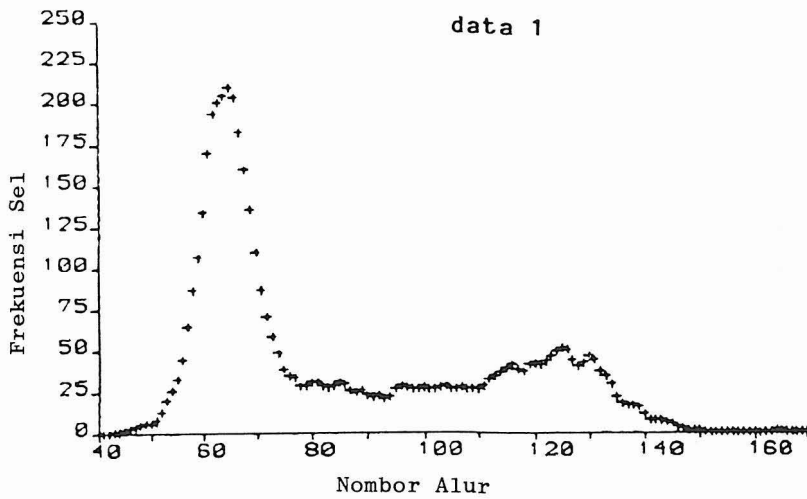
$$g(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5) = g(\theta_1) \cdot g(\theta_2, \theta_3) \cdot g(\theta_4) \cdot g(\theta_5) \quad [7]$$

yang

$$\begin{aligned} g(\theta_1) &\propto \text{pemalar}; \\ g(\theta_2, \theta_3) &\propto \theta_2; \theta_3 (1 - \theta_2, \theta_3); \\ g(\theta_4) &\propto \text{eksp}(\theta_4); \\ g(\theta_5) &\propto \frac{1}{\Pi(1 + \theta_5)} \end{aligned}$$

**PENGANGGARAN PARAMETER**

Merujuk kepada persamaan (2), masalah penganggaran Bayes adalah masalah menentukan ketumpatan posterior  $h(\theta | \underline{x})$ . Kemudian posterior min diambil sebagai anggaran titik bagi parameter anu. Oleh sebab fungsi kebolehhajian melibatkan fungsi tak linear  $g(y)$  dan kekangan tak linear, maka penghitungan ketumpatan posterior dijalankan dengan kaedah berangka (Naylor dan Smith 1982).



Rajah 2. Data DNA FCM barah buah dada

**JADUAL 1**  
Analisis Bayes pada tiga set data simulasi

Parameter	$x_1$	$x_2$	$\rho_1$	$\rho_3$	$\gamma$	$\alpha$
<b>Data 1,</b>						
Input:	40.00	80.00	0.550	0.350	0.050	0.050
Bayes (s.r.)	39.96 (0.05)	80.42 (0.29)	0.505 (0.041)	0.418 (0.002)	0.049 (0.001)	0.033 (0.023)
KM (s.r.)	40.00 (0.05)	80.12 (0.24)	0.549 (0.017)	0.350 (0.028)	0.049 (0.001)	0.054 (0.020)
<b>Data 2,</b>						
Input:	40.00	80.00	0.450	0.350	0.075	0.200
Bayes (s.r.)	40.07 (0.09)	80.26 (0.24)	0.468 (0.020)	0.350 (0.031)	0.075 (0.001)	0.190 (0.011)
KM (s.r.)	40.01 (0.10)	80.31 (0.28)	0.457 (0.025)	0.366 (0.039)	0.075 (0.001)	0.104 (0.064)
<b>Data 3,</b>						
Input:	40.00	80.00	0.350	0.450	0.075	0.200
Bayes (s.r.)	39.92 (0.17)	80.51 (0.35)	0.320 (0.038)	0.495 (0.060)	0.074 (0.001)	0.160 (0.141)
KM (s.r.)	40.07 (0.13)	80.22 (0.25)	0.356 (0.023)	0.438 (0.034)	0.075 (0.001)	0.233 (0.153)

Dengan menggunakan prior dan fungsi kebolehdian yang telah ditakrifkan, kaedah penganggaran telah diuji pada tiga set data simulasi berasaskan kepada pelbagai gabungan input parameter anu seperti dalam Jadual 1. Menurut kajian penulis, input 0.05 dan 0.20 adalah dua nilai hampir kepada ekstrim bawah dan atas parameter  $\alpha$ . Input  $\gamma$  pula diberikan dua nilai dari bahagian tengah julat parameter itu; dan input  $\rho_1$  dan  $\rho_3$  adalah gabungan yang lumrah bagi sel dari populasi berasinkroni. Input parameter  $x_2$  sebagai dua kali input  $x_1$  adalah bersesuaian dengan tabii model  $t$  yang dibina. Gabungan kepelbagaian nilai input membolehkan ujian dilakukan dengan begitu rapi, prihatin dan berkesan bagi mewujudkan keyakinan yang tinggi terhadap kestabilan dan kemantapan model serta kaedah analisis yang digunakan. Hal ini penting diperolehi sebelum diaplikasikan ke atas data nyata; lebih-lebih lagi apabila data nyata seperti dalam *Rajah 2* itu unik bentuknya dengan sebahagian besarnya tidak malar. Penganalisan data seperti ini memang sukar untuk menghasilkan anggaran yang tinggi kepersisannya.

JADUAL 2  
Analisis Bayesan pada dua set data barah

Parameter	$x_1$	$x_2$	$\rho_1$	$\rho_3$	$\gamma$	$\alpha$
<b>Anggaran data 1,</b>						
Watson:	63.62	126.07	0.477	0.336	0.074	–
M. Kidin:	63.39	125.01	0.490	0.322	0.080	0.006
(s.r.)	(0.03)	(0.07)	(0.0022)	(0.0001)	(0.0001)	(0.0001)
<b>Anggaran data 2,</b>						
Watson:	63.62	135.32	0.270	0.451	0.093	–
M. Kidin:	67.16	135.89	0.164	0.670	0.111	0.038
(s.r.)	(0.51)	(1.00)	(0.051)	(0.082)	(0.003)	(0.028)

Hasil analisis data simulasi diberikan dalam Jadual 1, dengan disertakan anggaran menggunakan kaedah kebolehdajian maksimum (KM) sebagai perbandingan. Dua data set nyata dengan ikhsan Dr. Watson J.V. dari Klinik Unit Onkologi, Medical School, Cambridge telah dianalisis semula. Data ini adalah data DNA FCM barah buah dada. Anggaran parameter Model Kidin bagi kedua-dua data diberikan dalam Jadual 2 beserta dengan anggaran yang diperolehi oleh Watson (1977) sebagai perbandingan.

### KEPUTUSAN DAN CADANGAN

Dalam analisis apabila kejahilan tentang ketumpatan prior berlaku, lazimnya, taburan seragam boleh diambil sebagai prior (Zellner 1971). Dalam kes yang dibincangkan, boleh dikatakan prior bagi kesemua parameter anu tidak diketahui. Dengan demikian, taburan seragam telah dianggap sebagai prior kepada parameter anu.

Daripada analisis data simulasi, walau pun dengan prior mudah, kaedah Bayes mampu menghasilkan anggaran yang tinggi kepersisannya terutamanya bagi parameter penting seperti  $\rho_1$  dan  $\rho_3$ . Selain daripada itu anggaran yang diperolehi juga setanding dengan anggaran yang diperolehi menggunakan kaedah kebolehdajian maksimum (KM). Sekiranya maklumat prior parameter penting khususnya boleh diperolehi dengan tepat dan taburan prior yang betul dapat dikenal pasti, saya percaya dan yakin kaedah analisis Bayes dapat menghasilkan anggaran yang sangat baik.

Daripada analisis data nyata pula, walau pun anggaran  $\gamma$  lebih besar dibanding dengan anggaran Watson, tetapi anggaran parameter penting yang lain adalah setanding; khususnya parameter  $\rho_2$  dan  $\rho_3$ . Malah dengan ralat piawai yang kecil anggaran Bayes wajar lebih tepat lagi. Walau

bagaimana pun, anggaran Watson bukanlah nilai sebenar parameter untuk dijadikan piawai perbandingan.

Untuk kajian lanjut, ketumpatan prior yang 'betul' harus dikenal pasti dengan terlebih dahulu mendapatkan maklumat sebenar daripada ahli ujikaji; untuk memperolehi anggaran yang lebih tepat atau lebih tinggi kepersisannya.

### RUJUKAN

- BAISCH, H. dan H.P. BECK 1978. *Comparison of Cell Kinetic Parameters Obtained by Flow and Autoradiography Biometrics and Cell Kinetics*. Amsterdam: North Holland Biomedical Press.
- BRUNI, C., G. KOCH and C. ROSSI. 1983. On the inverse problem in flow cytometry recovering DNA distribution from fluoromikroflorometrik (FMK) data. *Cell Biophysics* **5**: 5-19.
- DEAN, P.N. 1980. A simplified method of DNA distribution analysis. *Cell Tissue Kinet.* **13**: 299-308.
- MOHD. KIDIN SHAHRAN. 1986. Pemodelan fungsi pendasar taburan fasa S dengan variasi kadar sintesis DNA. *Prosiding Simposium Kebangsaan Sains Matematik ke-2*. UKM: 351-363.
- MOHD. KIDIN SHAHRAN. 1992. Taburan pendarflor mewakili taburan DNA sitomertrik alir. *Prosiding Simposium Kebangsaan Sains - Matematik ke-V*, UTM: 175-182.
- MOHD. KIDIN SHAHRAN. 1993. Tinjauan model matematik untuk analisis data DNA FCM, *laporan teknik FSMK/September 1993/LT 30*. (19 halaman).
- NAYLOR, J.C. dan A.F.M. SMITH. 1982. Application of a method for the efficient of posterior distributions. *JRSS C.* **31(3)**: 214-225.
- WATSON, J.V. 1977. The application of age distribution theory in the analysis cytofluorometric DNA data. *Cell Tissue Kinet.* **10**: 157-169.
- ZELLNER, A. 1971. *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley.