Summary about dissertation of Zhang (2008) and Vercruysse (2022)Thesis work

May 13, 2025

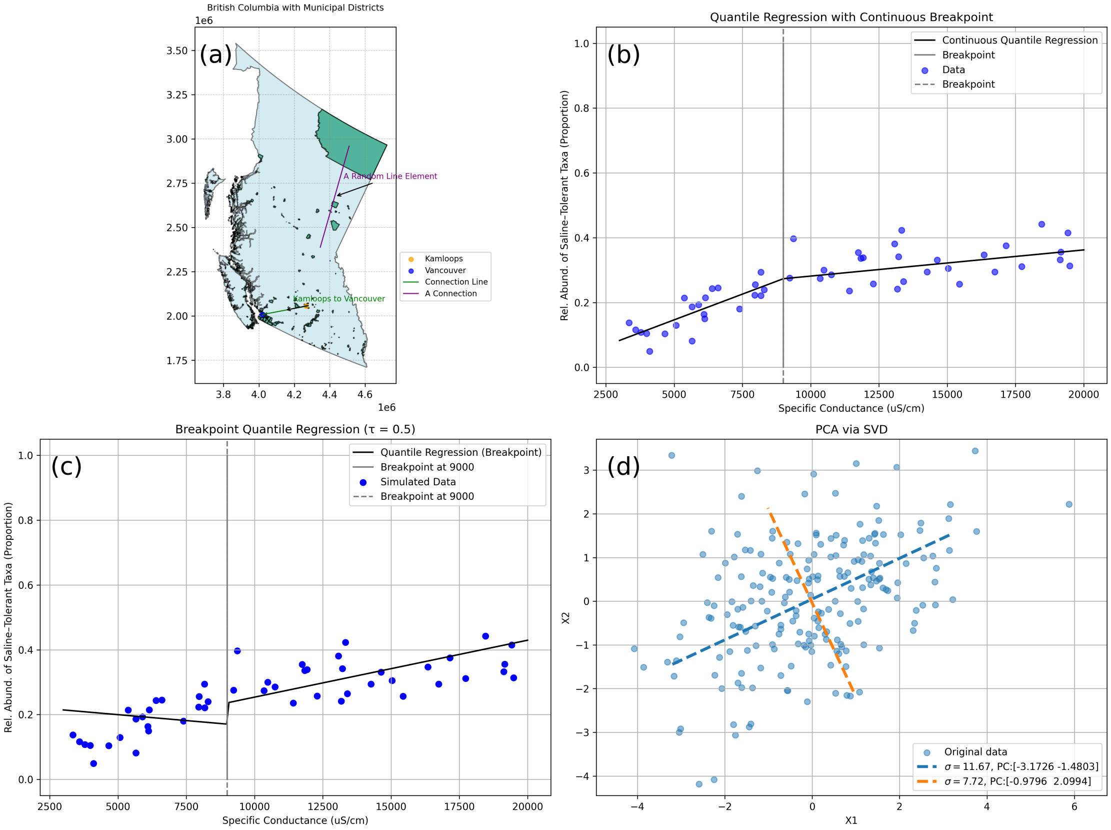
Abstract

In Part 1 and 2 of this document, i summarized the work of Zhang (2008) and Vercruysse et al. (2017) by listing the data, methods and results presented in their work. It can be skipped if the summary is not of interest.

In part 3, i did some coding with simualted data to see how can i quickly reproduce similar results and how could i extend the work. These ideas are tentative and not fully developed yet, but i think it is a good start and worth to be included in the report.

The major exploration i did can be concentrated into the figure below, which shows the geographic/spatial exploration(a), breakpoint quantile regression(b, c) and PCA by SVD(d).

The code and supportive files can be found in this **GitHub repository**: *src_test(source of code)*.



Contents

1 Zhang's work (2008)

The following notes summarize the work of Zhang (2008) on sediment contaminants and their effects on benthic communities in the Great Lakes region. The study focuses on the Detroit River and Lake St. Clair, examining the relationship between sediment contamination and zoobenthic community structure.

1.1 Data sources (Chapter 2)

1. Chemical Variables (Sediment Contaminants)

Grouped by Principal Components (PCA)

- **PC1: Trace and Minor Metals**
  - Aluminum (Al), Manganese (Mn), Cobalt (Co), Nickel (Ni), Iron (Fe), Copper (Cu), Chromium (Cr)
- **PC2: Trace Metals and PCBs**
  - Lead (Pb), Cadmium (Cd), Zinc (Zn), Mercury (Hg), Sum of PCBs
- **PC3: Organochlorine Compounds**
  - DDE (p,p-DDE), OCS (Octachlorostyrene)
- **PC4: Individual Variable**
  - Arsenic (As)

A composite sediment contamination score called **SumRel** was calculated by standardizing and summing all PC scores.

2. Environmental Variables (Habitat Features)

- Water depth (m)
- Water temperature (°C)
- Dissolved oxygen concentration (mg/L)
- Median particle size (phi scale)
- Total organic carbon (% LOI)
- Geographic coordinates (latitude, longitude)
- Site type (lake or river)
- Near-bottom water velocity (available only for Detroit River sites)

3. Biological Variables (Zoobenthic Taxa)

Depositional (Soft Substrate) Taxa

- Oligochaeta (Tubificidae)
- Chironomidae
- Ephemeridae (e.g., *Hexagenia*)
- Nematoda
- Gastropoda
- Acari

Erosional (Hard Substrate) Taxa

- Amphipoda (e.g., *Gammarus*, *Echinogammarus*)
- Dreissena (zebra mussels)
- Trichoptera (Hydropsychidae, Psychomyiidae, Polycentropodidae)
- Hydrozoa (e.g., *Hydra*, *Cordylophora*)

These taxa were used in cluster analysis, Bray-Curtis ordination, and to construct the Zoobenthic Condition Index (ZCI).

Data sampling and processing approaches

There are other data sampling and processing procedures used in the study, they are not listed here. But it is worth to menttion that to make the reliability and accuracy of the data.

1.2 Analysis Methods (Chapter 2)

Sediment Contaminant Analysis – PCA

Data used: Chemical variables

Principal Component Analysis (PCA) was applied to 16 sediment chemical variables to identify major contaminant groupings (e.g., metals, PCBs, organochlorines) and reduce dimensionality. Each site was assigned a composite sediment contamination score called **SumRel**, calculated by summing scaled (0–1) principal component scores.

Site Classification

Data used: Chemical variables

Sites were classified based on their SumRel scores. The 62 sites with the lowest contamination were identified as **Reference (REF)** sites, and the 62 sites with the highest contamination were classified as **Degraded (DEG)** sites. These served as the endpoints for defining a reference–degraded biological gradient.

Zoobenthic Community Grouping – Cluster Analysis

Data used: Biological variables

Ward's hierarchical clustering was performed on octave-transformed relative abundances of 16 dominant zoobenthic taxa. This analysis identified two major clusters:

- Cluster C1: Depositional (soft-substrate taxa)
- Cluster C2: Erosional (hard-substrate taxa)

Environmental Prediction – Discriminant Function Analysis (DFA)

Data used: Habitat variables, Biological variables

Discriminant Function Analysis used habitat data including depth, dissolved oxygen, temperature, organic content, particle size, and site type (lake or river) to classify sites into either Cluster C1 or C2. The DFA model was applied to predict habitat-associated biological community type for all 311 sites.

Gradient Analysis – Bray-Curtis Ordination

Data used: Biological variables

Bray-Curtis ordination was applied to zoobenthic community data to arrange sites along a biological similarity axis. This allowed construction of the **Zoobenthic Condition Index (ZCI)**, which ranges from 1.0 (reference-like) to 0.0 (degraded-like) based on community composition.

Statistical Modeling – Quantile Regression(The major part that i need to reproduce and try to extend)

Data used: Chemical variables, Biological variables

Quantile regression[1] was used to model the relationship between ZCI (response) and SumRel scores (predictor). Regressions were fit to the 10th, 50th (median), and 90th percentiles. The results revealed significant negative relationships, showing that biological condition declines with increasing sediment contamination.

Detroit River Case Study – Validation

Data used: Habitat variables, Biological variables

1.3 Findings (Chapter 2)

- PCA of 16 chemical variables revealed four main contaminant groupings and provided a composite contamination index (**SumRel**) used to rank site condition.
- Sites were successfully classified into **Reference** and **Degraded** categories based on SumRel scores.
- Cluster analysis of zoobenthic communities identified two distinct biological assemblages:
  - **Cluster C1**: Depositional habitat taxa
  - **Cluster C2**: Erosional habitat taxa
- Discriminant Function Analysis (DFA) using habitat variables accurately predicted community type (C1 or C2) for all 311 sites.
- Bray-Curtis ordination enabled creation of the **Zoobenthic Condition Index (ZCI)**, which captured each site's biological status on a 0–1 scale from degraded to reference.
- Quantile regression revealed a significant **negative relationship** between ZCI and SumRel, confirming that higher contamination is linked to degraded benthic community condition.
- A case study of the Detroit River showed that including **near-bottom water velocity** as an environmental predictor improved DFA classification, especially in depositional–erosional transitional zones.

1.4 Data sources (Chapter 3)

- **Biological data:** Chironomid larvae were collected from 113 sites during the 2004–2005 Lake Huron–Lake Erie Corridor survey. Larvae were examined for mouthpart deformities in the mentum or ligula.
- Only common genera with at least 40 individuals in more than one zone were analyzed (e.g., *Chironomus*).
- **Geographic scope:** Sampling covered 12 zones:
  - St. Clair River (4 zones)
  - Lake St. Clair (4 zones)
  - Detroit River (4 zones)

[1] Quantil regression in Wikipedia

1.5 Analysis Methods(Chapter 3)

- Larvae were preserved in ethanol, mounted on slides, and examined using a compound microscope.
- Deformities assessed included extra or missing teeth and abnormal Kohn gaps; damaged or worn mouthparts were excluded.
- The proportion of deformed larvae was calculated per genus and zone, with standard error estimated using a binomial formula.
- **Statistical analysis:** Replicated G-statistic tests were used to test for:
  - Spatial variation in deformity frequency across zones
  - Taxonomic variation among different genera

1.6 Findings (Chapter 3)

- Deformity rates varied significantly by genus; *Chironomus* had the highest deformity incidence.
- Spatial variation showed elevated deformities in zones not previously classified as degraded (e.g., Walpole Island, Canadian side of the St. Clair River, Belle Isle).
- These results suggest that mouthpart deformities are sensitive indicators of low-level or undetected contamination.
- **Conclusion:** Mouthpart deformities offer complementary information to community composition and can reveal stress not captured by broader biological indicators.

The full analytical approach was re-applied specifically to Detroit River sites to test its validity. Inclusion of near-bottom water velocity (a habitat variable available only for this region) improved the accuracy of DFA classification, especially in sites with mixed depositional–erosional conditions.

2 Vercruysse's work (2022)

The following notes summarize the work of Vercruysse (2022) on the environmental gradients in saline fens of Alberta. The study focuses on the relationship between morphometry and water chemistry in saline wetlands, particularly in relation to the distribution of aquatic vegetation.

2.1 Data Source (Chapter 2)

Data were collected from 52 waterbodies in a saline fen complex in Alberta during 2020, categorized into three morphometry types: flark (n = 38), flark/pond (n = 9), and pond (n = 5).

Environmental variables collected:

- **Water quality:** Temperature (°C), pH, dissolved oxygen (mg/L), specific conductance (μS/cm), redox potential (mV)
- **Nutrients and ions:** Chloride, sulfate (SO_4-S), phosphate (PO_4-P), ammonium (NH_4-N), total organic nitrogen (TON-N)
- **Cations and elements:** Calcium (Ca), Magnesium (Mg), Sodium (Na), Potassium (K), Aluminum (Al), Boron (B), Barium (Ba), Iron (Fe), Manganese (Mn), Lithium (Li), Strontium (Sr), Silicon (Si), Sulfur (S)
- **Physical site info:** Maximum depth (cm), northing and easting coordinates

All chemical analyses were conducted at the NRAL facility at the University of Alberta.

2.2 Analysis Methods (Chapter 2)

- **Dixon's Q-test:** Used to identify outlier sites in the environmental dataset, which were removed before multivariate analysis.
  Variables used: All environmental variables listed below were included:
  - pH, temperature, dissolved oxygen, specific conductance, redox potential
  - Nutrients: phosphate (PO_4), ammonium (NH_4), nitrate (NO_3), total organic nitrogen (TON)
  - Major ions and elements: Na, K, Ca, Mg, Cl, SO_4, Al, Fe, Mn, Sr, Si, Li, B, Ba, S
- **Pearson's correlation:** Tested relationships between morphometry class (coded as flark = 1, flark/pond = 2, pond = 3) and each environmental variable.
  Variables used: Same as above (individual environmental variables vs. morphometry index).
- **Holm's correction:** Applied to adjust p-values for multiple comparisons in the Pearson correlation analysis.
  Variables used: Adjusted p-values from all Pearson tests involving morphometry and environmental variables.
- **Principal Component Analysis (PCA):**
  - Conducted on water chemistry and environmental variables to reduce dimensionality and detect structure in gradients.
  - Morphometry class was excluded as a variable to avoid bias.
  - Varimax rotation was used to enhance interpretability of component loadings.

  Variables used: Full set of chemical and environmental variables (same as above), excluding morphometry.

2.3 Findings (Chapter 2)

- **Depth** was the only variable moderately correlated with morphometry (Pearson's r = 0.512), though it was not statistically significant after Holm correction.
- Several variables had weak correlations (r = 0.3–0.5) with morphometry, including: specific conductance, chloride, sulfate, sodium, calcium, magnesium, sulfur, strontium, boron, and northing.
- Phosphate concentration showed the most notable difference in means among morphometry types, highest in flark/ponds and lowest in ponds.
- PCA revealed no clear grouping by morphometry — suggesting overlapping environmental conditions across wetland types

2.4 Data Source (Chapter 3)

Samples were collected from 52 waterbodies in a boreal saline fen in Alberta between September 6–8, 2020. Wetlands were categorized into three morphometry types:

- Flark (n = 38)
- Flark/pond (n = 9)
- Pond (n = 5)

Biological data:

- Aquatic invertebrates sampled using CABIN protocol (D-frame net, 250 μm mesh).
- 20 jabs per site, with organisms sieved through 4.00, 1.00, 0.50, and 0.25 mm mesh sizes.
- Two replicate samples per site were preserved in ethanol and identified to the lowest taxonomic level.

Water chemistry data:

- Parameters included: specific conductance, pH, dissolved oxygen, temperature, redox potential.
- Ion and nutrient concentrations: Na, K, Ca, Mg, Cl, SO_4, PO_4, NH_4, TON, Fe, Mn, Sr, etc.

2.5 Analysis Methods (Chapter 3)

- **Community Composition Assessment:**
  - Calculated total invertebrate abundance and family richness from site replicates.
- **Non-metric Multidimensional Scaling (NMDS):**
  - Applied to genus-level community composition.
  - Hellinger transformation used on abundance data.
  - Bray-Curtis dissimilarity used for ordination.
  - Environmental vectors (e.g., conductance, depth, pH) fitted to NMDS axes.
- **Indicator Species Analysis (IndVal):**
  - Tested for taxa associated with wetland types (flark, flark/pond, pond).
  - No taxa showed statistically significant indicator values.
- **Linear Regression:**
  - Regressed total invertebrate abundance and family richness against log-transformed specific conductance.
  - Tested the hypothesis that richness decreases and abundance increases with salinity.

2.6 Findings (Chapter 3)

- Community composition did not differ significantly among morphometry types.
- NMDS showed overlapping site groupings; no distinct clustering by wetland type.
- Indicator species analysis revealed no statistically significant indicator taxa.
- Depth was the only variable marginally associated with morphometry.
- Conductance and major ion concentrations showed weak, non-significant relationships with wetland type.
- Regression analyses did not strongly support hypothesized trends in richness or abundance relative to conductivity.

3 How can i reproduce and extend the work?

3.1 Geographic visualization and spitial analysis by coding

I found in several places in the papers, the authors used geographic maps to show the location of survery sites and did spitial analysis on them. To might make the visualization more clear and the analysis more efficient, i can use librarys like *geopandas* and *geopy* in python to do the geographic visualization and spatial analysis. They may provide more possibilities to explore and extend the work of Zhang (2008) and Vercruysse et al. (2017) and might be a powerful tool for spitial data.

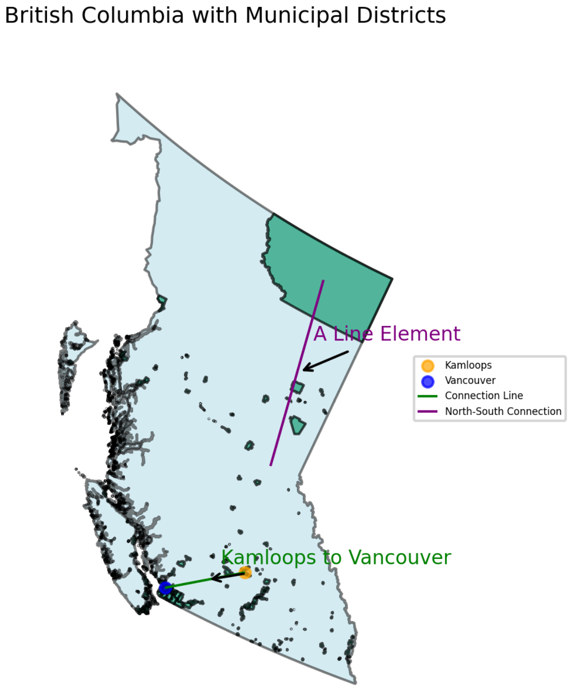


Figure 1: *Example of a map showing the location of Kamloops and Vancouver in BC. 4 points and two lines were added to the map.*

Figure 1 shows an example of a map showing the location of Kamloops and Vancouver in BC. I downloaded the geographic data from government website, including: *open.canada.ca* and *catalogue.data.gov.bc.ca*. They provide the provincial and municipal boundaries of BC.

I managed to add basic geographic elements to the map, like points, lines and polygons and it worked well. If this part is considered worth to be included, i would spcifically try to create maps for Ontario Province with the Lake Huron-Lake Erie area and the Alberta Province with the fen complex distribution.

3.2 Breakpoint quantile regression

The detailed math part of breakpoint quantile regression is not discussed here, it could be left to work on later. I tried to reproduce a similar result by using simulated data to validate the implementation with Python and my own code.

Continuous Breakpoint Regression: In continuous breakpoint regression, the relationship between the independent variable x and the dependent variable y is modeled such that the function remains continuous at the breakpoint c. For example:

$$y = \begin{cases} \beta_0 + \beta_1 x & \text{if } x \leq c \\ \beta_0 + \beta_1 c + \beta_2 (x - c) & \text{if } x > c \end{cases}$$

Discontinuous Breakpoint Regression: In discontinuous breakpoint regression, the relationship between x and y is allowed to have a jump at the breakpoint c. For example:

$$y = \begin{cases} \beta_0 + \beta_1 x & \text{if } x \leq c \\ \beta_2 + \beta_3 x & \text{if } x > c \end{cases}$$

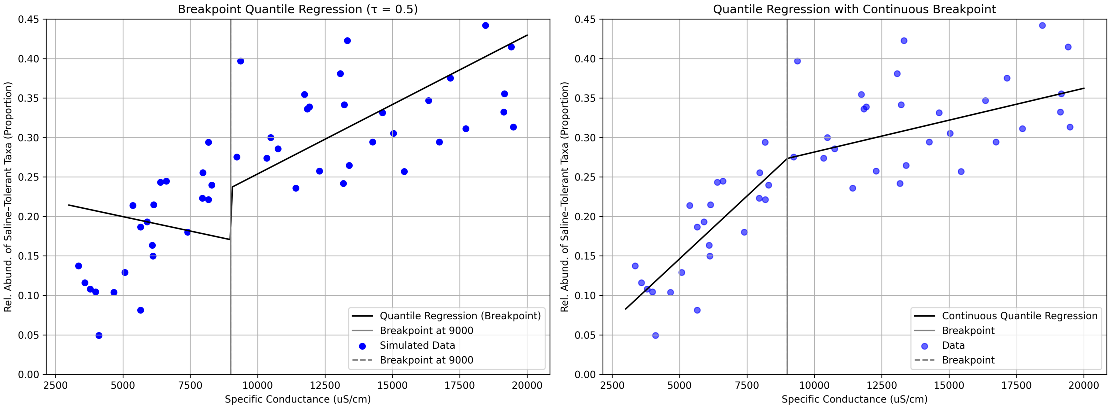


Figure 2: *Example of breakpoint quantile regression*

It looks great, similar to one of the results by Vercruysse's work.

Later, i will try to incorporate this method with Bayesian formula and other methods to try to get a better result.

3.3 PCA by Singular Value Decomposition (SVD)

PCA is a method that was used in both Zhang's and Vercruysse 's work, i am also interested in this method.

These days i am reading the book *Introduction to Linear Algebra, Fifth Edition(Gilbert Strang)* and very interested in the part about singular value decomposition(SVD) and its importance in achieving PCA. Therefore, i tried to implement PCA using SVD by linear algebra with Python.

The SVD of X_c is given by:

$$X_c = U \Sigma V^T$$

where: - U is an $m \times m$ orthogonal matrix, - Σ is an $m \times n$ diagonal matrix with singular values on the diagonal, - V is an $n \times n$ orthogonal matrix.

The transformed data in the reduced space is given by:

$$X_{\text{PCA}} = X_c V_k$$

where V_k contains the top k principal components.

The reason of using SVD is that i want to try to adjust the PCA method and incoporate it with other methods to make it more flexible and specific for the data on hand.

For example, figure 3 shows the PCA result by SVD, which is upon a simulated data. One tentative idea could be to find a way to use less data(samples) to get relatively accurate singular values and vectors, which can be used to save time and effort in collecting data and computation.

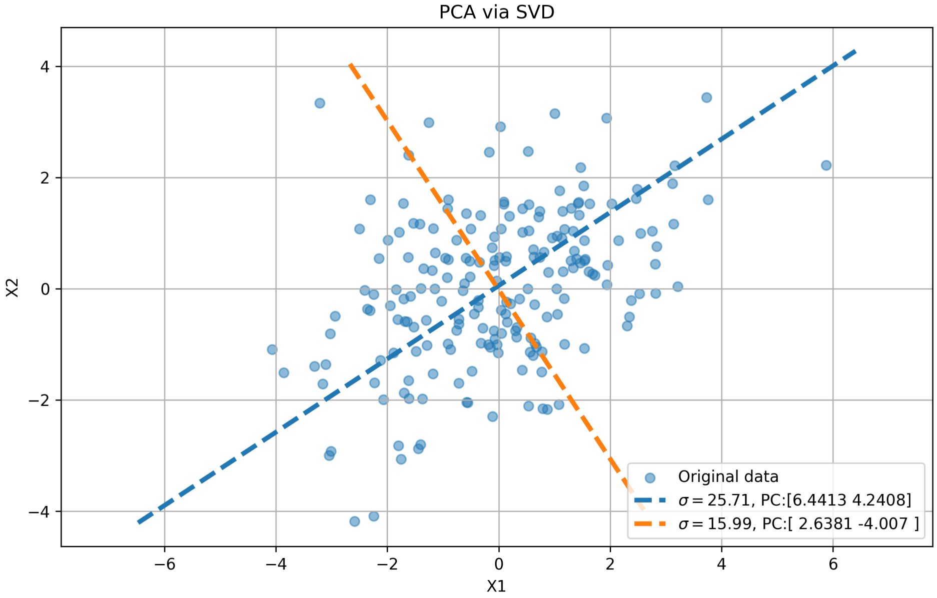


Figure 3: *Example of PCA and SVD on the simulated data*

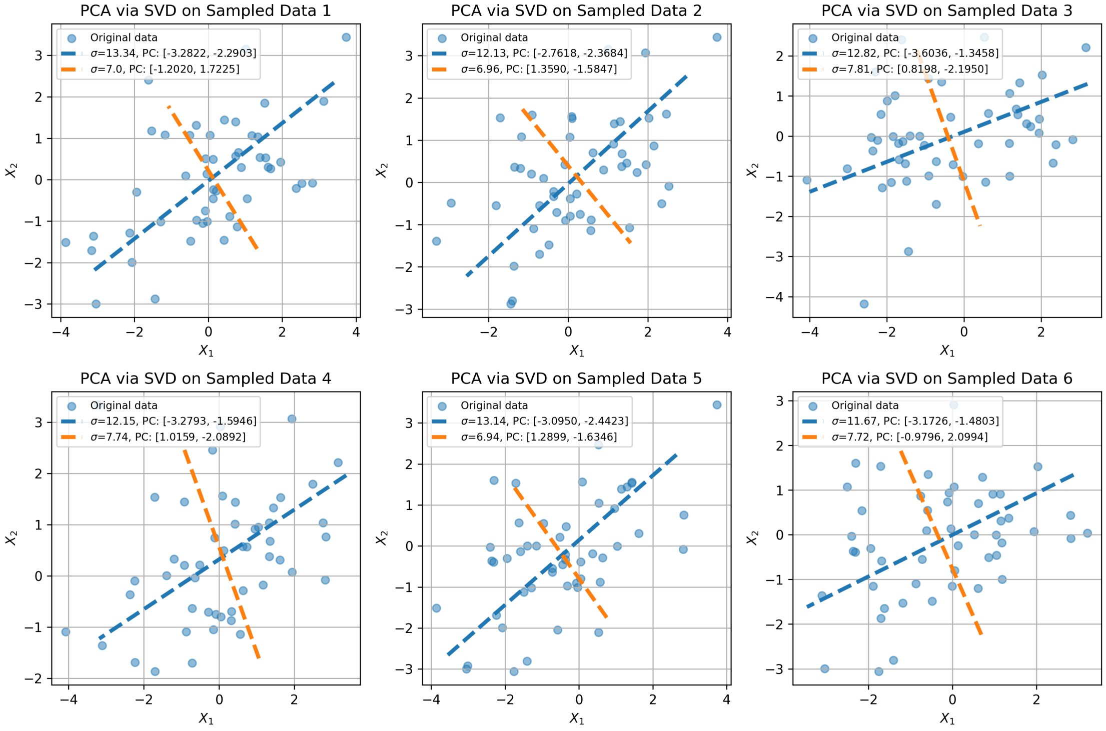


Figure 4: *Example of PCA and SVD on sampled data from the simulated data*

Figure 4 shows the PCA result by SVD, which is upon different sampled sub-data from the original simulated data. The result is not as good as the one in figure 3, which uses the whole data. But this can be an aspect to explore along the work.