UIN IMAM BONJOL PADANG

MAp Journal
Mathematics & Applications

Web : ejournal.uinib.ac.id/jurnal/index.php/MAp
Email : map_journal@uinib.ac.id

METODE *SCREENING* KOLMOGOROV-SMIRNOV UNTUK DATA SURVIVAL BERDIMENSI TINGGI

KOLMOGOROV-SMIRNOV SCREENING METHOD FOR HIGH-DIMENSIONAL SURVIVAL DATA

Syarto Musthofa[1§], Danardono[2]

[1]Universitas Islam Negeri Imam Bonjol Padang [Email: syartom@uinib.ac.id]
[2]Universitas Gadjah Mada Yogyakarta [Email: danardono@ugm.ac.id]
[§]*Corresponding Author*

Received Mei 2021; Accepted Juni 2021; Published Juni 2021;

Abstrak

Ada banyak metode *screening* variabel yang bisa menangani data berdimensi tinggi. Beberapa dari metode tersebut bisa mengurangi dimensi data secara efektif dan menjamin semua variabel aktif tetap muncul dengan probabilitas tinggi. Namun, kebanyakan prosedur *screening* yang ada saat ini dikembangkan hanya untuk data lengkap berdimensi tinggi dan tidak layak diterapkan pada data survival dengan informasi tersensor. Metode *Screening* Kolmogorov-Smirnov dapat dimodifikasi untuk mengatasi masalah ini dengan mengganti fungsi distribusi kumulatif dengan fungsi survival yang diestimasi dengan estimator Kaplan-Meier. Metode ini dapat bekerja dengan berbagai tipe kovariat baik itu kontinu, diskrit, maupun kategorikal. Performa dari metode ini diukur berdasarkan studi simulasi. Suatu contoh data riil mengenai ekspresi gen juga digunakan sebagai aplikasi dari metode ini.

Kata Kunci: Metode *screening*, data berdimensi tinggi, data survival.

Abstract

There are numerous variable screening methods available for high-dimensional data. Some of the methods can effectively reduce the dimensionality while ensuring that all the active variables can be retained with high probability. However, most existing screening procedures are developed for high-dimensional complete data and cannot be applicable to censored survival data. The Kolmogorov-Smirnov Screening Method could be modified to overcome this problem by replacing the cummulative distribution function with survival function which estimated by Kaplan-Meier estimator. This method can work witmany types of covariates including continuous, discrete, and categorical variables. The performance of this method presented via simulation study. A real data example of gene expression is used to illustrate the application of the method.

Keywords*: Screening method, high-dimensional data, survival data.*

1. Pendahuluan

Berbicara mengenai data berdimensi tinggi berarti berbicara mengenai data yang memuat informasi kovariat yang sangat banyak yang jauh melampaui banyaknya observasi yang dimiliki. Dalam analisis data tentunya tidak mungkin melibatkan semua kovariat yang ada karena akan menyebabkan kerumitan yang luar biasa. Disamping itu pada prinsipnya sebagian besar kovariat pada data berdimensi tinggi dipandang tidak berpengaruh pada variabel independen. Keadaan tersebut menuntun para peneliti dalam mengembangkan metode-metode yang bertujuan untuk mereduksi dimensi data.

Penelitian mengenai penyaringan fitur telah banyak dilakukkan dengan tujuan untuk mengurangi kompleksitas kovariaat yang begitu tinggi. Ada berbagai penelitian yang telah dilakukan yang mengembangkan metode pemilihan variabel seperti LASSO [1] , *smoothly clipped absolute deviation* [2], *the adaptive* LASSO [3], *Danzig selector* [4], dan minimax concave penalty [5]. Namun, menurut Liu, et.al. [6] metode-metode yang telah disebutkan tersebut tidak bekerja dengan baik ketika diterapkan pada data dengan kovariat yang sangat banyak.

Untuk mengurangi kompleksitas kovariat yang ada secara efektif, beberapa metode yang berbasis model dan yang bebas model (bebas distribusi) juga telah dikembangkan. Sebagai contoh [7] dan [8] menawarkan Metode *Sure Independence Screening* (SIS) yang berbasis model regresi linear. Kemudian Mai dan Zou [9] [10] mengembangkan metode penyaringan bebas model dengan *fused* Kolmogorov flter.

Menurut Liu, et.al [6], apabila berbicara data berdimensi sangat tinggi maka kompleksitas kovariat yang sangat besar akan menjadikan metode yang dibangun atas dasar asumsi distribusi lebih sulit diterapkan. Mereka menyatakan prosedur penyaringan yang bebas distribusi lebih robust karena syarat yang tetap bisa dipenuhi meski dengan kondisi yang lebih lemah (artinya asumsi distribusi tidak terpenuhi atau sulit diterapkan). Selain itu pemilihan model pendekatan statistik setelah penyaringan selesai akan lebih feksibel dengan menggunakan metode bebas distribusi.

Diantara beberapa model nonparametrik yang telah disebutkan diatas, Liu, et.al [6] memandang bahwa model yang ditawarkan oleh Mai dan Zou [9] [10] lebih superior dibanding yang lain. Sehingga metode ini dinilai efsien untuk diterapkan pada data berdimensi sangat tinggi. Namun, metode tersebut hanya baik diterapkan untuk data dengan pengamatan lengkap dan tidak bekerja dengan baik apabila diterapkan pada data yang mengandung informasi tersensor. Untuk itu dilakukan pengembangan yang mengacu pada metode [10] sehingga bisa menangani data tersensor pada analisis survival.

2. Prosedur Penyaringan Variabel

Metode penyaringan variabel ini menjadi sangat penting ketika kita dihadapkan dengan

permasalahan statistika yang mempertimbangkan eksistensi variabel respon Y dan kovariat $Z = (Z_1, \cdots, Z_p)^T \in \mathbb{R}^p$dengan p yang sangat besar. Asumsi yang populer dalam hal ini adalah asumsi kekosongan (sparsity asumption) yang berarti hanya ada sebuah subset kecil dari kovariat tersebut yang benar-benar berpengaruh terhadap Y. Oleh karena itu asumsi kekosongan menyatakan bahwa $|\mathcal{A}| \ll p$. Tujuan dari penyaringan variabel adalah untuk menemukan sebuah subset $\mathcal{B}$ sehingga $\mathcal{A} \subset \mathcal{B}$ [10].

2.1. Penyaringan Berbasis Statistik Kolmogorov-Smirnov

Untuk mengetahui mengapa statistik KS ini sangat bermanfaat dalam penyaringan variabel dapat dilihat konsep Kolmogorov-flter biner sebagaimana yang dinyatakan oleh Mai dan Zou [9]. Apabila variabel responnya biner, katakan $Y = 1,2$, maka variabel Zdikatakan independen terhadap Y jika dan hanya jika fungsi distribusi bersyarat Zdengan diberikan $Y = 1$ atau $Y = 2$ dalah fungsi yang identik. Berdasarkan hal ini mereka menawarkan untuk menggunakan:

$$K_j = \sup_z \left| F_j(z|Y = 1) - F_j(z|Y = 2) \right| \qquad (1)$$

untuk mengukur independensi antara X_j dan Y, dengan F_j merupakan CDF variabel X_j. Apabila diberikan data hasil pengamatan, maka versi empiris dari K_j didefinisikan sebagai:

$$\widehat{K}_j = \sup_z \left| \hat{F}_j(z|Y = 1) - \hat{F}_j(z|Y = 2) \right|$$

Dari metode Kolmogorov-filter biner tersebut Mai dan Zou [10] mengembangkan lingkup pembahasan dari yang semula hanya menangani variabel respon biner menjadi bisa menangani variabel respon kontinu atau diskrit. Untuk itu mereka memodifikasi (1) menjadi:

$$K_j^* = \max_{y_1, y_2} \sup_z \left| F_j(z|Y = 1) - F_j(z|Y = 2) \right| \qquad (2)$$

K_j^* adalah generalisasi dari K_j, oleh karena itu $K_j^* = 0$ jika dan hanya jika Z_j independen dari Y. Sebagaimana K_j, K_j^* juga memiliki versi empiris yang dinyatakan sebagai:

$$\widehat{K}_j^* = \max_{y_1, y_2} \sup_z \left| \hat{F}_j(z|Y = 1) - \hat{F}_j(z|Y = 2) \right|$$

Langkah untuk menentukan nilai K_j^* trivial apabila diterapkan pada kasus respon biner. Namun akan menjadi lebih sulit ketika Y merupakan variabel random dengan kemungkinan nilai yang tak berhingga, karena membutuhkan pengetahuan tentang $F_j(z|Y = y)$ untuk semua kemungkinan nilai y. Oleh karena itu perlu ditempuh suatu cara untuk menemukan aproksimasi dari K_j^* dengan melakukan pengirisan (*slicing*) pada variabel respon dengan mendefinisikan partisi:

$$\boldsymbol{G} = \left\{ \begin{array}{c} [a_l, a_{l+1}) : l = 0, \dots, G-1 \\ \text{dan } \cup_{j=1}^{G-1} [a_l, a_{l+1}) \backslash \{a_0\} = \mathbb{R} \end{array} \right\} \qquad (3)$$

dengan $a_0 = -\infty$ dan $a_G = \infty$. Masing-masing $[a_l, a_{l+1})$ disebut *slice*. Kemudian didefinisikan variabel random $H \in \{1, \dots, G\}$ sedemikian sehingga $H = l + 1$ jika dan hanya jika y berada di *slice* ke-l. Secara khusus, apabila Y diskrit sebagaimana dalam permasalahan *multiclass*, yakni $Y = 1, \dots, G$, maka dapat ditentukan $H = Y$. Untuk itu alternatif dari (2) menjadi:

$$K_j^{\boldsymbol{G}} = \max_{l,m} \sup_z \left| F_j(z|H = l) - F_j(z|H = m) \right| \qquad (4)$$

Jelas bahwa Z_j dikatakan independen terhadap Y jika dan hanya jika $K_j^G = 0$ ketika Y mengambil nilai yang finite.

3. Metode Penyaringan KS untuk Data Survival Berdimensi Tinggi

Misalkan $\{X_i, \Delta_i, Z_i \equiv (Z_{i1}, \cdots, Z_{ip})^T : i = 1, \dots, n\}$ adalah observasi-observasi independen data survival berdimensi tinggi $\{X, \Delta, \boldsymbol{Z} = (Z_1, \cdots, Z_p)^T\}$, dengan $\boldsymbol{Z}$ adalah kovariat-kovariat vektor berdimensi-p, dan $X = \min(T \leq C)$. Diasumsikan mekanisme penyensorannya adalah random, yang berarti waktu survival T dan waktu tersensor C independen terhadap $\boldsymbol{Z}$. Kemudian dietapkan τ sebagai akhir masa studi.

Diperhatikan fungsi survival bersyarat $S(t|\boldsymbol{Z}) = P(T > t|\boldsymbol{Z})$ dengan diberikan $\mathbf{Z}$. Dalam permasalahan dimensi tinggi, dimensionalitas p jauh melampaui ukuran sampel n. Sebagaimana yang telah disebutkan sebelumnya asumsi kekosongan menekankan bahwa hanya ada sebuah subset kecil kovariat yang benar-benar berkontribusi pada fungsi survival kondisional $S(t|\boldsymbol{Z})$. Untuk mengidentifikasi yang aktif berkontribusi dari p kovariat terhadap fungsi survival, didefinisikan dahulu set aktif kovariatnya sebagai:

$\mathcal{A} = \{j : S(t|\boldsymbol{Z})$ bergantung pada $Z_j, j = 1, \dots, p\}$

Tujuan penyaringan ini adalah untuk mengungkap set aktif A setepat mungkin meski dengan adanya observasi tersensor.

Sebagaimana yang dilakukan oleh Mai dan Zou [10], untuk mengakomodasi kovariat-kovariat yang bertipe kontinu atau diskrit, Liu, et.al [6] menggunakan ide pengirisan (*slicing*) dan mengkonstruksi statistik Kolmogorov-Smirnov untuk mengukur dependensi antara masing-masing kovariat dengan waktu survival. Untuk setiap kovariat Z_j, didefinisikan partisi:

$$\boldsymbol{\Lambda_j} = \begin{Bmatrix} [a_l^j, a_{l+1}^j) : l = 0, \dots, \Lambda_j - 1 \\ \text{dan } \bigcup_{l=0}^{\Lambda_j - 1} [a_l^j, a_{l+1}^j) \backslash \{a_0^j\} = \mathbb{R} \end{Bmatrix} \quad (5)$$

Dengan $a_0^j = -\infty$, $a_{\Lambda_j}^j = \infty$, dan $j = 1, \dots, p$. Masing-masing $[a_l^j, a_{l+1}^j)$ disebut *slice*. Kemudian didefinisikan sebuah variabel random $I_j \in \{1, \dots, \Lambda_j\}$ sedemikian sehingga $I_j = l + 1$ jika dan hanya jika $Z_j \in [a_l^j, a_{l+1}^j)$. Statistik Kolmogorov-Smirnov untuk kasus survival ini adalah:

$$K_j^{\boldsymbol{\Lambda_j}} = \max_{l_1, l_2} \sup_{0 \leq t \leq \tau} |S_j(t|I_j = l_1) - S_j(t|I_j = l_2)| \quad (6)$$

Dengan $S_j(t|I_j = l_1) = P(T > t|I_j = l_1)$ dan $S_j(t|I_j = l_2) = P(T > t|I_j = l_2)$.

Jelas bahwa ketika $Z_j (j = 1, \dots, p)$ merupakan nilai-nilai yang finit sehingga masing-masing nilai membentuk *slice*, T dikatakan independen terhadap Z_j jika dan hanya jika $K_j^{\boldsymbol{\Lambda_j}} = 0$. Jika Z_j kontinu atau diskrit umum, maka $K_j^{\boldsymbol{\Lambda_j}}$ tetap dapat ditentukan untuk meninjau kebergantungan T terhadap Z_j dengan ketentuan:

(i) T independen terhadap Z_j jika dan hanya jika $K_j^{\boldsymbol{\Lambda_j}} = 0$ untuk setiap kemungkinan dalam partisi $\boldsymbol{\Lambda}_j$

(ii) T tidak independen terhadap Z_j jika dan hanya jika ada bagian dari partisi $\boldsymbol{\Lambda}_j$ sedemikian sehingga $K_j^{\boldsymbol{\Lambda_j}} \neq 0$.

Bila diperhatikan perbandingan metode yang ditawarkan [10] dengan yang dilakukan [6] dapat dilihat adanya persamaan konsep dasar namun perbedaan dalam teknis pengerjaan. Untuk persamaannya dapat dilihat bahwa yang

dilakukan [6] sebenarnya adalah pengembangan metode [10] sehingga konsep *slicing* variabel tetap digunakan oleh mereka. Disamping itu keduanya juga mendefinisikan set aktif dengan cara yang sama meski menggunakan fungsi berbeda. Perbedaannya dapat dilihat secara teknis ketika mendefinisikan partisi. Mai dan Zou [10] mendefinisikan partisi untuk variabel respon, dengan kata lain hanya ada satu partisi dalam proses penyaringan yang mereka lakukan sebagaimana persamaan (3). Sementara Liu, et.al [6] mendefinisikan partisi untuk kovariat sehingga ada sebanyak *p* partisi yang mewakili masing-masing kovariat seperti yang mereka lakukan pada (5). Perbedaan berikutnya terlihat pada fungsi yang digunakan. Ketika [10] menggunakan fungsi distribusi kumulatif untuk statistik pada persamaan (4) sebagaimana teori dasar statistik Kolmogorov-Smirnov, sementara itu [6] mengganti fungsi distribusi kumulatif tersebut dengan fungsi survival sebagaimana pada persamaan (6).

Estimator untuk $K_j^{\mathbf{\Lambda}_j}$ adalah berikut ini:

$$\widehat{K}_j^{\mathbf{\Lambda}_j} = \max_{l_1,l_2} \sup_{0\le t\le\tau} \left|\hat{S}_j(t|I_j = l_1) - \hat{S}_j(t|I_j = l_2)\right| \quad (7)$$

Dimana $\hat{S}_j(t|I_j = l)$ adalah estimator Kaplan-Meier untuk $S_j(t|I_j = l)$ berdasarkan sampel $\{X_i, \Delta_i, \boldsymbol{Z}_i : i \in D_{lj}\}$ dengan $D_{lj} = \{i : Z_{ij} \in [a_l^j, a_{l+1}^j), i = 1, \dots, n\}$. Estimator $\hat{S}_j(t|I_j = l)$ dinyatakan sebagai:

$$\hat{S}_j(t|I_j = l) = \prod_{i\in D_{lj}} \left\{1 - \frac{1}{\sum_{k\in D_{lj}} I(X_k \ge X_i)}\right\}^{\Delta_i I(X_i \le t)}$$

Untuk menentukan $\widehat{K}_j^{\mathbf{\Lambda}_j}$ [9] melakukan pengirisan seragam (uniform slicing) terhadap partisi data menjadi sebanyak Λ_j *slice*, dengan ketentuan:

(a) Jika Z_j adalah variabel bertipe ordinal dengan level $1, \dots, \Lambda_j$ atau variabel bertipe diskrit dengan nilai berhingga (*finite possible values*) $1, \dots, \Lambda_j$ maka $I_j = Z_j$.

(b) Jika Z_j adalah variabel bertipe diskrit, yakni 1,2,3, ..., maka $I_j = Z_j$ apabila $Z_j < \Lambda_j$ dan $I_j = \Lambda_j$ apabila $Z_j \ge \Lambda_j$.

(c) Jika Z_j adalah variabel bertipe kontinu, maka partisi $\mathbf{\Lambda}_j$ ibagi ke dalam interval-interval yang dibatasi oleh kuantil sampel $\frac{1}{\Lambda_j}$ dari Z_j, dengan $l = 0,1, \dots, \Lambda_j$.

3.1 Perpaduan Statistik KS-SM (*the fused* KS-SM)

Pada bagian ini digunakan ide perpaduan (*fusion*) yang diperkenalkan oleh [11] yang juga diterapkan oleh [10] dan [6] dalam rangka meningkatkan efisiensi statistik Kolmogorov-Smirnov yang digunakan. Dimisalkan Z_j seperti pada kasus (b) dan (c), berdasarkan ide *fusion* tersebut dipunyai sebanyak N_j partisi yang berbeda untuk $\mathbf{\Lambda}_{kj}$ (dengan $k = 1, \dots, N_j$), yang berarti masing-masing partisi untuk sebuah Z_j (yakni $\mathbf{\Lambda}_{kj}$) emuat interval sebanyak Λ_{kj}. Sehingga estimator statistik untuk Z_j diganti dari persamaan (7) menjadi hasil jumlahan estimator-estimator statistik tersebut dengan partisi berbeda, yaitu:

$$\widehat{K}_j = \sum_{k=1}^{N_j} \widehat{K}_j^{\mathbf{\Lambda}_{kj}} \quad (8)$$

yang merupakan estimator untuk $K_j = \sum_{k=1}^{N_j} K_j^{\mathbf{\Lambda}_{kj}}$. Kemudian sebagaimana yang disarankan oleh [10], banyaknya interval pada masing-masing $\mathbf{\Lambda}_{kj}$ adalah tidak lebih dari

$\lceil \log n \rceil$, artinya $\Lambda_{kj} \leq \lceil \log n \rceil$ untuk $k = 1, \dots, N_j$, sehingga tetap terdapat ukuran sampel yang memadai pada masing-masing *slice* yang dibuat. Notasi $\lceil x \rceil$ menandakan nilai yang diambil adalah integer terkecil yang tidak kurang dari x, misalnya $\lceil 4{,}285 \rceil = 5$. Dengan demikian $\Lambda_{kj} = 3, \dots, \lceil \log n \rceil$ untuk masing-masing partisi $\boldsymbol{\Lambda}_{kj}$. Di sisi lain jika kasusnya seperti pada poin (a) maka $N_j = 1$, yang berarti untuk kasus ini persamaan (7) akan sama dengan persamaan (8). Akhirnya berdasarkan $\widehat{K}_j$ tersebut dengan $j = 1, \dots, p$. Didefinisikan estimator untuk set aktif $\mathcal{A}$ sebagai berikut:

$$\mathcal{A}(d_n) = \left\{ \begin{array}{c} 1 \leq j \leq p : \widehat{K}_j \text{ terbesar yang} \\ \text{berada pada posisi } d_n \text{ pertama} \end{array} \right\}$$

dengan d_n adalah ukuran model yang menjadi batas banyaknya kovariat terpilih. Prosedur ini dinamakan metode penyaringan dengan perpaduan statistik Kolmogorov-Smirnov (*the fused Kolmogorov-Smirnov statistic-based Screening Method*).

Secara teoritis apabila statistik Kolmogorov-Smirnov pada (8) bernilai mendekati nol pada kovariat Z_j untuk suatu $j = j_0$, maka kovariat dengan indeks j_0 tersebut dipandang tidak berpengaruh terhadap waktu survival. Dengan kata lain kovariat tersebut akan tereliminasi dan j_0 dianggap tidak masuk ke dalam set aktif. Apabila proses tersebut diulang sebanyak p kali untuk semua kovariat yang ada, maka akan diperoleh suatu set yang memuat set aktif yang dicari. Sedangkan pada praktiknya (yang tentunya melibatkan komputasi) semua kovariat Z_j akan diberi bobot berdasarkan estimator statistik pada persamaan (8) tersebut, kemudian diurutkan dari yang paling besar sampai paling kecil lalu diambil estimasi set aktif sebanyak d_n kovariat yang nilai statistiknya paling besar, dengan $d_n = \lceil n/\log n \rceil$ sebagai mana pada [10].

3.2 Studi Simulasi

Studi simulasi merupakan langkah penelitian yang cukup penting dilakukan dalam rangka menilai kinerja (performance) suatu metode yang dikembangkan. Dalam suatu simulasi, data dibangkitkan dengan model yang diinginkan dan relevan dengan studi yang dilakukan. Dengan demikian pola hubungan antar variabel yang ada dalam data telah diketahui berdasarkan model yang dipilih. Pola hubungan yang telah diketahui itulah yang akan dijadikan sebagai informasi untuk melihat kinerja metode yang dikembangkan tersebut.

Simulasi yang akan dilakukan dalam studi ini akan menggunakan Model Regresi Cox (*Cox Proportional Hazards Model*) sebagaimana dalam dan Model AFT loglinear sebagaimana dalam [12] sebagai berikut:

- Konfigurasi Simulasi 1

  Dimisalkan waktu survival T mengikuti Model Regresi Cox dengan fungsi sebagai berikut:

  $$h(t|\boldsymbol{Z}) = h_0(t) \exp[\boldsymbol{\beta}^T \boldsymbol{Z}]$$

  dengan fungsi baseline hazard-nya berdistribusi Weibull yakni $h_0(t) = \lambda v t^{v-1}$, dengan $\lambda = 0{,}00001$ upakan parameter skala dan $v = 8{,}9$ merupakan parameter bentuk. Kemudian kovariat berdimensi tinggi $\boldsymbol{Z} = \{Z_1, \dots, Z_p\}$ mengikuti distribusi normal multivariat dengan mean $\mathbf{0}$ dan matriks korelasi $\boldsymbol{\Sigma} = \left(0{,}5^{|i-j|}\right)$ untuk $i, j = 1, \dots, p$.

Ditetapkan juga parameter koefisien $\boldsymbol{\beta} = (0{,}85; 0{,}85; 0{,}85; 0{,}85; 0{,}85; 0; \dots; 0)^T$ yang berarti hanya lima kovariat pertama yang aktif. Informasi tersensornya ditentukan dengan membangkitkan data berdistribusi $\text{Binomial}(n; 0{,}7)$ yang berarti 30% dari keseluruhan observasi itu tersensor. Kemudian waktu survival untuk setiap observasi yang bersesuaian dengan informasi tersensor dikurangi dengan suatu bilangan random berdistribusi $\text{Uniform}(0, T_i/4)$ sehingga waktu survival tersebut berubah status menjadi tersensor. Data waktu survival dengan model tersebut sebagaimana yang dijelaskan di dalam [13] dibangkitkan dengan $T = \left(\frac{\log U}{\lambda \exp[\boldsymbol{\beta}^T \boldsymbol{Z}]}\right)^{\frac{1}{v}}$.

- Konfigurasi Simulasi 2

  Dimisalkan waktu survival T mengikuti model AFT log-linear berikut:

  $$\log T = \mu + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_p Z_p + \sigma_\epsilon \epsilon$$
  $$= \mu + \boldsymbol{\beta}^T \boldsymbol{Z} + \sigma\epsilon$$

  dengan $\mu = -0{,}5$ adalah intersep dan $\sigma_\epsilon = 1$ adalah parameter skala. Error ϵ mengikuti distribusi normal standar. Kovariat $\boldsymbol{Z} = \{Z_1, \dots, Z_p\}$ juga mengikuti distribusi normal multivariat dengan mean $\mathbf{0}$ dan matriks korelasi $\boldsymbol{\Sigma} = \left(\sigma^{|i-j|}\right)$ untuk $i, j = 1, \dots, p$. Nilai $\boldsymbol{\beta}$ dan mekanisme penyensoran data juga mengikuti bentuk sebagaimana yang diberikan pada simulasi 1.

Pada masing-masing konfigurasi simulasi tersebut ditentukan dua jenis korelasi yakni $\sigma = 0{,}5$ dan $\sigma = 0{,}2$, kemudian ditetapkan tiga jenis ukuran sampel yakni $n = 50{,}100{,}150$, dan banyaknya kovariat adalah $p = 1000$. Simulasi dilakukan sampai 100 kali untuk setiap ukuran sampel pada setiap korelasi dalam kedua konfigurasi di atas. Sehingga total repetisi simulasi dilakukan 1200 kali.

Untuk mengukur kinerja dari metode penyaringan tersebut dilakukan dengan tiga cara sebagaimana yang diberikan [14]. Cara pertama dengan melihat ukuran minimum model yang memuat semua kovariat aktif. Dengan demikian jika semakin kecil ukuran modelnya maka semakin baik kinerja dari metode yang diberikan. Dari 100 kali repetisi diambil median dari keseluruhan ukuran minimum model. Cara kedua dengan melihat proporsi sebuah kovariat aktif bisa terpilih dalam 100 kali pengulangan yang disimbolkan dengan $\mathcal{P}_e$. Dan cara ketiga dengan melihat proporsi kelima kovariat aktif tersebut termuat dalam output secara bersamaan yang disimbolkan dengan $\mathcal{P}_a$. Dalam hal ini, pada cara kedua dan ketiga apabila semakin besar proporsinya berarti semakin baik kinerjanya.

Analisis dilakukan dengan bantuan software R-3.6.1 dan *script* programnya ditulis di R-Studio agar pengerjaannya lebih interaktif dan mudah dikelola. Setelah dilakukan repetisi membangkitkan data dan melakukan penyaringan variabel sebanyak 1200 kali, diperoleh hasil pengukuran kinerja berdasarkan ketiga kriteria yang telah disebutkan sebelumnya dan disajikan dalam pada Tabel 1.

Dari Tabel 1 dapat dilihat bahwa semakin besar ukuran sampel maka semakin kecil nilai median serta semakin besar proporsi $\mathcal{P}_e$ dan $\mathcal{P}_a$ yang berarti semakin efektif metode tersebut

bekerja. Hal itu sesuai dengan ketentuan *sure screening property* pada [7]. Berdasarkan ketiga kriteria itu juga dapat dilihat bahwa korelasi antar kovariat aktif yang semakin tinggi juga mempermudah metode *screening* Kolmogorov-Smirnov ini bekerja dengan baik. Secara keseluruhan metode ini bekerja cukup efektif untuk melakukan seleksi kovariat aktif pada suatu data survival berdimensi tinggi yang mengandung observasi tersensor.

Tabel 1. Median, $\mathcal{P}_e$, $\mathcal{P}_a$dalam dalam 100 kali repetisi membangkitkan data untuk dua jenis korelasi antar-kovariat dan tiga jenis ukuran sampel pada kedua konfigurasi.

Model	σ	n	Median	$\mathcal{P}_e$					$\mathcal{P}_a$
				Z_1	Z_2	Z_3	Z_4	Z_5	
Konf.1	0.5	50	76	0,47	0,79	0,85	0,73	0,44	0,11
		100	6	0,90	1,00	1,00	1,00	0,94	0,84
		150	5	1,00	1,00	1,00	1,00	1,00	1,00
	0.2	50	329	0,26	0,42	0,42	0,36	0,24	0,01
		100	68	0,61	0,81	0,83	0,79	0,66	0,20
		150	14	0,92	0,96	0,97	0,98	0,88	0,73
Konf.2	0.5	50	66	0,48	0,79	0,81	0,79	0,50	0,10
		100	7	0,92	0,99	0,99	0,99	0,90	0,80
		150	5	1,00	1,00	1,00	1,00	0,99	0,99
	0.2	50	384	0,25	0,41	0,35	0,50	0,29	0,02
		100	46	0,73	0,84	0,87	0,83	0,60	0,25
		150	8	0,92	0,98	0,97	0,95	0,94	0,78

3.3 Studi Kasus

Studi kasus pada penelitian ini dilakukan dengan menerapkan metode penyaringan Kolmogorov-Smirnov tersebut pada data pasien *Mantle Cell Lymphoma* yang dapat diakses melalui *http://llmpp.nih.gov/MCL*. Data ini mengandung 8810 cDNA dari 92 pasien yang terdeteksi menderita *Mantle Cell Lymphoma* (sejenis kanker) berdasarkan kriteria morfologi dan imunofenotip. Masing-masing cDNA tersebut memiliki *unique identification* (UNIQID) sebagai keterangan identitas kovariat dalam bentuk angka disamping keterangan lain. Selama pasien-pasien tersebut berada dalam pengamatan, sebanyak 64 orang meninggal dan sebanyak 28 orang lainnya tersensor (tidak meninggal sampai masa studi berakhir).

Tujuan utama dari penerapan metode penyaringan ini adalah untuk mengidentifikasi gen-gen yang memberikan pengaruh besar terhadap kemampuan bertahan pasien dari kematian. Dapat dilihat bahwa studi kasus ini melibatkan jumlah prediktor yang jauh lebih banyak dibandingkan ukuran sampel yang ada. Proses penyaringan variabel dalam hal ini merupakan langkah persiapan untuk mereduksi dimensi sebelum melakukan pemodelan lebih lanjut. Hasil penyaringan tersebut berupa indeks-indeks kovariat yang diurutkan dari bobot statistik (*fused kolmogorov-smirnov*) paling besar sampai paling kecil.

Di dalam data yang disajikan Rosenwald et.al [15] gen-gen yang berkorelasi tinggi dengan prediksi kemampuan bertahan pasien MCL tersebut adalah sebagai berikut:

Tabel 2. Hasil yang diperoleh [15]

No	UNIQID	No	UNIQID	No	UNIQID
1	16555	8	27057	15	29897
2	**16587**	9	**27095**	16	**30142**
3	24610	10	27762	17	30157
4	24719	11	28581	18	30620
5	24723	12	28640	19	**30898**
6	**24734**	13	28990	20	**32699**
7	26191	14	29357	21	**34790**

Berikutnya diterapkan metode *wrapper sequential forward selection* sebagaimana dijelaskan dalam [16] terhadap hasil *screening* untuk mengungkap kovariat aktif dengan signifikansi yang lebih baik. Metode *wrapper sequential forward selection* yang digunakan berbasis regresi Cox Proportional Hazard. Karena *wrapper forward selection* bekerja melibatkan model (dalam hal ini regresi Cox P.H.) maka estimasi bisa dilakukan apabila banyaknya kovariat yang dipertimbangkan tidak lebih dari $n = 92$. Karena data mengandung *missing value* pada baris-baris observasi tertentu, maka baris-baris tersebut dihilangkan terlebih dahulu sehingga diperoleh $n = 74$ observasi lengkap (tanpa *missing value*). Dengan demikian pada metode *wrapper forward selection* dipertimbangkan 74 kovariat teratas dari hasil *screening* yang telah dilakukan sebelumnya. Terdapat 7 kovariat hasil yang diperoleh Rosenwald et.al [15] beririsan dengan 74 kovariat teratas hasil penyaringan tersebut, yakni gen-gen dengan UNIQID 30898, 34790, 16587, 30142, 27095, 24703, dan 32699. Oleh karena itu 7 kovariat tersebut akan dipertimbangkan sebagai kovariat awal yang tetap digunakan sebelum menambah ukuran model dengan kovariat lainnya dalam proses *wrapper forward selection*. Output yang diberikan adalah indeks-indeks kovariat yang terseleksi sebagai prediktor paling berpengaruh. Ukuran awal model ditetapkan dengan *hard treshold* (fungsi sederhana terhadap n) dan mengacu pada yang dilakukan Mai dan Zou [10] yakni $d_n = \lceil n/\log n \rceil$. Karena $n = 74$ berarti ukuran modelnya adalah 17. Setelah dievaluasi kembali dari 17 kovariat yang diperoleh ada sebanyak 16 kovariat yang signifikan di dalam model.

Sama halnya dengan simulasi di atas, proses penyaringan variabelnya dilakukan dengan bantuan software R-3.6.1 dan *script* programnya ditulis di R-Studio. Hasil yang diperoleh setelah melalui proses *screening* Kolmogorov-Smirnov dan metode *wrapper* tersebut adalah sebagai berikut:

Tabel 3. Hasil *Screening* dan *Wrapper* pada Kasus MCL

No	Indeks Variabel	UNIQID	No	Indeks Variabel	UNIQID
1	8723	34790	9	2462	24832
2	5533	30142	10	220	16118
3	3522	27095	11	1783	23826
4	2371	24734	12	3610	27203
5	7297	32699	13	5545	30157
6	5659	30334	14	7847	33531
7	3850	27496	15	5647	30319
8	3691	27310	16	3404	26950

Pada Tabel 3 terdapat informasi indeks variabel dan UNIQID. Indeks variabel adalah indeks dari kovariat yang diberikan penulis berurutan sesuai urutan kovariat pada data, yaitu $j = 1, \ldots, 8810$, dalam rangka memudahkan

looping pada saat melakukan proses komputasi. UNIQID adalah ID yang terdapat pada data sebagai pengidentifikasi terhadap masing-masing gen. Hasil yang diperoleh penulis menunjukkan 5 dari 7 gen yang sama dengan yang diperoleh Rosenwald et.al [15] tetap signifikan di dalam model.

4. Kesimpulan Dan Saran

Berdasarkan pembahasan serta simulasi dan studi kasus yang dilakukan dapat ***disimpulkan*** beberapa hal yakni, (1) untuk menangani permasalahan data survival berdimensi tinggi yang mengandung observasi tersensor menggunakan metode *Screening* Kolmogorov-Smirnov dilakukan dengan mengganti fungsi distribusi kumulatif kovariat bersyarat variabel respon menjadi fungsi survival variabel respon bersyarat kovariat yang diestimasi dengan estimator fungsi survival Kaplan-Meier; (2) metode *Screening* Kolmogorov-Smirnov dapat digunakan untuk menangani data dengan kovariat bertipe nominal, ordinal, diskrit, maupun kontinu, dan prosedurnya tidak bergantung pada asumsi model tertentu; (3) hasil *screening* akan semakin akurat seiring bertambahnya ukuran sampel dan semakin besarnya korelasi antar kovariat aktif; dan (4) hasil *screening* yang diperoleh bersifat fleksibel untuk ditangani lebih lanjut dengan berbagai model (tidak memiliki kecenderungan pada model tertentu) karena proses *screening*nya bebas model.

Metode *screening* Kolmogorov-Smirnov yang dibahas dalam artikel ini menggunakan perpaduan (jumlahan) statistik KS yang mengacu pada banyaknya sampel untuk setiap kovariat. Apabila ada kovariat tertentu yang mengandung *missing value* cukup banyak maka pembobotan statistik untuk kovariat tersebut tidak akan sebanding dengan kovariat lain dengan *missing value* lebih sedikit atau dengan kovariat lain yang tidak mengandung *missing value*. ***Disarankan*** untuk mengembangkan model ini agar pembobotan masing-masing kovariat menjadi sebanding meski dengan perbedaan banyaknya *missing value* di dalamnya.

5. Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberi saran untuk penyempurnaan penulisan artikel ini.

Daftar Pustaka

[1] R. Tibshirani, “Regression shrinkage and selection via the lasso,” J. R. Stat. Soc. Ser. B, vol. 58, no. 1, pp. 267–268, 1996, doi: 10.1017/s0272503700054525.

[2] J. Fan and R. Li, “Variable Selection via Nonconcave Penalized,” J. Am. Stat. Assoc., vol. 96, no. 456, pp. 1348–1360, 2001.

[3] H. Zou, “The adaptive lasso and its oracle properties,” J. Am. Stat. Assoc., vol. 101, no. 476, pp. 1418–1429, 2006, doi: 10.1198/016214506000000735.

[4] E. Candes and T. Tao, “The Dantzig selector: Statistical estimation when p is much larger than n,” *Ann. Stat.*, vol. 35, no. 6, pp. 2313–2351, 2007, doi: 10.1214/009053606000001523.

[5] C. H. Zhang, *Nearly unbiased variable selection under minimax concave penalty*, vol. 38, no. 2. 2010.

[6] Y. Liu, J. Zhang, and X. Zhao, “A new nonparametric screening method for ultrahigh-dimensional survival data,” *Comput. Stat. Data Anal.*, vol. 119, pp. 74–85, 2018, doi: 10.1016/j.csda.2017.10.003.

[7] J. Fan and J. Lv, “Sure independence screening for ultrahigh dimensional feature

space," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 70, no. 5, pp. 849–911, 2008, doi: 10.1111/j.1467-9868.2008.00674.x.

[8] J. Fan and R. Song, "Sure independence screening in generalized linear models with NP-dimensionality," *Ann. Stat.*, vol. 38, no. 6, pp. 3567–3604, 2010, doi: 10.1214/10-AOS798.

[9] Q. Mai and H. Zou, "The Kolmogorov filter for variable screening in high-dimensional binary classification," *Biometrika*, vol. 100, no. 1, pp. 229–234, 2013, doi: 10.1093/biomet/ass062.

[10] Q. Mai and H. Zou, "The fused Kolmogorov filter: A nonparametric model-free screening method," *Ann. Stat.*, vol. 43, no. 4, pp. 1471–1497, 2015, doi: 10.1214/14-AOS1303.

[11] R. D. Cook and X. Zhang, "Fused estimators of the central subspace in sufficient dimension reduction," *J. Am. Stat. Assoc.*, vol. 109, no. 506, pp. 815–827, 2014, doi: 10.1080/01621459.2013.866563.

[12] Danardono, *Analisis Data Survival*. Yogyakarta: FMIPA UGM, 2012.

[13] R. Bender, T. Augustin, and M. Blettner, "Generating survival times to simulate Cox proportional hazards models," *Stat. Med.*, vol. 24, no. 11, pp. 1713–1723, 2005, doi: 10.1002/sim.2059.

[14] R. Li, W. Zhong, and L. Zhu, "Feature screening via distance correlation learning," *J. Am. Stat. Assoc.*, vol. 107, no. 499, pp. 1129–1139, 2012, doi: 10.1080/01621459.2012.695654.

[15] A. Rosenwald *et al.*, "The proliferation gene expression signature is a quantitative integrator of oncogenic events that predicts survival in mantle cell lymphoma," *Cancer Cell*, vol. 3, no. 2, pp. 185–197, 2003, doi: 10.1016/S1535-6108(03)00028-X.

[16] R. Panthong and A. Srivihok, "Wrapper Feature Subset Selection for Dimension Reduction Based on Ensemble Learning Algorithm," *Procedia Comput. Sci.*, vol. 72, pp. 162–169, 2015, doi: 10.1016/j.procs.2015.12.117.