STATISTIKA NONPARAMETRIK: PENGGABUNGAN DIAGRAM POHON POLYA YANG BERHINGGA

Asri Ode Samura
Tadris Matematika, IAIN Ternate, *Asri22samura@gmail.com*

Abstrak

Sebuah pendekatan tentang nonparametrik bayes, penggabungan diagram pohon polya (MPT). MPT menggunakan sebuah partisi dari dukungan kepadatan distribusi aslinya. Kepadatan umumnya mempertahankan bentuk distribusi asli pada tiap-tiap rangkaian partisi dan menambahkan parameter baru dari peluang-peluang bersyarat. Model diagram pohon polya dapat diaplikasikan secara luas dan mudah diprogramkan dengan diberikannya skema MCMC untuk penyesuaian model parametrik aslinya.

Kata Kunci. Bayes, Model gabungan liniear tergeneralisasi, MGLT

1. Pendahuluan

Analisis Bayes membutuhkan probabilitas prior untuk parameter-parameter distribusi sampling. Akan bisa sulit memilih prior yang masuk akal dalam masalah-masalah dengan banyak parameter, sehingga parameter-parameter tersebut disini dipecah menjadi dua kelompok yang tepat: (1) yang berhubungan dengan keluarga parametrik asli, dan (2) yang berhubungan dengan generalisasi keluarga asli tersebut. Metode standard dapat digunakan untuk membentuk sebuah prior bagi parameter asli; lihat Bedrick, Christensen, dan Johnson (1996) untuk sebuah pendekatan dalam model linier tergeneralisasi. Prior acuan yang tepat ada untuk parameter-parameter dari suatu generalisasi. Diagram pohon Polya adalah sebuah distribusi probabilitas random. Bersyarat pada sebuah anggota himpunan parametrik asli yang ditetapkan (contohnya distribusi normal dengan rata-rata μ dan ragam σ^2), himpunan distribusi tergeneralisasi bersama dengan prior acuan merupakan diagram pohon Polya yang berhingga. Integrasi prior terhadap parameter himpunan asli (misalnya μ dan σ^2) membentuk sebuah gabungan diagram pohon Polya.

MPT sudah sangat luas digunakan akan tetapi pada contoh kita akan melibatkan penggunaan model gabungan linier tergeneralisasi (GLMM), lihat Breslow dan Clayton (1993), dan keduanya melibatkan generalisasi distribusi normal yang biasanya diasumsikan bersama dengan keduanya. GLMM memberikan sebuah

kerangka kerja yang sangat dikenal untuk analisis ukuran-ukuran longitudinal dan data berkelompok yang muncul dalam banyak bidang seperti pertanian, biologi, epidemiologi, ekonomi, dan geofisika. Model-model tersebut menghitung korelasi antar pengamatan berkelompok dengan memasukkan efek random ke dalam komponen prediktor linier dari model tersebut. Meskipun penyesuaian GLMM secara khusus bersifat kompleks, intersep random standar dan model intersep/slop random dengan efek random berdistribusi normal sekarang dapat secara rutin disesuaikan dalam paket-paket software komersil seperti SAS dan Stata. Model-model tersebut cukup fleksibel dalam mengakomodasi perilaku heterogen, akan tetapi mereka memiliki kekurangan yang sama yaitu kurangnya ketahanan terhadap tujuan dari asumsi distribusi seperti model statistika lain yang berdasarkan distribusi Gauss.

Suatu strategi untuk menjaga terhadap asumsi normalitas yang tidak sesuai adalah dengan memasukkan asumsi distribusi yang lebih fleksibel untuk efek random ke dalam model. Sehingga tampaknya diperlukan pengembangan nonparametrik dari GLMM parametrik. Kita menggambarkan perluasan tersebut beserta akibat dari penggunaan yang salah untuk asumsi model tradisional dalam GLMM dengan menggunakan pada contoh kehidupan nyata kita. Demi tujuan perbandingan, kita menyesuaikan semua model berdasarkan kedua asumsi efek random berdistribusi normal dan generalisasi.

2. Landasan Teori

2.1. Theorema Bayes

Misal $y' = (y_1, \ldots, y_n)$ adalah sebuah vektor dari n pengamatan yang distribusi peluangnya $p(y|\theta)$ bergantung pada nilai k parameter $\theta' = (\theta_1, \ldots, \theta_k)$. Dimisalkan bahwa θ sendiri memiliki distribusi peluang $p(\theta)$. Maka

$$P(y|\theta)P(\theta) = P(y,\theta) = P(\theta|y)P(y) \qquad (2.1)$$

Dengan diberikan data pengamatan y, distribusi bersyarat dari θ menjadi

$$P(\theta|y) = \frac{P(y|\theta)P(\grave{e})}{P(y)}. \qquad (2.2)$$

Maka

$$P(y) = EP(y|\grave{e}) = C^{-1} = \begin{cases} \int P(y|\grave{e})P(\grave{e})d\grave{e}, jika\ \grave{e}\ kontinu \\ \sum P(y|\grave{e})P(\grave{e}), jika\ \grave{e}\ diskrit \end{cases} \qquad (2.3)$$

dimana sigma atau integralnya diambil pada rentang dari θ yang dapat diterima, dan dimana $E[f(\theta)]$ adalah ekspektasi matematika dari $f(\theta)$ sehubungan dengan distribusi $P(\theta)$. Oleh karena itu kita dapat menuliskan sebagai berikut

$$P(\grave{e}|y) = cP(y|\grave{e})P(\grave{e}). \qquad (2.4)$$

2.2. Distribusi Posterior

Distribusi posterior è diberikan x (atau disingkat posterior) akan dinotasikan ð(è|x), dan didefinisikan sebagai distribusi bersyarat dari è, diberikan x sampel pengamatan. Dengan memperhatikan bahwa è dan X memiliki (subjektif) kerapatan

$$h(x|\text{è}) = \text{ð}(\text{è})f(x|\text{è}), \tag{2.5}$$

dan bahwa X telah marjinal (tanpa syarat) kerapatan

$$m(x) = \begin{cases} \int f(x|\text{è})\text{ð}(\text{è})d\text{è}, jika\ \text{è}\ kontinu \\ \sum_{\text{è}} f(x|\text{è})\text{ð}(\text{è}), jika\ \text{è}\ diskrit \end{cases} \tag{2.6}$$

jelas bahwa (dengan memberikan $m(x) \neq 0$)

$$\text{ð}(\text{è}|x) = \frac{h(x|\text{è})}{m(x)}. \tag{2.7}$$

dengan menentukan estimator bayes, maka persamaan yang dipakai

$$E\big(\text{ð}(\text{è}|x)\big) = \begin{cases} \int \text{è}f(x|\text{è})\text{ð}(\text{è})d\text{è} jika \text{è} kontinu \\ \sum_{\text{è}} \text{è}f(x|\text{è})\text{ð}(\text{è}) jika \text{è} diskrit \end{cases} \tag{2.8}$$

Adapun dapat menentukan estimator dengan cara Box-Tiao. Dari persamaan (2.1) dapat tulis dalam bentuk

$$\text{ð}(\text{è}|x) \propto f(x|\text{è})\text{ð}(\text{è}) \tag{2.9}$$

Konstanta kesebandingan dipilih sedemikian hingga ð(è|x) merupakan densitas.

Teorema : 2.1 : *Jika $x_1, x_2, \dots, x_n$ adalah sampel random dari distribusi $N(\text{ì}, \text{ë}^{-1})$ dengan presisi* ë, (ë > 0) *diketahui, misalkan distribusi prior untuk rata-rata* ì *adalah distribusi $N\big(\text{ì}_0, (n_0\text{ë}^{-1})\big)$, maka distribusi posterior* ì *bersyarat bilamana diberikan nilai $X_i = x_i$, untuk $i = 1,2, \dots, n$ adalah berdistribusi $N(\text{ì}_n, \text{ë}_n^{-1})$ dimana*

$$\text{ë}_n = (n_0 + n)\text{ë} \quad \text{dan} \quad \text{ì}_n = \frac{n\bar{x}+n_0\text{ì}_0}{n_0+n} \tag{2.10}$$

Bukti:

Diberikan $X_i = x_i$, untuk $i = 1,2, \dots, n$ karena ë diketahui, dengan mensubtitusi fungsi likelihood $L\big(\text{ì}|\underline{x}\big) \propto Exp\{-n\text{ë}(\text{ì} - \bar{x})^2/2\}$ dan fungsi densitas prior ì pada ì(ì|ë) $\propto Exp\{-n_0\text{ë}(\text{ì} - \text{ì}_0)^2/2\}$ kedalam persamaan

$$\text{ð}\big(\text{ì}|\underline{x}\big) \propto L\big(\text{è}|\underline{x}\big)\text{ð}(\text{è}) \tag{2.11}$$

Sehingga diperoleh

$$\text{ð}\big(\text{ì}|\underline{x}\big) \propto L(\text{ì}|x)\text{ð}(\text{ì})$$

$$\propto Exp\left\{-\frac{1}{2}[n\text{ë}(\text{ì} - \bar{x})^2 + n_0\text{ë}(\text{ì} - \text{ì}_0)^2]\right\} \tag{2.12}$$

Ekspresi$n\text{ë}(\text{ì} - \bar{x})^2 + n_0\text{ë}(\text{ì} - \text{ì}_0)^2$ pada ruas kanan (2.12) diuraikan menjadi

$$n\ddot{e}(\grave{\imath}-\bar{x})^2+n_0\ddot{e}(\grave{\imath}-\grave{\imath}_0)^2=(n_0+n)\ddot{e}(\grave{\imath}-\grave{\imath}_n)+\frac{\ddot{e}nn_0(\bar{x}-\grave{\imath}_0)^2}{n_0+n} \qquad (2.13)$$

Dimana $\grave{\imath}_n$ adalah sebagaimana pada persamaan (2.10), karena suku terakhir pada ruas kanan persamaan (2.13) adalah konstan terhadap ì, maka suku tersebut dapat diabaikan, dan setelah itu hasilnya disubtitusikan kembali kedalam persamaan (2.12) sehingga diperoleh fungsi densitas posterior

$$\eth(\grave{\imath}|\underline{x})\propto Exp\left\{-\left(\frac{1}{2}\right)n\ddot{e}(\grave{\imath}-\grave{\imath}_n)^2\right\} \qquad (2.14)$$

Yang merupakan fungsi densitas dari distribusi normal $N(\grave{\imath}_n,\ddot{e}_n^{-1})$ dimana $\grave{\imath}_n$ dan $\ddot{e}_n$ adalah seperti pada persamaan (2.10) jadi teorema terbukti. ■

Teorema: 2.2 :*Jika $x_1,x_2,\dots,x_n$ adalah sampel random dari distribusi $N(\grave{\imath},\ddot{e}^{-1})$; untuk* ì *dan* ë *kedua-duanya tidak diketahui. Jika distribusi prior bersama dari* $(\grave{\imath},\ddot{e})$ *adalah distribusi* $NG(\grave{\imath},\ddot{e};\grave{\imath}_0,\ddot{e}_0,\acute{a}_0,\hat{a}_0)$*; $-\infty<\grave{\imath}_0<\infty$, $n_0>0$, $\ddot{e}>0$, $\acute{a}>0$, $\hat{a}>0$, maka distribusi posterior bersama* $(\grave{\imath},\ddot{e}|x)$ *adalah distribusi* $NG(\grave{\imath},\ddot{e}|\grave{\imath}_n,\ddot{e}_n,\acute{a}_n,\hat{a}_n)$ *dimana*

$$\grave{\imath}_n=(n\bar{x}+n_0\grave{\imath}_0)/(n_0+n)\ \text{ dan }\ \ddot{e}_n=(n_0+n)\ddot{e} \qquad (2.15)$$

Dan

$$\acute{a}_n=\acute{a}_0+{}^{n}/_{2}\ \text{ dan }\ \hat{a}_n=\hat{a}_0+\tfrac{1}{2}\textstyle\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{nn_0(\bar{x}-\grave{\imath}_0)^2}{2(n_0+n)} \qquad (2.16)$$

Bukti:

Diberikan $X_i=x_i i=1,2,3,\dots,n,$ karena ì dan ë kedua-duanya tidak diketahui, disubtitusi fungsi likelihood

$$L(\grave{\imath},\ddot{e}|\underline{x})\propto\ddot{e}^{n/2}Exp\{-(\ddot{e}/2)\textstyle\sum_{i=1}^{n}(x_i-\bar{x})^2\}exp\{-n\ddot{e}(\grave{\imath}-\bar{x})^2/2\} \quad (2.17)$$

Dan fungsi densitas prior normal gamma pada

$$\eth(\grave{\imath},\ddot{e}|n_0,n_1,\dots,n_n)=\eth(\grave{\imath},\ddot{e}|n_0,\grave{\imath}_0,\acute{a}_0,\hat{a}_0)$$

$$\propto\left\{\ddot{e}^{1/2}Exp\left\{-\frac{n_0\ddot{e}}{2}(\grave{\imath}-\grave{\imath}_0)^2\right\}\right\}\left(\ddot{e}^{\acute{a}_0-1}Exp\{-\hat{a}_0\ddot{e}\}\right) \qquad (2.18)$$

Kedalam persamaan $\eth(\grave{e}|x)\propto L(\grave{e}|x)\eth(\grave{e})$ dimana $\grave{e}=(\grave{e}_1,\grave{e}_2)=(\grave{\imath},\ddot{e})$, sehingga diperoleh

$$\eth(\grave{\imath},\ddot{e}|\underline{x})\propto\ddot{e}^{(n/2)+\acute{a}_0-1}\ddot{e}^{1/2}Exp(-\hat{a}_0\ddot{e})$$
$$\cdot Exp\left\{-\left(\frac{\ddot{e}}{2}\right)[\textstyle\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\grave{\imath}-\bar{x})^2+n_0(\grave{\imath}-\grave{\imath}_0)^2]\right\} \quad (2.19)$$

Karena $\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\grave{\imath}-\bar{x})^2+n_0(\grave{\imath}-\grave{\imath}_0)^2$ dapat diuraikan menjadi

$$(n_0+n)(\grave{\imath}-\grave{\imath}_n)^2+\textstyle\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{nn_0(\bar{x}-\grave{\imath}_0)^2}{n_0+n} \qquad (2.20)$$

Dengan $\grave{\imath}_n$ seperti pada persamaan (2.15) dan persamaan (2.19) sehingga menjadi

$$\eth(\grave{\imath},\ddot{e}|\underline{x})\propto\ddot{e}^{(n/2)+\acute{a}_0-1}\ddot{e}^{1/2}$$

$$\cdot Exp\left\{-\frac{\ë}{2}(n_0+n)(\ì-\ì_n)^2+\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{nn_0(\bar{x}-\ì_0)^2}{n_0+n}\right\}$$

$$\cdot Exp(-\â_0\ë)\propto \ë Exp\{-\ë^{1/2}(n_0+n)(\ì-\ì_n)^2/2\}$$

$$\cdot \ë^{(n/2)+\á_0-1}Exp\left\{-\left[\â_0+\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2+\frac{1}{2}\frac{nn_0(\bar{x}+\ì_0)^2}{n_0+n}\right]\ë\right\}$$

Atau setara dengan

$$\ð(\ì,\ë|\underline{x})\propto\{\ë^{1/2}Exp[-\ë_n(\ì-\ì_n)^2/2]\}\left(\ë^{\á_n-1}Exp(-\â_n\ë)\right) \qquad (2.21)$$

Ini adalah fungsi densitas dari distribusi normal gamma $NG(\ì,\ë;\ì_n,\ë_n,\á_n,\â_n)$ dengan $\ì_n,\ë_n,\á_n,\â_n$ sebagaimana di dalam persamaan (2.15) dan (2.16) ■

Teorema; 2.5.3: *Diberikan* $x_1,x_2,\dots,x_n$ *dari distribusi* $N(\ì,\ë^{-1})$*; dengan* ì *dan* ë *kedua-duanya tidak diketahui. Jika digunakan distribusi prior-informatif* $\ð(\ì,\ë)=\ë^{-1}$ *maka distribusi posterior bersama* $(\ì,\ë|\underline{x})$ *adalah distribusi* $NG(\ì'_n,\ë'_n,\á'_n,\â'_n)$*, dimana*

$$\ì'_n=\bar{x} \quad \text{dan} \quad \ë'_n=n\ë \qquad (2.22)$$

Dan

$$\á'_n=(n-1)/2 \quad \text{dan} \quad \â'_n=\textstyle\sum_{i=1}^{n}(x_i-\bar{x})^2/2 \qquad (2.23)$$

Bukti:

Diberikan $X_i=x_i$ dimana $i=1,2,\dots,n,$ dengan ì dan ë kedua-duanya tidak diketahui maka dapat kita pakai persamaan $\ð(\è|\underline{x})\propto L(\è|x)\ð(\è)$ dapat digunakan untuk $\underline{\è}=(\è_1,\è_2)=(\ì,\ë)$ kedalam persamaan $\ð(\è|\underline{x})\propto L(\è|x)\ð(\è)$ ini disebut fungsi likelihood pada persamaan $L(\ì|\underline{x})\propto Exp\{-n\ë(\ì-\bar{x})^2/2\}$, dan fungsi diensitas prior tak informatif pada persamaan $\ð(\ì,\ë)=\ë^{-1}$ maka diperoleh

$$\ð(\ì,\ë|\underline{x})\propto L(\ì,\ë|\underline{x})\ð(\ì,\ë)$$

$$\propto \ë^{-1}\ë^{n/2}Exp\left\{-\left(\frac{\ë}{2}\right)\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}\frac{Exp\{-n\ë(\ì-\ë)\}^2}{2}$$

$$\propto n^{1/2}\ë^{1/2}Exp\{-n\ë(\ì-\bar{x})^2/\}\cdot\left((n-1)/2\right)^{-1}$$

$$\cdot Exp\left\{-(\ë/2)\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}$$

$$\ð(\ì,\ë|\underline{x})\propto(n\ë)^{1/2}Exp\{-n\ë(\ì-\bar{x})^2/2\}\cdot\ë^{\á'_n-1}\cdot Exp(-(\â'_n/2)\ë) \qquad (2.24)$$

Ini menunjukan distribusi normal-Gamma $NG(\ì,\ë|\ì'_n,\ë'_n,\á'_n,\â'_n)$, dengan $\ì'_n,\ë'_n,\á'_n,\â'_n$ sebagaimana pada persamaan (2.22) dan (2.23) ■

2.3. Pohon Polya

Definisi. Probabilitas random P pada R mempunyai distribusi pohon Polya dengan parameter $(\Pi, \mathcal{A})$ terdapat variabel random $Y = \{Y_0, Y_{00}, Y_{10}, \ldots\}$ jika memenuhi:

1. Variabel-variabel random padaY independen;
2. Untuk setiap $\epsilon \in E^\star, Y_{\epsilon 0} \sim Beta(\acute{a}_{\epsilon 0}, \acute{a}_{\epsilon 1})$;
3. Untuk setiap $m = 1,2, \ldots$ dan untuk setiap $\epsilon \in E^m$, maka:

$$P(B_\epsilon) = \left[\prod_{\{\epsilon_j=0\}}^{m_{j=1}} Y_{\epsilon_1,\ldots\epsilon_j 0}\right]\left[\prod_{\{\epsilon_j=1\}}^{m_{j=1}} 1 - Y_{\epsilon_1\ldots\epsilon_{j-1}0}\right] \quad (2.25)$$

Yang dinotasikan

$$P \sim PT(\Pi, \mathcal{A}). \quad (2.26)$$

Varibel random $W_1, \ldots, W_n$ dikatakann sampel berukuran n dari P jika:

$$\begin{aligned} P[W_i \in B_\epsilon] &= EP[W_i \in B_\epsilon | P] \\ &= E[P(B_\epsilon)] \\ &= E[P(B_{\epsilon 1})P(B_{\epsilon 1 \epsilon 2}|B_{\epsilon 1}) \ldots P(B_\epsilon | B_{\epsilon 1 \ldots \epsilon m-1})] \\ &= \frac{\acute{a}_{\epsilon 1}}{\acute{a}_0 + \acute{a}_1} \ldots \frac{\acute{a}_{\epsilon 1 \ldots \epsilon m-1}}{\acute{a}_{\epsilon 1 \ldots \epsilon m-1} 0 + \acute{a}_{\epsilon 1 \ldots \epsilon m-1} 1}. \end{aligned} \quad (2.27)$$

2.4. Rantai Markov Monte Carlo (MCMC).

Rantai Markov dapat digunakan untuk berbagai keperluan dalam perhitungan dan optimasi. MCMC algoritma yang paling dikenal adalah Gibbs sampler dan algoritma Metropolis-Hastings (M-H).

Rantai Markov

a. Rantai markov waktu diskrit adalah proses stokastik dengan keadaan diskrit yang memenuhi:
$P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0\} = P\{X_{n+1} = j | X_n = i\}$ (2.28)

Sehingga

$$P_{ij} = P\{X_{n+1} = j | X_n = i\} \quad (2.29)$$

b. Rantai markov waktu kontinu adalah proses stokastik yang memiliki distribusi bersyarat dari masa yang akan datang dengan $X(t+s)$, masa sekarang $X(s)$, dan masa lalu $X(u)$, $0 \le u < s$, tergantung pada masa sekarang dan independen dari masa lalu, Yaitu:

$$\begin{aligned} &P\{X(t+s) = j | X(s) = i,\ X(u) = x(u),\ 0 \le u < s\} \\ &\quad = P\{X(t+s) = j | X(s) = i\} \end{aligned} \quad (2.30)$$

Dimana $s, t \ge 0$

Proses $\{X(t),\ t \ge 0\}$ memiliki probabilitas transisi stasioner atau homogen jika

$$P[X(t+s)=j|X(s)=i]=P_{ij}(t) \tag{2.31}$$

2.5. Metode Simulasi Monte Carlo

Simulasi Monte Carlo dapat diartikan sebagai metode simulasi statistik, di mana simulasi statistik didefinisikan sebagai istilah umum untuk semua metode simulasi yang menggunakan rangkaian bilangan acak. Simulasi Monte Carlo adalah suatu metode untuk mengevaluasi suatu model yang melibatkan bilangan acak sebagai salah satu input.

Simulasi Monte Carlo dapat diartikan juga sebagai metode untuk menganalisa perambatan ketidakpastian, di mana tujuannya adalah untuk menentukan bagaimana variasi acak atau error mempengaruhi sensitivitas, dari sistem yang sedang dimodelkan. Simulasi Monte Carlo digolongkan sebagai metode sampling karena input yang dibangkitkan secara acak dari suatu distribusi probabilitas untuk proses sampling dari suatu populasi nyata.

2.6. Gibbs Sampling

Gibbs sampling adalah algoritma yang didasarkan pada generasi yang berurutan dari kepadatan bersyarat penuh $p(\grave{e}_i|\grave{e}_{\neq i}, y)$, yaitu, kepadatan posterior pada elemenke-idari $\grave{e} \doteq (\grave{e}_1, \ldots, \grave{e}_d)'$, diberikan semua elemen lainnya, di mana elemen è dapat berupa skalar atau sub-vektor.

Menghitung vektor dalam beberapa urutan sebagai vektor 1,2,...,s dan mengidentifikasi vektor j dengan keadaan j pada rantai Markov dengan probabilitas transisi. Jika vektor-vektor i dan j berbeda lebih dari satu dalam komponen, dimana $p_{ij}=0$. Jika mereka berbeda dalam maksimal satu komponen, misalkan, menjadi konkrit, berbeda dalam komponen pertama $(i \neq j)$. Dengan vektor i adalah $(y_1, y_2, \ldots, y_k)$ dan vektor j sebagai $(y_1^*, y_2, \ldots, y_k)$. Maka

$$p_{ij} = prob\left(Y_1 = y_1^* | Y_2 = y_2, Y_3 = y_3, \ldots, Y_k = y_k\right)$$

$$= \frac{Prob(Y_1=y_1^*, Y_2=y_2, Y_3=y_3, \ldots, Y_k=y_k)}{Prob(Y_2=y_2, Y_3=y_3, \ldots, Y_k=y_k)} \tag{2.32}$$

Probabilities pada pembilang dan penyebut dihitung dengan menggunakan $P_Y(y)$, kami mengklaim bahwa rantai Markov tereduksi dan aperiodik, dan selanjutnya memiliki distribusi stasioner $P_y(y)$.

3. Pembahasan

3.1. Penggabungan Diagram Pohon Polya yang Berhingga

Dengan menggunakan definisi yang relatif sederhana dari diagram pohon Polya, kita dapat menggambarkan proses generalisasi sebuah keluarga distribusi parametrik dengan menggunakan keluarga N(μ, σ^2). Keluarga parametrik lain digeneralisasi dengan cara yang sama.

Generalisasi tersebut melalui sejumlah tahapan, sebut saja J. Pada tiap tahapan kita memperkenalkan parameter baru untuk menggeneralisasi tahapan sebelumnya. Pada tahap pertama, kita membagi garis bilangan riil, yaitu pendukung distribusi normal, menjadi dua interval yang terbagi oleh median μ. Selanjutnya kita membiarkan perubahan yang terjadi dalam probabilitas menjadi di bawah atau di atas μ, tetapi kita mempertahankan bentuk kepadatan normal baik di bawah μ dan di atas μ.

Parameter baru yang ada pada tahap pertama adalah θ_{11} probabilitas yang tidak lebih besar dari μ, dan θ_{12} probabilitas yang berada di atas μ. Secara formal, diketahui X_1 memiliki distribusi tahap pertama, maka

$$\text{è}_{11} \equiv P[X_1 \leq \text{ì}], \tag{3.1}$$

dan

$$\text{è}_{12} \equiv P[X_1 > \text{ì}] = 1 - \text{è}_{11}. \tag{3.2}$$

Karena kita mempertahankan bentuk normal pada kedua himpunan, jika a ≤ μ dan Y~ N(μ, σ^2), secara kondisional kita akan mendapatkan

$$P[X_1 \leq a | X_1 \leq \text{ì}] = \frac{P[Y \leq a]}{0.5}, \tag{3.3}$$

Dimana

$$\begin{aligned} P[Y \leq a] &= P\left(\frac{\text{y} - \mu}{\text{ó}} < \frac{a - \mu}{\text{ó}}\right) \\ &= P\left(Z < \left(\frac{a - \mu}{\text{ó}}\right)\right) \\ &= \Phi\left(\frac{a - \mu}{\text{ó}}\right) \end{aligned}$$

Sehingga

$$\begin{aligned} P[X_1 \leq a | X_1 \leq \text{ì}] &= \frac{\Phi\left(\frac{a-\mu}{\text{ó}}\right)}{1/2} \\ &= 2\Phi[(a - \text{ì})/\text{ó}] \end{aligned} \tag{3.4}$$

dimana Φ(·) merupakan fungsi kepadatan bersyarat (cdf) dari suatu normal standar. Begitu pula, jika b > μ,

$$\begin{aligned} P[X_1 > b | X_1 > \text{ì}] &= \frac{P[Y > b]}{1/2} \\ &= 2P[Y > b], \end{aligned} \tag{3.5}$$

Dengan

$$\begin{aligned} P[Y > b] &= P\left(\frac{y + \text{ì}}{\text{ó}} > \frac{b - \text{ì}}{\text{ó}}\right) \\ &= P\left((1 - Z)\left(\frac{b - \text{ì}}{\text{ó}}\right)\right) \\ &= 1 - \Phi[(b - \text{ì})/\text{ó}] \end{aligned}$$

Maka

$$P[X_1 > b|X_1 > \text{ì}] = 2P[Y > b] = 2\{1 - \Phi[(b - \mu)/\acute{o}]\} \quad (3.6)$$

Dengan cara lain kita bisa menuliskan

$$P[X_1 \le a] = P[Y \le a]2\grave{e}_{11} \quad (3.7)$$
$$P[X_1 > b] = P[Y > b]2\grave{e}_{12}. \quad (3.8)$$

Dengan $I_A(x)$ fungsi indikator dari A, kepadatan tahap distribusi 1 adalah

$$f(x_1|\text{ì}, \acute{o}^2, \grave{e}_{11}, \grave{e}_{12}) = \frac{1}{\sqrt{2\delta}\acute{o}} e^{-(x_1-\text{ì})^2/2\acute{o}^2} \text{x } 2^1\left[\grave{e}_{11} I_{(-\infty,\text{ì}]}(x_1) + \grave{e}_{12} I_{(\hat{\text{i}},\infty]}(x_1)\right]. \quad (3.9)$$

Ditentukan X_2 memiliki distribusi tahap kedua, kita memperkenalkan parameter baru, θ_{21}, θ_{22}, θ_{23}, θ_{24}, yang didefinisikan sebagai peluang bersyarat relatif terhadap himpunan yang digunakan pada tahap 1:

$$\grave{e}_{21} = P[X_2 \le q_1|X_2 \le \text{ì}] \quad (3.10)$$
$$\grave{e}_{22} = P[q_1 < X_2 \le \text{ì}|X_2 \le \text{ì}] \quad (3.11)$$
$$\grave{e}_{23} = P[\text{ì} < X_2 \le q_3|X_2 > \text{ì}] \quad (3.12)$$
$$\grave{e}_{24} = P[q_3 < X_2|X_2 > \text{ì}]. \quad (3.13)$$

Catat bahwa $\theta_{21} = 1 - \theta_{22}$ dan $\theta_{23} = 1 - \theta_{24}$. Dengan tanpa syarat, keempat himpunan memiliki peluang

$$P[X_2 \le q_1] = \grave{e}_{11}\grave{e}_{21} \quad (3.14)$$
$$P[q_1 < X_2 \le \text{ì}] = \grave{e}_{11}\grave{e}_{22} \quad (3.15)$$
$$P[\text{ì} < X_2 \le q_3] = \grave{e}_{12}\grave{e}_{23} \quad (3.16)$$
$$P[q_3 < X_2] = \grave{e}_{12}\grave{e}_{24}. \quad (3.17)$$

Dalam tiap himpunan, kita menggunakan bentuk kepadatan normal asli sehingga, sebagai contoh, jika $\mu < a < b \le q_3$ dan $Y \sim N(\mu, \sigma^2)$,

$$\begin{aligned} &P[a < X_2 \le b] \\ &= P[a < X_2 \le b|\text{ì} < X_2 \le q_3]P[\text{ì} < X_2 \le q_3] \\ &= P[a < Y \le b|\text{ì} < Y \le q_3]P[\text{ì} < X_2 \le q_3] \\ &= \frac{P[a < Y \le b]}{P[\text{ì} < Y \le q_3]} P[\text{ì} < X_2 \le q_3] \\ &= P[a < Y \le b]\frac{\grave{e}_{12}\grave{e}_{23}}{0.25}. \end{aligned} \quad (3.18)$$

Pada umumnya, kepadatan pada distribusi tingkat 2 adalah

$$\begin{aligned} &f(x_2|\text{ì}, \acute{o}^2, \grave{e}_{11}, \grave{e}_{12}, \grave{e}_{21}, \grave{e}_{22}, \grave{e}_{23}, \grave{e}_{24}) \\ &= \frac{1}{\sqrt{2\delta}\acute{o}} e^{-(x_2-\text{ì})^2/2\acute{o}^2} 2^2 \left[\grave{e}_{11}\grave{e}_{21} I_{(-\infty,q_{1]}}(x_2) + \grave{e}_{11}\grave{e}_{22} I_{(q_1,\text{ì}]}(x_2) + \grave{e}_{12}\grave{e}_{23} I_{(\hat{\text{i}},q_3]}(x_2) + \right. \\ &\left. \grave{e}_{12}\grave{e}_{24} I_{(q_3,\infty)}(x_2)\right] \end{aligned} \quad (3.19)$$

Untuk menampilkan analisis Bayes dengan distribusi sampling ini, kita perlu sebuah distribusi prior gabungan pada parameter θ_{js}. θ_{js} sangat mudah

diinterpretasikan, informasi prior yang sangat berarti bisa muncul pada mereka. Sebuah kasus ekstrim memilih $\theta_{11} = \theta_{12} = 0{,}5$ dengan peluang prior 1. Hal ini memberikan peluang prior satu untuk median menjadi μ untuk suatu J. Namun, terdapat terlalu banyak sekali parameter untuk memilih sebuah distribusi yang menggambarkan informasi prior yang berarti pada semua θ_{js}, sehingga prior acuan juga dimasukkan. Secara khusus, informasi prior yang berarti akan dibatasi untuk parameter dari beberapa tahap j yang pertama.

Dengan selalu berhubungan, distribusi prior dan posterior yang fokus pada probabilitas yang tinggi di area sekitar $\theta_{js} = 0{,}5$ untuk semua js, akan berperilaku seperti distribusi normal. Hal ini terjadi ketika c besar dalam $\alpha_{js} = cp(j)$. Dengan $\varrho(j)$ yang meningkat, nilai j yang besar mengimplikasikan bahwa probabilitas prior yang tinggi diletakkan pada θ_{js} di dekat 0,5. Hal ini adalah sifat yang telah disebutkan dalam pendahuluan yang memperbolehkan jumlah parameter yang besar sehubungan dengan jumlah pengamatan independennya.

Di lain pihak, ketika c kecil, distribusi tersebut lebih "nonparametrik". Ditentukan A_J merupakan sebuah himpunan dalam pembagian level ke-J. Ketika c kecil, sebuah pengamatan dalam A_J memiliki pengaruh yang besar pada semua distribusi beta posterior dari θ_{js} sehubungan dengan A_J, sehingga menyebabkan probabilitas yang tinggi untuk A_J dalam distribusi posterior. Karena A_J adalah sebuah himpunan dalam pembagian terbaik yang ditentukan, sehingga menyebabkan perilaku yang menonjol yang memperkirakan diskrit dalam posterior.

Distribusi sampling tergeneralisasi tahap ke-J, sebut G, tergantung pada θ_{js}. G bersama dengan prior pada θ_{js}, menentukan diagram pohon Polya yang berhingga, dimana kita tulis

$$G \sim PT_j\left(c, \tilde{n}, N(\grave{\imath}, \acute{o}^2)\right). \qquad (3.20)$$

Sebuah prior pada (μ, σ) mengimplikasikan bahwa median μ, kuartil, oktil, dan sebagainya, adalah random. Hal ini memiliki pengaruh pada proses smoothing.

Kepadatan sampling random yang diperoleh dengan membangkitkan himpunan $\{\theta_{js}\}$ berdasarkan prior mereka, tetapi dirata-ratakan atas sebuah prior pada (μ, σ) diistilahkan sebagai sebuah gabungan diagram pohon Polya, dan ditulis

$$G \sim \int PT_j\left(c, \tilde{n}, N(\grave{\imath}, \acute{o}^2)\right) p(d\grave{\imath}, d\acute{o}^2). \qquad (3.21)$$

Hanson (2006) menunjukkan bahwa untuk prior khusus, kepadatan MPT randomnya halus. Diagram pohon Polya dan versi lain dari penggabungan diagram pohon Polya tidak perlu memiliki sifat ini; lihat Barron et al. (1999), Paddock (1999), dan Berger dan Guglielmi (2001).

3.2. Perhitungan Posterior

Sangat penting dalam menggunakan metode bayes nonparametrik dibandingkan dengan analisis parametrik karena kemampuan untuk memasukkan lebih banyak ketidakpastian tentang distribusi sampling. Akan tetapi, fleksibilitas ini meningkatkan kompleksitas perhitungan analisis. Banyak perkembangan model Bayes nonparametrik yang telah menjadi hasil langsung dari bantuan-bantuan dalam simulasi yang didasarkan pada metode perhitungan, khususnya metode MCMC. Pengenalan metode MCMC dalam bidang tersebut dimulai dengan penelitian Escobar (1994) untuk penggabungan proses Dirichlet. Dalam bagian ini kita membahas beberapa aspek perhitungan dari sampling posterior. Modelnya diberikan sebagai

$$X_1, \dots, X_n | G \overset{iid}{\sim} G,$$
$$G | c, \grave{\imath}, \acute{o} \sim PT_j\left(c, \tilde{n}, N(\grave{\imath}, \acute{o}^2)\right), \quad (3.22)$$

dan

$$(\grave{\imath}, \acute{o}^2) \sim p(\grave{\imath}) p(\acute{o}^2).. \quad (3.23)$$

Disini c ditentukan tetapi dapat juga diperlakukan sebagai random.

4. Simulasi

Pendekatan yang digunakan untuk menganalisa data yang bersifat ordinal dengan menyesuikan serangkaian model logistik seperti rasio kontinuitas atau logit kumulatif. Dengan secara khusus, ditentukan $Y_{ij} = 1$ jika indifidu i mengalami kesulitan pada waktu j, dengan $Y_{ij} = 0$ jika tidak mengalami kesulitan atau ringan dan dapat menentukan model,

$$logit\{P(Y_{ij} = 1 | \hat{a}, \tilde{a}_i)\}$$
$$= \tilde{a}_i + \hat{a}_1 Trt_i + \hat{a}_2 Time_{ij} + \hat{a}_3 Trt_i x Time_{ij} \quad (4.1)$$

Dimana i = 1, …, n, j = 1, …, N_i. Disini y_i merupakan pengaruh random untuk masing-masing subyek, dan $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ adalah parameter regresi yang berhubungan dengan Trt, indikator perlakuan biner, waktu dalam periode, dan interaksi Trt x Waktu. Secara khusus, y_i akan diasumsikan independen $N(\mu, \sigma^2)$. Kita mengganti asumsi normalitas denagn prior MPT,

$$\tilde{a}_1, \dots, \tilde{a}_n | G \overset{iid}{\sim} G,$$
$$G | \grave{\imath}, \acute{o}^2 \sim PT_4\left(c, j^2, N(\grave{\imath}, \acute{o}^2)\right). \quad (4.2)$$

Metodologi dalam tulisan ini memberikan asumsi kenormalan yang pengaruh random dan implikasi kesalahan spesifikasi model normal yang mungkin. Prior acuan yang berbeda untuk θ_{js} ditentukan dengan menggunakan tiga nilai c, yaitu $c = 0{,}1$, $c = 1$, dan $c = 10$, untuk menggambarkan peningkatan tingkat kepercayaan dalam normalitas untuk pengaruh random. Analisis yang lebih awal didasarkan pada asumsi normal menampilkan ketidakkonsistenan antara marginal

dan pengaruh khusus subyek yang membuat dalam menduga validitas analisis berdasarkan asumsi kenormalan.

Analisis Bayes membutuhkan prior baik untuk parameter regresi $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)'$ maupun parameter μ dan σ^2 dari keluarga normal asli yang digeneralisasi oleh MPT. Kita menggunakan normal konjugat independen – prior gamma invers yang menampilkan informasi prior yang lemah.

Adapun simulasi dijalankan dengan menggunakan softwer dan program R, dimana data idnr, trt, dan time digeneret, dan ditentukan model dengan menggunakan persamaan (4.1) dan (4.2). parameter diasumsikan bahwa $\hat{a}_0$ = trt, $\hat{a}_1$ = time, dan $\hat{a}_3$ = trt x time dengan parameter lainnya yaitu ì dan $ó^2$. Untuk menentukan model mana yang terbaik, cukup kita melihat nilai (Deviance Information Criterion) DIC dan (Log Pseudo Marginal Likelihood) LPML yang terkecil.

Untuk menentukan model masing-masing dihitung dengan nilai c. Hasil simulasi untuk tiap-tiap model yang diberikan dalam tabel sebagai berikut:

Tabel 1. Hasil simulasi untuk posterior padamodelMPT dengan ukuran $c = 0.1$

				95%HPD	
Parameter	Mean	Median	Std. Dev	Lower	Upper
(Intercept)	-1.8328	-1.8228	0.4552	-2.6822	-0.9701
Trt	0.2922	0.2412	0.4913	-0.6572	1.2069
Time	-0.3544	-0.3524	0.0411	-0.4291	-0.2756
Trt*time	-0.1268	-0.1253	0.0637	-0.2404	0.0017
ì	-0.7987	-0.8020	0.0589	-0.9292	0.6969
$ó^2$	8.4236	8.5307	1.0476	6.2112	10.1113

Untuk hasil simulasi pada posterior model MPT dengan $c = 0{,}1$ memperoleh (Deviance Information Criterion) DIC = 930.4 dan (Log Pseudo Marginal Likelihood) LPML = - 477.6

Tabel 2. Hasil simulasi untuk posterior padamodelMPT dengan ukuran $c = 1$

				95%HPD	
Parameter	Mean	Median	Std. Dev	Lower	Upper
Intercept	-0.462823	-0.490515	0.530045	-1.382388	0.670885
Trt	0.136144	0.128177	0.717310	-1.250292	1.472916
Time	-0.177666	-0.166746	0.096298	-0.368031	0.006302
Trtxtime	0.029648	0.030614	0.121340	-0.191784	0.276181
ì	-0.52133	-0.55759	0.68342	-1.89266	0.79928
$ó^2$	0.68683	0.42639	0.85987	0.06532	2.17872

Untuk hasil simulasi pada posterior model MPT dengan $c = 1$ memperoleh (Deviance Information Criterion) DIC = 924.2 dan (Log Pseudo Marginal Likelihood) LPML = -468.0.

Tabel 3. Hasil simulasi untuk posterior padamodelMPT dengan ukuran $c = 10$

				95%HPD	
Parameter	Mean	Median	Std. Dev	Lower	Upper
Intercept	-1.7524	-1.7491	0.4429	-2.5020	-0.8428
Trt	-0. 1527	-0. 1743	0.5711	-1. 2850	0. 9228
Time	-0. 3931	-0.3924	0. 0448	-0. 4898	-0.3142
Trtxtime	-0.1349	-0.1357	0.0705	-0.2730	-0.0076
ì	-1.9454	-1.9418	0.6870	-3. 3357	-0. 7314
$ó^2$	16. 8547	16. 2695	3. 9546	10. 4916	24. 8731

Untuk hasil simulasi pada posterior model MPT dengan $c = 10$ memperoleh (Deviance Information Criterion) DIC = 955.3 dan (Log Pseudo Marginal Likelihood) LPML = -481.3.

Tabel 4. Cara posterior, kriteria perbandingan model,dan interval95% HPD untuk model gabungan linier tergeneralisasi (GLMM) parameter

		MPT	
Parameter	á = 10	á = 1	á = 0,1
Intercept	-1.7524	-2.2108	-1.8328
Trt	-0. 1527	0.2310	0.2922
Time	-0. 3931	-0.3841	-0.3544
Trtxtime	-0.1349	-0.1301	-0.1268
ì	-1.9454	-1.2101	-0.7987
$ó^2$	16. 8547	22.7622	8.4236
DIC	955.3	924.2	930.4
LPML	-481.3	-468.0	-477.6
		Posterior Interval	
Intercept	-2.5020, -0.8428	-3.3212, -1.2262	-2.6822, -0.9701
Trt	-1. 2850, 0.9229	-0.7307, 1.0223	-0.6572, 1.2069
Time	-0. 4898, 0.3142	-0.4735, -0.3076	-0.4291, -0.2756
Trt*time	-0.2730, 0.0076	-0.2634, -0.0020	-0.2404, 0.0017
ì	-3. 3357, 0.7314	-3.0559, 0.0173	-0.9292, 0.6969
$ó^2$	10. 4916, 24.8731	8.8015, 44.5144	6.2112, 10.1113

Hasil simulasi di atas yang ditunjukkan pada tabel 4.4 dapat kita simpulkan bahwa model yang dipakai atau yang terbaik yaitu pada c sama dengan 1 dengan nilai DIC dan LPML nya adalah model penyesuaian terbaik. Model ini lebih baik dibandingkan dengan model c = 0.1 dan c = 10.

5. Kesimpulan dan Saran

5.1. Kesimpulan

Kemampuan untuk menyelesaikan model yang rumit selalu digunakan analisis bayes. Hal ini bergantung pada kemampun untuk menggunakan metode Rantai Markof Monte Carlo (MCMC) yang dapat memprediksi distribusi posterior. Dengan kata lain analisi bayes mampu memberikan fleksibilitas yang besar akan tetapi tidak menggunakan fleksibilitas itu jika data tersebut benar-benar membutuhkannya.

Diagram pohon polya merupakan sebuah distribusi probabilitas random, dengan mempelajari MPT kita dapat mengambarkan sebuah proses generalisasi distribusi parametrik yang normal. Generalisasi tersebut dengan beberapa tahapan dapat kita memperkenalkan parameter baru untuk menggeneralisasi tahapan sebelumnya, parameter yang dipakai berupa è $atau$ â

Berdasarkan simulasi dengan menggunakan tiga nilai acuan c = 0,1, c =1 serta c = 10, diperoleh hasil model penyesuaian terbaik pada c = 1 didasarkan pada nilai DIC dan LPMLnya.

5.2. Saran

Sangat menarik untuk penulisan lebih lanjut merupakan konstruksi diagram pohon polya untuk ruang multidimensional, dan memperluas diagram pohon polya untuk menentukan model probabilitas atas distribusi yang berhubungan, yaitu diagram pohon polya dependen dalam waktu, ruang, atau nilai kovariant silang. Diagram pohon polya dapat juga diadaptasi untuk menghasilkan realisasi dependen yang sesuai untuk memodelkan data kurva pertumbuhan. Merupakan pengaruh random yang mudah dirubah, atau aplikasi lain yang membutuhkan evolusi kompleks sebuah kepadatan dalam waktu, ruang atau dengan kovariatnya.

Daftar Pustaka

Bain, L.J., and Max Engelhardt (1992) Introduction to Probability and Mathematical Statistics, $2^{nd} edn$, Duxbury Press, California.

DeGroot, M.H. (1970) Optimal Statistical Decision, McGraw-Hill.

Rohatgi, V.K. (1976) An Introduction to Probability Theory and Mathematical Statistics, John Wiley-Sons.

W.J.,Conover. (1971) Practical Nonparametric Statistics, John Wiley-Sons, New York-London-Sydney-Toronto.

George E.P.Box and George C.Tiao. (1973) Bayesian Inference In Statistical Analyssis, Addison-Wesley Publishing Company.

James O. Berger (1980) Statistical Decision Theory Foundations, and Methods, Springer-Verlag, New York Heidelberg Berlin.

Mauldin,R.D.,Sudderth, W.D., and Williams, S.C. (1992). "Polya Trees and Random Distributions," *The Annals of Statistics*, 20, 1203-1221.

Reich, B.J., and Fuentes, M. (2007), "Multivariate Semiparametric Bayesian Spatial Modeling Framework for Hurricane Surface Wind Fields," *Annals of Applied Statitics*, 1, 249-264.

Dey,D., Muller, P., and Sinha, D. (1998), Partical Nonparametric and Semiparametric Bayesian Statistics, New York: Springer.

Paddock, S.M.(1999), "Randomized Polya Trees: Bayesian Nonparametrics for Multivariate Data Analysis," Unpublished doctoral thesis, Institute of Statistics and Decision Sciences, Duke University.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Van der Line, A. (2000),"Bayesian Measures of Model Complexity and Fit, "*Journal of the Royal Statistical Society, Series B*, 64, 583-639.

Paddock, S.M., Ruggeria, F., Lavine M., and West, M. (2003), "Randomized Polya Tree Models for Nonparametric Bayesian inference," *Statistica Sinica*, 13, 443-460.

Walker, S.G., and Mallick, B.K. (1999), "Semiparametric Accelerated Life Time Model," *Biometrics*, 55, 477-483.

Warren, J., Ewens, and Gregory, R., Grant. (2005), "Statisical Methods in Bioinformatics: An Introduction, Springer.