GRAFIK PENGENDALI NON PARAMETRIK EMPIRIK

Oleh : Rukun Santoso
Program Studi Statistika FMIPA UNDIP

Abstract

Shewhart control chart is constructed base on the normality assumption of process. If the normality is fail then the empirical control chart can be an alternative solution. This means that the control chart is constructed base on empirical density estimator. In this paper the density function is estimated by kernel method. The optimal bandwidth is selected by *leave one out Cross Validation* method. The result of empirical control chart will be compared to ordinary Shewhart chart.

Key words : Control chart, Kernel, Cross Validation

1. PENDAHULUAN

Diagram pengendali rata-rata $\overline{X}$ untuk melakukan pengawasan terhadap tanda-tanda tak terkendali telah digunakan sejak tahun 1941 yaitu ketika diperkenalkan oleh Shewhart. Oleh karenanya kemudian dikenal sebagai grafik pengendali Shewhart. Tanda tak terkendali diperoleh jika ada titik jatuh di luar batas pengendali 3σ, baik untuk data kontinu maupun attribut. Asumsi normalitas proses diperlukan dalam menyusun diagram pengendali Shewhart untuk data kontinu.

Diagram pengendali dapat dibangun secara nonparametrik berdasarkan densitas peluang empirik. Penggunaan densitas empirik dengan pendekatan deret Fourier untuk membangun diagram pengendali telah dibahas oleh Rukun Santoso (2007). Jika asumsi normalitas proses dipenuhi maka pengendali nonparametrik dengan pendekatan deret Fourier memberikan hasil yang sama dengan diagram Shewhart. Namun jika asumsi normalitas proses tidak dipenuhi maka diagram pengendali nonparametrik memberikan batas-batas pengendali yang lebih realistik.

Diagram pengendali nonparametrik yang dibahas dalam tulisan ini adalah disusun dengan pendekatan kernel, yaitu densitas peluang empirik dibangun dengan metode kernel. Estimasi densitas dengan pendekatan deret fourier mengasumsikan bahwa fungsi densitas f adalah anggota dari ruang Hilbert $L_2[-\pi, \pi]$ sehingga data pengamatan dalam range $(-\infty, \infty)$ harus ditransformasikan ke dalam range $[-\pi, \pi]$. Pendekatan kernel dibangun berdasarkan kenyataan bahwa fungsi densitas f adalah anggota dari ruang Hilbert $L_2(R)$ sehingga dapat dilakukan pada setiap data pengamatan berharga riil tanpa melalui proses transformasi.Beberapa rujukan tentang pendugaan densitas dengan metode kernel antara lain Hardle[1] dan Ogden [3].

Karena metode ini memerlukan banyak perhitungan numerik maka untuk memudahkan pekerjaan dan mendapatkan hasil yang memuaskan diperlukan bantuan komputer. Simulasi komputer dikerjakan dengan paket S-Plus yang memungkinkan memadukan antara pemrograman, perhitungan statistik dan komputer grafis[4].

2. FUNGSI DALAM $L_2(R)$

Diberikan f fungsi terukur yang didefinisikan pada himpunan terukur $E \subset R$. Fungsi f dikatakan terintegral kuadrat (Lebesgue) jika f^2 terintegral Lebesgue pada E. Himpunan semua fungsi terukur yang terintegral kuadrat pada E dinotasikan dengan $L_2(E)$

$$L_2(E)=\left\{ f : \int_E f^2 < \infty \right\}$$

merupkan ruang linier. Lebih lanjut terhadap norma $\|\bullet\|$ dengan aturan jika $f \in L_2(E)$ didefinisikan $\|f\| = \left\{ \int_E f^2 \right\}^{1/2}$ maka $L_2(E)$ merupakan ruang Banach.

Jika $L_2(E)$ diperlengkapi dengan *inner product* $\langle .,. \rangle$ dengan aturan jika $f,g \in L_2(E)$ didefinisikan $\langle f, g \rangle = \int_E fg$ maka $L_2(E)$ merupakan ruang pre Hilbert. Lebih lanjut ruang pre Hilbert $L_2(E)$ terhadap norma $\|\bullet\|$ di atas merupakan ruang Hilbert.

Definisi 2.1

Dua fungsi $f,g \in L_2(E)$ dikatakan saling ortogonal jika $\langle f, g \rangle = 0$

Definisi 2.2

Barisan fungsi $\{f_n\} \subset L_2(E)$ dikatakan ortonormal jika untuk setiap indek i berlaku $\|f_i\| = \sqrt{\langle f_i, f_i \rangle} = 1$ dan $\langle f_i, f_j \rangle = 0$ untuk $i \neq j$

Definisi 2.3

Barisan fungsi $\{f_n\} \subset L_2(E)$ dikatakan sistem ortonormal lengkap (Complete Orthonormal System=CONS) jika $\{f_n\}$ ortonormal dan jika $g \in L_2(E)$ sedemikian hingga $\langle f_i, g \rangle = 0$ untuk setiap indek i, maka g adalah fungsi nol.

Teorema 2.1

Jika $\{f_n\} \subset L_2(E)$ merupakan sistem ortonormal lengkap maka untuk setiap $f \in L_2(E)$ dapat dinyatakan sebagai

$$f = \sum_{i=i}^{\infty} \langle f, f_i \rangle f_i$$

Bukti :

Diketaui $\{f_n\}$ CONS berarti jika $g \in L_2(E)$ dan $\langle f_i, g \rangle = 0$ untuk setiap i maka $g=\mathcal{O}$.

Ambil $g = f - \sum_{i=i}^{\infty} \langle f, f_i \rangle f_i$ maka untuk sebarang indek k berlaku $\left\langle f - \sum_{i=1}^{n} \langle f, f_i \rangle f_i, f_k \right\rangle = 0$ dengan kata lain $f = \sum_{i=i}^{\infty} \langle f, f_i \rangle f_i$. ■

3. FUNGSI DENSITAS EMPIRIK

Jika F(x) menyatakan fungsi distribusi kumulatif (CDF) dari random variabel X maka peluang suatu observasi sama dengan atau lebih kecil dari x adalah P(X≤x)=F(x). Karena fungsi densitas f(x) didefinisikan sebagai turunan dari F(x) maka dapat dituliskan sebagai :

$$f(x) = \lim_{\lambda \to 0} \frac{1}{2\lambda} \left(F(x+\lambda) - F(x-\lambda) \right)$$

Fungsi densitas ini dapat ditaksir dengan fungsi densitas empirik

$$\hat{f}(x) = \lim_{\lambda \to 0} \frac{1}{2\lambda} \left(\tilde{F}(x+\lambda) - \tilde{F}(x-\lambda) \right)$$

$$= \frac{1}{2n\lambda} .\#x$$

#x menyatakan banyaknya data yang berada dalam interval (x-λ, x+λ]

Jika didefinisikan fungsi kernel

$$K(x) = \begin{cases} 1/2 & -1 < x \le 1 \\ 0 & \text{yang lain} \end{cases}$$

maka fungsi densitas empirik di atas dapat dituliskan sebagai

$$(3.1) \quad \hat{f}_\lambda(x) = \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{x - X_i}{\lambda}\right)$$

dengan X_i=sampel ke-i, i=1,2,...,n dan λ= lebar *bandwidth*

Beberapa fungsi kernel yang terkenal antara lain Kernel Gausian, Kernel Uniform, Kernel Triangle, dan Kernel Epanechnikov. Setiap fungsi kernel mempunyai sifat sebagai fungsi densitas dan simetri terhadap garis x=0. Secara analitis telah dibuktikan bahwa setiap fungsi kernel dapat digunakan untuk memberikan pendekatan terhadap persamaan 3.1.

4. PENDUGA DENSITAS TERBAIK

Kebaikan penduga densitas kernel ditentukan dua hal penting yaitu pemilihan fungsi kernel dan lebar *bandwidth.* namun yang paling menentukan adalah pemilihan bandwidth yang tepat (optimal). Salah satu metode memilih bandwidth optimal adalah menggunakan metode *Least Squares Cross Validation*. Dibentuk persamaan jarak antara fungsi densitas f dan fungsi penduga $\hat{f}_\lambda$ dinyatakan sebagai

$$d_f(\lambda) = \int (\hat{f}_\lambda - f)^2(x)dx$$

$$= \int \hat{f}_\lambda^2(x)dx - 2\int (\hat{f}_\lambda f)(x)dx + \int f^2(x)dx$$

$$= A + B + C$$

Bagian A dapat dihitung dari data dan bagian C merupakan nilai konstan yang tidak tergantung kepada λ, sehingga meminimalkan $d_f(\lambda)$ adalah identik dengan meminimalkan

$$(4.1).... \; d_f(\lambda) - \int f^2(x)dx = \int \hat{f}_\lambda^2(x)dx - 2\int (\hat{f}_\lambda f)(x)dx \; .$$

Bagian B adalah bentuk dari $2E_X[f_\lambda(X)]$ yang harus diduga dari data. Dengan menggunakan metode *leave one out cross-validation* diperoleh

$$E_X[\hat{f}_\lambda(X)] = n^{-1} \sum_{i=1}^{n} \hat{f}_{\lambda,i}(X_i)$$

dengan

$$\hat{f}_{\lambda,i}(X_i) = (n-1)^{-1}\lambda^{-1}\sum_{j\neq i} K(\frac{X_i - X_j}{\lambda})$$

Sehingga memilih λ yang meminimalkan (4.1) dapat didekati secara numerik dengan memilih λ yang meminimalkan

$$(4.2)... CV(\lambda) = \int \hat{f}_{\lambda}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{\lambda,i}(X_i)$$

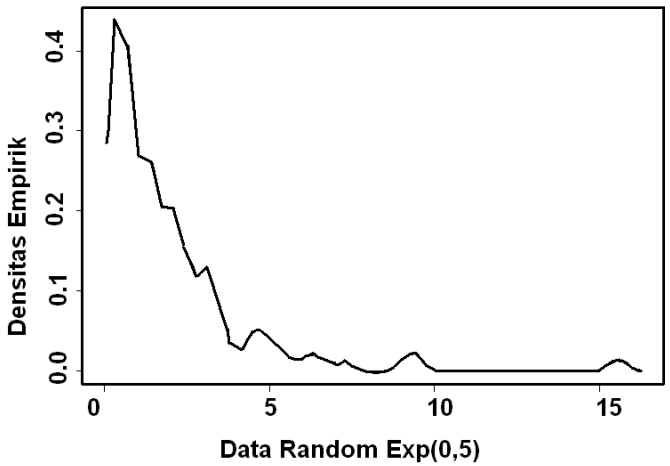
Untuk membantu penyusunan program komputer dapat memngikuti algorithma berikut

1. Tentukan beberapa nilai awal untuk λ yaitu $\lambda_1<\lambda_2<...<\lambda_k$
2. Hitung $CV(\lambda_i)$, i=1,2,...,k
3. Terdapat 3 kemungkinan
   a. $CV(\lambda_1)< CV(\lambda_2)<...< CV(\lambda_k)$ berarti λ optimum yang sebenarnya berada di sebelah kiri λ_1. Ulangi langkah 1 untuk nilai-nilai λ di sebelah kiri λ_1
   b. $CV(\lambda_1)> CV(\lambda_2)>...> CV(\lambda_k)$ berarti λ optimum yang sebenarnya berada di sebelah kanan λ_k. Ulangi langkah 1 untuk nilai-nilai λ di sebelah kanan λ_k
   c. Terdapat indek i sehingga $CV(\lambda_{i-1})> CV(\lambda_i)=CV(j)< CV(\lambda_{i+1})$ dengan j menyatakan tingkat iterasi, berarti λ optimum yang sebenarnya berada di sekitar λ_i. Ulangi langkah 1 untuk nilai-nilai λ di sekitar λ_i. Iterasi dihentikan jika telah diperoleh $|CV(j)-CV(j+1)|<\varepsilon$

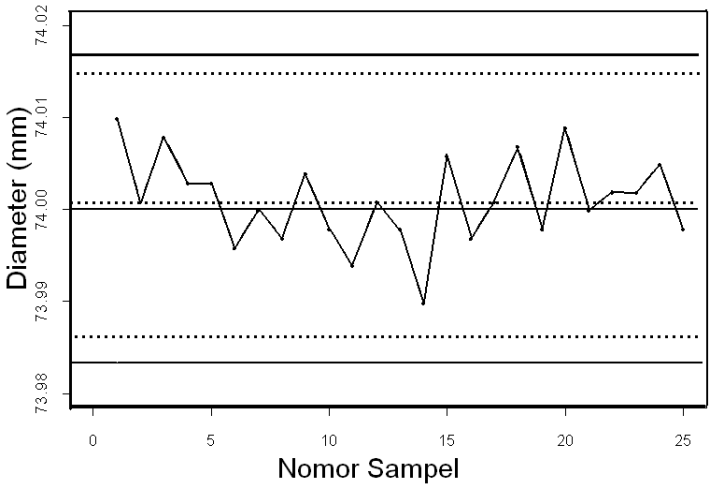
5. HASIL DAN PEMBAHASAN

Untuk membandingkan bentuk diagram pengendali berdasarkan fungsi densitas hampiran (nonparametrik) dan berdasarkan asumsi kenormalan digunakan data diameter cicncin piston dari Montgomery (*Introduction to Quality Control*, 2001, halaman 213) yang telah diyakini berasal dari proses berdistribusi normal. Bentuk diagram pengendali 3σ dengan metode Shewhart (asumsi normalitas proses) dan metode nonparametrik tersaji dalam gambar 5.2. Kedua metode memberikan batas-batas pengendali yang sama, perbedaan kecil mungkin terjadi sebagai akibat pembulatan angka, sehingga kedua metode memberikan penafsiran yang sama.

Gambar 5.3 menggambarkan grafik pengendali 3σ untuk $\overline{X}$ dari suatu proses yang berdistribusi eksponensial dengan λ=0.5. Data percobaan diambil secara random melalui simulasi komputer dengan jumlah ulangan 25 kali dan masing-masing berukuran 5. Densitas hampiran dengan metode Kernel dari data random yang dihasilkan divisualisasikan pada Gambar 5.1 Terdapat perbedaan batas-batas grafik pengendali dengan asumsi normalitas dan perhitungan secara nonparametrik. Asumsi normal memberikan estimasi varian yang terlalu rendah dibandingkan dengan varian hasil perhitungan berdasarkan fungsi densitas empirik. Akibatnya batas-batas kendali dengan asumsi normalitas memberikan kisaran yang lebih sempit. Akibatnya pada sampel ke-15 terjadi satu titik di luar batas kendali, sedangkan pada grafik pengendali empirik titik tersebut masih dalam kategori terkendali.



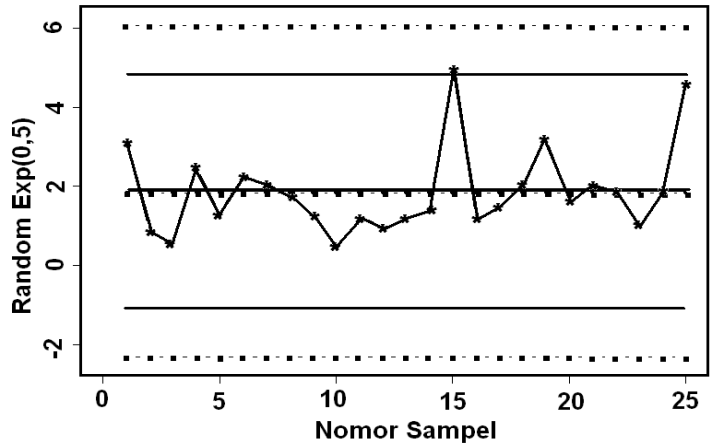
Gambar 5.1 Densitas Empirik Data Random Exponensial(0,5)



Keterangan
—— hampiran Kernel
UCL = 74.01673
Center = 74
LCL = 73.98327

..... Asumsi Normal
UCL = 74.014
Center = 74.001
LCL = 73.988

Gambar 5.2. Diagram Pengendali $\overline{X}$ Data Diameter Cincin Piston



Keterangan
—— Asumsi Normal
UCL = 4.227402
Center = 1.872352
LCL = -0.4826982

…… hampiran Kernel
UCL = 6.064632
Center = 1.872182
LCL = -2.320268

Gambar 5.3 Diagram Pengendali $\overline{X}$ Proses Random Exponensial(0.5)

6. KESIMPULAN

Jika asumsi normalitas dari variabel proses dipenuhi maka diagram pengendali Shewhart dan diagram pengendali nonparametrik memberikan hasil yang sama, namun jika asumsi normalitas tersebut tidak dipenuhi maka kedua metode memberikan hasil yang berbeda. Perbedaan tersebut dikarenakan prosedur yang berlaku pada asumsi normal memberikan penduga varian proses dengan bias yang lebih besar dibandingkan dengan prosedur nonparametrik.

Daftar Pustaka

[1] Hardle, W., *The Smoothing Techniques with Implementation in S,* Springer, 1990.
[2]. Montgomery, D.C., *Introduction to Statistical Quality Control,* John Wiley, 2005.
[3]. Ogden, R.Todd, *Essential Wavelets for Statistical Applications and Data Analysis,* Birkhäuser: Berlin, 1997
[4]. StatSci Division, *S-PLUS User Guide* Math Soft, Inc. Seattle, 1995.
[5]. Walter,G.G , *Wavelets and Other Orthogonal Systems with Applications,* CRCPress: Boca Raton, Florida, 1994.

Lampiran

Program S-Plus untuk menghitung batas-batas kendali empirik dan menggambar grafik pengendalinya

```
function(a)
{
     x <- apply(a, 1, mean)
     s2 <- apply(a, 1, var)
     n <- length(x)
     x1 <- sort(x)
     s1 <- sqrt(mean(s2))
     x.bar <- mean(x)
     r.bar <- mean(apply(a, 1, jangkauan))
     densitas <- density(x)
     runs <- c(1:n)
     h <- hker.opt(x)[[3]]
     ycv <- rep(0, n)
     m1 <- rep(0, n)
     m2 <- rep(0, n)
     v1 <- rep(0, n)
     v2 <- rep(0, n)
     for(i in 1:n) {
          ycv[i] <- kde(x1[ - i], h, x1[i])
     }
     for(i in 1:(n - 1)) {
          m1[i] <- x1[i] * ycv[i] * (x1[i + 1] - x1[i])
          m2[i] <- x1[i] * ycv[i + 1] * (x1[i + 1] -
x1[i])
          v1[i] <- (x1[i]^2) * ycv[i] * (x1[i + 1] -
x1[i])
```

```
          v2[i] <- (x1[i]^2) * ycv[i + 1] * (x1[i + 1] -
x1[i])
     }
     std <- sqrt(sum(densitas$y * (densitas$x -
mean(densitas$x))^2) *
           (densitas$x[4] - densitas$x[3]))
     m <- sum(densitas$y * densitas$x) * (densitas$x[4]-
densitas$x[3])
     lbx <- m - 3.1 * std
     ubx <- m + 3.1 * std
     bpa <- m + 3 * std
     bpb <- m - 3 * std
     bak1 <- x.bar + 1.427 * s1
     bbk1 <- x.bar - 1.427 * s1
     runbpa <- rep(bpa, n)
     runbpb <- rep(bpb, n)
     runx.bar <- rep(x.bar, n)
     runm <- rep(m, n)
     runbak <- rep(bak1, n)
     runbbk <- rep(bbk1, n)
     plot(runs, runbpa, type = "b", xlim = c(0, n), ylim
= c(lbx, ubx))
     par(new = T)
     plot(runs, runm, type = "p", xlim = c(0, n), ylim =
c(lbx, ubx))
     par(new = T)
     plot(runs, runbpb, type = "b", xlim = c(0, n), ylim
= c(lbx, ubx))
     par(new = T)
     plot(runs, x, type = "b", xlim = c(0, n), ylim =
c(lbx, ubx))
     par(new = T)
     plot(runs, runbak, type = "l", xlim = c(0, n), ylim
= c(lbx, ubx))
     par(new = T)
     plot(runs, runx.bar, type = "l", xlim = c(0, n),
          ylim = c(lbx, ubx))
     par(new = T)
     cat("beres", "\n")
     plot(runs, runbbk, type = "l", xlim = c(0, n), ylim
= c(lbx, ubx))
     cat("beres", "\n")
     return(m, bpa, bpb, std, x.bar, bak, bbk, s1)
}
```