BAB II

LANDASAN TEORI

Bab ini akan membahas landasan teori yang menjadi dasar dalam pendeteksian cluster dengan metode semi-parametrik, yaitu data spasial, model rasio densitas semi-parametrik, kelas eksponensial dari pdf, estimasi maksimum likelihood, uji rasio likelihood, metode Langrange, dan *Bonferroni Correction*.

2.1 Data Spasial

Data spasial adalah suatu hasil pengukuran yang memuat informasi mengenai lokasi dari pengukuran. Data spasial adalah data dependen, karena berasal dari lokasi spasial yang berbeda yang mengindikasikan ketergantungan antara nilai pengukuran dengan lokasi.

2.2 Model Rasio Densitas Semi-parametrik

Metode semi-parametrik yang digunakan dalam tugas akhir ini didasarkan pada model rasio densitas semi-parametrik yang dipelajari oleh Fokianos (2001).

Perhatikan $m = q + 1$ sampel yang saling bebas,

$$\begin{aligned} x_1 &= (x_{11}, x_{12}, \ldots, x_{1n_1})' \sim g_1(x) \\ x_2 &= (x_{21}, x_{22}, \ldots, x_{2n_2})' \sim g_2(x) \\ &\vdots \\ x_m &= (x_{m1}, x_{m2}, \ldots, x_{mn_m})' \sim g_m(x) \end{aligned}$$

dimana $g_j(x)$ adalah p.d.f dari x_{ji}, $j = 1, \ldots, m$, $i = 1, \ldots, n_j$. Pilih sampel ke-m sebagai sampel acuan dan $g_m(x)$ sebagai densitas acuan. Digunakan model rasio densitas semi-parametrik sebagai berikut:

$$\frac{g_j(x)}{g_m(x)} = \exp\{(\alpha_j + \boldsymbol{\beta}_j'\boldsymbol{h}(x))\}, \quad j = 1, \ldots q, \quad q = m - 1 \qquad (2.3.1)$$

dimana $\boldsymbol{h}(x)$ adalah fungsi dari x yang dapat berupa skalar, misalnya x, x^2, atau $\log x$, atau berupa vektor, misalnya $(x,\ x^2)'$ atau $(x,\ \log x)'$, dan lainnya. α_j berupa skalar, sedangkan $\boldsymbol{\beta}_j$ dapat berupa skalar atau vektor tergantung pada $\boldsymbol{h}(x)$. Terlihat bahwa jika $\boldsymbol{\beta}_j = \mathbf{0}$ maka $\alpha_j = 0$. Hal ini dikarenakan dengan mensubstitusi $\boldsymbol{\beta}_j = \mathbf{0}$ ke persamaan (2.3.1) diperoleh:

$$\begin{aligned} g_j(x) &= \exp(\alpha_j) \cdot g_m(x) \\ \int_{-\infty}^{\infty} g_j(x)dx &= \int_{-\infty}^{\infty} \exp(\alpha_j) \cdot g_m(x)dx \\ \int_{-\infty}^{\infty} g_j(x)dx &= \exp(\alpha_j) \int_{-\infty}^{\infty} g_m(x)dx \end{aligned}$$

karena $g_m(x)$ merupakan densitas maka $\int_{-\infty}^{\infty} g_m(x)dx = 1$, sehingga diperoleh

$\int_{-\infty}^{\infty} g_j(x)dx = \exp(\alpha_j)$. Dikarenakan $g_j(x)$ juga merupakan densitas maka haruslah $\int_{-\infty}^{\infty} g_j(x)dx = 1$. Oleh karena itu, $\exp(\alpha_j)$ haruslah sama dengan 1 atau $\alpha_j = 0$.

Hipotesis $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_q = \mathbf{0}$ menyatakan bahwa semua *m* sampel berasal dari suatu distribusi dengan densitas probabilitas $g_m(x) \equiv g(x)$.

2.3 Kelas Eksponensial dari pdf

2.3.1 Kasus dengan Satu Parameter

Pandang suatu famili pdf $\{f(x;\theta : \theta \in \Omega\}$ dimana Ω adalah himpunan interval $\Omega = \{\theta : \gamma < \theta < \delta\}$, dimana γ dan δ adalah konstanta-konstanta yang diketahui, dan dimana

$$\begin{aligned} f(x;\theta) &= \exp[p(\theta)K(x) + S(x) + q(\theta)], \quad \mathrm{a} < x < \mathrm{b} \\ &= 0, \text{untuk lainnya} \end{aligned}$$

Suatu pdf yang berbentuk seperti di atas dikatakan anggota dari kelas eksponensial dari pdf berjenis kontinu.

Suatu pdf

$$\begin{aligned} f(x;\theta) &= \exp[p(\theta)K(x) + S(x) + q(\theta)], \; x = \mathrm{a}_1, \mathrm{a}_2, \mathrm{a}_3, \ldots \\ &= 0, \text{untuk lainnya} \end{aligned}$$

dikatakan merupakan anggota dari kelas eksponensial dari pdf berjenis diskrit.

2.3.2 Kasus dengan Beberapa Parameter

Misalkan X_1, X_2, ..., X_n, $n > m$, menyatakan sampel random dari suatu distribusi yang bergantung pada m parameter dan mempunyai p.d.f sebagai berikut:

$$\begin{aligned} f(x;\theta_1,\theta_2,\ldots,\theta_m) &= \exp\left[\sum_{j=1}^{m} p_j(\theta_1,\theta_2,\ldots,\theta_m)K_j(x) + S(x) + q(\theta_1,\theta_2,\ldots,\theta_m)\right], a < x < b \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad , \text{lainnya} \end{aligned}$$

Suatu p.d.f dengan bentuk seperti di atas dikatakan anggota dari kelas eksponensial dari p.d.f berjenis kontinu.

Suatu p.d.f

$$\begin{aligned} f(x;\theta_1,\theta_2,\ldots,\theta_m) &= \exp\left[\sum_{j=1}^{m} p_j(\theta_1,\theta_2,\ldots,\theta_m)K_j(x) + S(x) + q(\theta_1,\theta_2,\ldots,\theta_m)\right], x = a_1, a_2, \ldots, \\ &= 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad , \text{lainnya} \end{aligned}$$

dikatakan anggota dari kelas eksponensial dari p.d.f berjenis diskrit.

2.4 Metode Lagrange

Metode Langrange merupakan suatu metode yang digunakan untuk memaksimumkan suatu fungsi yang terdiri dari beberapa variabel dengan satu atau lebih kendala. Misalkan akan dimaksimumkan suatu fungsi $f(p)$ terhadap kendala $g_i(p) = 0$. Selanjutnya didefinisikan suatu fungsi Lagrange sebagai berikut:

$$L(p,\lambda_1,\lambda_2,\cdots,\lambda_n) = f(p) - \lambda_1 g_1(p) - \lambda_2 g_2(p) - \cdots - \lambda_n g_n(p)$$

dimana $g_i(p) = 0$, $i = 1, \ldots, n$ merupakan kendala dan $\lambda_1, \lambda_2, \ldots, \lambda_n$ merupakan pengali Lagrange. Lalu selesaikan sistem persamaan berikut secara bersamaan:

$$\text{grad}(L(p, \lambda_1, \lambda_2, \cdots, \lambda_n)) = 0,$$

$$g_i(p) = 0,\ i = 1, \ldots, n.$$

Setiap titik p yang memenuhi sistem persamaan di atas disebut titik kritis untuk masalah nilai ekstrem dengan kendala.

2.5 Estimasi Maksimum Likelihood

Maksimum likelihood adalah suatu metode yang digunakan untuk melakukan penaksiran titik dari suatu parameter yang tidak diketahui dalam suatu fungsi probabilitas.

Misalkan X adalah peubah acak yang mempunyai bentuk fungsi probabilitas tertentu tetapi fungsi probabilitasnya bergantung kepada suatu parameter θ yang tidak diketahui. Fungsi probabilitas X dapat ditulis sebagai $f(x;\theta), \theta \in \Omega$, dengan Ω adalah ruang parameter.

Definisi 2.4.1

Misalkan $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$ adalah sampel acak berukuran n dari populasi yang memiliki fungsi pdf, $f(x;\theta)$, dengan $\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ merupakan kumpulan parameter pdf sebanyak n, yang akan diestimasi.

Fungsi likelihood didefinisikan sebagai pdf bersama dari observasi x_i yang ditulis sebagai

$$L(\theta_1,\theta_2,\ldots,\theta_n;x_1,x_2,\ldots,x_n)=\prod_{i=1}^{n} f(x_i,\theta_1,\theta_2,\ldots,\theta_n) \qquad (2.4.1)$$

akan dicari taksiran parameter yaitu $\hat{\Theta}=\{\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_n\}$ yang memaksimumkan persamaan (2.4.1).

Secara matematis, estimasi maksimum likelihood lebih mudah dilakukan dengan memanipulasi persamaan (2.4.1) menjadi logaritma fungsinya. Memaksimumkan fungsi likelihood ekivalen dengan memaksimumkan logaritma fungsi likelihood. Didapatkan bentuk logaritma fungsi likelihood untuk persamaan (2.4.1)

$$\ln L=\sum_{i=1}^{n}\ln(f(x_i,\theta_1,\theta_2,\ldots,\theta_n)) \qquad (2.4.2)$$

Untuk mencari nilai taksiran parameter $\hat{\Theta}=\{\hat{\theta}_1,\hat{\theta}_2,\ldots,\hat{\theta}_n\}$ yang memaksimumkan (2.4.2), maka pendekatan yang paling sering digunakan adalah menentukan turunan parsial dari (2.4.2) untuk setiap parameter lalu menyamakan dengan nol,

$$\frac{\partial \ln(L)}{\partial \theta_j}=0 \quad \text{untuk } j = 1, 2, \ldots, k$$

Berdasarkan persamaan di atas, akan diperoleh persamaan sebanyak parameter yang tidak diketahui. Taksiran ini dapat diselesaikan secara bersamaan. Jika penyelesaian dari fungsi turunan parsial tidak dapat ditemukan, maka pendekatan numerik dilakukan.

2.6 Uji Rasio Likelihood

Uji rasio likelihood adalah suatu metode yang digunakan untuk menguji hipotesis H_0 : $\theta \in \omega$ terhadap hipotesis alternatif H_1 : $\theta \in \Omega$, dengan Ω adalah ruang parameter keseluruhan dan ω adalah ruang parameter dalam H_0, $\omega \subset \Omega$.

Misalkan X_1, X_2, …, X_n dinotasikan sebagai n variabel random yang saling bebas dengan masing-masing, dan secara berurutan pdf-nya adalah $f_i(x_i, \theta_1, \theta_2, ..., \theta_m)$, i = 1, 2, …, n. Himpunan yang terdiri dari seluruh *point* parameter $(\theta_1, \theta_2, ..., \theta_m)$ dinotasikan oleh Ω, yang disebut sebagai ruang parameter. Misalkan ω sebagai suatu subset dari ruang parameter Ω. Ingin dilakukan uji hipotesis H_0 : $(\theta_1, \theta_2, ..., \theta_m) \in \omega$ terhadap seluruh hipotesis alternatif. Didefinisikan fungsi likelihood, sebagai berikut:

$$L(\omega) = \prod_{i=1}^{n} f_i(x_i, \theta_1, \theta_2, \ldots, \theta_m), \qquad (\theta_1, \theta_2, \ldots, \theta_m) \in \omega$$

dan

$$L(\Omega) = \prod_{i=1}^{n} f_i(x_i, \theta_1, \theta_2, \ldots, \theta_m), \qquad (\theta_1, \theta_2, \ldots, \theta_m) \in \Omega$$

Definisi 2.5.1 Rasio Likelihood (Hog & Craig, 1995)

Misalkan $L(\hat{\omega})$ dan $L(\hat{\Omega})$ fungsi likelihood yang memaksimumkan dan diasumsikan kedua fungsi likelihood ini ada. Rasio dari $L(\hat{\omega})$ terhadap $L(\hat{\Omega})$ disebut ***likelihood ratio*** dan dinotasikan sebagai berikut

$$\Lambda = \frac{L(\hat{\omega})}{L(\hat{\Omega})} \tag{2.5.1}$$

Dalam menggunakan pengujian likelihood rasio, kita harus mengetahui distribusi dari statistik Λ di bawah kondisi H_0 benar. Dalam beberapa kasus, Λ merupakan suatu fungsi dari statistik lain yang distribusinya diketahui. Sedangkan dalam beberapa kasus lainnya, kita dapat mengaproksimasi distribusi dari Λ dengan suatu distribusi tertentu seperti distribusi Chi-Square. Dengan teorema berikut ini, kita dapat mengaproksimasi distribusi dari Λ dengan distribusi Chi-Square.

Teorema 2.5.1

Misalkan $X_1, \ldots, X_n$ i.i.d variabel random dengan suatu densitas yang memenuhi kondisi B1 sampai B6 yang terdapat pada Lampiran 1 dengan $I(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$ dimana $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Jika MLE $\hat{\boldsymbol{\theta}}_n$ memenuhi

$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N(0, I^{-1}(\boldsymbol{\theta}))$ maka LR statistik $\Lambda_n = \prod_{i=1}^{n} \frac{f(X_i; \hat{\boldsymbol{\theta}}_n)}{f(X_i; \hat{\boldsymbol{\theta}}_{n0})}$ untuk menguji $H_0 : \theta_1 = \theta_{10}, \ldots, \theta_r = \theta_{r0}$ memenuhi

$$2\ln(\Lambda_n) \xrightarrow{d} V \sim \chi_r^2$$

di bawah kondisi H_0 benar.

Pembuktian teorema di atas dapat dilihat pada Lampiran 2.

2.7 *Bonferroni Correction*

Bonferroni correction merupakan suatu metode yang dapat digunakan untuk mengatasi permasalahan yang terjadi jika dilakukan banyak pengujian secara simultan. Dengan *Bonferroni correction*, tingkat signifikansi yang digunakan pada setiap pengujian yaitu α / n dimana n menyatakan banyaknya pengujian, sehingga tingkat signifikansi untuk keseluruhan pengujian yaitu α .