Analysis of multiple data sequences with different distributions: defining common principal component axes by ergodic sequence generation and multiple reweighting composition

Ikuo Fukuda and Kei Moritsugu

1 Graduate School of Simulation Studies, University of Hyogo, Kobe 650-0047, Japan
2 Graduate School of Medical Life Science, Yokohama City University, Yokohama 230-0045, Japan

* Author to whom correspondence should be addressed.

E-mail: ifukuda@sim.u-hyogo.ac.jp

Keywords: principal component analysis, multiple data sequences, molecular dynamics, ergodic sampling, reweighting, Boltzmann–Gibbs distribution, statistical analysis

Abstract

Principal component analysis (PCA) defines a reduced space described by PC axes for a given multidimensional-data sequence to capture the variations of the data. In practice, we need multiple data sequences that accurately obey individual probability distributions and for a fair comparison of the sequences we need PC axes that are common for the multiple sequences but properly capture these multiple distributions. For these requirements, we present individual ergodic samplings for these sequences and provide special reweighting for recovering the target distributions.

1. Introduction

Principal component analysis (PCA) is a statistical analysis that defines a framework, i.e., a reduced space determined by the PC axes, for a given multidimensional-data sequence to properly capture the varieties of the data [1]. Our target is a sequence of \( m \)-dimensional \( N_{\text{seq}} \) data, \( \mathcal{X} \equiv \{ x^{11}, \ldots, x^{N_{\text{seq}}} \} \subset \mathbb{R}^m \), generated by a dynamical system or computer simulation such as molecular dynamics (MD) or Monte Carlo (MC) [2].

Our purpose is, first, to generate a data sequence that enables to completely describe a specific probability distribution \( P \), which determines the variety of the data. The Boltzmann–Gibbs (BG) (or canonical) distribution, for example, is applicable useful as the physicochemical distribution since it enables realistic comparisons with experiments performed at constant temperature [3].

Our second purpose concerns with two or more given data sequences, say, \( \mathcal{Y} \equiv \{ y^{11}, \ldots, y^{N_{\text{seq}}} \} \subset \mathbb{R}^m \) in addition to \( \mathcal{X} \), and to constitute PC axes for a joint system described by \( \mathcal{X} \cup \mathcal{Y} \subset \mathbb{R}^m \). Namely, instead of seeking independently the PC axis for \( \mathcal{X} \) and that for \( \mathcal{Y} \), we seek a common set of PC axes for both sequences \( \mathcal{X} \) and \( \mathcal{Y} \), enabling us to fairly compare them in a unified framework. Such a requirement occurs, for example, in a computational study for biomolecular research [4]. Suppose two molecules that share common atomic components, for which we should compare 3D structures of the two common components. Specifically, \( x^{(v)} \in \mathbb{R}^m \) is coordinates of \( m/3 \) C\(_{\alpha}\)-atoms of a protein \( X \), and \( X = \{ x^{11}, \ldots, x^{N_{\text{seq}}} \} \subset \mathbb{R}^m \) is a trajectory sequence for these coordinates. We are also interested in other protein \( Y \) that has the same C\(_{\alpha}\)-atom group, describing \( y^{(v)} \in \mathbb{R}^m \) with yielding \( \mathcal{Y} = \{ y^{11}, \ldots, y^{N_{\text{seq}}} \} \), but has mutations in side chains [5]. To measure this effect on the C\(_{\alpha}\)-atom structures, comparison between the data sequences \( \mathcal{X} \) and \( \mathcal{Y} \) is required. In these situations, we should compare them on a common PC space defined by the common PC axes. This is because the comparison by two individual PC axes, i.e., PC axes for \( \mathcal{X} \) and other PC axes for \( \mathcal{Y} \), may lead incorrect interpretation such that different behavior between \( \mathcal{X} \) and \( \mathcal{Y} \) may be observed as similar behavior for example.

Our purpose is thus (i) to generate two (or more if needed) sequences \( \mathcal{X} \) and \( \mathcal{Y} \) that can accurately reproduce distributions \( P \) and \( P' \), respectively; and (ii) to seek for \( \mathcal{X} \cup \mathcal{Y} \) unique PC axes that duly capture the individual varieties for \( \mathcal{X} \) and those for \( \mathcal{Y} \). MD or MC protocol has been usually used for a practical purpose to generate the
BG distribution, whereas the accurate BG distribution is not generated in general due to broken ergodicity or sampling insufficiency. We also seek a desired distribution $P$, not limited to the BG distribution. These problems can be overcome by an enhanced sampling method that can generate a modified distribution $\tilde{P}$ to enhance the ergodicity [6], with help of reweighting to reproduce $P$. Although this will be a solution to (i), it is far from a solution to (ii). This is because we need two different reweightings for $P$ and $P'$, which will not be easily compatible with the notion of the composition of the two data sequences. We present a scheme to satisfy both (i) and (ii) with providing common PC axes. Furthermore, we introduce a scheme to seek distributions $P$ and $P'$ on the resultant common PCA space.

2. Basics

We give a probable setting of the data, as encountered in MD, and treat only two data sequences $X$ and $Y$ to simplify the notations.

2.1. Data sequences

Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ represent coordinates of a given physical system (‘system 1’) with $n_1$ degrees of freedom, and

$$\{x(\nu \Delta t) \in \mathbb{R}^n | \nu = 1, \ldots, N_{\text{step}}\}$$

a coordinate sequence, from time $\Delta t$ to $N_{\text{step}} \Delta t$, generated for this system. Instead of all $n_1$ coordinates $x(\nu \Delta t) = \{x_1(\nu \Delta t), \ldots, x_n(\nu \Delta t)\}$, our interest is in its $m$ parts, $\{x_{k_1}(\nu \Delta t), \ldots, x_{k_m}(\nu \Delta t)\} = \pi(x(\nu \Delta t)) \in \mathbb{R}^m$ for every time $\nu \Delta t$. Here we denote a projection map for $x \in \mathbb{R}^n$ by

$$\pi(x) = (\pi_i(x))_{i=1,\ldots,m} = (x_{k_i})_{i=1,\ldots,m} \in \mathbb{R}^m.$$  

We thus describe each member in $X = \{x^{(1)}, \ldots, x^{[N_{\text{step}}]}\} \subset \mathbb{R}^m$ using a component index $i$ and time index $\nu$ such that

$$x_i^{(\nu)} \equiv x_{k_i}(\nu \Delta t) = \pi_i(x(\nu \Delta t)) \in \mathbb{R}, \quad i = 1, \ldots, m.$$  

We also consider $\{y(\nu \Delta t) \in \mathbb{R}^n | \nu = 1, \ldots, N'_{\text{step}}\}$, a sequence of coordinates of $n_2$ degrees of freedom, generated by other physical system (‘system 2’), and are interested in $m$ parts $\{y_{l_1}(\nu \Delta t), \ldots, y_{l_m}(\nu \Delta t)\} = \pi'(y(\nu \Delta t)) \in \mathbb{R}^m$ (where $\pi: \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $\pi': \mathbb{R}^n \rightarrow \mathbb{R}^m$ are projections into the same dimensional space $\mathbb{R}^m$),

$$y_i^{(\nu)} \equiv y_{l_i}(\nu \Delta t) = \pi'_i(y(\nu \Delta t)) \in \mathbb{R}, \quad i = 1, \ldots, m,$$

giving a sequence $Y \equiv \{y^{(1)}, \ldots, y^{[N'_{\text{step}}]}\} \subset \mathbb{R}^m$.

2.2. PCA: review

PCA defines a linear map from the target data space $\mathbb{R}^m$ into a reduced space, $\varphi: \mathbb{R}^m \rightarrow \mathbb{R}^l$, where $l \equiv \dim \varphi(\mathbb{R}^m)$ is less than $m$ (typically 2 or 3) [11]. For system 1, this map is designed to capture the variety of data sequence $X = \{x^{(1)}, \ldots, x^{[N_{\text{step}}]}\} \subset \mathbb{R}^m$ and represent them on the reduced space $\mathbb{R}^l$. The map $\varphi$ can be constructed via the $m \times m$ symmetric covariance matrix

$$S = \frac{1}{N_{\text{step}}} \sum_{\nu=1}^{N_{\text{step}}} (x^{(\nu)} - \bar{x}) \odot (x^{(\nu)} - \bar{x}) = \frac{1}{N_{\text{step}}} \sum_{\nu=1}^{N_{\text{step}}} (x^{(\nu)} - \bar{x})(x^{(\nu)} - \bar{x}))_{i=1,\ldots,m},$$

where $\bar{x}$ is the average of the data,

$$\mathbb{R}^m \ni x \equiv (x_i)_{i=1,\ldots,m} = \left(\frac{1}{N_{\text{step}}} \sum_{\nu=1}^{N_{\text{step}}} x^{(\nu)}\right)_{i=1,\ldots,m}.$$ 

That is, by seeking $l$ eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_l$ and the corresponding (normalized) eigenvectors $u_1, u_2, \cdots, u_l \in \mathbb{R}^m$ for $S$, the map $\varphi$ is defined by a projection into $\langle u_1, u_2, \cdots, u_l \rangle$, which is isomorphic to $\mathbb{R}^l$, such that

$$\varphi: \mathbb{R}^m \rightarrow \mathbb{R}^l, \quad x \mapsto ((\langle x | u_k \rangle)_{k=1,\ldots,l},$$

where $\langle \cdot | \cdot \rangle$ is the inner product of $\mathbb{R}^m$. Here, $u_1$ is interpreted to indicate the direction to which the variety of the data in $X$ takes the maximum, $u_2$ the second, and so on. Hence the average $\bar{x}$ and matrix $S$ are key quantities to determine the PC axes. Similarly, the average and matrix are defined for sequence $Y \equiv \{y^{(1)}, \ldots, y^{[N'_{\text{step}}]}\}$ for system 2.
3. Method for solution

3.1. Strategy
Suppose that there exist ideal time series for the two systems, i.e.,
\[ \{ \dot{x}(\nu t) \in \mathbb{R}^n | \nu = 1, \ldots, N_{\text{step}}^1 \} \]
that exactly obeys a distribution \( P \) for system 1 and
\[ \{ \dot{y}(\nu t) \in \mathbb{R}^n | \nu = 1, \ldots, N_{\text{step}}^2 \} \]
that exactly obeys a distribution \( P' \) for system 2 (we use \( \dot{\cdot} \) to represent the ideal), wherein \( N_{\text{step}}^1 \) and \( N_{\text{step}}^2 \) should be sufficiently large. This ideal situation will directly satisfy requirement (i) in section 1. Furthermore, requirement (ii) can be fulfilled by constructing a PC map \( \mathbb{R}^m \rightarrow \mathbb{R}^m \) using a simple sum of coordinates for the two systems
\[
\frac{1}{N_{\text{Tot}}^1} \sum_{i=1}^{N_{\text{Tot}}^1} \dot{x}^{[i]} - \tilde{x} \rightleftharpoons \frac{1}{N_{\text{Tot}}^2} \sum_{j=1}^{N_{\text{Tot}}^2} \dot{y}^{[j]} - \tilde{y} \quad \tilde{x}, \tilde{y} \in \mathbb{R}^m
\]
along with a simple sum of the covariance matrices for the two systems
\[
\frac{1}{N_{\text{Tot}}^1} \sum_{i=1}^{N_{\text{Tot}}^1} (\dot{x}^{[i]} - \tilde{x}) \otimes (\dot{x}^{[i]} - \tilde{x}) + \frac{1}{N_{\text{Tot}}^2} \sum_{j=1}^{N_{\text{Tot}}^2} (\dot{y}^{[j]} - \tilde{y}) \otimes (\dot{y}^{[j]} - \tilde{y}) \quad \tilde{T} \in \text{End} \mathbb{R}^m,
\]
where \( N_{\text{Tot}} = N_{\text{step}}^1 + N_{\text{step}}^2 \).

Remark—The sum of the first and the second terms in equation (8) and those in equation (9) are mathematically well defined (because projections \( \pi \) and \( \pi' \) are into \( \mathbb{R}^m \)) and the most natural expressions to represent the composed sequence \( \{ \dot{x}^{[1]}, \ldots, \dot{x}^{[N_{\text{step}}^1]} \} \cup \{ \dot{y}^{[1]}, \ldots, \dot{y}^{[N_{\text{step}}^2]} \} \). These simple sums can be generalized: e.g. for equation (8), it can be replaced into a weighted sum such as \( w_1 \sum_{i=1}^{N_{\text{Tot}}^1} \dot{x}^{[i]} + \sum_{j=1}^{N_{\text{Tot}}^2} \dot{y}^{[j]} \) or a form such as \( g(\sum_{i=1}^{N_{\text{Tot}}^1} \dot{x}^{[i]}, \sum_{j=1}^{N_{\text{Tot}}^2} \dot{y}^{[j]}) \) using a certain function \( g \) (\( m \) can be changed in system 2) if necessary and meaningful in the context of, e.g., chemical or physical sense.

As will be discussed below, however, generation of ideal time series (6) and (7) is nontrivial. Despite this fact, our purpose is to have accurate \( \tilde{x} \) and \( \tilde{T} \), which are described by the ideal time series. Our strategy is to meet this seemingly contradictory demand by deriving quantities that are equivalent to equations (8) and (9).

3.2. Solution to requirement (i): ergodic sequence generation
Generation of ideal time series obeying an arbitrary distribution within a practical \( N_{\text{step}} \) is hard in general. For example, data \( \{ x(\nu t) \in \mathbb{R}^n | \nu = 1, \ldots, N_{\text{step}} \} \) generated by a conventional canonical simulation for system 1 does not accurately obey the BG distribution and often becomes uncontrollable, due to broken ergodicity and/or sampling inefficiency. Thus, we cannot meet requirements (i) and (ii) with a conventional method. Even if an accurate sampling method exists that can directly generate the distribution, its accurate generation is significantly time consuming due to the feature of the distribution, i.e., the exponential damping with respect to the physical system energy.

Hereafter, we assume a distribution with the form of \( P = \rho(x, p) \, dp \, dx \) and set it as the BG distribution, viz, \( \rho(x, p) \propto \rho_{\text{BG}}(x, p ; \beta) \equiv \exp(-\beta E(x, p)) \), considering the utility and a challenge to the faced problem, although \( \rho \) can be generalized. Here, \( p = (p_1, \ldots, p_n) \in \mathbb{R}^n \) and \( x = (x_1, \ldots, x_n) = \{ x^{[1]} = (x_1^{[1]}, \ldots, x_n^{[1]}), \ldots, x^{[N_{\text{Tot}}]} = (x_1^{[N_{\text{Tot}}]}, \ldots, x_n^{[N_{\text{Tot}}]}), \ldots \} \). A similar setting applies to system 2, where \( P' = \rho'(y, q) \, dq \, dy \) \( \propto \rho_{\text{BG}}'(y, q ; \beta') \equiv \exp(-\beta E'(y, q)) \), \( E'(y, q) = \mathbb{U}'(y) + K'(p) \).

In our method, (i) will be fulfilled by an indirect method, where a reweighting technique is used for producing sequences (6) and (7). Now, the ideal time series for a suitably defined distribution, \( \rho_{\text{BG}} \), can be generated by specialized dynamics [6] such as coupled Nosé-Hoover (cNH) equation [7]. The cNH realizes the equality
\[
\dot{A} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau A(x(t), p(t)) \, dt = \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \, dx \, dp \, \rho_{\text{BG}}(x, p)
\]
for any physical quantity \( A : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R} \) and any trajectory for systems 1 under the ergodic condition [7], wherein a delocalized density \( \rho_{\text{BG}} = \int \rho_{\text{BG}}(x, p ; \beta) f(\beta) \, d\beta \) is utilized with a suitable function \( f \) to efficiently cover the target region for \( \rho_{\text{BG}} \) and enhance the phase-space sampling [8]. Equation (10) enables reweighting to the target density \( \rho_{\text{BG}} [7]:\)
Thus, \( \{ x(\nu \Delta t) \in \mathbb{R}^n | \nu = 1, \ldots, N_{\text{step}} \} \) generated by the cNH for system 1 satisfies (i). We explicitly see this, by defining a weight

\[
 w(x, p) := \frac{\rho_{BG}(x, p)/\rho_R(x, p)}{\sum_{\nu=1}^{N_{\text{step}}} (\rho_{BG}/\rho_R)(x(\nu \Delta t), p(\nu \Delta t))},
\]

such that for any function \( B: D \rightarrow \mathbb{R} \),

\[
 \sum_{\nu=1}^{N_{\text{step}}} w(x(\nu \Delta t), p(\nu \Delta t))B(x(\nu \Delta t)) \\
 \approx B\rho_{BG}/\rho_R \rho_{BG}/\rho_R \\
 = \int BdP / \int dP.
\]

For systems 2, a weight

\[
 w'(y, q) := \frac{\rho'_{BG}(y, q)/\rho'_R(y, q)}{\sum_{\nu=1}^{N_{\text{step}}'} (\rho'_{BG}/\rho'_R)(y(\nu \Delta t), q(\nu \Delta t))}
\]

yields

\[
 \sum_{\nu=1}^{N_{\text{step}}'} w'(y(\nu \Delta t), q(\nu \Delta t))B'(y(\nu \Delta t)) \\
 \approx \int B'dP' / \int dP'
\]

for any \( B' \), satisfying (i). Instead of the cNH, a method suffices if it ensures equation (10) for any \( A \) and any trajectory and if the relationship (11) works for the target BG distribution for any systems.

### 3.3. Solution to requirement (ii): multiple reweighting composition

Based on the above results, requirement (ii) can be satisfied as follows. By substituting \( B \equiv \pi_i \) in equation (13) and \( B' \equiv \pi'_j \) in equation (14), we have

\[
 z^{W}_i := \frac{N_{\text{step}}}{N_{\text{Tot}}} \sum_{\nu=1}^{N_{\text{step}}} w(x(\nu \Delta t), p(\nu \Delta t))x_i^{[\nu]} + \frac{N_{\text{step}}'}{N_{\text{Tot}}'} \sum_{\nu=1}^{N_{\text{step}}'} w'(y(\nu \Delta t), q(\nu \Delta t))y_i^{[\nu]} \\
 \approx \frac{N_{\text{step}}}{N_{\text{Tot}}} \int \pi_i(x) dP / \int dP + \frac{N_{\text{step}}'}{N_{\text{Tot}}'} \int \pi'_j(y) dP' / \int dP' \\
 \approx \frac{N_{\text{step}}}{N_{\text{Tot}}} \sum_{\nu=1}^{N_{\text{step}}} \bar{x}_i^{[\nu]} + \frac{N_{\text{step}}'}{N_{\text{Tot}}'} \sum_{\nu=1}^{N_{\text{step}}'} \bar{x}_j^{[\nu]} \\
 = \bar{z}_i \text{ for } i = 1, \ldots, m,
\]

where the third line comes from the fact that the ideal time series (6) and (7) obey the distributions \( P \) and \( P' \), respectively. Thus, one of the target quantities, equation (8), is obtained by calculating \( z^{W}_i \). For the other quantity, we have
using \( \bar{z}_i \simeq z^{W}_j \) \( (i = 1, \ldots, m) \) concluded in equation (15). Therefore, these procedures for obtaining \( \bar{x} \) and \( \bar{T} \) by calculating \( z^W \) and \( T^W \) ensure the satisfaction of (ii).

### 3.4. BG distribution on the PCA space

Hence, we have a PCA space \( \mathbb{R}^d \) defined by map (5), \( \phi \equiv \phi^W \), constructed from the covariance matrix \( T^W \). The BG distribution on PCA space \( \mathbb{R}^d \) for system 1 (similarly for system 2) is formulated as an induced probability measure for \( P \) via a map \( \phi: D \times \mathbb{R}^n \rightarrow \mathbb{R}^d \), \( (x, p) \mapsto \phi(x, p) \), where \( \pi \) is projection (2). That is,

\[
P_{\phi}\colon B^d \supset B \mapsto d \int_{\phi^{-1}(B) \times \mathbb{R}^n} \rho_{BG}(x, p)\, dx\, dp.
\]

Thus, \( P_{\phi}(B) \) can be evaluated by counting weight (12) if the \( I \) PC-coordinates of \( x^{[i]} \) fall into bin \( B \):

\[
\int \sum_{(i^{[i]}_{[1]}, \ldots, i^{[i]}_{[n]}) \in B} w(x(\nu\Delta t), p(\nu\Delta t))
\]

\[
\simeq \frac{X_B \rho_{BG}/\rho_R/\rho_{BG}/\rho_R}{\int D \times \mathbb{R}^n \rho_{BG}(x, p)\, dx\, dp},
\]

where

\[
\hat{x}_B \colon D \rightarrow \mathbb{R}^d \ni x \mapsto \begin{cases} 1 & \text{if } \phi(x) \in B \\ 0 & \text{otherwise} \end{cases}
\]

### 4. Numerics

To illustrate our method, it has been applied to ‘system 1’ and ‘system 2’ modeled with four degrees of freedom \( (n_1 = n_2 = 4) \) described by potential function \( U(x) = \sum_{i=1}^{10} (x_i^2 - b_i^2)^2 + \sum_{i=1}^{3} (x_i - x_{i+1})^2 \). The difference between the two systems is only in the values of “intra” parameters \( b_1 \) and \( b_2 \) \( (b_1 = 6, b_2 = 1) \) for system 1; \( b_1 = 1, b_2 = 4 \) for system 2; while \( b_3 = b_4 = 0.4, k = 10^{-3} \) for both systems). Figure 1(a) shows plots for the sequence \( \mathcal{X} \cup \mathcal{Y} \subset \mathbb{R}^3 \), where \( \mathcal{X} = \{ x^{[i]} \}_{i=1}^{N_{\text{Seq}}} \subset \mathbb{R}^3 \) was obtained by a cNH simulation (detailed in [7]) of system 1 along with a projection \( \pi: \mathbb{R}^4 \rightarrow \mathbb{R}^3 \), \( (x_0, x_0, x_0, x_0) \mapsto (x_0, x_0, x_0) \), and \( \mathcal{Y} = \{ y^{[i]} \}_{i=1}^{N_{\text{Seq}}} \) was that for system 2 with \( \pi' = \pi \) (viz, \( m = 3 \)). The accuracies were evaluated by marginal distributions of the reweighted BG distributions, where the errors from the exact values in 2-dim distributions for major variables \( (x_i, x_0) \) were \( 5.6 \times 10^{-5} \) and \( 1.9 \times 10^{-5} \) (with s.d. \( 5.5 \times 10^{-3} \) and \( 9.4 \times 10^{-3} \)), which are sufficiently small [7], for systems 1 and 2, respectively.

Figures 1(b) and (c) show the current PCA results with \( l = 2 \) for systems 1 and 2, respectively, which were obtained from the unique common PC axes determined by equations (15) and (16) and from reconstruction of
the BG distribution via equation (18). The current method adequately describes the difference between the two systems. This is because the raw data (figure 1(a)) suggest the role conversion between the first and the second degrees of freedom (i.e., system 1 has the largest variations for $x_1$ and the second largest variations for $x_2$, while system 2 has the largest for $y_2$ and the second largest for $y_1$), and because the current PCA results capture the role conversion between the two degrees of freedom via PC1 and PC2, as clearly seen by the difference between figures 1(b) and (c), owing to the fact that PC1 and PC2 axes are common for the two systems.

In contrast, procedures without data jointing, i.e., PCA for system 1 by $\mathcal{X}$ and independent PCA for system 2 by $\mathcal{Y}$, resulted in misleading results, as shown in figures 1(d) and (e), respectively. Namely, these individual PCA results conclude that the two systems are similar. Although such judgment whether PCA results are reasonable or misleading is possible in these simple model systems, it is not for general systems. Thus, the method using independent PCA procedures, which does not satisfy requirement (ii), may lose the important information of the original systems and lead to incorrect conclusions, even if it holds requirement (i). Hence, it is critical to meet requirement (ii), which is to seek for $\mathcal{X}' \cup \mathcal{Y}'$ unique PC axes that duly capture system 1 with distribution $\mathcal{P}$ and system 2 with $\mathcal{P}'$.

Figures 1(f) and (g) show the PCA results utilizing $\mathcal{X}' \cup \mathcal{Y}'$ composed by (not shown) sequences $\mathcal{X}_c$ and $\mathcal{Y}_c$ obtained in conventional canonical MD simulations for systems 1 and 2, respectively. The results show less accuracy due to dissatisfaction of requirement (i) by inefficient sampling with local traps. Thus, (i) is also critical to get the proper information of the systems.

5. Conclusion

As confirmed from the numerical study, a conventional method that satisfies either requirement (i) or (ii) often leads to incorrect conclusion. Both (i) and (ii) are the key to succeed PCA to capture the difference/similarity of two systems. The proposed method simply satisfies both (i) and (ii) by combining ergodic sequence generation and multiple reweighting composition. The current treatment of two data sequences $\mathcal{X}$ and $\mathcal{Y}$ can be easily generalized into multiple data sequences. Furthermore, the data sequence and the distributions, which are herein supposed in use in MD, can be generalized or transformed into other context without any difficulty.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

ORCID iDs

Ikuo Fukuda @ https://orcid.org/0000-0001-8668-7197
References

[1] Wold S, Esbensen K and Geladi P 1987 Chemom. Intell. Lab. Sys. 2 37
[2] Allen M and Tildesley D 2002 Computer Simulation of Liquids (New York: Clarendon)
[3] Hoover W G 1991 Computational Statistical Mechanics (Amsterdam: Elsevier)
[4] Schlick T 2006 Molecular Modeling and Simulation (Berlin: Springer)
[5] Moritsugu K et al 2021 J. Chem. Inf. Model. 61 1921
[6] Fukuda I and Moritsugu K 2015 J. Phys. A 48 455001
[7] Fukuda I and Moritsugu K 2017 J. Phys. A 50 015002
[8] Fukuda I and Moritsugu K 2020 J. Phys. A 53 375004