BAB III

METODE CROSS-VALIDATION PENGOPTIMALAN BANDWIDTH DALAM PENGHALUSAN FUNGSI DENSITAS

Estimator densitas kernel adalah metode yang cukup fleksibel untuk mengestimasi fungsi densitas. Teknik penghalusan estimator densitas kernel sangat diperlukan agar memperoleh pendekatan yang optimal terhadap fungsi densitas yang belum diketahui.

Sebagaimana dijelaskan pada bab terdahulu, salah satu faktor yang sangat berpengaruh dalam penghalusan estimator densitas kernel adalah parameter penghalus atau biasa disebut bandwidth. Oleh karena itu pemilihan bandwidth sangat penting untuk memperoleh pendekatan yang optimal.

Pada bab ini akan dibahas mengenai pemilihan bandwidth dengan metode Cross-validation khususnya menggunakan algoritma direct. Pembahasan akan diawali dengan aturan dan batasan yang bisa digunakan sebagai acuan dalam pemilihan bandwidth.

3.1. Batasan Pemilihan Bandwidth

Mengingat begitu luasnya kemungkinan mengenai besar bandwidth optimal, maka perlu adanya batasan mengenai bandwidth yang akan diseleksi. Dengan batasan ini, akan diketahui bahwa bandwidth optimal berada dalam suatu interval tertentu.

Bandwidth yang terlalu besar menimbulkan kurva pendekatan terhadap fungsi densitas mengalami 'oversmoothing'. Apabila besar bandwidth tertentu mengakibatkan oversmoothing, maka penghalusan tidak akan mencapai optimal untuk bandwidth yang lebih besar lagi karena justru akan semakin terjadi oversmoothing. Oleh karena itu, keadaan ini bisa dijadikan tolok ukur sebagai batas atas bandwidth yang akan diseleksi.

Terrel (1990) memberikan suatu formula mengenai besar bandwidth yang menimbulkan oversmoothing, yaitu :

$$h_{os} = 3 \left[\frac{\|K\|_2^2}{35\ \mu_2^4(K)} \right]^{\frac{1}{5}} \sigma\ n^{-\frac{1}{5}}$$

Keterangan :

1. n = jumlah observasi
2. σ = standard deviasi, dengan rumus :

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

3. $\mu_2(K)$ dan $\|K\|_2^2$ dihitung menggunakan integrasi numerik dengan aturan trapesium.

   Dalam integrasi numerik dengan aturan trapesium, untuk menghitung $I = \int_a^b f(x)\, dx$, interval dari a sampai b dibagi ke dalam p interval yang sama, masing-masing berukuran $h = \frac{b-a}{p}$. Rumus umum untuk integrasi ini adalah :

$$\int_a^b f(x)\,dx = \frac{h}{2}\,[f(x_0)+2f(x_1)+\ldots+2f(x_{p-1})+f(x_p)]$$

dimana h adalah beda untuk tiap interval.

Sehingga penerapan rumus di atas dalam perhitungan $\mu_2(K)$ dan $\|K\|_2^2$ adalah :

$$\mu_2(K) = \frac{h}{2}\,[u_0^2K(u_0)+2u_1^2K(u_1)+\ldots+2u_{p-1}^2K(u_{p-1})+u_pK(p)$$

$$\|K\|_2^2 = \frac{h}{2}\,[K^2(u_0)+2K^2(u_1)+\ldots+2K^2(u_{p-1})+u_pK^2(u_p)$$

dengan $u_0, u_1, \ldots u_p$ adalah suatu nilai yang didapat dari :

$u_0 = -1$ dan $u_p = 1$, karena untuk jenis kernel selain Gaussian batas interval adalah [-1,1].

$$\{u\}_{i=1}^{p-1} = (-1) + hi$$

Dengan demikian algoritma dari perhitungan rumus-rumus di atas sebagai berikut:

1. Perhitungan standard deviasi

```
Rata←  jumdata/n
hasil← 0
for i←  1 to n do
begin
   selisih ← sqr(data[i]-rata)
   hasil ←  hasil+selisih
end
deviasi ←  sqrt(1/n*hasil)
```

2. Perhitungan $\mu_2(K)$

```
hasil← 0
for i← 1 to p-1 do
begin
   nilai ← (-1)+beda*i
   hasil← hasil+2*sqr(kernel(nilai))
end
hasil← hasil+sqr(kernel(-1))+sqr(kernel(1))
hasil← hasil+beda/2
K2← hasil;
```

3. Perhitungan $\|K\|_2^2$

```
hasil← 0
for i← 1 to p-1 do
begin
   nilai ← (-1)+beda*i
   hasil← hasil+2*sqr(nilai)*(kernel(nilai))
end
hasil← hasil+*sqr(-1)*kernel(-1)+sqr(1)*kernel(1)
hasil← hasil+beda/2
Miu← hasil
```

4. Perhitungan h_{os}

```
hasilnya ← K2/(35*sqr(sqr(Miu))*n)
hos ←3* exp(1/5*ln(hasilnya))* deviasi
```

3.2. Metode Cross-validation

Pemilihan bandwidth merupakan masalah utama dalam estimasi densitas kernel, karena parameter ini sangat mempengaruhi bentuk estimator densitas yang dihasilkan. Oleh karena itu besar banwidth tertentu akan memberikan aspek yang tertentu pula terhadap suatu struktur data.

Cross-validation adalah metode yang digunakan untuk memilih bandwidth optimal dalam penghalusan fungsi densitas. Teknik ini terdiri dari tiga bentuk yaitu :

1. Maximum likelihood cross-validation
2. Least squared cross-validation
3. Biased cross-validation

Metode Cross-validation menggunakan kriteria kesalahan estimasi densitas sebagai acuan untuk mendapatkan parameter penghalus. Dengan acuan tersebut diperoleh suatu fungsi yang digunakan untuk menemukan bandwidth optimal. Dari sekumpulan bandwidth yang akan diseleksi didapatkan nilai fungsi yang berbeda-beda.

Mengenai bandwidth yang akan diseleksi, sebagai keadaan awal kita tentukan bandwidth minimal, bandwidth maksimal dan jumlah bandwidth yang diinginkan untuk diseleksi. Penentuan Bandwidth minimal tergantung dari bentuk metode cross-validation yang digunakan. Bandwidth maksimal diambil dari besar h_{os}. Mengenai jumlah bandwidth yang akan diseleksi tidak ada pedoman khusus dalam penentuannya. Akan tetapi semakin banyak jumlah bandwidth yang ditentukan untuk diseleksi akan semakin mendekati

optimal.

Agar di dalam sederetan bandwidth yang akan diseleksi memiliki penambahan atau beda yang sama maka algoritma berikut sangat diperlukan untuk menentukan besar bandwidth ke-i dari sejumlah bandwidth yang akan diseleksi :

```
hstep ⟵ (hmax-hmin)/(m-1)
for k ⟵ 0 to m-1 do
    h[k]⟵ hmin+hstep * k
```

dengan :

hstep = penambahan h

m = jumlah bandwidth yang akan diseleksi

hmax = bandwidth maximal

hmin = bandwidth minimal

hmax, hmin, dan m terlebih dahulu ditentukan sesuai dengan petunjuk yang telah dijelaskan di atas.

3.2.1. Maximum Likelihood Cross-validation

Sebagai permulaan, untuk h tertentu dilakukan hipotesa terhadap estimator densitas kernel $\hat{f}_h$:

$$\hat{f}_h(x) = f(x) \quad \text{vs} \quad \hat{f}_h(x) \neq f(x)$$

Uji rasio likelihood ini didasarkan pada uji statistik $\frac{f(x)}{\hat{f}(x)}$. Untuk bandwidth yang optimal, statistik ini akan mendekati 1. Dengan kata lain $E_x[\log \left(f/\hat{f}_h\right)$ mendekati 0.

Efek optimasi dari kriteria Kullback-Leibler adalah merupakan kriteria kesalahan suatu estimator. Didefinisikan suatu kriteria Kullback-leibler :

$$d_{KL}(f,\hat{f}_h) = \int \log \left[f/\hat{f}_h \right](x) f(x)dx$$

dengan

$f(x)$ = fungsi densitas yang diestimasi

$\hat{f}_h(x)$ = estimator densitas kernel.

Dengan demikian bandwidth yang optimal akan memberikan nilai minimal terhadap $d_{KL}(f,\hat{f}_h)$.

Metode maksimum likelihood Cross-validation menggunakan kriteria Maximum Likelihood untuk meminimalkan $d_{KL}(f,\hat{f}_h)$. Jika h dipilih untuk memaksimalkan $\prod_i \hat{f}_h(X_i)$ yaitu

$$\prod_{i=1}^{n} \hat{f}_h(X_i) = \prod_{i=1}^{n} \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{X_i - X_j}{h}\right)$$

maka terlihat bahwa jika $h \longrightarrow 0$ akan menimbulkan nilai likelihood infinite untuk i sama dengan j. Oleh karena itu digunakan estimator leave-one-out yang merupakan estimator dengan observasi n-1.

Didefinisikan suatu estimator leave-one out :

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

Dari estimator ini didapatkan :

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

Fungsi Maximum Likelihood Cross-validation adalah :

$$CV_{KL}(h) = n^{-1} \log \prod_{i=1}^{n} \hat{f}_{h,i}(X_i)$$

$$CV_{KL}(h) = n^{-1} \sum_{i=1}^{n} \log \left[\hat{f}_{h,i}(X_i)\right]$$

$$CV_{KL}(h) = n^{-1} \sum_{i=1}^{n} \log\left[\sum_{i \neq j} K\left(\frac{X_i - X_j}{h}\right) \right] - \log[(n-1)h]$$

$CV_{KL}(h)$ adalah fungsi Maximum likelihood cross-validation. Bandwidth optimal didefinisikan dengan

$$\hat{h}_{KL} = \arg \max_{h} CV_{KL}(h)$$

yang berarti nilai h yang memberikan nilai maksimal terhadap $CV_{KL}(h)$.

Algoritma direct

Apabila akan diseleksi sejumlah bandwidth $h_1, h_2, \ldots, h_m$, maka komputasi dari nilai $CV_{KL}(h)$ mengikuti algoritma :

```
for k ← — 1 to m do
    for i ← — 1 to n do
```

```
        score←— 0
        for j←— 1 to n do
            score←— score+K((X[i]-X[j])/h[k]
        endfor j
        cv.kl[k]←— cvkl[k]+log(score-K(0))
    endfor i
    cvkl[k]←— cvkl[k]/n-log((n-1)*h[k])
endfor k
```

Sedangkan algoritma untuk menemukan nilai h optimal adalah sebagai berikut :

```
Max←— cvkl[1]
for k ←— 2 to m do
begin
    if cvkl[k]>Max then
      begin
        Max ←— cvkl[k]
        hopt←— h[k]
      end
end
```

Untuk menghindari timbulnya logaritma yang bernilai nol, maka dalam metode ini, pengambilan bandwidth minimal harus lebih besar dari selisih terbesar dari dua observasi yang telah diurutkan. Ini berlaku untuk jenis kernel selain Gaussian.

Dalam kaitannya dengan peminimalan efek optimasi Kullback-Leibler, pemaksimalan fungsi maksimum likelihood cross-validation di atas identik dengan meminimalkan $d_{KL}(f,\hat{f}_h)$.

Karena X_i berdistribusi identik, maka $\log \hat{f}_{h,i}(X_i)$ juga berdistribusi identik sehingga :

$$E\left[CV_{KL}(h)\right] = E\left[\log \hat{f}_{h,i}(X_i)\right]$$

Dengan mengabaikan efek leave-one-out, persamaan di atas bisa dinyatakan dengan :

$$E\left[CV_{KL}(h)\right] \approx E\left[\log \hat{f}_h(X_i)\right]$$

$$\approx \int f(x) \log \hat{f}_h(x)\, dx$$

$$\approx \int f(x) \log \frac{f(x)}{\frac{f(x)}{\hat{f}_h(x)}}\, dx$$

$$\approx \int f(x) \log f(x)\, dx$$

$$- \int \log \left[f/\hat{f}_h \right](x) f(x) dx$$

$$E\left[CV_{KL}(h)\right] \approx \int f(x) \log f(x)\, dx - d_{KL}(f,\hat{f}_h)$$

Karena $\int f(x) \log f(x)\, dx$ tidak tergantung dari h maka dapat disimpulkan bahwa pemaksimalan $CV_{KL}(h)$ akan meminimalkan $d_{KL}(f, \hat{f}_h)$.

3.2.1. Least-Squares Cross-validation

Metode Least-Squares Cross-validation didasarkan atas peminimalan kriteria kesalahan estimasi densitas yaitu Integrated Squares Error (ISE). ISE adalah integrasi dari selisih kuadrat antara fungsi densitas yang akan diestimasi yaitu f(x) dengan estimator densitas kernel yaitu $\hat{f}_h(x)$.

ISE didefinisikan dengan :

$$ISE = \int (\hat{f}_h - f)^2 dx$$

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

Dari formula di atas terlihat bahwa :

- $\int \hat{f}_h^2(x)$ dapat dihitung dari data
- $\int f^2(x)dx$ tidak dipengaruhi oleh besar h, karena f(x) adalah fungsi densitas yang diestimasi.

- $\int (\hat{f}_h f)(x)dx$ harus diestimasi dari data.

Karena $\int f^2(x)\, dx$ tidak dipengaruhi oleh bandwidth h maka dapat dibawa ke ruas kiri sehingga persamaan di atas

menjadi :

$$ISE - \int f^2(x)dx = \int \hat{f}_h(x)dx - \int (\hat{f}_h f)(x)dx$$

Menurut Rudemo (1982) $\int(\hat{f}_h f)(x)dx = E_x[\hat{f}_h(x)]$, dimana ekspektasi ini dihitung menggunakan estimator leave-one-out sebagai berikut :

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

Dengan estimasi tersebut, dapat ditentukan bandwidth optimal melalui peminimalan $ISE - \int f^2(x)\,dx$ atau fungsi Least-Squares Cross-validation sebagai berikut :

$$CV(h) = \int \hat{f}_h^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} f_{h,i}(X_i)$$

dimana $\int \hat{f}_h^2(x)dx$ diperoleh sebagai berikut :

$$\int \hat{f}_h^2(x)dx = (nh)^{-2}\int\left[\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\right]^2 dx$$

$$= (nh)^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K\left(\frac{x-X_i}{h}\right) K\left(\frac{x-X_j}{h}\right) dx$$

$$= n^{-2}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K(s)K\left(\frac{X_i-X_j}{h} + s\right)ds$$

$$= n^{-2}h^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K(s)K\left(\frac{X_j-X_i}{h} - s\right)ds$$

$$= n^{-2}h^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} K * K\left(\frac{X_j - X_i}{h}\right)$$

$$CV(h) = n^{-2}h^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} K * K\left(\frac{X_j - X_i}{h}\right) - \frac{2}{n} \sum_{i=1}^{n} \hat{f}_{h,i}(X_i)$$

$$CV(h) = \frac{2}{n^2 h}\left[\frac{n}{2} K*K(0) + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} K*K\left(\frac{X_j - X_i}{h}\right) - \frac{2n}{n-1} K\left(\frac{X_j - X_i}{h}\right)\right.$$

Dengan demikian diperoleh fungsi CV(h) yang merupakan fungsi Least-Squared Cross-validation.

Bandwidth optimal diperoleh dengan memilih nilai h yang memberikan nilai minimal terhadap CV(h). Secara matematis dinyatakan dengan :

$$\hat{h}_{CV} = \arg\min_{h} CV(h)$$

Perhitungan $K*K\left(\frac{X_j - X_i}{h}\right)$ menggunakan teori konvolusi sebagai berikut :

$$K * K(x) = \int K(x-u)\, K(u)\, du$$

Sebagai contoh, operasi konvolusi untuk kernel Gaussian sebagai berikut :

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right)$$

$$K(x-u) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-u)^2\right]$$

$$K*K(x) = \int K(x-u)\, K(u)\, du$$

$$= \int \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-u)^2 \right] \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}u^2 \right] du$$

$$= \frac{1}{2\pi} \int \exp\left[-\frac{1}{2}\,(x^2 - 2xu + u^2 + u^2) \right] du$$

$$= \frac{1}{2\pi} \exp\left[-\frac{1}{2}\,x^2 \right] \int \exp\left[-\frac{1}{2}\,(2u^2 - 2xu\,) \right] du$$

$$= \frac{1}{2\pi} \exp\left[-\frac{1}{2}\,x^2 \right] \int \exp\left[-(u^2 - xu) \right] du$$

$$= \frac{1}{2\pi} \exp\left[-\frac{1}{2}\,x^2 \right] \int \exp\left[-((u - \frac{1}{2}x)^2 - \frac{1}{4}x^2) \right] du$$

$$= \frac{1}{2\pi} \exp\left[-\frac{1}{4}\,x^2 \right] \int \exp\left[-((u - \frac{1}{2}x)^2 \right] du$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2}} \exp\left[-\frac{1}{4}x^2 \right] \int \frac{1}{\sqrt{2\pi}} \sqrt{2} \exp\left[-\frac{1}{2} \left(\frac{u - \frac{1}{2}x^2}{\frac{1}{\sqrt{2}}} \right) \right] du$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2}} \exp\left[-\frac{1}{4}x^2 \right] .\ 1$$

Jadi untuk jenis kernel Gaussian :

$$K * K(x) = \frac{1}{2\sqrt{\pi}} \exp\left[-\frac{1}{4}x^2 \right]$$

Algoritma direct

Algoritma untuk perhitungan nilai CV(h) dari sekumpulan bandwidth $h_1, h_2, \ldots, h_m$ yang akandiseleksi adalah sebagai berikut :

```
for k←— 1 to m do
  for i←— 1 to n-1 do
   for j←— i+1 to n do
    cv[k]←— cv[k]+kernel.conv((X[j]-X[i])/h[k])
    cv[k]←— cv[k]-kernel((X[j]-X[i])/h[k])*2*n/(n-1)
   endfor j
  endfor i
  cv[k]←— cv[k]+kernel.conv(0)*n/2
  cv[k]←— cv[k]*2/(n^2*h[k])
endfor k
```

Algoritma pemilihan bandwidth optimal $\hat{h}_{cv}$ adalah :

```
Min←— cv[1]
     for k←— 2 to m do
     begin
          if cv[k]<Min then
          begin
               Min←— cv[k]
               hopt←— h[k]
          end
        end
```

3.2.3. Biased Cross-validation

Metode ini dikemukakan oleh Scoot dan Terrel (1987) yang didasarkan pada estimasi A-MISE[$\hat{f}_h$].

Dalam estimasi densitas kernel, $MSE[\hat{f}_h(x)]$ dinyatakan dengan :

$$MSE\hat{f}_h(x)] = \frac{1}{nh} f(x)\|K\|_2^2 + \frac{h^4}{4} (f''(x)\mu_2(K))^2$$

$$+o((nh)^{-1}) + o(h^4)$$

dengan

$$\mu_2(K) = \int u^2 K(u)\, du$$

$MSE[\hat{f}_h(x)]$ tidak bisa dihitung mengingat f(x) dan f"(x) tidak diketahui. Ada suatu kriteria yang lain yaitu MISE(Mean Integrate Squares Error).

$$MISE\ [\hat{f}_h] = E\ [ISE\ ]$$

$$= E\left[\int \left[\hat{f}_h(x) - f(x) \right]^2 dx \right]$$

$$= \int E\ [\hat{f}(x) - f(x)]^2\ dx$$

$$= \int MSE\left[\hat{f}_h(x) \right] dx$$

$$= (nh)^{-1}\|K\|_2^2 \int f(x)dx + \frac{h^4}{4} (\mu_2(K))^2 \int (f''(x))^2 dx$$

$$+ o(n^{-1}h^{-1}) + o(h^4)$$

Karena $\int f(x)dx = 1$ maka :

$$MISE[\hat{f}_h] = \frac{1}{nh} \|K\|_2^2 + \frac{h^4}{4} (\mu_2(K))^2 \|f''\|_2^2$$

$$+ o(n^{-1}h^{-1}) + o(h^4) \ , \ h \to 0, \ nh \to \infty$$

Sehingga diperoleh

$$A\text{-}MISE[\hat{f}_h] = \frac{1}{nh} \|K\|_2^2 + \frac{h^4}{4} (\mu_2(K))^2 \|f''\|_2^2$$

dengan $\|f''\|_2^2 = \int (f''(x))^2 \, dx$.

Karena f tidak diketahui maka $\|f''\|_2^2$ juga tidak dapat dihitung. Untuk mengatasi hal ini $\|f''\|_2^2$ harus diestimasi. Penggunaan estimator untuk $\|f''\|_2^2$ ke dalam $A\text{-}MISE[\hat{f}_h]$ menghasilkan suatu fungsi $BCV_1(h)$ sebagai berikut:

$$BCV_1(h) = \frac{1}{nh} \|K\|_2^2 + \frac{h^4}{4} (\mu_2(K))^2 \|\hat{f}''\|_2^2$$

Untuk mengetahui apakah $BCV_1(h)$ ini bisa digunakan dalam pendekatan $MISE[\hat{f}_h]$ maka perlu diselidiki terlebih dahulu dahulu $Var[\hat{f}''(x)]$ yaitu :

$$Var[\hat{f}_h''(x)] = Var\left[n^{-1} \sum_{i=1}^{n} K_h'' [x - X_i]\right]$$

$$= n^{-2} \quad \mathrm{Var}\left[\sum_{i=1}^{n} K_h'' [x - X_i] \right]$$

$$= n^{-2} \sum_{i=1}^{n} \mathrm{Var} \left[K_h'' [x - X_i] \right]$$

$$= n^{-1} \mathrm{Var} \left[K_h'' [x - X] \right]$$

$$= n^{-1} \left\{ E [K_h''^2 (x - X)] - (E [K_h''(x - X)])^2 \right\}$$

$$= n^{-2} h^{-6} \left[\int K''^2 \left(\frac{x - u}{h} \right) f''(u)\, du - (f''(x) + o(h))^2 \right]$$

Dengan subtitusi $\frac{x - u}{h} = s$ diperoleh :

$$\mathrm{Var}[\hat{f}_h''(x)] = n^{-1} \left[h^{-5} \int K''^2(s) f''(x+sh)\, ds - (f''(x) + o(h))^2 \right]$$

$$= n^{-1} \left[h^{-5} \| K'' \|_2^2 (f''(x) + o(h)) - (f''(x) + o(h))^2 \right]$$

$$\sim n^{-1} h^{-5} \| K'' \|_2^2$$

Karena $\mathrm{Var}[\hat{f}_h''(x)]$ tidak konvergen ke 0 maka untuk $\hat{f}_h''(x)$ tidak bisa dipakai untuk pendekatan terhadap $\| f'' \|_2^2$ dalam peminimalan A-MISE $[\hat{f}_h]$.

Scoot dan Terrel memberikan suatu formula untuk ekspektasi $\|\hat{f}_h''\|_2^2$ dengan kernel K dan fungsi densitas adalah kontinu differensiabel kedua sebagai berikut :

$$E[\|\hat{f}_h''\|_2^2] = \|f''\|_2^2 + \frac{1}{nh^5}\|K\|_2^2 + O(h^2)$$

Dengan demikian estimator unbias asymptotic untuk $\|f''\|_2^2$ adalah :

$$\|\hat{f}''\|_2^2 = \|\hat{f}_h''\|_2^2 - \frac{1}{nh^5}\|K\|_2^2$$

Subtitusi estimator di atas ke dalam $BCV_1(h)$ menghasilkan suatu formula untuk Biased cross-validation :

$$BCV(h) = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}(\mu_2(K))^2\left[\|\hat{f}_h''\|_2^2 - \frac{1}{nh^5}\|K\|_2^2\right]$$

Bandwidth optimal dengan metode ini ditentukan dengan memilih h yang memberikan nilai minimal terhadap BCV(h). Secara matematis dinyatakan dengan :

$$\hat{h}_{BCV} = \arg\min_h BCV(h)$$

Untuk perhitungan nilai BCV(h), $\|\hat{f}_h''\|_2^2$ dijabarkan sebagai berikut :

$$\|\hat{f}_h''\|_2^2 = \int (\hat{f}_h''(x))^2 dx$$

$$= (n^{-2}h^{-6})\int \left[\sum_{i=1}^{n} K''\left(\frac{x-X_i}{h}\right)\right]^2 dx$$

$$= (n^{-2}h^{-6})\sum_{i=1}^{n}\sum_{j=1}^{n} \int K''\left(\frac{x-X_i}{h}\right) K''\left(\frac{x-X_j}{h}\right) dx$$

$$= n^{-2}h^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n} \int K''(s)K''\left(\frac{X_i-X_j}{h} + s\right)ds$$

$$= n^{-2}h^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n} \int K''(s)K''\left(\frac{X_j-X_i}{h} - s\right)ds$$

$$= n^{-2}h^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n} K''* K''\left(\frac{X_j-X_i}{h}\right)$$

$$BCV(h) = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}(\mu_2(K))^2\left[\|\hat{f}_h''\|_2^2 - \frac{1}{nh^5}\|K\|_2^2\right]$$

$$= \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}(\mu_2(K))^2\left[n^{-2}h^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n} K''* K''\left(\frac{X_j-X_i}{h}\right)\right.$$

$$\left. - \frac{1}{nh^5}\|K\|_2^2 + \right]$$

$$= \frac{1}{nh}\|K\|_2^2 + \frac{1}{4}n^{-1}h^{-1}(\mu_2(K))^2\left[\frac{2}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} K''* K''\left(\frac{X_j-X_i}{h}\right)\right.$$

$$\left. + K''*K''(0) - K''*K''(0)\right]$$

$$= \frac{1}{nh}\|K\|_2^2 + \frac{1}{2}n^{-2}h^{-1}\mu_2^2(K)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} K''* K''\left(\frac{X_j-X_i}{h}\right)$$

$$BCV(h) = \frac{1}{nh}\|K\|_2^2 + \frac{1}{2}n^{-2}h^{-1}\mu_2^2(K)\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} K''* K''\left(\frac{X_j-X_i}{h}\right)$$

Algoritma direct

Algoritma untuk perhitungan BCV(h) adalah sebagai berikut :

```
for k ←— 1 to m do
  score ←— 0
     suku1 ←— norm(K)/(n*h[k])
     suku2 ←— moment(K^2)/(2*n^2*h[k]
     for i←— 1 to n-1
       forj←— i+1 to n
       distance←— (X[j]-X[i])/h[k]
         score←—  score+kernel2der.conv(distance)
       endfor j
     endfor i
     BCV[k]←— suku1+suku2*score
endfor k
```

Algoritma untuk pemilihan $\hat{h}_{BCV}$ adalah :

```
Min←— BCV[1]
hopt ←— h[1]
     for k←— 1 to m do
          if BCV[k]<Min then
               Min ←— BCV[k]
               hopt←— h[k]
          endif
     endfor k
```

Metode ini hanya bisa diterapkan untuk jenis kernel yang kontinu differensiabel yaitu kernel Quartic, Triweight dan Gaussian.

3.3. Implementasi Metode Cross-validation dengan Turbo Pascal

Dengan metode Cross-validation menggunakan algoritma direct yang diterapkan pada suatu data maka dapat diketahui parameter-parameter yang mempengaruhi nilai bandwidth h. Algoritma tersebut diimplementasikan ke dalam Turbo Pascal sehingga akan diperoleh nilai bandwidth optimal dari ketiga bentuk metode Cross-validation yaitu Maximum Likelihood Cross-validation, Least-Squares Cross-validation dab Biased Cross-validation. Untuk keperluan tersebut digunakan data input yaitu data Buffalo Snowfall (lihat lampiran I).

Sebagai input lain yaitu :

a. Bandwidth minimal yang akan diseleksi (hmin)
b. Banyaknya bandwidth yang akan diseleksi (m)
c. Jenis kernel yang diinginkan

Penentuan bandwidth minimal untuk metode Maximum Likelihood Cross-validation berbeda dengan penentuan bandwidth minimal untuk metode Least-Squares Cross-validation dan Biased Cross-validation. Dalam metode Maximum Likelihood Cross-validation, untuk jenis kernel selain Gaussian, bandwidth minimal harus lebih

besar atau sama dengan selisih maksimal dari dua observasi yang telah diurutkan. hal ini dimaksudkan agar nilai score tidak sama dengan nol, sehingga nilai ln tidak infinite. Sedangkan dalam metode Least Squares Cross-validation dan Biased Cross-validation bisa diambil bilangan positif yang kecil.

Untuk penerapan metode Maximum Likelihood Coss-validation dan Least-Squares Cross-validation pada data Buffalo Snowfall digunakan jenis kernel Epanechnicov (3) dan jenis kernel Gaussian. Jenis kernel Epanechnicov dipilih karena menghasilkan nilai $\mu(K)\|K\|_2^2$ paling kecil dibanding jenis-jenis kernel yang lain sehingga dapat meminimalkan nilai A-MISE dari estimator kernel densitas. Jenis kernel Gaussian atau disebut juga kernel Normal lebih baik untuk pendekatan pada data Buffalo Snowfall mengingat distribusi frekuensi dari data ini adalah mendekati suatu bentuk normal (lihat lampiran III).

Bandwidth optimal untuk jenis kernel Epanechnicov tidak bisa dicari dengan metode Biased Cross-validation karena jenis kernel ini tidak kontinu differensiabel. Oleh karena itu untuk penerapan metode Biased Cross-validation hanya diambil jenis kernel Gaussian saja.

Output dari hasil penggunaan ketiga bentuk metode Cross-validation di atas dapat dilihat pada lampiran II.

3.4. Analisa hasil

Dari hasil output sebagaimana diperlihatkan pada lampiran II dihasilkan tabel sebagai berikut :

1. Metode Maximum Likelihood Cross-validation

JENIS KERNEL	hmin	hos	m	$CV_{KL}(h)$	$\hat{h}_{KL}$
Epanechnicov (3)	14.80	25,995	15	-2.00100	18,799
	14,80	25,995	40	-2,00094	18,532
Gaussian (6)	0,01	11,743	15	-2,00385	9,229
	0,01	11,743	40	-2,00383	9,637

Tabel 3.1

Pada tabel 3.1 terlihat bahwa pengambilan hmin untuk jenis kernel Epanechnicov adalah 14,80 yang merupakan selisih terbesar dari dua observasi dalan data Buffalo Snowfall. Dengan jumlah m yang berbeda diperoleh $\hat{h}_{KL}$ yang berbeda pula, baik pada jenis kerenl Epanechnicov maupun Gaussian. Apabila ditinjau dari nilai $CV_{KL}(h)$, maka untuk jenis kernel Epanechnicov $\hat{h}_{KL}$= 18,532 memberikan nilai $CV_{KL}(h)$ lebih besar dibanding $\hat{h}_{KL}$=18,799 sehingga $\hat{h}_{KL}$=18,532 lebih optimal. Demikian pula untuk jenis kernel Gaussian, $\hat{h}_{KL}$=9,637 lebih optimal dibanding $\hat{h}_{KL}$=9,229.

2. Metode Least-Squares Cross-validation

JENIS KERNEL	hmin	hos	m	CV(h)	$\hat{h}_{cv}$
Epanechnicov (3)	0,01	25,995	15	-0.01105	16,715
	0,01	25,995	40	-0,01105	16,668
Gaussian (6)	0,01	11,743	15	-0,01099	9,229
	0,01	11,743	40	-0,01099	9,336

Tabel 3.2

Pada tabel 3.2 terlihat bahwa dengan pengambilan m=15 dan m=40 menghasilkan nilai CV(h) yang sama untuk masing-masing jenis kernel. Akan tetapi $\hat{h}_{cv}$ yang dihasilkan berbeda dengan karena perbedaan m menimbulkan penambahan h yang berbeda pula. Meskipun demikian dapat diambil kesimpulan bahwa besar bandwidth optimal adalah 16,668 mengingat lebih banyaknya bandwidth yang diseleksi.

3. Biased Cross-validation

JENIS KERNEL	hmin	hos	m	CV(h)	$\hat{h}_{cv}$
Gaussian (6)	0,01	11,743	15	-0.00111	11,743
	0,01	11,743	40	-0,00111	11,743

Tabel 3.3

Pada tabel 3.3 terlihat bahwa bandwidth optimal ($\hat{h}_{BCV}$) yang diperoleh sama dengan h_{OS} yaitu 11,743. Hal ini disebabkan karena dari output pada lampiran II menunjukkan bahwa nilai BCV(h) semakin mengecil dengan

penambahan h, sehingga jika h terus bertambah maka secara matematis dinyatakan dengan :

$$\lim_{h \to \infty} BCV(h) = 0$$

Oleh karena itu bandwidth optimal yang diperoleh merupakan batas minimal oversmoothed bandwidth (h_{os}).

Pengambilan jumlah bandwidth yang diseleksi (m) juga mempengaruhi besar bandwidth optimal yang dihasilkan meskipun tidak menimbulkan perbedaan yang sangat jauh. Perbedaan pengambilan m menimbulkan perbedaan output penambahan nilai h yang diseleksi meskipun tidak menutup kemungkinan dihasilkan bandwidth optimal yang berniali sama. Jika jumlah bandwidth yang diseleksi (m) bisa mencakup seluruh kemungkinan mengenai besar bandwidth dalam interval antara hmin dan h_{os} maka dapat diketahui bandwidth yang benar-banar optimal. Untuk mencapai hal ini perlu m yang besar apabila h_{os} besar.

Perbedaan metode yang digunakan juga bisa menghasilkan perbedaan bandwidth optimal. Sebagai contoh, dengan metode Maximum Likelihood Cross-validation diperoleh bandwidth optimal 18,532 dan dengan metode Least-Squares Cross-validation diperoleh bandwidth optimal 16,668 untuk jenis kernel Epanechnicov. Perbedaan ini disebabkan karena masing-masing metode mempunyai karakteristik yang berbeda-beda. Metode Maximum

Likelihood didasarkan atas pemilihan $d_{KL}(f,\hat{f}_h)$ atau efek optimasi Kullback-Leibler. Metode Least-Squares Cross-validation didasarkan atas peminimalan ISE dan metode Biased Cross-validation didasarkan atas peminimalan A-MISE. Akan tetapi tidak menutup kemungkinan dihasilkan nilai bandwidth yang sama dengan metode yang berbeda. Hal ini dapat dilihat pada tabel 3.1 dan tabel 3.2 di mana $\hat{h}_{KL}$ dan $\hat{h}_{CV}$ adalah sama dengan pengambilan m = 15.