Penggunaan Kernel PCA Gaussian dalam Penyelesaian Plot Multivariat Non Linier

Bernhard M. Wongkar[1], John S. Kekenusa[2], Hanny A.H. Komalig[3]

[1]**Program Studi Matematika, FMIPA, UNSRAT Manado, bernhard.wongkar2011@gmail.com**
[2]**Program Studi Matematika, FMIPA, UNSRAT Manado, johnkekenusa@yahoo.com**
[3]**Program Studi Matematika, FMIPA, UNSRAT Manado, hanoy07@yahoo.com**

Abstrak

Analisis Komponen Utama merupakan salah satu analisis peubah ganda yang digunakan untuk mentransformasi data secara linier sehingga terbentuk sistem koordinat baru. Untuk menyelesaikan masalah data yang tidak linier digunakan Kernel PCA. Penelitian dilakukan untuk menyelesaikan masalah plot multivariate non linier menggunakan Kernel PCA dengan fungsi Gaussian, terutama masalah data non linier yang berkelompok secara melingkar dalam dua dimensi. Hasil penelitian menunjukkan bahwa Kernel PCA dengan fungsi Gaussian dengan menggunakan parameter 0,1; 0,5; 1; 1,5 dapat melakukan pengelompokkan pada data tersebut, yang mana tidak dapat dikelompokkan PCA linier

Kata kunci : Gaussian, Kernel PCA, Pengelompokkan, Plot Multivariat

The Use of Gaussian PCA Kernel in Solving Non Linier Multivariate Plot

Abstract

Principal Component Analysis is one of Multivariate Analysis that are used to linearly transform data to form new coordinate system. To solve non linear data problem, PCA Kernel were used. The research was conducted to solve non linear multivariate plot problem using PCA Kernel with Gaussian function, especially problem of non linear data that circularly groued in two dimensions. Research result showed that PCA Kernel with Gaussian function using parameters 0,1; 0,5; 1; 1,5 can group the data which can not be grouped by linear PCA.

Keywords*: Gaussian, PCA Kernel, Grouping, Multivariate Plot*

1. Pendahuluan

Dalam penyajian data statistik, terdapat dua cara yang sering digunakan yaitu dengan tabel dan diagram. Dalam penggunaanya kerapkali data yang disajikan dalam bentuk tabel sulit untuk dipahami, lain halnya jika data disajikan dalam bentuk diagram maka data dapat lebih cepat dipahami.

Diagram adalah gambar yang menyajikan data secara visual yang biasanya berasal dari tabel yang telah di buat. Dalam gambar kita dapat melihat hubungan dari data – data objek dari sebaran plotnya. Salah satu analisis statistika yang dapat memvisualisasikan data ialah Analisis Peubah Ganda.

Analisis peubah ganda (*multivariate analysis*) merupakan metode dalam melakukan penelitian terhadap lebih dari dua variable secara bersamaan. Dengan analisis peubah ganda, kita dapat menganalisis pengaruh beberapa variabel terhadap variabel lainnya dalam waktu yang bersamaan. Secara umum analisis peubah ganda di gunakan dalam penelitian dengan tujuan untuk mereduksi data dan menyederhanakan struktur, pemilahan dan pengelompokan, pengamatan mengenai ketergantungan di antara peubah, peramalan, serta pembentukan dan pengujian hipotesis. Beberapa analisis yang termasuk jenis analisis peubah ganda, yaitu Analisis Komponen Utama (*Principal Component Analysis*), Analisis gerombol (*Cluster Analysis*), Analisis Faktor (*Factor Analysis*), *Multidimension Scalling*, dan *Correspondence Analysis*[1].

Analisis Komponen Utama atau *Principal Component Analysis* yang biasa disingkat PCA, merupakan analisis tertua dalam Analisis Peubah Ganda 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].

PCA dapat menyederhanakan suatu data, dengan cara mentransformasi data secara linier sehingga terbentuk suatu koordinat baru. Namun PCA tidak dapat memodelkan data yang memiliki hubungan tidak linier antar peubah. Untuk memodelkan data yang tidak linier maka digunakan metode kernel PCA.

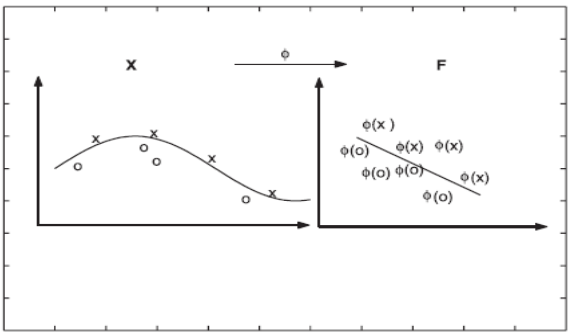
Kernel PCA merupakan perluasan dari PCA yang berguna untuk menyelesaikan masalah data yang tak linier (*non linear*) dan mengklasifikasikan obyek ke dalam kelompok untuk mendapatkan kesalahan klasifikasi terkecil. Pada Kernel PCA, data yang ada dipetakan ke ruang fitur, namun tidak semua hasil pemetaanya diketahui. Sehingga nilai eigen dan vektor eigen hanya dapat diperoleh dari matriks dua hasil pemetaan tersebut di ruang fitur.

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. Tujuan penelitian ini Dapat menyelesaikan masalah plot multivariat non linier menggunakan Kernel PCA dengan fungsi Gaussian, terutama yang berhubungan dalam pengelompokkan.

2. Kernel PCA

Metode kernel adalah salah satu cara untuk mengatasi kasus-kasus yang tidak linier. Dengan metode kernel suatu data $\boldsymbol{x}$ di *input space* dipetakan ke *feature space* dengan dimensi yang lebih tinggi melalui pemetaan $\boldsymbol{\Phi}$ sebagai berikut $\boldsymbol{\Phi} : \boldsymbol{x} \mapsto \boldsymbol{\Phi}(\boldsymbol{x})$. Karena itu data $\boldsymbol{x}$ di *input space* menjadi $\boldsymbol{\Phi}(\boldsymbol{x})$ di *feature space*.

Sering kali fungsi $\boldsymbol{\Phi}(\boldsymbol{x})$ 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 $\boldsymbol{\Phi}(\boldsymbol{x})$ mungkin tidak diketahui, *dot product* $\langle \boldsymbol{\Phi}(\boldsymbol{x}_i), \boldsymbol{\Phi}(\boldsymbol{x}_j) \rangle$ masih bisa dihitung di *feature space*. Suatu fungsi kernel $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, bisa untuk menggantikan *dot product* $\langle \boldsymbol{\Phi}(\boldsymbol{x}_i), \boldsymbol{\Phi}(\boldsymbol{x}_j) \rangle$. 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 menjadi lebih mudah [3].



Gambar 1. Pemetaan kernel mengubah masalah yang tidak linier menjadi linier dalam *space* baru

Kernel PCA merupakan PCA yang diaplikasikan pada input data yang telah ditransformasikan ke *feature space*. Misalkan $\boldsymbol{\Phi}: \mathbb{R}^n \mapsto \boldsymbol{F}$ fungsi yang memetakan semua input data $\boldsymbol{x}_i \in \mathbb{R}^n$, berlaku $\boldsymbol{\Phi}(\boldsymbol{x}_i) \in \boldsymbol{F}$. Berdasarkan transformasi ini, terlihat bahwa *feature space* dibangun oleh vektor-vektor $\{\boldsymbol{\Phi}(\boldsymbol{x}_1), \boldsymbol{\Phi}(\boldsymbol{x}_2), \dots, \boldsymbol{\Phi}(\boldsymbol{x}_m)\}$. Sehingga semua vektor di *feature space* dapat dinyatakan sebagai kombinasi linier dari vektor-vektor $\{\boldsymbol{\Phi}(\boldsymbol{x}_1), \boldsymbol{\Phi}(\boldsymbol{x}_2), \dots, \boldsymbol{\Phi}(\boldsymbol{x}_m)\}$.

PCA menemukan sumbu utama dengan mendiagonalkan matriks peragam

$$\boldsymbol{C} = \frac{1}{m} \sum_{j=1}^{m} \boldsymbol{x}_j \boldsymbol{x}_j^T \quad (1)$$

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

$$\boldsymbol{C}v = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{x}_j \boldsymbol{x}_j^T \, v = \lambda v \tag{3}$$

sehingga

$$v \; = \frac{1}{m\lambda}\sum_{j=1}^{m} \boldsymbol{x}_j \boldsymbol{x}_j^T \, v \;\; = \frac{1}{m\lambda}\sum_{j=1}^{m} (\boldsymbol{x}_j \, . \, v)\boldsymbol{x}_j \tag{4}$$

ditunjukkan bahwa $(\boldsymbol{x}\boldsymbol{x}^T)v = (\boldsymbol{x} \, . \, v)\boldsymbol{x}$

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

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

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

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

tapi $(\boldsymbol{x}.v)$ hanya skalar, jadi ini berarti bahwa semua solusi v dengan $\lambda \neq 0$ terletak pada rentang $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$, yaitu

$$v = \sum_{i=1}^{m} a_i \boldsymbol{x}_i \tag{6}$$

dengan demikian, matriks peragam di *feature space* untuk vektor $\{\boldsymbol{\Phi}(\boldsymbol{x}_1), \boldsymbol{\Phi}(\boldsymbol{x}_2), \dots, \boldsymbol{\Phi}(\boldsymbol{x}_m)\}$ dapat dituliskan sebagai

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

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 $\boldsymbol{\Phi}(\boldsymbol{x}_1), \dots, \boldsymbol{\Phi}(\boldsymbol{x}_m)$, yaitu

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

dimana

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

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

$$\lambda(\boldsymbol{\Phi}(\boldsymbol{x}_k) \cdot \sum_{i=1}^{m} a_i \boldsymbol{\Phi}(\boldsymbol{x}_i)) \; = \left(\boldsymbol{\Phi}(\boldsymbol{x}_k) \cdot \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{\Phi}(\boldsymbol{x}_j)\boldsymbol{\Phi}(\boldsymbol{x}_j)^{\mathrm{T}} \sum_{i=1}^{m} a_i \boldsymbol{\Phi}(\boldsymbol{x}_i)\right)$$

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

dimana

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\Phi}(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\Phi}(\boldsymbol{x}_j) \tag{12}$$

3. Gaussian Kernel

Kernel Gaussian adalah contoh radial fungsi dasar kernel,

$$k(x, y) = \exp(-\frac{||x-y||^2}{2\sigma^2}) \tag{13}$$

dimana : x adalah data X1 yang sudah distandarisasi
y adalah data X2 yang sudah distandarisasi
σ adalah parameter

Atau bisa juga diaksanakan dengan menggunakan :

$$k(x, y) = \exp\left(-\gamma\left||x - y|\right|^2\right)$$

Parameter σ memainkan peran utama dalam kinerja kernel, dan harus hati-hati disetel untuk masalah yang dihadapi. Jika berlebihan, eksponensial akan berperilaku hampir linear dan proyeksi dimensi yang lebih tinggi akan mulai kehilangan daya non-linier. Jika terlalu kecil, fungsi akan tidak teratur.

4. Metodologi Penelitian

4.1. Data Penelitian

Penelitian ini menggunakan data sekunder yaitu, gambar plot multivariat non linier yang diambil dari masalah analisis gerombol (*cluster analysis*) dalam buku "***Multivariate Statistical Methods A Primer***", Bryan F.J Manly 1986, halaman 105 gambar F [4]. Plot sebaran data pada gambar F belum dapat dipisahkan, maka dipakai sebagai data penelitian.

4.2. Prosedur Penelitian

1. Gambar plot yang ada difoto kembali, kemudian diperbesar dan dicetak pada kertas bergaris.
2. Ditarik sumbu untuk menentukan titik koordinat plot data, untuk mendapatkan data X1 dan X2.
3. Data X1 dan X2 di standarisasi.
4. Digunakan PCA untuk menampilkan data hasil standarisasi.
5. Dilakukan kernel PCA Gaussian untuk mentransformasi data hasil standarisasi.
6. Setelah itu, dicari *score component* dari data yang sudah ditransformasi.
7. Matriks plot dari *score komponent* data yang distandarisasi dan data yang ditransformasi
8. Interpretasi gambar *matrix plot* yang diperoleh.

5. Hasil dan Pembahasan

5.1. Pengambilan Data

Data penelitian diambil dari gambar plot Analisis Gerombol (*Cluster Analysis*) dalam buku pustaka [4]. Gambar data tersebut difoto kembali, diperbesar, kemudian diprint pada kertas bergaris. Selanjutnya ditarik 2 sumbu yang tegak lurus pada plot data tersebut, kemudian dinamakan X1 dan X2. Setelah koordinat dibuat, diperoleh data X1 dan X2 yang akan menjadi data penelitian. Dimana data X1 yaitu titik yang sejajar sumbu x horizontal dan X2 titik yang sejajar sumbu y vertikal.

Tabel 1. Data X1 dan X2

X1	X2
36	66
37	46
40	90
43	107
50	30
54	114
64	118
68	25
78	120
87	25
94	115
103	35
110	101
111	57
116	84
68	79
69	71
75	61
77	76
80	68
81	83
87	78

5.2. Standarisasi Data

Standarisasi data $dilakukan$ bertujuan untuk mendapatkan variabel-variabel data X1 dan X2 memiliki μ = 0, dan σ = 1. Standarisasi dilakukan dengan menggunakan rumusan :

$$X = \frac{x_i - \bar{x}}{S}$$

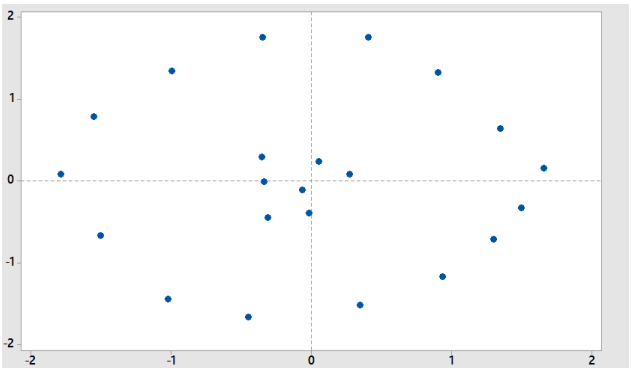
dimana : x_i adalah data ke i

$\bar{x}$ adalah nilai rata-rata dari data

S adalah simpangan baku, $S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

5.3. Penyelesaian dengan PCA

Data yang sudah distandarisasi akan diselesaikan dengan PCA, untuk dijadikan pembanding dengan hasil Kernel PCA. Hasil PCA dari data X1 dan X2 ditampilkan dalam *score plot* seperti disajikan pada Gambar 2.



Gambar 2. *Score plot* data

Pada Gambar 2, dapat dilihat bahwa sebaran plot data yang ada tidak memungkinkan ditarik satu garis lurus untuk memisahkan sebaran kedua kelompok data terebut. Jadi PCA tidak dapat mengelompokkan plot multivariate tersebut.

5.4. Penyelesaian dengan Kernel PCA fungsi Gaussian

Untuk penyelesaian Kernel PCA dengan fungsi Gaussian, data X1 dan X2 yang sudah distandarisasi, ditransformasi dengan fungsi Kernel Gaussian. Rumusan yang digunakan seperti pada persamaan (13).

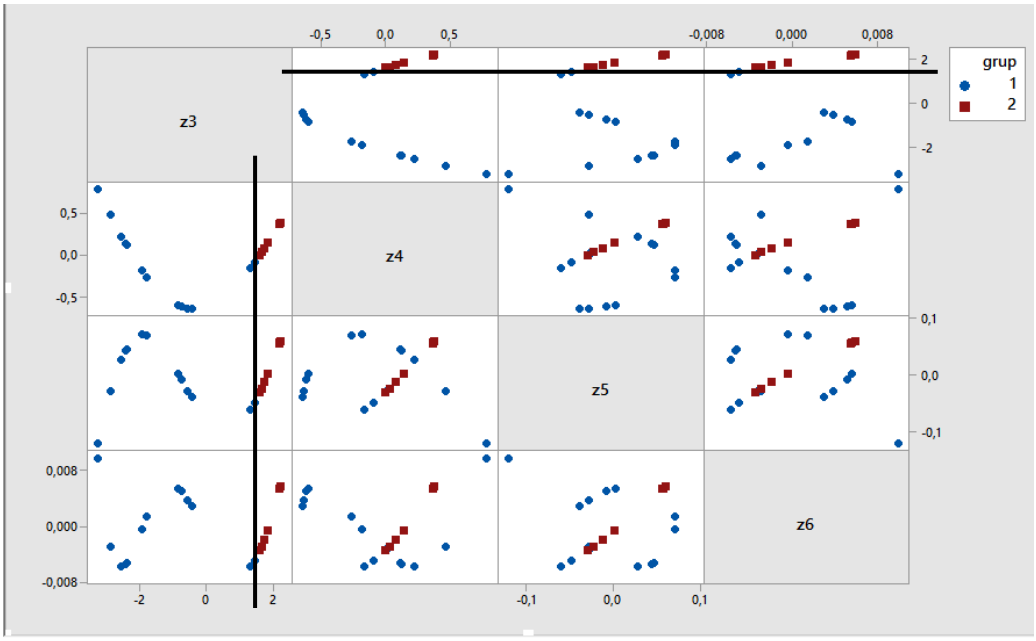
Transformasi kernel PCA Gaussian ini dilakukan menggunakan 4 nilai parameter, yaitu σ =0,1 ; 0,5 ; 1 ; 1,5. Keempat nilai parameter tersebut didapat setelah dilakukan beberapa kali percobaan degan menggunakan beberapa parameter, yaitu σ = 0,001 ; 0,005 ; 0,01 ; 0,05 ; 0,1 ; 0,5; 1 ; 1,5 ; 5 ; 10, didapati hasil gambar terbaik dalam pengelompokkan, yaitu menggunakan paramaeter σ =0,1 ; 0,5 ; 1 ; 1,5.

Setelah nilai transformasi dengan fungsi Gaussian yang didapat, kemudian ditentukan *score component* nya untuk mendapatkan peubah baru dari Kernel PCA fungsi Gaussian. Peubah baru yang didapat, hasilnya ditampilkan pada *matrix plot* seperti pada Gambar 3.

Dilihat pada gambar 3 ada beberapa gambar yang jika ditarik garis pemisah dapat membagi sebaran plot menjadi dua kelompok, kelompok 1 (warna biru) dan kelompok 2 (warna merah).

Pengelompokkan sebaran data belum maksimal karena masih ada data Tumpang tindih pada gambar di atas. Tumpang tindih data menandakan adanya kesamaan sifat umum beberapa data kelompok 1 dengan kelompok 2. Karena kecilnya jumlah data yang tumpang tindih, menandakan bahwa kedua kelompok data memiliki perbedaan sifat yang besar.

Meskipun masih adanya data yang tumpang tindih, pengelompokkan data mengunakan Kernel PCA dengan fungsi Gaussian dapat dikatakan berhasil karena dapat merubah pola sebaran data dan mengelompokkannya.



Gambar 3. *Matrix plot* skor komponen data

6. Kesimpulan

PCA tidak dapat digunakan dalam penyelesaian plot multivariat non linier, sebaliknya penggunaan Kernel PCA (*Kernel Principal Component Analysis*) dengan fungsi Gaussian sangat membantu dalam menyelesaikan masalah plot multivariat yang belum dapat dikelompokkan dengan garis linier sebagai pemisah.

7. Daftar Pustaka

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

[2] Johnson, R.A. and D.W. Wichern. 2002. *Applied Multivariate Analysis* 5th Ed. Prentice Hall Inc, New Jersey.

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

[4] Manly, B.F.J. 1986. *Multivariate Statistical Methods A PRIMER*. Chapman and Hall, New York.