Plot *Multivariate* Menggunakan *Kernel Principal Component Analysis* (KPCA) dengan Fungsi *Power Kernel*

Vitawati Bawotong[1], Hanny Komalig[2], Nelson Nainggolan[3]

[1]Program Studi Matematika, FMIPA, UNSRAT, vbawotong@gmail.com
[2]Program Studi Matematika, FMIPA, UNSRAT, hanoy07@yahoo.com
[3]Program Studi Matematika, FMIPA, UNSRAT, bapaivana@yahoo.com

Abstrak

Kernel PCA merupakan PCA yang diaplikasikan pada input data yang telah ditransformasikan ke *feature space*. Misalkan $\Phi: R^n \rightarrow F$ fungsi yang memetakan semua input data $x_i \in R^n$, berlaku $\Phi(x_i) \in F$. Salah satu dari banyak fungsi kernel adalah *power kernel*. Fungsi *power kernel* $K(x_i, x_j) = -\| x_i - x_j \|^{\beta}$ dengan $0 < \beta \leq 1$. Tujuan dari penelitian ini yaitu mempelajari penggunaan Kernel PCA (KPCA) dengan fungsi *Power Kernel* untuk membantu menyelesaikan masalah plot *multivariate* nonlinier terutama yang berhubungan dalam pengelompokan. Hasil menunjukkan bahwa Penggunaan KPCA dengan fungsi *Power Kernel* sangat membantu dalam menyelesaikan masalah plot *multivariate* yang belum dapat dikelompokan dengan garis pemisah yang linier.

Kata kunci : *Kernel Principal Component Analysis* (KPCA), Plot *Multivariate*, *Power Kernel*

Multivariate Plot Using Kernel Principal Component Analysis (KPCA) with Power Kernel Functions

Abstract

Kernel PCA is PCA which applied to the input data that has been transformed to feature space. Let $\Phi: R^n \rightarrow F$ *function that maps all data input* $x_i \in \mathbb{R}^n$, *applies* $\Phi(x_i) \in F$. *One of many kernel functions is the power kernel. Power kernel function* $K(x_i, x_j) = -\|x_i - x_j\|^{\beta}$ *with* $0 < \beta \leq 1$. *The purpose of this research is to study the use of Kernel Principal Component Analysis (KPCA) with Power Kernel functions to help solve the problem of multivariate nonlinear plot mainly dealing in the grouping. Results showed that the use of KPCA with Kernel Power function is very helpful in solving the problem of multivariate plot that can not be grouped with the dividing line is linear.*

Keywords*: Kernel Principal Component Analysis* (KPCA), *Multivariate Plot, Power Kernel*

1. Pendahuluan

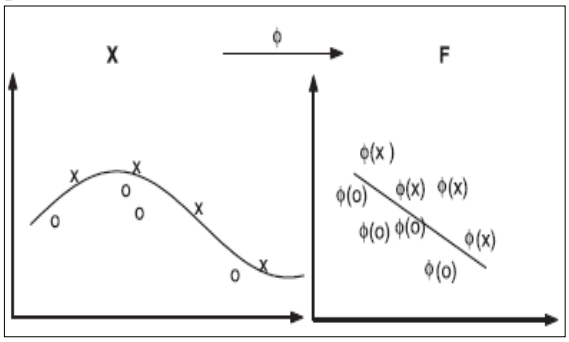
Untuk menyampaikan suatu data atau informasi, lebih menarik jika ditampilkan dalam bentuk gambar, dan juga menampilkan data-data dari suatu objek. Supaya lebih mudah dibaca oleh pengguna informasi, posisi jarak relatif objek-objek berdasarkan data yang ada dapat ditampilkan dalam sebuah plot. Salah satu contoh plot data dalam dua dimensi disebut diagram pencar (*scatter plot*). Analisis yang dapat memvisualisasikan data adalah Analisis Peubah Ganda. Pada tahun 2005 [1], analisis peubah ganda (*multivariate analysis*) merupakan metode dalam melakukan penelitian terhadap lebih dari dua peubah secara bersamaan. Metode ini dilakukan pada data yang memiliki karakteristik lebih dari satu peubah bebas dan/atau lebih dari satu peubah tak bebas atau terikat. Beberapa jenis analisis yang masuk dalam kategori analisis peubah ganda, diantaranya : Analisis Komponen Utama (*Principal Component Analysis*), Analisis Gerombol (*Cluster Analysis*), Analisis Faktor (*Factor Analysis*), Korelasi Kanonik, Analisis Biplot, Analisis Diskriminan (*Discriminant Analysis*), dan Penskalaan Dimensi Ganda (*Multidimension Scalling*).

Analisis Komponen Utama (*Principal Component Analysis*), merupakan analisis tertua dalam APG yang diperkenalkan oleh Karl Pearson tahun 1901, yang biasanya digunakan untuk: (1) identifikasi peubah baru yang mendasari data peubah ganda, (2) mereduksi jumlah himpunan peubah yang banyak dan saling berkorelasi menjadi peubah-peubah baru yang tidak berkorelasi dengan mempertahankan sebanyak mungkin keragaman data tersebut, dan (3) menghilangkan peubah-peubah asal yang tidak memberi informasi yang penting [2]. Namun, PCA tidak dapat

memodelkan data yang kompleksitasnya tinggi dengan hubungan tidak linier antar peubah. Untuk menyelesaikan persoalan tersebut maka digunakanlah metode Kernel PCA (KPCA) dengan fungsi *Power Kernel*. Fungsi kernel memetakan data ke dimensi yang lebih tinggi dan membangun fungsi pemisah dalam ruang yang terpisahkan. Hal ini dilakukan dengan menghitung fungsi kernel yang memberikan nilai hasil kali dalam pada *feature space* tanpa menunjukkan pemetaan secara eksplisit. Menurut [3] Kernel PCA sebagai metode berbasis memori, yaitu jika *x* merupakan suatu objek maka menemukan skor untuk objek tersebut dapat menggunakan nilai eigen dan vektor eigen dari data asal. Karena dalam mengklasifikasikan suatu objek ke dalam suatu kelompok diperlukan beberapa peubah penciri yang dapat membedakan antara satu kelompok dengan kelompok yang lainnya, maka atas dasar inilah Kernel PCA dapat digunakan dalam menyelesaikan pengklasifikasian suatu objek dalam suatu kelompok.

2. Kernel PCA

Metode kernel adalah salah satu cara untuk mengatasi kasus-kasus yang tidak linier. Dengan metode kernel suatu data x di *input space* dipetakan ke *feature space* dengan dimensi yang lebih tinggi melalui pemetaan $\boldsymbol{\Phi}$ sebagai berikut $\boldsymbol{\Phi} : x \mapsto \boldsymbol{\Phi}(x)$. Karena itu data x di *input space* menjadi $\boldsymbol{\Phi}(x)$ di *feature space*. Sering kali fungsi $\boldsymbol{\Phi}(x)$ tidak tersedia atau tidak bisa dihitung, tetapi *dot product* dari dua vektor dapat dihitung baik di dalam *input space* maupun di *feature space*. Dengan kata lain, sementara $\boldsymbol{\Phi}(x)$ mungkin tidak diketahui, *dot product* $\langle \boldsymbol{\Phi}(x_i), \boldsymbol{\Phi}(x_j) \rangle$ masih bisa dihitung di *feature space*. Suatu fungsi kernel $K(x_i, x_j)$, bisa untuk menggantikan *dot product* $\langle \boldsymbol{\Phi}(x_i), \boldsymbol{\Phi}(x_j) \rangle$. Kemudian di *feature space*, kita bisa membuat suatu garis pemisah yang linier yang mewakili fungsi nonlinier di *input space*. Gambar 1 mendeskripsikan suatu contoh *feature mapping* dari ruang dua dimensi ke *feature space* dua dimensi. Dalam *input space*, data tidak bisa dipisahkan secara linier, tetapi kita bisa memisahkan di *feature space* menjadikan tugas klasifikasi lebih mudah [4].



Gambar 1. Ilustrasi pemetaan kernel mengubah masalah yang non linier menjadi linier dalam *space* baru

PCA menemukan sumbu utama dengan mendiagonalkan matriks peragam

$$\boldsymbol{C} = \frac{1}{m}\sum_{j=1}^{m} x_j x_j^T \tag{1}$$

dan dengan demikian dapat didiagonalkan dengan nilai eigen non negatif

$$\lambda v = \boldsymbol{C} v \tag{2}$$

di mana v adalah vektor eigen. Dengan mensubstitusi persamaan (1) ke dalam persamaan (2), sehingga

$$\boldsymbol{C} v = \frac{1}{m}\sum_{j=1}^{m} x_j x_j^T \; v = \lambda v \tag{3}$$

sehingga

$$\begin{aligned} v &= \frac{1}{m\lambda}\sum_{j=1}^{m} x_j x_j^T \; v \\ &= \frac{1}{m\lambda}\sum_{j=1}^{m} (x_j \,.\, v) x_j \end{aligned} \tag{4}$$

Ditunjukkan bahwa $(xx^T)v = (x \,.\, v)x$

$$(xx^T)v = \begin{pmatrix} x_1x_1 & x_1x_2 & \dots & x_1x_M \\ x_2x_1 & x_2x_2 & \dots & x_2x_M \\ \vdots & \vdots & \ddots & \vdots \\ x_Mx_1 & x_Mx_2 & \dots & x_Mx_M \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_M \end{pmatrix}$$

$$= (x_1v_1 + x_2v_2 + \cdots + x_Mv_M)\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_M \end{pmatrix}$$

$$= (x.v)x \tag{5}$$

Tapi $(x.v)$ hanya skalar, jadi ini berarti bahwa semua solusi v dengan $\lambda \neq 0$ terletak pada rentang $x_1, \dots, x_m$, yaitu

$$v = \sum_{i=1}^{m} a_i x_i \tag{6}$$

Dengan demikian, matriks peragam di *feature space* untuk vektor $\{\boldsymbol{\Phi}(x_1), \boldsymbol{\Phi}(x_2), \dots, \boldsymbol{\Phi}(x_m)\}$ dapat dituliskan sebagai

$$\boldsymbol{C} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{\Phi}(x_j)\boldsymbol{\Phi}(x_j)^{\mathrm{T}} \tag{7}$$

dan masalah *eigen-value* di ruang *feature* $\boldsymbol{F}$ dapat dinyatakan sebagai

$$\lambda v = \boldsymbol{C}v \tag{8}$$

Sekarang akan ditunjukkan bahwa semua solusi v dengan $\lambda \neq 0$ terletak pada rentang $\boldsymbol{\Phi}(x_1), \dots, \boldsymbol{\Phi}(x_m)$, yaitu

$$\lambda(\boldsymbol{\Phi}(x_k) \cdot v) = (\boldsymbol{\Phi}(x_k) \cdot \boldsymbol{C}v) \quad ; k = 1, \dots, m \tag{9}$$

dimana

$$v = \sum_{i=1}^{m} a_i \boldsymbol{\Phi}(x_i) \tag{10}$$

substitusi persamaan (7) dan (10) ke dalam persamaan (9), maka

$$\lambda(\boldsymbol{\Phi}(x_k) \cdot \sum_{i=1}^{m} a_i \boldsymbol{\Phi}(x_i)) = \left(\boldsymbol{\Phi}(x_k) \cdot \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{\Phi}(x_j)\boldsymbol{\Phi}(x_j)^{\mathrm{T}} \sum_{i=1}^{m} a_i \boldsymbol{\Phi}(x_i)\right)$$

$$m\lambda \sum_{j=1}^{m} a_j \boldsymbol{\Phi}(x_i) = \sum_{i=1}^{m}\sum_{j=1}^{m} a_j \boldsymbol{\Phi}(x_i) K(x_i, x_j) \tag{11}$$

dimana

$$K(x_i, x_j) = \boldsymbol{\Phi}(x_i)^{\mathrm{T}} \boldsymbol{\Phi}(x_j) \tag{12}$$

Ada beberapa kernel yang sudah dikenal, antara lain :

- Gauss $K(x_i, x_j) = exp\left(-\frac{1}{2}\left(\frac{\|x_i - x_j\|}{\sigma^2}\right)^2\right)$
- Polinom $K(x_i, x_j) = \left({x_i}^T x_j + h_0\right)^p$
- *Power* $K(x_i, x_j) = -\|x_i - x_j\|^\beta$

3. *Power Kernel*

Umumnya transformasi $\boldsymbol{\Phi}$ ini tidak diketahui, dan sangat sulit untuk dipahami, maka perhitungan *dot product* tersebut sesuai teori Mercer dapat digantikan dengan fungsi kernel $K(x_i, x_j)$ yang mendefinisikan secara implisit transformasi $\boldsymbol{\Phi}$. Hal ini disebut sebagai "*Kernel Trick"* yang dirumuskan sebagai berikut :

$$K(x_i, x_j) = \boldsymbol{\Phi}(x_i)^T \boldsymbol{\Phi}(x_j) \tag{13}$$

Power kernel adalah salah satu trik kernel. Fungsi *power kernel* yang terbentuk adalah sebagai berikut :

$$K(x_i, x_j) = -\|x_i - x_j\|^\beta \qquad 0 < \beta \leq 1 \tag{14}$$

Dengan trik kernel ini cukup mengetahui fungsi kernel yang dipakai, dan tidak perlu mengetahui wujud dari fungsi non linier $\boldsymbol{\Phi}$ [5].

4. Metodologi Penelitian

Yang menjadi data penelitian ini adalah data sekunder, yang merupakan gambar plot *multivariate* non linear yang diambil dari masalah analisis gerombol (*cluster analysis*) pada buku

"***Multivariate Statistical Methods A PRIMER***" di halaman 105. Data yang pakai dalam penelitian ini hanya gambar plot data bagian c, d, e, dan f. Dikarenakan gambar plot data bagian a, dan b sudah terselesaikan yaitu dapat dipisahkan dua kelompok dari data tersebut, sedangkan plot data c digunakan sebagai pembanding, dan plot data d, e, dan f, belum dapat dipisahkan oleh garis linier maka dipakai sebagai data penelitian.

Langkah-langkah dalam metode analisis :

1) Gambar plot yang ada dalam buku "***Multivariate Statistical Methods A PRIMER***" di halaman 105 difoto kembali, kemudian diperbesar dan dicetak pada kertas *millimeter block*.
2) Buat sumbu koordinat X_1 dan X_2 untuk menentukan titik-titik koordinat dari X_1 dan X_2.
3) Titik koordinat yang diperoleh dari sumbu X_1 digunakan sebagai data X_1, begitupun dengan X_2.
4) Setelah data diperoleh, dilakukan standarisasi dengan rumus sebagai berikut.

$$Standarisasi\ X = \frac{x - \bar{x}}{s}$$

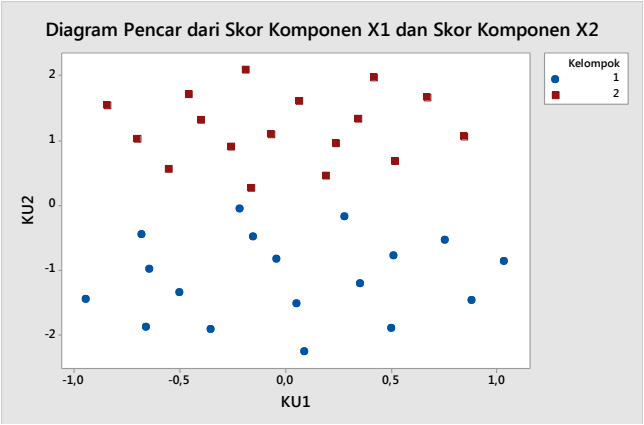
5) Dilakukan analisis komponen utama untuk menampilkan *score component* dari hasil standarisasi X_1 dan X_2.
6) Dilakukan perhitungan fungsi *power kernel* dengan pangkat berbeda yang akan digunakan menjadi data X_3, X_4, X_5, X_6, X_7.
7) Setelah itu, dicari *score component* dari hasil standarisasi X_1, X_2 dan hasil fungsi *power kernel*.
8) Ditampilkan matriks plot dari *score component* data yang distandarisasi dan *score component* hasil fungsi *power kernel*.
9) Analisa plot dari *matrix plot* yang diperoleh.

5. Hasil dan Pembahasan

Hasil yang didapat dalam penelitian ini yaitu plot-plot *multivariate* nonlinier dapat dipisahkan oleh garis pemisah yang linier menggunakan fungsi *power kernel*.

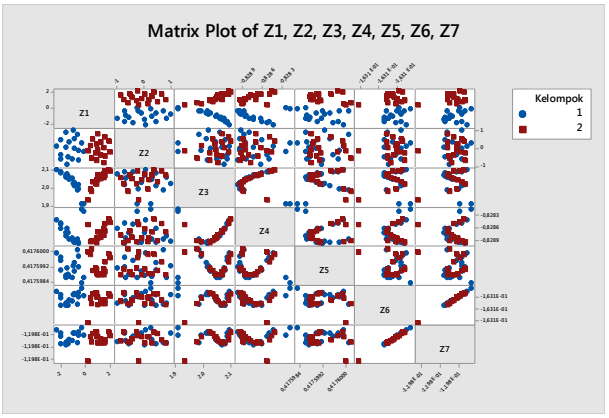
5.1 Plot Data-a

Setelah diperoleh data X_1 dan X_2 dari titik-titik koordinat pada sumbu koordinat X_1 dan X_2 untuk plot-a selanjutnya ditentukan kelompok dari plot tersebut, kemudian akan diolah menggunakan *software* statistika. Data X_1, X_2 diinput lalu distandarisasi. Hasil standarisasi dari X_1 dan X_2 kemudian dilakukan analisis komponen utama untuk melihat plot dari skor komponen pada data-a yang dihasilkan. Plot sebaran skor komponen data hasil standarisasi X_1 dan X_2 terlihat masih sama seperti sebaran plot awal.

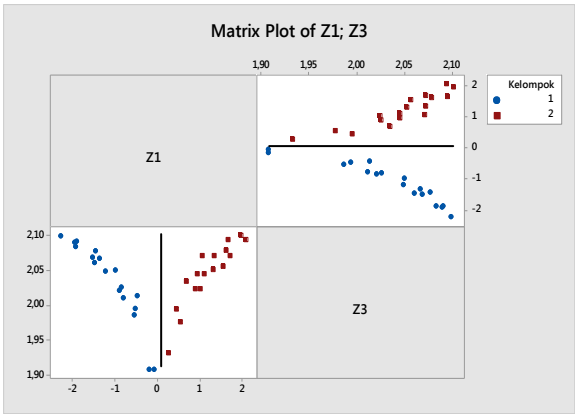


Gambar 2. Diagram pencar dari skor komponen pada data-a

Kemudian dilanjutkan dengan analisis komponen utama kernel dengan fungsi *power kernel*. Fungsi *power kernel* dilakukan pada *software* statistika. Setelah diperoleh data hasil analisis komponen utama kernel fungsi *power*, ditampilkan skor komponennya. Hasil dari skor komponen kemudian diplot dalam matriks plot. Hasil matriks plot yang diperoleh dapat dilihat pada Gambar 3 dan 4.



Gambar 3. Matriks plot skor komponen data-a

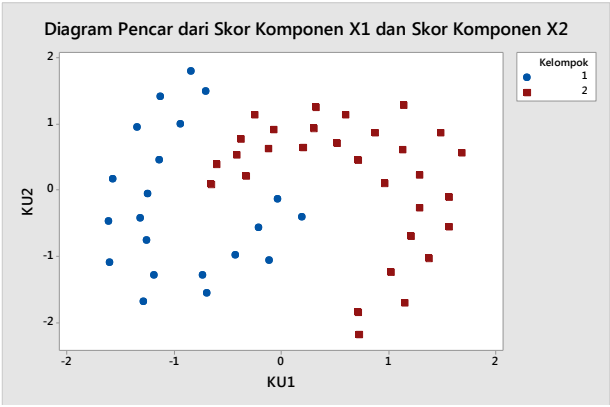


Gambar 4. Matriks plot komponen pertama dengan komponen ketiga pada data-a

Dari plot sebarannya, dapat dibuat suatu garis linier yang dapat memisahkan sebaran-sebaran individu kedalam dua kelompok yang berbeda. Pada plot, suatu garis linier dapat digunakan untuk diskriminasi. Pada Gambar 4 komponen utama pertama dengan komponen hasil fungsi *power kernel* berpangkat β = 0,01 dapat memisahkan dengan baik antara kelompok satu (berwarna biru) dengan kelompok dua (berwarna merah).

5.2 Plot Data-b

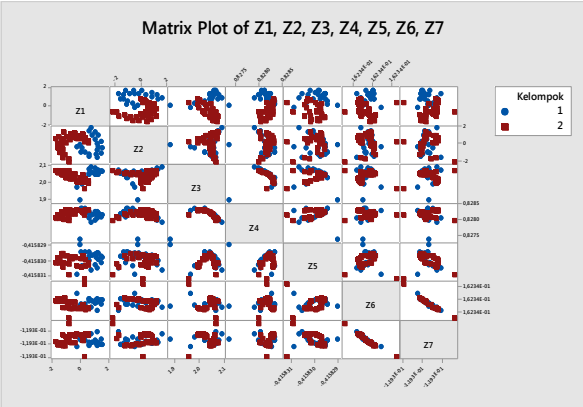
Dengan tahapan yang sama pada plot-a, diperoleh plot sebaran skor komponen data hasil standarisasi X_1 dan X_2 terlihat seperti gambar berikut.



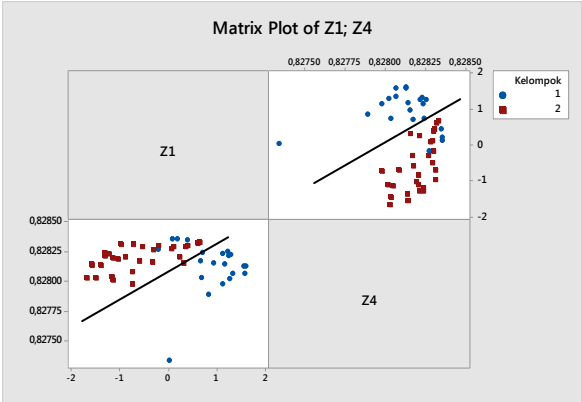
Gambar 5. Diagram pencar dari skor komponen pada data-b

Plot hasil analisis komponen utama masih menyerupai plot awal yang belum dapat dipisahkan antara kelompok yang satu dengan yang lain, maka dilanjutkan dengan analisis komponen utama kernel fungsi *power*. Hasil matriks plot dari skor komponen yang diperoleh dapat dilihat pada Gambar 6 dan 7.

Dari plot sebarannya, pada komponen utama pertama dengan komponen hasil fungsi *power kernel* berpangkat β = 0,02 terlihat bahwa masih ada beberapa individu dari kelompok satu yang belum terpisah, ini dikarenakan jarak antara kelompok satu dengan kelompok yang lain berdekatan pada plot awal data-b. Sejumlah individu memerlukan ketelitian khusus untuk memprediksi hasil diskriminasi tersebut.



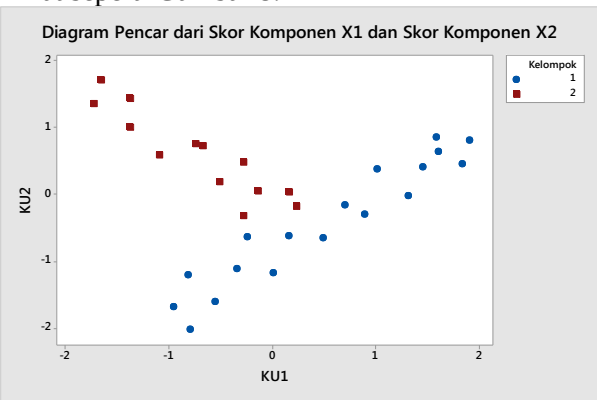
Gambar 6. Matriks plot skor komponen data-b



Gambar 7. Matriks plot komponen pertama dengan komponen keempat pada data-b

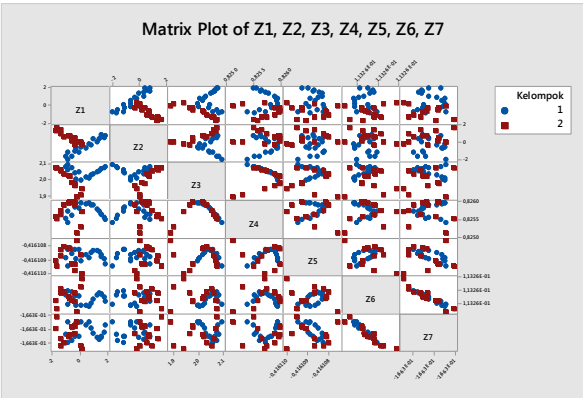
5.3 Plot Data-c

Dengan tahapan yang sama pada plot-a, diperoleh plot sebaran skor komponen data hasil standarisasi X_1 dan X_2 terlihat seperti Gambar 8.

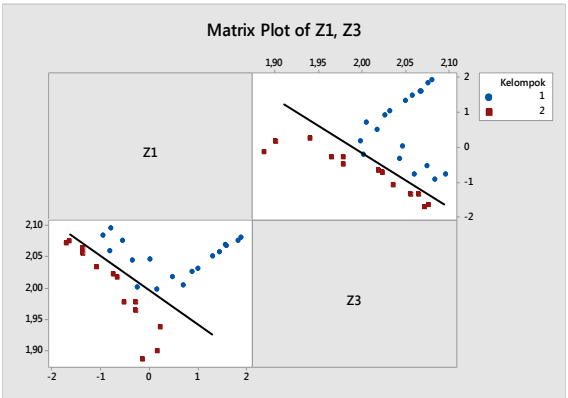


Gambar 8. Diagram pencar dari skor komponen pada data-c

Plot hasil analisis komponen utama masih menyerupai plot awal yang belum dapat dipisahkan antara kelompok yang satu dengan yang lain, maka dilanjutkan dengan analisis komponen utama kernel fungsi *power*. Hasil matriks plot dari skor komponen yang diperoleh dapat dilihat pada Gambar 9 dan 10.



Gambar 9. Matriks plot skor komponen data-c

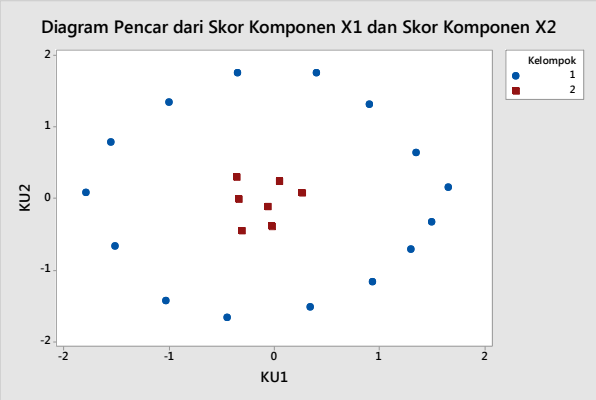


Gambar 10. Matriks plot komponen pertama dengan komponen ketiga pada data-c

Dari plot sebarannya, dapat dibuat suatu garis linier yang dapat memisahkan sebaran-sebaran individu kedalam dua kelompok yang berbeda. Pada plot, suatu garis linier dapat digunakan untuk diskriminasi. Komponen utama pertama dengan komponen hasil fungsi *power kernel* berpangkat $\beta = 0{,}01$ dapat memisahkan dengan baik.

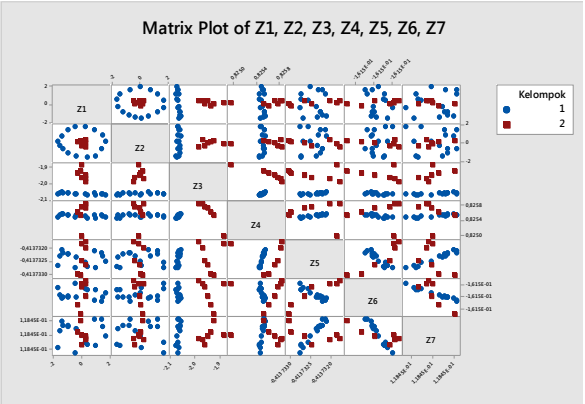
5.4 Plot Data-d

Dengan tahapan yang sama pada plot-a, diperoleh plot sebaran skor komponen data hasil standarisasi X_1 dan X_2 terlihat seperti Gambar 11.

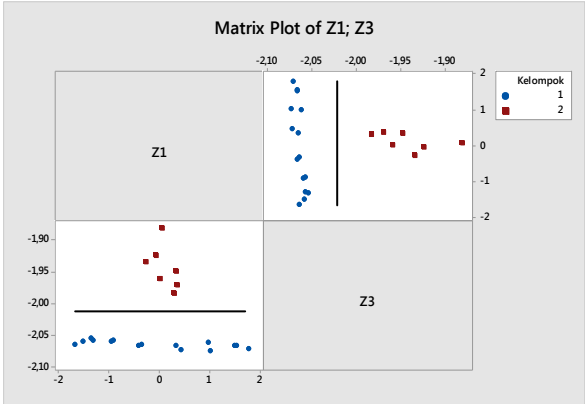


Gambar 11. Diagram pencar dari skor komponen pada data-d

Plot hasil analisis komponen utama masih menyerupai plot awal yang belum dapat dipisahkan antara kelompok yang satu dengan yang lain, maka dilanjutkan dengan analisis komponen utama kernel fungsi *power*. Hasil matriks plot dari skor komponen yang diperoleh dapat dilihat pada Gambar 12 dan 13.



Gambar 12. Matriks plot skor komponen data-d



Gambar 13. Matriks plot komponen pertama dengan komponen ketiga pada data-d

Dari plot sebarannya, dapat dibuat suatu garis linier yang dapat memisahkan sebaran-sebaran individu kedalam dua kelompok yang berbeda. Pada plot, suatu garis linier dapat digunakan untuk diskriminasi. Komponen utama pertama dengan komponen hasil fungsi *power kernel* berpangkat $\beta = 0{,}01$ dapat memisahkan dengan baik.

6. Kesimpulan

Penggunaan *Kernel Principal Component Analysis* (KPCA) dengan fungsi *Power Kernel* sangat membantu dalam menyelesaikan masalah plot *multivariate* yang belum dapat dikelompokkan dengan garis pemisah yang linier. Dengan menggunakan fungsi *Power Kernel*, plot data yang sebelumnya tidak dapat dipisahkan maka sudah dapat dipisahkan antar kelompok yang satu dengan kelompok yang lain.

- Plot data-a dapat dipisahkan dengan baik oleh fungsi *power kernel* berpangkat $\beta = 0{,}01$.
- Plot data-b sudah dapat dipisahkan hanya saja masih ada beberapa data kelompok satu yang berada pada kelompok dua.
- Plot data-c dapat dipisahkan dengan baik oleh fungsi *power kernel* berpangkat $\beta = 0{,}01$.

- Plot data-d juga dapat dipisahkan dengan baik oleh fungsi *power kernel* berpangkat $\beta = 0{,}01$.
- Fungsi *power kernel* dapat mengklasifikasikan data dengan baik jika jarak antar kelompok tidak berdekatan.

7. Daftar Pustaka

[1] Simamora, B. 2005. Analisis Multivariate Pemasaran. Gramedia Pustaka Utama, Jakarta.

[2] Siswadi dan Suharjo. 1997. Analisis Eksplorasi Data Peubah Ganda. Jurusan Matematika FMIPA IPB, Bogor.

[3] Kharismahadi, H. 2014. Klasifikasi Data Menggunakan Analisis Komponen Utama Kernel Dengan Fungsi Isotropik [skripsi]. FMIPA IPB, Bogor.

[4] Scholkopf B, and A.J. Smola. 2002. *Learning with Kernels*. The MIT Press, Cambridge, Massachusetts.

[5] Boolchandani D and V. Sahula. 2011. Exploring Efficient Kernel Functions for Support Vector Machine Based Feasibility Models for Analog Circuits. *International Journal of Design, Analysis and Tools for Circuits and Systems*. 1(1): 1-8.