Decomposition of a set of distributions in extended exponential family form for distinguishing multiple oligo-dimensional marker expression profiles of single-cell populations and visualizing their dynamics

Daigo Okada[1], Ryo Yamada[1*]

1 Unit of Statistical Genetics, Center for Genomic Medicine Graduate School of Medicine, Kyoto University, Kyoto, Japan

* ryamada@genome.med.kyoto-u.ac.jp

Abstract

Single-cell expression analysis is an effective tool for studying the dynamics of cell population profiles. However, the majority of statistical methods are applied to individual profiles and the methods for comparing multiple profiles simultaneously are limited. In this study, we propose a nonparametric statistical method, called Decomposition into Extended Exponential Family (DEEF), that embeds a set of single-cell expression profiles of several markers into a low-dimensional space and identifies the principal distributions that describe their heterogeneity. We demonstrate that DEEF can appropriately decompose and embed sets of theoretical probability distributions. We then apply DEEF to a cytometry dataset to examine the effects of epidermal growth factor stimulation on human breast epithelial cell line. It is shown that DEEF can describe the complex dynamics of cell population profiles using two parameters and visualize them as a trajectory. The two parameters identified the principal patterns of the cell population profile without prior biological assumptions. As a further application, we perform a dimensionality reduction and a time series reconstruction. DEEF can reconstruct the distributions based on the top coordinates, which enables the creation of an artificial dataset based on an actual single-cell expression dataset. Using the coordinate system assigned by DEEF, it is possible to analyze the relationship between the attributes of the distribution sample and the features or shape of the distribution using conventional data mining methods.

Introduction

Single-cell expression analysis is an effective tool for studying the dynamics of cell population profiles [1–3]. Cytometry data, a type of single-cell expression data, quantify the amount of protein marker expression in each of a large number of randomly selected cells. Single-cell RNA sequencing (scRNA-seq) data, another type of single-cell expression data, has recently become popular. This type of data allows comprehensive quantification of the amount of mRNA expression for genome-wide genes in single cells. Such single-cell expression data can be used to quantify or identify specific cell subsets based on the biomarkers. For example, specific lymphocyte subset (e.g. T cell and B cell subset) have been defined by the expression patterns of several cell surface protein markers [4, 5]. When many cells are sampled from a donor and their expression profiles are obtained, the expression data can be

regarded as an observation of an unknown probability distribution of the cells. The expression profile of each cell can be viewed as a sample from a multidimensional distribution, where the number of dimensions is the number of markers.

Several computational methods developed for single-cell data analysis, such as spanning-tree progression analysis for density-normalized events (SPADE), monocle, and Wanderlust, have been used to investigate various phenomena [6–9]. Most of these methods focus on the diversity of multiple cells or the mutual phylogenic relationship among them within a set of cells sampled to identify subtypes of cells or to visualize their heterogeneity. Expression profile analyses such as cytometry and scRNA-seq are applied to many samples, each of which consists of many cells from individual donors. In recent years, demand for a computational method for heterogeneous multiple samples in the form of distribution has been increasing, and actually a method that integrates multiple expression profiles together and to identify subpopulation in data-driven manner was proposed [10]. These expression profiles take the form of multidimensional distributions, which have to be statistically investigated using methods such as clustering, case-control comparison, and chronological pattern analysis. Multiomics studies analyze phenotypes, transcriptomes, and cytometry data from hundreds or thousands of individuals [11–13]. In these studies, the distributions of cytometry profiles should be statistically analyzed with other datasets from different platforms. However, conventional statistical methods do not take distributions as inputs and thus cell population profiles in the form of distributions have to be modified into a suitable form, such as cell subtype fractions, via gating procedures. This modification of flow cytometry distribution data into multi-categorical fractions loses information. Therefore, the method used to convert the density information of a cell population into a form that can be handled by regular statistical procedures is very important.

Computational methods for extracting feature statistics from data in multidimensional distribution form can be classified into two types, namely parametric and nonparametric. A representative parametric method is the Gaussian mixture model [14]. This method is mainly used for the automation of the manual gating of cytometric data, which is of interest in computational cytometry. However, it is know that in many cases, the Gaussian mixture model, along with other parametric approaches such as t-mixture models [14], is too simple to represent the complexity of the distributions of a cell population profile.

Some nonparametric methods for embedding single-cell expression data or other kinds of distribution-type data into a low-dimensional space have been proposed [15–17]. Most of these methods are based on multidimensional scaling (MDS) [18]. MDS-based methods first estimate a population distribution based on samples using a nonparametric probability density estimation method such as the kernel density estimation method or the k-nearest neighbor (kNN) method [19, 20]. Then, the symmetric distance is defined between two distributions based on information theory. Finally, MDS-based methods generate a distance matrix for a set of distributions and embed the individual distributions in a low-dimensional Euclidean coordinate space that maintains these distance relationships as much as possible. Although this approach is simple and powerful for the visualization of samples from different donors, the embedding into Euclidean coordinate space is essentially non-precise and imperfect because the definition of distance based on information theory is non-Euclidean [21].

Information geometry is a field of statistics that deals with the geometry of probability distributions [21]. In this research, based on the idea of information geometry, we propose a method, called Decomposition into Extended Exponential Family (DEEF), for embedding sample distributions into a low-dimensional coordinate

space. The DEEF method finds an exponential-family-like formula for an arbitrary set of distributions and component distributions to describe the set of distributions and gives the coordinates and potential function value for each distribution. The only difference between an extended exponential family (EEF) and the exponential family itself is that the potential function of a regular exponential family is convex whereas that of an EEF is not. DEEF estimates the inner products of distribution pairs and assigns coordinates θ to each distribution based on the eigenvalue decomposition of a matrix related to the inner products. The coordinate system contains imaginary coordinates, as in Minkowski space [22]. The coordinates θ can always recover the distributions without loss of information and in many cases, θ from only a limited number of principal axes can recover the original distributions with negligible residuals. This overcomes the drawbacks of the conventional MDS-based method.

In this paper, we define an EEF and discuss the theoretical aspects of the log-linear decomposition of the probability matrix $\mathbf{P}$ into exponential-like representations. We apply the DEEF method to a set of theoretical probability distributions and show that it can be used for data-driven extraction and the visualization of potential parameter structures of the dataset. We then apply DEEF to a cytometry dataset to examine the effects of epidermal growth factor (EGF) stimulation on human breast epithelial cell line. It is shown that DEEF can extract parameters that identify the principal patterns of the cell population profile and describe the complex dynamics of cell population profiles as a trajectory. In addition, DEEF can be used to perform a dimensionality reduction for this dataset and a time-series reconstruction, which enables the creation of an artificial cytometry dataset based on the properties of the actual data.

Results

Method outline

We propose a statistical method called DEEF (Fig 1). An exponential family is a set of probability distributions whose probability density/mass functions are expressed in the form

$$\log P(x, \theta) = C(x) + \sum_{k=1} F_k(x)\theta_k - \psi(\theta) \tag{1}$$

where $C(x)$, $F_k(x)$, and $\psi(\theta)$ are known functions ($\psi(\theta)$ should be convex), and θ is the parameters that specify distribution instances. Many parametric probability distributions, such as the normal distribution and the binomial distribution, are included in the exponential family. Some probability distributions are not included in the exponential family, such as the mixture normal distribution. The details are given in S1 Text.

The distributions, one dimensional or multidimensional, in life sciences and other field, including expression profiles, are sometimes too complex to fit to simple parametric distribution. Some of them can be adequately described as a mixture of multiple parametric distributions. Actually, mixture of multiple distributions such as a mixture normal distribution or a mixture t distribution is commonly used in the parametric model for cytometry data [11]. And further complicated distributions can be fitted to only non-parametric distribution. Choosing the appropriate parametric model is difficult because it depends on the situation. While the exponential family can represent many simple probability distributions, it cannot represent most mixture distributions or more complex distributions often used in single-cell expression analysis.

We define an EEF as:

$$\log P(x,\theta) = C(x) + \sum_{k=1} F_k(x)\theta_k - \psi'(\theta) \tag{2}$$

$$\psi'(\theta) = \sum_{k=1} h_k \theta_k^2 \quad where \quad h_k = -1 \quad or \quad 1 \tag{3}$$

An EEF is almost identical to Eq 1, but with the potential function $\psi(\theta)$ modified as shown in Eq 3. We loosened the restriction that $\psi(\theta)$ should be convex so that a set of arbitrary distributions can fit the formula. We also modified $\psi(\theta)$ as shown in Eq 3. θ represents the coordinates of each distribution, where the inner product of the θ coordinates between two distributions is defined as half the logarithm of the inner product of density/mass functions. Using this definition of θ, $C(x)$ and $F_k(x)$ are solvable when a set of distributions P(x, θ) is given.

We obtain a set of multidimensional probability distributions from the experimental results. We divide the space into grid cells and estimate the probability mass functions P for the grid cells, which makes the dimensions of Eq 2 and Eq 3 finite and makes the estimation of EEF forms a linear algebraic calculation.

A matrix-operation-based simple algorithm can be constructed for log-linear decomposing probability matrix $\mathbf{P}$ into $\mathbf{C} + \mathbf{\Theta F} - \mathbf{\Psi}$, where $\mathbf{C}$, $\mathbf{\Theta F}$, and $\mathbf{\Psi}$ are the discretized representations of EEF forms for multiple distributions (details given in Method section). Then, we can obtain the EEF representation of any distribution set. The input is only the probability matrix $\mathbf{P}$, whose rows represent the probability mass function. DEEF can be applied to distribution sets to embed each distribution in the defined EEF space by considering $\mathbf{\Theta}$ as the feature statistics of the distributions. Because the θ coordinate is calculated from eigenvalue decomposition, a few coordinates with the top eigenvalues contain a lot of the information of the probability distribution set. In addition, the $\mathbf{F}$ matrix provides principal compositional distributions in the original space. DEEF extracts the compositional distribution F_i to a data driven manner. The θ_i coordinate indicates how much each sample has F_i. This is an interpretation of θ coordinate space, where hold difference between samples. A detailed description of the theory are given in the Appendix in S1 Text.

Simulation data analysis

First, we applied DEEF to a normal distribution set that consisted of 900 instances of a normal distribution, with the mean ranging from -1 to 1 and the standard deviation (sd) ranging from 2 to 4 at a fixed interval of 0.069 for each (Fig 2(a)). We called these parameters defined in the specific parametric models as original parameters. And, a space using these original parameters as coordinate axes is called an original parameter space. We compared the DEEF method and a conventional MDS-based method [15] using this normal distribution set.

We compared the θ coordinate spaces with the top three absolute eigenvalues (θ_{last}, θ_1, θ_2) (Fig 2(b)) and the top three MDS coordinate spaces (MDS1, MDS2, MDS3) (Fig 2(c)). The θ coordinate is denoted θ_i in decreasing order of eigenvalues. θ_{last} is the coordinate corresponding to the lowest eigenvalue, whose absolute value is largest in this case. Although both methods displayed a two-dimensional manifold in three-dimensional space, the two-dimensional manifold for DEEF was much simpler than that for MDS. The colors in Fig 2 indicate the Kullback-Leibler (KL) divergence from the distribution in the center of the mean-sd parametric grid (indicated by a black dot). Because the two-dimensional manifolds of DEEF and MDS were curved surfaces, it was not appropriate to use the Euclidean distance between points as a measure of divergence between two distributions. However, the simpler manifold for

DEEF seems to be intuitively better for visualizing divergence. The number of total extracted coordinates for the MDS-based method was 445 because the decomposed matrix was not positive definite and some information was missing; the number of total extracted coordinates for DEEF was 900.

The normal distribution can usually be characterized by two parameters, mean and sd, on the original parameter space. However, they are also allowed to be expressed in different two parameters. While parameterization by mean and sd is only possible under the assumption that it is a normal distribution, the θ coordinates calculated by DEEF can be assigned to the distribution without any assumptions. In both original parameters and θ coordinates, information about the difference between distributions is represented by the same number of parameters. In fact, when the distributions are generated sufficiently densely, it is visualized in Fig 2 that the topological relation among the distributions is maintained.

We apply DEEF to multiple normal distribution sets with different parameter structures, namely a mixture normal distribution set and an exponential distribution set, in S1 Text. Here, we apply the DEEF method to a set of theoretical probability distributions and show that it can be used for data-driven extraction and the visualization of the potential parameter structures of the dataset. DEEF successfully embedded these distributions in the θ coordinate space. The distributions could be recovered without loss of information and in many cases θ from only a limited number of principal axes could recover the original distributions with negligible residuals.

EGF stimulation cytometry data analysis

Cytometry data can be considered as an unknown multidimensional probability distribution of cells, where the number of dimensions is the number of markers. We applied DEEF to a cytometry dataset.

We used mass cytometry data from a study on the effect of EGF stimulation on an human breast epithelial cell line [23]. In the experiment, measurements were made at 10 time points (0, 0.5, 1, 3, 6, 10, 15, 30, 60, and 120 minutes) in two replicates, one each after EGF stimulation and under control conditions. We picked four marker proteins, namely pAKT, pERK, pPLCγ2, and pS6, which were shown to respond to EGF stimulation in the original study. The pre-processed marker expression data for each time point after EGF stimulation for Replicate1 and Replicate2 are shown in Fig 3. We applied the DEEF method to the four marker single-cell expression datasets. Unlike for the simulation data, the population distribution was unknown and thus a sample set was obtained. Then, we estimated the probability matrix $\mathbf{P}$ of the single-cell expression dataset before we applied DEEF, as described below. Each single-cell expression dataset was a sample set from an unknown population distribution in the number-of-markers-dimensional space (four-dimensional space in this case). First, we decided the range of each marker. For each sample, we calculated the α percentile and the 1 - α percentile for each marker expression. We used the range of each marker between the minimum α percentile value and the maximum 1 - α percentile value among all samples so that all samples contained the expression range between the α and 1 - α percentiles for cells. In this case, we used $\alpha = 0.05$. Next, we separated this range into equally spaced m points (m=20), where m is a defined parameter. The number of grids was m^4. For the determined grids, we estimated the probability density using the kNN method (k=800). The row vector P, representing the kNN-based densities of m^4 grids, was standardized so that its total value was 1. We applied the DEEF method to $\mathbf{P}$ built using the above procedure and calculated the corresponding θ coordinates. θ_{last} corresponded to a negative eigenvalue, and θ_1, θ_2, and θ_3 corresponded to positive eigenvalues (S1 Fig). The boxplot of error shows that the performance of the distribution reproduction increases with increasing number of θ

coordinates but at a slower rate than that for the simulation distribution set (S1 Fig).

We embedded all cell population profiles into a low-dimensional coordinate space and visualized them using the DEEF method. θ_1, θ_2, and θ_3 accounted for 69.6%, 13.9%, and 8.9% of the sum of positive eigenvalues, respectively. Fig 4(a) shows scatter plots of the top positive θ coordinates derived from the DEEF method. θ_1 and θ_2 give common trajectories during EGF stimulation between Replicate1 and Replicate2 but θ_3 gives a different trajectory. After EGF stimulation, the cell population profile moved on the θ_1 and θ_2 coordinate space and then returned to the region near the baseline. We then used θ_1 and θ_2 to parameterize the cell population dynamics after EGF stimulation which is common between Replicates1 and Replicate2.

F_1 and F_2, which correspond to θ_1 and θ_2, respectively, show the type of cell population profile change represented by the trajectory. Fig 4(b) shows F_1 and F_2 for pAKT and pS6. F_1 explains the number of cells with high pAKT expression and high pS6 expression and F_2 explains the number of cells with low pAKT and high pS6 expression. An increase in θ_1 and a decrease in θ_2 correspond to the initial response. This change can be well expressed as a synthesis of the patterns of the three underlying cell population profiles. The increase in θ_2 that occurs in the second half corresponds to the increase in pS6, which arose later than that of pAKT. The density plots of F_1 and F_2 for all four markers are shown in S2 Fig.

S3 Fig shows a scatter plot of all samples for MDS1 and MDS2 derived by applying the MDS-based method to this dataset. The dynamics after EGF stimulation have a trajectory pattern similar to that obtained with DEEF. However, we cannot get further information from this analysis.

To visualize F(x) as a four-dimensional function all at once, we performed SPADE analysis and described F_1 and F_2 on the SPADE tree. SPADE is a computational cytometry method that automatically clusters cells for multiple cytometry datasets and creates one consensus tree of the cell clusters. We applied SPADE to all 40 samples to create a SPADE tree that consisted of ten cell clusters (Fig 5(a)). Each SPADE cluster can be characterized by the four-marker expression pattern (Fig 5(b)). Fig 5(c) shows SPADE trees with F_1 and F_2 values. Each cluster was assigned F_1 and F_2 values of the grid to which the representative location of the cluster belongs. In F_1 on the SPADE tree, Cluster 9 has the highest positive F_1 values. This result is reasonable because Cluster 9 showed high expression for all four markers. This result corresponds to the fact that all marker expressions increase after EGF stimulation. Cluster 8 has the highest negative F_1 value, which is reasonable because this cluster showed low expression for all four markers. F_2, which corresponds to a different trajectory pattern from that for F_1, shows a different pattern on the SPADE tree. Cluster 3, which has the highest positive F_2 values, showed high expression for pS6 and pPLCγ2. These two markers are expressed later than pAKT and pERK. Interestingly, Cluster 2, which showed low expression for pERK and pS6, has the highest negative F_2 value. Using the table of the representative values for each cluster (S1 Table), this subset can be confirmed on the density plot of samples obtained 6 minutes after stimulation (Fig 5(d)). The DEEF method can provide insight into patterns that are difficult to detect using conventional methods.

Dimension reduction and time-course reconstruction using EGF stimulation dataset

In the previous section, we showed that DEEF works well with a real cytometry dataset. In this section, as further applications of DEEF for biological research, we describe dimensionality reduction and time-course reconstruction.

DEEF can reconstruct a distribution using only the coordinates with the top

absolute eigenvalues. To reduce the dimensionality of a cell population profile, we expressed the cell population profile using only the synthetic sum of the main patterns; other differences were considered to be noise. A dimension reduction of the EGF stimulation dataset using the top θ coordinates was conducted. The panels in the first column of Fig 6(a) shows the change in the median marker intensity in the raw data along the time course for the four markers. The expression levels of pAKT and pERK increased first, followed by those of pS6 and pPLCγ2. This is consistent with the results in the original study. The panels in the second column of Fig 6 show the change in the median of marker intensity calculated from the reconstructed distribution using θ_1, θ_2, and θ_{last}, corresponding to top three highest absolute eigenvalues (K=3). These results suggest that the cell population profile reconstructed using only the main patterns well captures the characteristics of the dynamics of the original data. Here, the patterns that have a small contribution to the difference among the sample set were eliminated. If DEEF can decompose the information into meaningful data and noise, reproduction using only principal functions would denoise the data.

Next, using this scheme, we conducted a time-course reconstruction of Replicate1's EGF stimulation dataset whose original time course contained 10 time points. The value of the θ coordinate at each time point was estimated by linearly interpolating and dividing the value of the θ coordinate between each time point into 10 equal parts, and reconstructing the θ coordinate at a total of 91 images. Fig 6(b) shows the 25th and 65th images of the 91 images as examples of the estimated cell population profiles between the measurements. Based on the estimated value, the distribution was reproduced at K=3. An animation of the cell population dynamics including the unmeasured time points is available (S1 Movie).

Discussion

In this study, we proposed a class of probability distributions called EEFs and a nonparametric decomposition method for probability distribution sets called DEEF (Fig 1). The DEEF method provides geometric coordinates for each distribution and obtains feature statistics for a sample set by estimating an exponential family-like representation for a multidimensional probability distribution set. DEEF can identify the parameters that well discriminate the difference among a distribution set as θ. In addition, the coordinates identified by DEEF have a biological meaning, as shown by $F_i(x)$. The log-linear decomposition did not lose the information in the original datasets and the original distributions could be reproduced. The DEEF method extracted the feature statistics of distributions as θ coordinates without loss of information, unlike similar methods such as the MDS-based method (Fig 2, S1 Text).

When the DEEF method was applied to a cytometry dataset obtained after EGF stimulation, as shown in Fig 3, it extracted the main underlying patterns from the probability distribution set, embedded them into the coordinate system, and indicated the quantitative differences among samples (Fig 4). We parameterized the dynamics after the EGF stimulation with two parameters and expressed them as trajectories. We could then visualize the $F(x)$ function on the SPADE tree (Fig 5). By using SPADE, information on the combination of multidimensional markers can be simultaneously visualized; this is not possible with a two-marker density plot. The characteristics of the response to EGF are useful for characterizing a subset of human mammary cells and are essential information for understanding the properties of epithelial cancers [23]. DEEF may provide new insights into such characteristics with consideration of not only the change of a single marker but also a combination of multiple markers.

As a further application of DEEF, we performed a dimension reduction and a

reconstruction of cell population profiles using highly contributing coordinates (Fig 6, S1 Movie). This method is considered to be effective for complementing cytometry data acquired along the time course. When cytometry data have an ordered structure such as a time series, complementary estimation of the state between measurements can be performed. In addition, DEEF can easily create an artificial dataset with a large sample size that conforms to the properties of the real data. This is useful in computational biology research.

In this study, cell population profiles were embedded into a low-dimensional space by applying the DEEF method to flow cytometry data. By treating the values of θ coordinates as a trait and performing an association analysis with genotype and transcriptome data, DEEF can identify genes and pathways related to the entire cell population profile and their dynamics. Multiomics analysis, which combines various types of large-scale omics data such as genomes, transcriptomes, and metabolomes, is widely used in various fields to study complex life systems [24–26]. Our research will make it easier to add single-cell data to multiomics analysis. In many biological fields, such as immunology and stem cell biology, the behavior of a whole cell population profile is very important for elucidating life phenomena. This behavior can be very complicated. A combination of the proposed method and omics analysis is expected to advance the understanding of these complex biological phenomena.

In recent years, high-dimensional single cell expression data such as scRNA-seq or CyTOF has become popular. Computational methods for such high-dimensional single cell expression data are also being actively developed [27]. On the other hand, DEEF is not suitable for handling genome-wide gene expression because the number of grids grows exponentially with the dimensionality and kNN estimation and the linear algebraic algorithm can't work well. However, by the novel theory and algorithm, DEEF provides high-resolution analysis for sample heterogeneity where the calculated coordinates and the original marker expression pattern are completely associated by F(x) function. In many case, cellular subsets, such as lymphocyte subset, have been defined by the expression patterns of several markers. From this perspective, DEEF are expected to provide a novel insight on the analysis of cell population profiles. Then, it is necessary to select only a few important markers for high-dimensional CyTOF and scRNA-seq data. Although choosing irrelevant markers would theoretically not have much effect on the results because DEEF treats each grid as independent, it would waste computational resources. One potential solution might be the combination of DEEF with dimension reduction method, such as t-SNE and Uniform Manifold Approximation and Projection (UMAP) [28], although it seems necessary to study the effect of the non-linear embedding on the DEEF ' s decomposition logic. Further investigations would be beneficial to overcome this drawback.

Several other improvements can be considered for the DEEF method. In its present form, DEEF handles grids independently; it does not consider the positional relationships among neighboring grid cells. Taking these relationships into account would make the functions C and F smoother, which may remove random errors and improve machine learning accuracy and the interpretability of results. Another possible improvement is the use of the kernel method to estimate $\mathbf{P}$ from raw data. In the present procedure, DEEF calculates the inner products between distributions discretely using kNN density estimation. This step could be improved by embedding the dataset into a reproducing kernel Hilbert space with infinite dimensions directly using the kernel method [29]. The introduction of the kernel method into DEEF might improve performance.

Conclusion

In this study, we developed a method called DEEF to analyze differences between cell population profiles using single-cell expression data. DEEF performs a log-linear decomposition of the probability matrix $\mathbf{P}$ to embed the distributions into a low-dimensional space. The DEEF method can extract the potential parameters of the probability distribution set and describe the meaning of the estimated parameters. Because single-cell expression data can be regarded as samples from an unknown population distribution, we can investigate the difference among cell population profile sets. DEEF can be used to examine and visualize the difference among single-cell expression datasets. DEEF can reconstruct the distributions from the top coordinates, which enables the creation of artificial datasets based on an actual single-cell expression dataset. Using the coordinate system assigned by DEEF, it is possible to analyze the relationship between the attributes of the distribution samples and the features or shape of the distribution using conventional data mining methods.

Method

1. DEEF method

First, we describe the theoretical basis of DEEF. An exponential family is a set of probability distributions whose probability density/mass functions are expressed in the form:

$$\log P(x,\theta) = C(x) + \sum_{k=1} F_k(x)\theta_k - \psi(\theta) \tag{4}$$

where $C(x)$, $F_k(x)$, and $\psi(\theta)$ are known functions ($\psi(\theta)$ should be convex) and θ is the parameters that specify distribution instances. Many parametric probability distributions, such as the normal distribution and the binomial distribution, are included in the exponential family. Some probability distributions are not included in the exponential family, such as the mixture normal distribution. We define an EEF as:

$$\log P(x,\theta) = C(x) + \sum_{k=1} F_k(x)\theta_k - \psi'(\theta) \tag{5}$$

$$\psi'(\theta) = \sum_{k=1} h_k\theta_k^2 \quad where \quad h_k = -1 \quad or \quad 1 \tag{6}$$

where an EEF is almost identical to Eq 4, but with the potential function $\psi(\theta)$ modified as shown in Eq 5. We loosened the restriction that $\psi(\theta)$ should be convex so that a set of arbitrary distributions can fit the formula. We also modified $\psi(\theta)$ as shown in Eq 6. $\psi'(\theta)$ does not become a convex function unless h_k is all 1. Therefore, an EEF can be defined as a probability distribution family that conditionally excludes rules on the convexity of the potential function from the definition of an exponential family.

Regardless of whether the potential function is convex or not, the functional inner product between exponentially expressed functions $P(x)$ and $Q(x)$ can be expressed as follows using only θ coordinates and the potential function (proof is shown in S1 Text, Appendix Theorem 1).

$$< P(x,\boldsymbol{\theta}^P), Q(x,\boldsymbol{\theta}^Q) >= \frac{e^{\psi(\boldsymbol{\theta}^P+\boldsymbol{\theta}^Q)}}{e^{\psi(\boldsymbol{\theta}^P)}e^{\psi(\boldsymbol{\theta}^Q)}} \tag{7}$$

If $P(x)$ and $Q(x)$ are both EEFs, the following simple relationship between $P(x)$ and $Q(x)$ is satisfied for their functional inner product and θ coordinates (proof is shown in S1 Text, Appendix Theorem 2).

$$\frac{1}{2}\log < P(x,\boldsymbol{\theta}^P), Q(x,\boldsymbol{\theta}^Q) > = \sum_{k=1} h_k \theta_k^P \theta_k^Q \tag{8}$$

Consider an $n \times n$ matrix $\mathbf{M}$, whose (i, j)-th element $m_{i,j}$ is identified as $\frac{1}{2}\log q_{i,j}$ where $q_{i,j}$ is the functional inner product between i-th and j-th distributions. Let the i-th eigenvalue of $\mathbf{M}$ be λ_i. Then, $\mathbf{M}$ can be represented by eigenvalue decomposition as follows:

$$\mathbf{M} = \mathbf{V}^T \boldsymbol{\Lambda} \mathbf{V} \tag{9}$$

where the i-th column of $\mathbf{V}$ represents the i-th eigenvectors of $\mathbf{M}$ and $\boldsymbol{\Lambda}$ is a diagonal matrix whose i-th diagonal elements are λ_i. Note that the eigenvalues of $\mathbf{M}$ contain negative values. Then, $\mathbf{M} = \mathbf{V}^T \boldsymbol{\Lambda}' \mathbf{S} \mathbf{V} = (\mathbf{V}\sqrt{\boldsymbol{\Lambda}'})^T \mathbf{S} (\mathbf{V}\sqrt{\boldsymbol{\Lambda}'})$, where $\mathbf{S}$, $\boldsymbol{\Lambda}'$ and $\sqrt{\boldsymbol{\Lambda}'}$ are $n \times n$ diagonal matrices whose i-th diagonal elements are $sign(\lambda_i)$, $|\lambda_i|$, and $\sqrt{|\lambda_i|}$, respectively. Therefore, when we take the θ coordinate matrix $\boldsymbol{\Theta}$ and h_i as follows, Eq 4 is completely satisfied.

$$\boldsymbol{\Theta} = \mathbf{V}\sqrt{\boldsymbol{\Lambda}'} \tag{10}$$

$$h_i = sign(\lambda_i) \tag{11}$$

where $\boldsymbol{\Theta}$ is the θ coordinate matrix whose (i,j)-th element represents the j-th coordinate value of the i-th distribution in the EEF expression. Because $\mathbf{M} = \boldsymbol{\Theta}^T \mathbf{S} \boldsymbol{\Theta}$, Eq 4 is completely satisfied.

The next step is the calculation of C(x) and $F_i(x)$. To treat this calculation discretely using a computer, the above expression must be expressed in matrix form as:

$$\mathbf{P}^{log} = \mathbf{C} + \boldsymbol{\Theta}\mathbf{F} - \boldsymbol{\Psi} \tag{12}$$

where $\mathbf{P}^{log}$ is an $n \times m$ matrix that represents a log-discretized probability mass function of m grids of n samples, $\mathbf{C}$ is an $n \times m$ matrix that corresponds to $C(x)$ and all of whose rows have the vector $\mathbf{c}$, $\boldsymbol{\Theta}$ is the $n \times n$ matrix obtained previously, $\mathbf{F}$ is an n × m matrix whose row vector corresponds to discretized $F_i(x)$, and $\boldsymbol{\Psi}$ is an $n \times m$ matrix whose column vector is the previously obtained $\sum_{k=1} h_k \theta_k^2 \mathbf{1}$. Then, this equation is rewritten as:

$$\mathbf{P}' = \boldsymbol{\Theta}'\mathbf{F}' \tag{13}$$

where $\mathbf{P}' = \mathbf{P}^{\mathbf{log}} + \boldsymbol{\Psi}$, $\mathbf{F}'$ is $[\mathbf{F}^T, \mathbf{c}]^T$, and $\boldsymbol{\Theta}'$ is $[\boldsymbol{\Theta}, \mathbf{1}]$. Therefore, $\mathbf{F}'$ can be obtained using the Moore-Penrose pseudo-inverse matrix $Ginv(\boldsymbol{\Theta}')$ as follows:

$$\mathbf{F}' = Ginv(\boldsymbol{\Theta}')\mathbf{P}' \tag{14}$$

Because $\mathbf{F}'$ is defined as $[\mathbf{F}^T, \mathbf{c}]^T$, all items necessary for the EEF expression of the distribution set can be obtained.

Based on the above theory, it is possible to construct a simple matrix-operation-based algorithm for decomposing probability matrix $\mathbf{P}$ to obtain the EEF representation of any distribution set. The input is probability matrix $\mathbf{P}$, whose rows represent the probability mass function. The first step is calculating matrix $\mathbf{M}$ from $\mathbf{P}$. The second step is the eigenvalue decomposition of $\mathbf{M}$. h_i are obtained to determine $\psi'(\theta)$ and an n sample × n coordinate matrix $\boldsymbol{\Theta}$ is obtained to embed all

samples. The third step is calculating $\mathbf{c}$ and $\mathbf{F}$ to determine all components of the EEF expression. The simulation data analysis method is described in S1 Text.

This method can be applied to distribution sets to embed each distribution in the defined EEF space by considering $\mathbf{\Theta}$ as the feature statistics of the distributions. Because the θ coordinate is calculated from eigenvalue decomposition, a few coordinates with the top eigenvalues have a lot of the information of the probability distribution set. In addition, the $\mathbf{F}$ matrix provides principal compositional distributions in the original space. The R package "deef" is available on GitHub (https://github.com/DaigoOkada/deef).

2. Distribution reproduction and performance evaluation

In DEEF, the distribution can be reproduced using any number of coordinates when C, F_i, θ_i, and h_i are obtained. We reproduced the distribution by reconstructing the probability mass function calculated by normalizing $exp(C(x) + \sum_{i=1}^{K} F_i(x)\theta_i - \psi(\theta)$, where the coordinates with the top K absolute eigenvalues were selected. In this study, performance was evaluated by Performance Index (PI) defined by the sum of the squared error between the true probability mass function and the reconstructed probability mass function. This value was calculated for each distribution included in the distribution set. A smaller squared error indicates better reproduction. In particular, if this value is zero, the original distribution and the reconstructed distribution are exactly the same.

3. Conventional MDS-based method

We embedded the distribution set using an MDS-based method using the following procedure. First, we calculated the distance matrix among samples. The distance between two distributions p_i and p_j is defined as $\frac{1}{2}(KL(p_i||p_j) + KL(p_j||p_i))$, and the coordinate values of each sample are calculated by applying MDS to the generated distance matrix. MDS was applied to this distance matrix to calculate the MDS coordinates of each sample. The coordinates are denoted MDS1, MDS2 and MDS3 in descending order of their eigenvalues.

4. Application of DEEF method to normal distribution set

We applied DEEF to a normal distribution set that consisted of 900 instances of a normal distribution, with the mean ranging from -1 to 1 and sd ranging from 2 to 4 at a fixed interval of 0.069 for each. The θ coordinate values and MDS were calculated using the theoretical value of the functional inner product or KL divergence defined by the mean and sd.

As the notation to distinguish the original parameter and θ coordinates, we named the original parameters using the alphabetic name used in the original parametric model. For example, in the case of normal distribution set, the original parameter is named as "mean" and "sd". On the other hand, θ coordinates are always named as θ_i using the Greek letter θ and the suffix number i.

5. Construction of probability matrix P from single-cell expression dataset

Unlike for the simulation data, the population distribution was unknown and thus a sample set was obtained. We estimated the probability matrix $\mathbf{P}$ of the single-cell expression dataset before we applied DEEF, as described below. Each single-cell expression dataset was a sample set from an unknown population distribution in

d-dimensional space, where d is the number of markers of the samples. First, we decided the range of each marker. For each sample, we calculated the α percentile and 1 - α percentile of each marker expression. We used the range of each marker between the minimum α percentile value and maximum 1 - α percentile value among all samples so that all samples contained the expression range between the α and 1 - α percentiles for cells. Next, we separated this range into equally spaced m points, where m is a defined parameter. The number of grids is m^d. For the determined grids, we estimated the probability density using the kNN method. The row vector P, representing the kNN-based densities of m^d grids, was standardized so that the sum of the vector was 1.

6. Application of DEEF method to EGF stimulation data

We used mass cytometry data from research on the effect of EGF stimulation on human breast epithelial cell line [23]. The data were obtained from the Flow Repository (ID: FR-FCM-ZYBC). In the experiments, measurements were made at 10 time points (0, 0.5, 1, 3, 6, 10, 15, 30, 60, and 120 minutes) in two replicates after EGF stimulation and control conditions, respectively. We picked four marker proteins, namely pAKT, pERK, pS6, and pPLCγ2, which were shown to respond to EGF stimulation in the original study. As preprocessing, the marker expression levels were converted using asinh (intensity/5), as done in the original study. The number of cells in this dataset was between 8,089 and 22,221. We constructed probability matrix $\mathbf{P}$ from the cytometry data. Each cell could be taken as a sample from the population distribution. The hyperparameters for constructing $\mathbf{P}$ were m = 20, $\alpha = 0.05$, and k = 800. Next, the DEEF method was applied to estimate $\mathbf{P}$. The coordinates are denoted θ_1, $\theta_2 \cdots \theta_{last}$ in descending order of their eigenvalues.

7. Visualization of F function with density plot and SPADE

We expressed F_i as a compositional distribution by standardizing $exp(F_i)$ so that its total value was 1. Then, from this distribution, we sampled 10,000 data points and drew the density plot using the matplotlib Python library.

To visualize the multimarker information simultaneously, we applied the SPADE algorithm to the EGF stimulation data [6]. The number of clusters was 10 and other hyperparameters were the same as those in the original article. In Creating minimum spanning tree step, we used the mst function of R package "ape". We used the complete linkage method in the clustering step. The representative marker expression was the median values of the cells belonging to each cluster on the consensus tree.

8. Dimension reduction and time-course reconstruction of EGF stimulation data

DEEF can reconstruct a distribution using only the top coordinates. To reduce the dimensionality of a cell population profile, we expressed the cell population profile using only the synthetic sum of the main patterns; other differences were considered to be noise. The reconstructed distributions (K=3) were obtained using the procedure described in Method section 2. For each of the four markers (pAKT, pERK, pS6, and pPLCγ2), we visualized the median expression value change for the original marker expression and the reconstructed marker expression. For the original marker expression, for each sample, we calculated the median value of each marker from the expression value of cells. For the reconstructed marker expression, we integrated the reconstructed distribution and eliminated all markers (three) except the one that we

focused on. Then, the 50th percentile value of the one marker expression was estimated as the median by linearly interpolating the values between the grids.

Next, we conducted the time-course reconstruction of Replicate1's EGF stimulation dataset whose original time course contained 10 time points. The value of the θ coordinate at each time point was estimated by linearly interpolating and dividing the value of the θ coordinate between each time point into 10 equal parts, and reconstructing the θ coordinate at a total of 91 time points. Based on the estimated value, the distribution was reproduced at K = 3.

Acknowledgments

We would like to thank Prof. Masaru Ishii, Dr. Takao Sudo, and Dr. Tetsuo Hasegawa, who are members of the Department of Immunology and Cell Biology, Osaka University Graduate School of Medicine.

References

1. Kunz DJ, Gomes T, James KR. Immune cell dynamics unfolded by single-cell technologies. Frontiers in immunology. 2018;9:1435.
2. Tyson DR, Garbett SP, Frick PL, Quaranta V. Fractional proliferation: a method to deconvolve cell population dynamics from single-cell data. Nature methods. 2012;9(9):923.
3. Heath JR, Ribas A, Mischel PS. Single-cell analysis tools for drug discovery and development. Nature reviews Drug discovery. 2016;15(3):204.
4. Walker LS, von Herrath M. CD4 T cell differentiation in type 1 diabetes. Clinical & Experimental Immunology. 2016;183(1):16–29.
5. Agematsu K, Hokibara S, Nagumo H, Komiyama A. CD27: a memory B-cell marker. Immunology today. 2000;21(5):204–206.
6. Qiu P, Simonds EF, Bendall SC, Gibbs Jr KD, Bruggner RV, Linderman MD, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nature biotechnology. 2011;29(10):886.
7. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology. 2014;32(4):381.
8. Bendall SC, Davis KL, Amir EaD, Tadmor MD, Simonds EF, Chen TJ, et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. Cell. 2014;157(3):714–725.
9. Saeys Y, Van Gassen S, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nature Reviews Immunology. 2016;16(7):449.
10. Barkas N, Petukhov V, Nikolaeva D, Lozinsky Y, Demharter S, Khodosevich K, et al. Joint analysis of heterogeneous single-cell RNA-seq dataset collections. Nature methods. 2019;16(8):695–698.
11. Tsang JS, Schwartzberg PL, Kotliarov Y, Biancotto A, Xie Z, Germain RN, et al. Global analyses of human immune variation reveal baseline predictors of postvaccination responses. Cell. 2014;157(2):499–513.

12. Nakaya HI, Wrammert J, Lee EK, Racioppi L, Marie-Kunze S, Haining WN, et al. Systems biology of vaccination for seasonal influenza in humans. Nature immunology. 2011;12(8):786.

13. Obermoser G, Presnell S, Domico K, Xu H, Wang Y, Anguiano E, et al. Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. Immunity. 2013;38(4):831–844.

14. Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytometry Part A: the journal of the International Society for Analytical Cytology. 2008;73(4):321–332.

15. Carter KM, Raich R, Finn WG, Hero III AO. Fine: Fisher information nonparametric embedding. IEEE transactions on pattern analysis and machine intelligence. 2009;31(11):2093–2098.

16. Gingold JA, Coakley ES, Su J, Lee DF, Lau Z, Zhou H, et al. Distribution Analyzer, a methodology for identifying and clustering outlier conditions from single-cell distributions, and its application to a Nanog reporter RNAi screen. BMC bioinformatics. 2015;16(1):225.

17. Nakamura N, Okada D, Setoh K, Kawaguchi T, Higasa K, Tabara Y, et al. LAVENDER: latent axes discovery from multiple cytometry samples with non-parametric divergence estimation and multidimensional scaling reconstruction. bioRxiv. 2019; p. 673434.

18. Mardia KV. Some properties of clasical multi-dimesional scaling. Communications in Statistics-Theory and Methods. 1978;7(13):1233–1241.

19. Fix E, Hodges Jr JL. Discriminatory analysis-nonparametric discrimination: consistency properties. California Univ Berkeley; 1951.

20. Parzen E. On estimation of a probability density function and mode. The annals of mathematical statistics. 1962;33(3):1065–1076.

21. Amari S. Information geometry. Contemporary Mathematics. 1997;203:81–96.

22. Walter S. The non-Euclidean style of Minkowskian relativity. The Symbolic Universe, Editor J Gray, Oxford University Press, Oxford. 1999; p. 91–127.

23. Knapp DJ, Kannan N, Pellacani D, Eaves CJ. Mass cytometric analysis reveals viable activated caspase-3+ luminal progenitors in the normal adult human mammary gland. Cell reports. 2017;21(4):1116–1126.

24. Silverbush D, Cristea S, Yanovich-Arad G, Geiger T, Beerenwinkel N, Sharan R. Simultaneous Integration of Multi-omics Data Improves the Identification of Cancer Driver Modules. Cell systems. 2019;8(5):456–466.

25. Okada D, Endo S, Matsuda H, Ogawa S, Taniguchi Y, Katsuta T, et al. An intersection network based on combining SNP coassociation and RNA coexpression networks for feed utilization traits in Japanese Black cattle. Journal of animal science. 2018;96(7):2553–2566.

26. Wang L, Xiao Y, Ping Y, Li J, Zhao H, Li F, et al. Integrating multi-omics for uncovering the architecture of cross-talking pathways in breast cancer. PloS one. 2014;9(8):e104282.

27. Chen G, Shi T. Single-cell RNA-seq technologies and related computational data analysis. Frontiers in genetics. 2019;10:317.

28. McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018;.

29. Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B, et al. Kernel mean embedding of distributions: A review and beyond. Foundations and Trends® in Machine Learning. 2017;10(1-2):1–141.

Supporting information

S1 Text Theory of DEEF and simulation data analysis.

S1 Fig Performance of DEEF for EGF stimulation data. (a) Eigenvalue plots for an EGF stimulation dataset. Left panel shows the absolute eigenvalues standardized so that its total value was 1, where black bars are positive eigenvalues and white bars are negative eigenvalues. Right panel shows the cumulative sum of absolute eigenvalues. (b) Performance boxplot of distributions reconstructed using only the top K coordinates with high absolute eigenvalues for the EGF stimulation dataset. The performance was evaluated by the Performance Index (PI) defined by the sum of the squared error between the true probability mass function and the reconstructed probability mass function. The overall performance increases with increasing value of K.

S2 Fig 4-by-4 density plot of F_1 and F_2 for EGF stimulation data.

S3 Fig Comparison of DEEF and MDS-based method with EGF stimulation data.(a) θ coordinate plot for coordinates θ_1 and θ_2 and (b) MDS coordinate plot for two coordinates MDS1 and MDS2 with the top eigenvalues.

S1 Table Representative marker expression values of ten clusters on the SPADE tree.

S1 Movie Animation of cell population dynamics for 91 time points after EGF stimulation for Replicate1. The reconstruction was done with θ_1, θ_2, and θ_{last} (K=3).

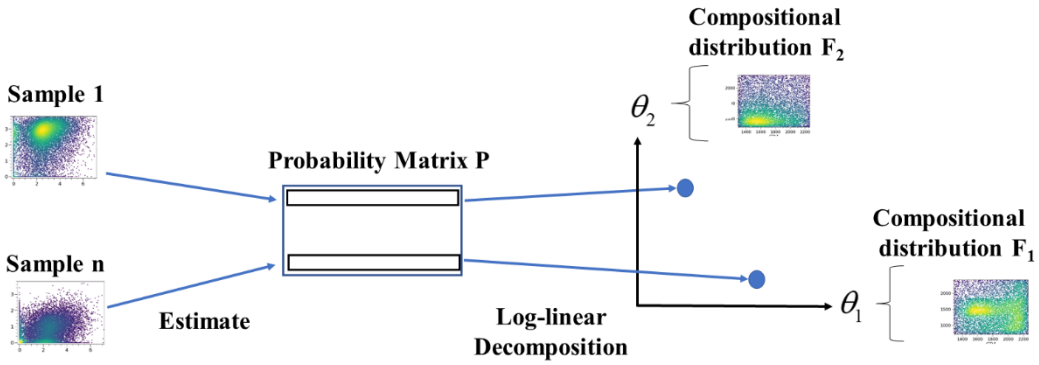


Fig1. The outline of DEEF for embedding data from multiple distributions in the θ coordinate space with its compositional distribution F.

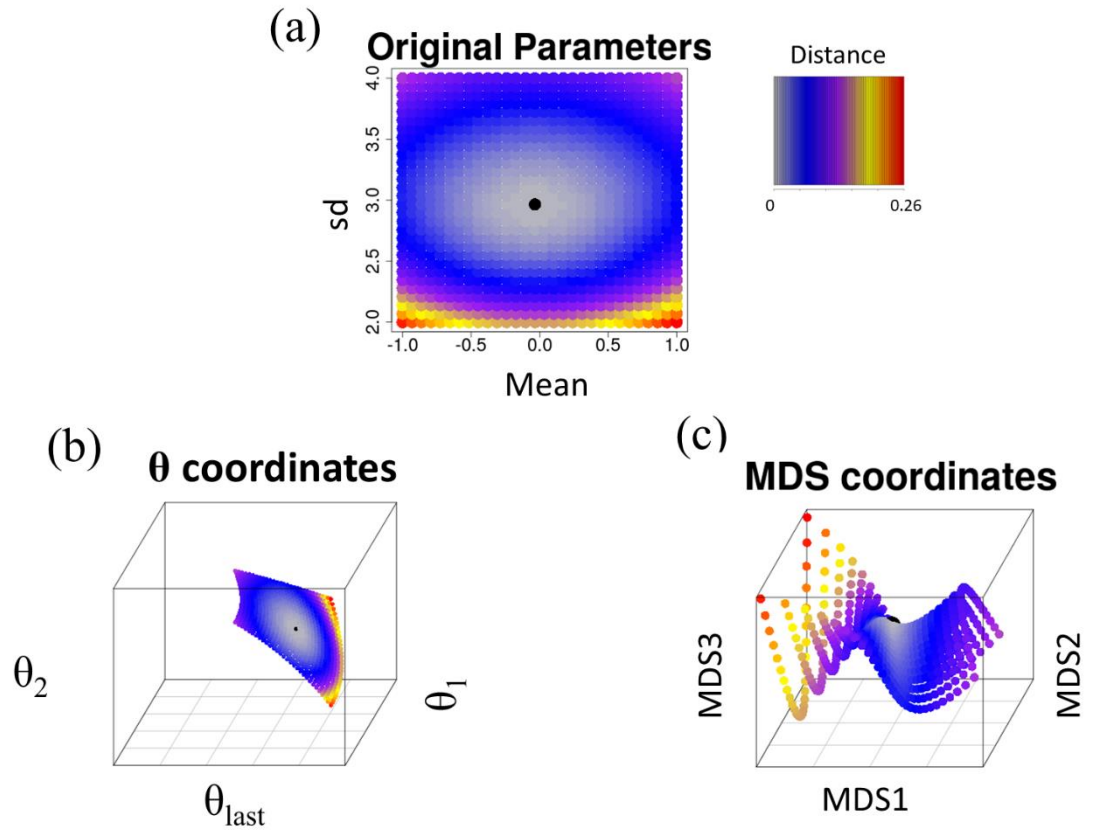


Fig 2. Comparison of (a) original parameter space, (b) θ coordinate space, and (c) MDS coordinate space in normal distribution set with the two parameters. The theoretical KL-divergence-based distance from one member distribution (black point) is visualized by the color scale. The Euclidean distance in the original parameter space does not match the KL-divergence-based distance. The Euclidean distance in the MDS space approximates the KL-divergence-based distance, but the parameter structure is broken, unlike the case when embedding in the θ coordinate space.

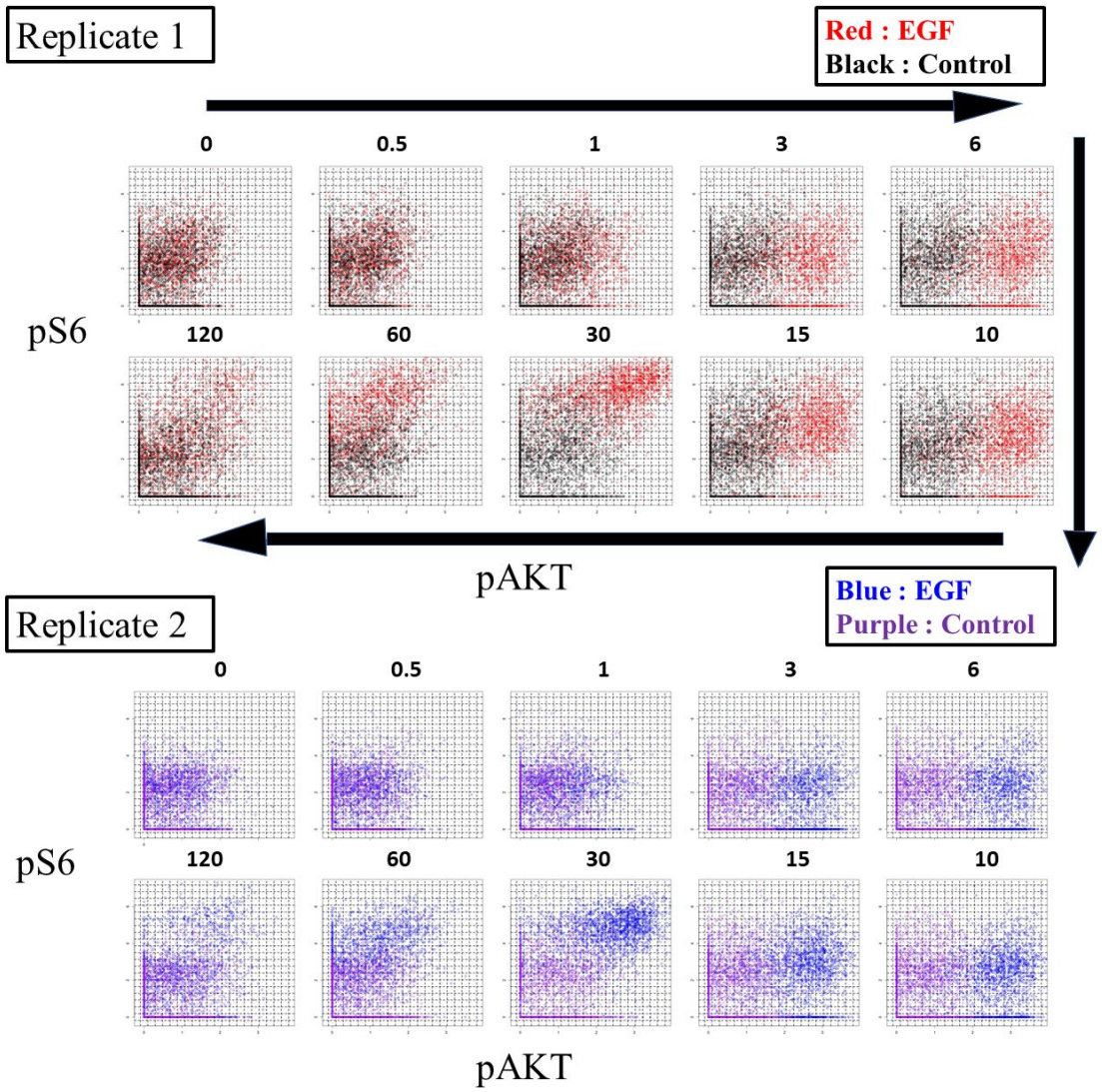


Fig 3. Scatter plot of pAKT and pS6 at 10 time points after EGF stimulation. For each replicate and condition, 2,000 randomly selected cells are plotted. The black dotted line represents the grids. The cell population profile changes dynamically after EGF stimulation but it is difficult to capture and evaluate this quantitatively using the raw data.

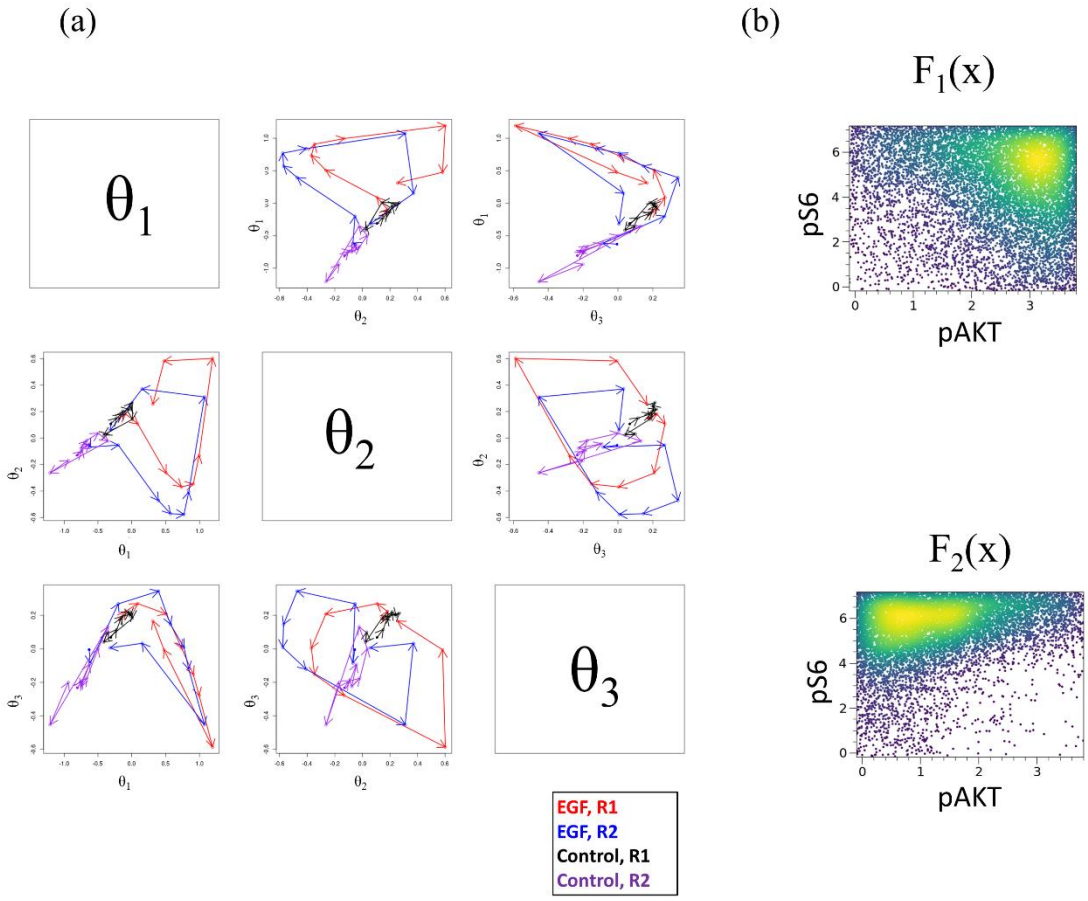


Fig 4. Application of DEEF to EGF stimulation data. The dynamics of the whole cell population profile are visualized and the dominant patterns that explain differences are extracted. (a) θ coordinate plot for coordinates θ_1, θ_2 and θ_3 (i.e., those with the top positive eigenvalues). (b) F_1 and F_2 in DEEF for pAKT and pS6. The density plot was generated from 10,000 randomly sampled data points from the standardized $\exp(F_i)$.

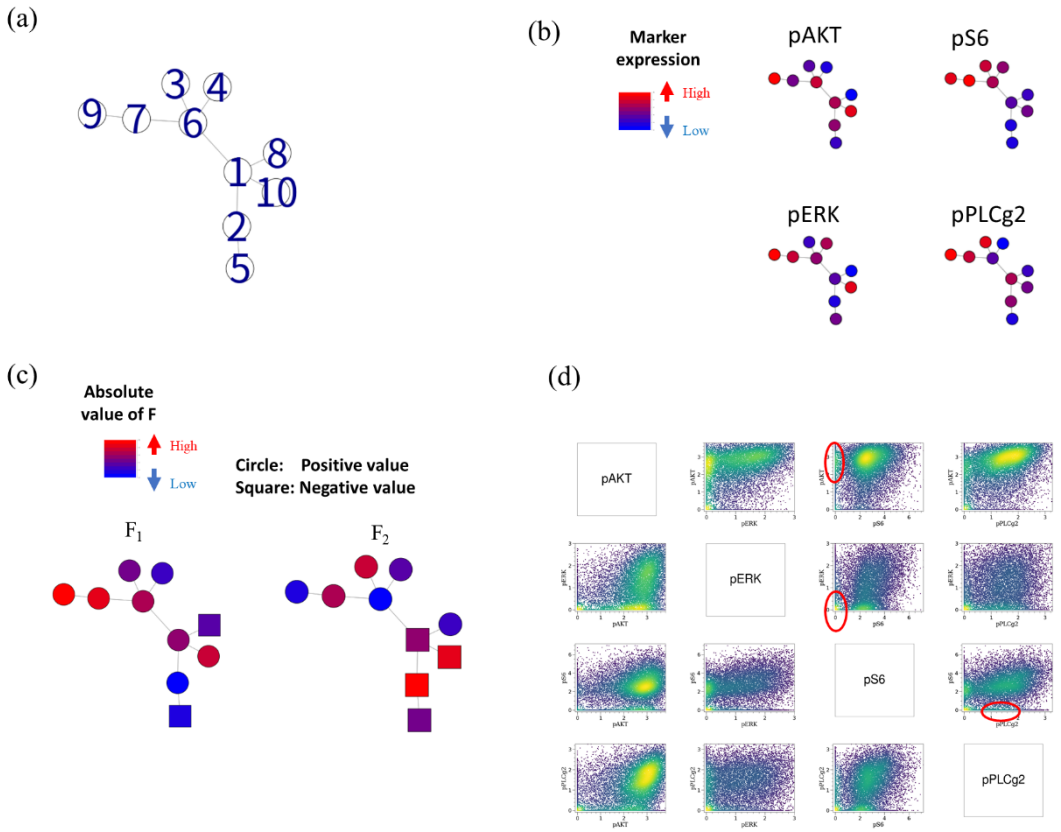


Fig 5. F_1 and F_2 of EGF stimulation data on SPADE tree. (a) Created SPADE tree with cluster number labels. (b) SPADE trees with four marker expression. The color represents each marker expression value. The colors are assigned according to the order of the values among the ten SPADE subsets. (c) SPADE trees with F_1 and F_2 values. Each cluster was assigned F_1 and F_2 values of the grid to which the representative location of the cluster belongs. The colors are assigned according to the order of the values among the ten SPADE subsets. (d) Region of Cluster 2 of SPADE tree of EGF stimulation data. The corresponding regions of SPADE Cluster 2 are shown by a red circles in the density plots of the four markers obtained 6 minutes after EGF stimulation for Replicate1.

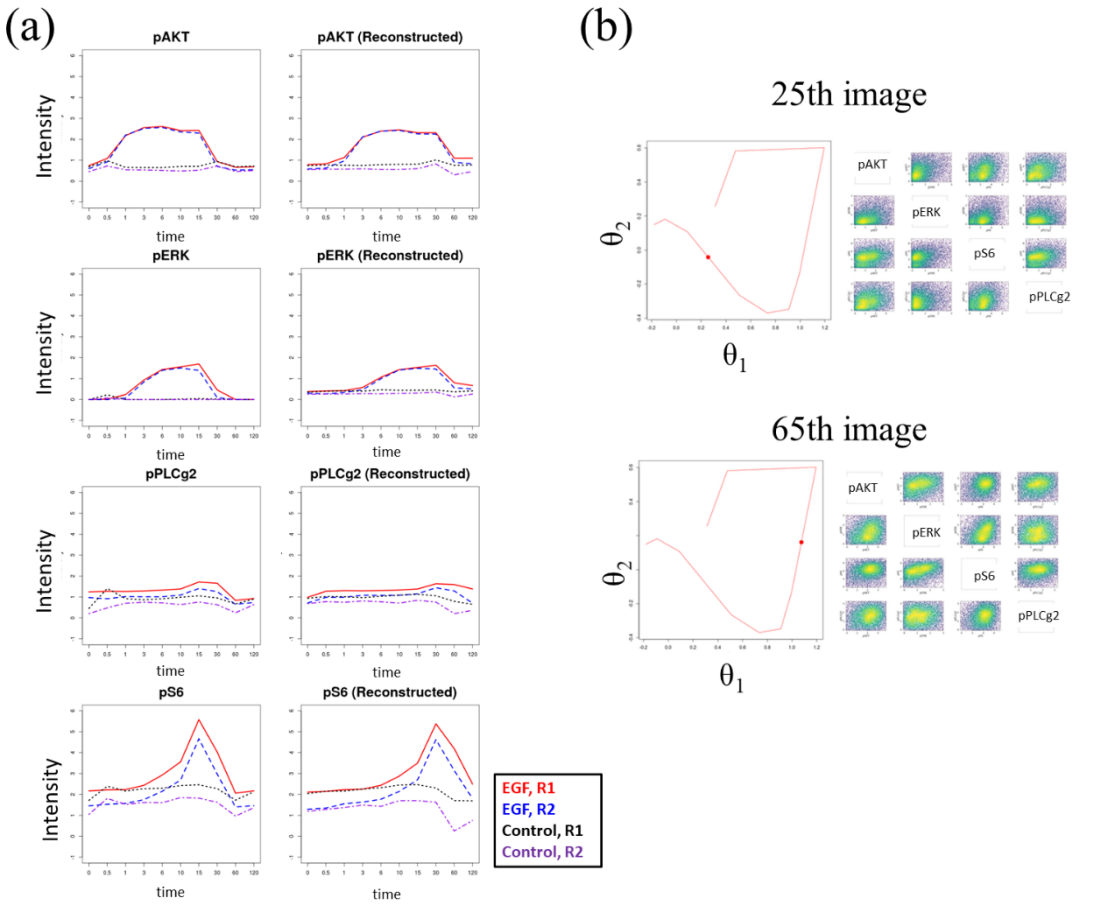
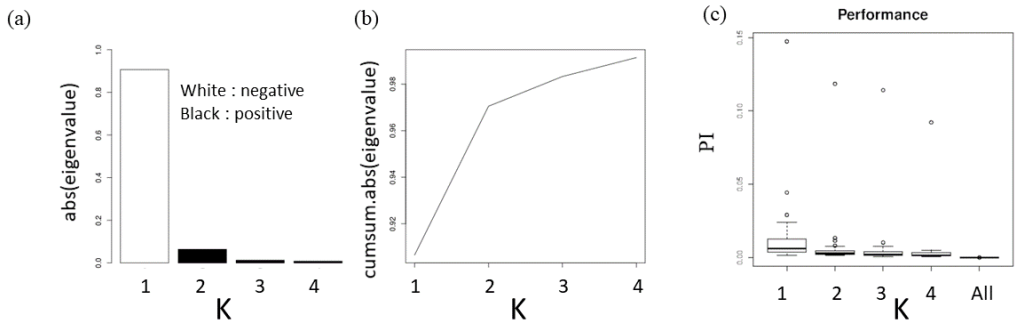
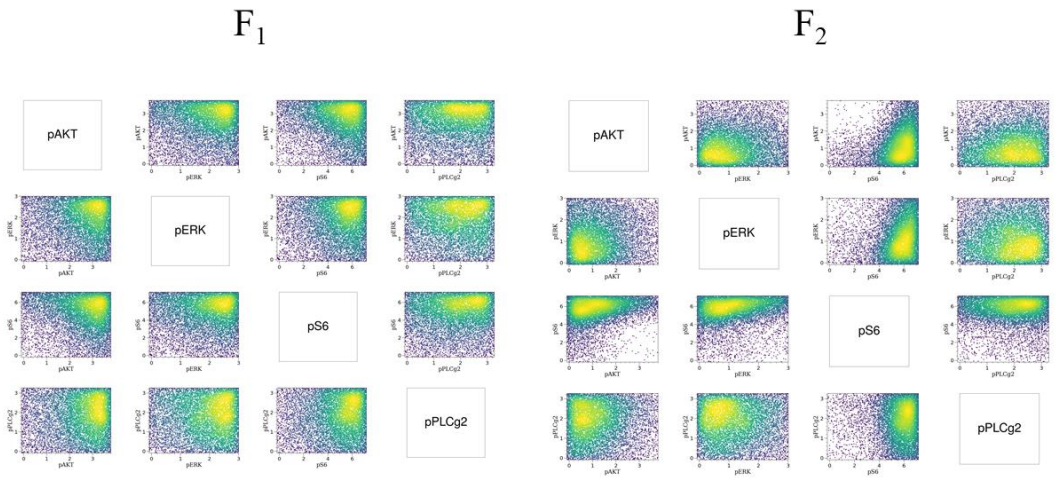


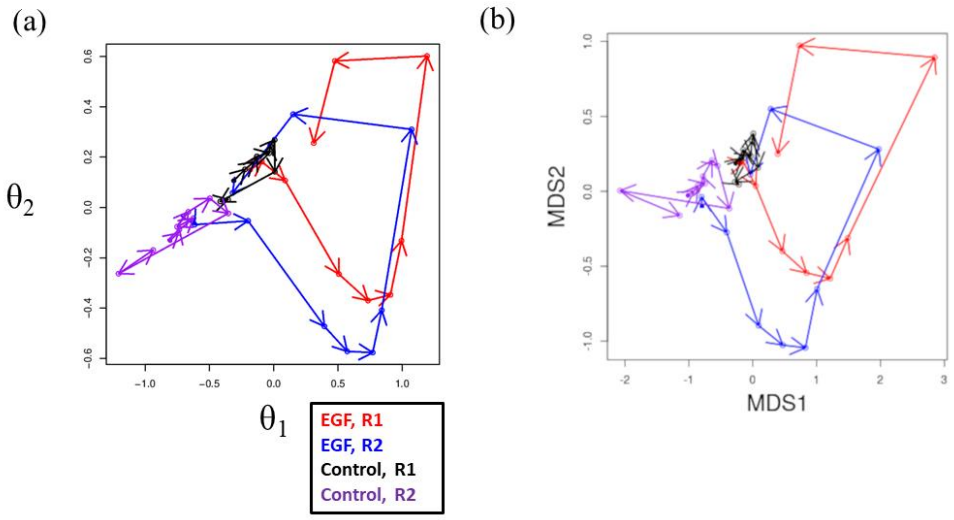
Fig 6. Results of dimension reduction of cell population profiles using the DEEF method. The reduction preserves the change in the marker expression along the time course for each marker (pAKT, pERK, pPLCγ2, and pS6). (a) Left panels are the median values for each marker expression, which match those in the original study. Right panels are the median values for the distribution reproduced using the top three θ coordinates, namely θ_{last} with the highest negative eigenvalue and θ_1 and θ_2 with the highest positive eigenvalues (K = 3). (b) 25th and 65th images of 91 images as examples of the reproduced distribution (K = 3) between the measurements of Replicate1 after EGF stimulation. The corresponding points in the θ coordinate space are indicated by red dots.



S1 Fig. Performance of DEEF for EGF stimulation data. (a) Eigenvalue plots for an EGF stimulation dataset. The panel shows the absolute eigenvalues standardized so that its total value was 1, where black bars are positive eigenvalues and white bars are negative eigenvalues. (b) The panel shows the cumulative sum of absolute eigenvalues. (c) Performance boxplot of distributions reconstructed using only the top K coordinates with high absolute eigenvalues for the EGF stimulation dataset. The performance was evaluated by the Performance Index (PI) defined by the sum of the squared error between the true probability mass function and the reconstructed probability mass function. The overall performance increases with increasing value of K.



S2 Fig. 4-by-4 density plot of F_1 and F_2 for EGF stimulation data.



S3 Fig. Comparison of DEEF and MDS-based method with EGF stimulation data. (a) θ coordinate plot for coordinates θ_1 and θ_2 and (b) MDS coordinate plot for two coordinates MDS1 and MDS2 with the top eigenvalues.

S1 Table Representative marker expression values of ten clusters on the SPADE tree.

The CSV file can be downloaded from the following URL.
https://doi.org/10.1371/journal.pone.0231250.s005

S1 Movie. Animation of cell population dynamics for 91 time points after EGF stimulation for Replicate1. The reconstruction was done with θ_1, θ_2, and θ_{last} (K = 3).

The GIF file can be downloaded from the following URL.
https://doi.org/10.1371/journal.pone.0231250.s006

Supplementary Text for DEEF

Daigo Okada and Ryo Yamada

October 24, 2020

1 Introduction to information geometry and exponential families

Research on information geometry has focused on exponential families and the coordinate space of probability distributions. An exponential family is a probability distribution that can be expressed in the following form:

$$\log P(x|\theta) = C(x) + \sum_{i=1} F_i(x)\theta_i - \psi(\theta)$$

where $P(x|\theta)$ is the probability density function, $C(x)$ is a function of x only, θ is the scalar value vector given for each distribution, θ_i is the i-th element of θ, $F_i(x)$ is the coefficient function of θ_i, and $\psi(\theta)$ is a potential function such that $P(x|\theta)$ satisfies the definition of a probability density distribution. Many probability distributions, including the standard normal distribution, can be expressed in this form and are included in the exponential family. If a distribution can be expressed as an exponential family, θ coordinates can be applied to it and it can be embedded in a low-dimensional space, which is a statistical manifold [1]. This space has two flat coordinate systems on which KL divergence can be calculated using each coordinate value [2]. In addition, the relation between the probability density/mass functions of θ coordinates is defined by $F_i(x)$. The mathematical nature of this space is well known in information geometry. However, some distributions are not included in the exponential family, such as the mixture normal distribution, which is commonly used in biology.

References

[1] Amari, S. I. "Information geometry." Contemporary Mathematics 203 (1997): 81-96.

[2] Nielsen, Frank, and Richard Nock. "Entropies and cross-entropies of exponential families." 2010 IEEE International Conference on Image Processing. IEEE, 2010.

2 Simulation analysis implementation

2.1 Embedding into θ coordinate space

First, we applied DEEF to a set of instances of the distribution in the exponential family and to a set of instances that are a parametric mixture of distributions in the exponential family to validate our theory. We generated four sets of simulation instances of a distribution using the monovariate normal distribution. The four sets are denoted 2D, Random, 1D, and Mixture. 2D consisted of 900 instances of a normal distribution, with the mean ranging from -1 to 1 and the sd ranging from 2 to 4 at a fixed interval of 0.069 for each. Random consisted of 50 instances randomly sampled from 2D. 1D was a normal distribution set that made a one-dimensional manifold in the same space as that of 2D. Mixture consisted of 900 instances that were a mixture of two normal distributions; one normal distribution was N(-1,1) and the other distribution had mean and sd ranging from 4 to 5 and 2 to 4 at fixed intervals of 0.034 and 0.069, respectively. The mixture ratio of the two distributions was 0.5 for all instances (Fig A). The number of grids was 10,000. The range for discretization was determined so that the section between the 0.5th percentile and the 99.5th percentile of all distributions was included.DEEF successfully extracted the parameter structure and reconstructed the distributions. The θ coordinate were calculated using the theoretical value of the functional inner product defined by the mean and sd.

The results of the application of our method to these four sets are shown in Fig A. The eigenvalues corresponding to each θ coordinate are shown in Fig B. For all distribution sets, the maximum eigenvalue is negative. The θ coordinate is denoted θ_i in decreasing order of eigenvalues.

θ_{last} is the coordinate corresponding to the lowest eigenvalue whose absolute value is largest. The eigenvalues calculated using this method always contain negative values (details given in the Appendix). Only the top two or three positive eigenvalues have meaningful contributions; the other positive eigenvalues have essentially no contribution. The number of parameters used to describe the heterogeneity of instances, or DoFs, for 2D is 2 (mean and sd). The DoFs for Random, 1D, and Mixture are 2, 1, and 2, respectively. These numbers correspond to the numbers of positive eigenvalues with meaningful absolute values.

We embedded all instance distributions into a three-dimensional space with the top three absolute eigenvalues (third column in Fig A). In the θ coordinate space, the original parameter structure, indicated by the color pattern, was maintained for all four sets. It can also be seen that the distributions were embedded on the manifold with the dimension of the original parameter structure.

Next, we investigated the $C(x)$ and $F(x)$ of the top coordinates of each distribution set. Fig C shows the calculated $C(x)$, $F_{last}(x)$, $F_1(x)$, and $F_2(x)$ for 2D, Random, 1D, and Mixture. For distribution set 2D, $C(x)$ has information about the average feature of the whole distribution set, as shown by the black curve (Fig C(a)). This function is convex, with a peak at the center of the x coordinate, which is the average pattern in distribution set 2D. θ_1 and θ_2 are

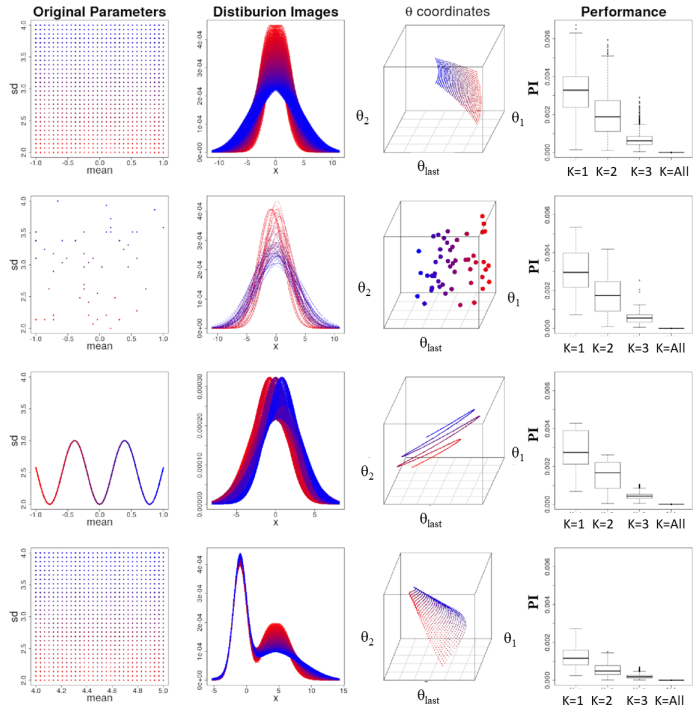


Fig A. Original parameter structure (first column), distribution (second column), θ coordinate mapping (third column), and boxplots of performance (fourth column) for four types of distribution set (2D, Random, 1D, Mixture). Each dot in the first and third column panels and each line in the second column panels together represent a distribution. Embedding in the θ coordinate space reproduces the original parameter structure with distortion. The fourth column panels show boxplots of the Performance Index (PI) defined by the sum of the squared error of distributions reconstructed using only the top K coordinates with high absolute eigenvalues for each distribution set. As K increases, the reconstructed distribution set approaches the original distribution set. When all θ coordinates are used, all distributions belonging to the reconstructed distribution set are identical to the original distributions.

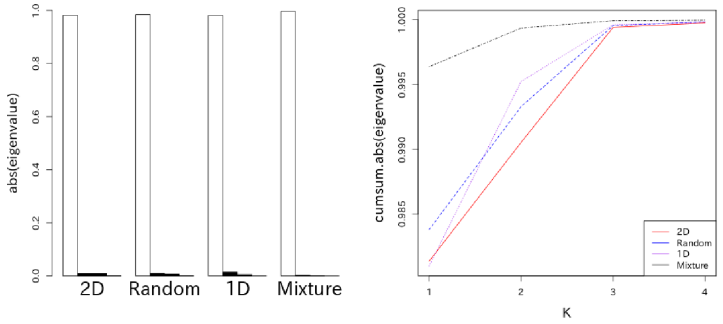


Fig B Eigenvalue plots for each distribution set. The left panel shows the absolute eigenvalues standardized so that its total value was 1, where the black bars are positive eigenvalues and the white bars are negative eigenvalues. For all distribution sets, the maximum eigenvalue was negative. The right panel shows the cumulative sum of eigenvalues. In distribution set 2D, the contribution increases by θ_1 and θ_2 are almost equal. This corresponds to a degree of freedom (DoF) of 2 for the parameter structure. This tendency also appears for Random and Mixture, although θ_1 has greater explanatory power. For distribution set 1D, the contribution of θ_2 is greatly reduced compared to that of θ_1.

the coordinates corresponding to the positive eigenvalues. $F_1(x)$ indicates that a larger value of θ_1 leads to a larger probability mass at both ends (Fig C(a), blue line). This is consistent with the fact that the distributions Normal(-1, 4) and Normal(1, 4) were embedded into the region with the largest θ_1 coordinate value in Fig A. $F_2(x)$ indicates that a larger value of θ_2 leads to a larger probability mass at the right end and a smaller probability mass at the left end (Fig C(a), purple line). The distribution with the maximum θ_2 value is Normal(1, 2), which is the member distribution with the highest mean and the lowest sd value in distribution set 2D in Fig A. These results suggest that $F_i(x)$, which corresponds to the positive eigenvalues, has information about what part of the difference each θ coordinate explains in the original distribution. $F_{last}(x)$ suggests that θ_{last}, the coordinate with the largest negative eigenvalue, is almost parallel to the x-axis and has little information about the distribution feature (Fig C(a), red line). This axis distorts the inner products and distances between the points on the manifold. For distribution set Random, a similar result was obtained but with slight distortion (Fig C(b)). Interestingly, the $F_1(x)$ and $F_2(x)$ for distribution set 1D are similar to $F_2(x)$ and inverted $F_1(x)$, respectively (Fig C(c)). For distribution set Mixture, extremely large values tended to be estimated in the edge region (Fig C (d)). Mixture(sub) is the magnified view of the central part of Mixture and shows that $C(x)$ captures the bimodality of the mixture normal distribution. Each $F_i(x)$ has a unique complex pattern in the distribution set, as is the case for the normal distribution set.

Finally, we reconstructed the distribution set using the top θ coordinates and

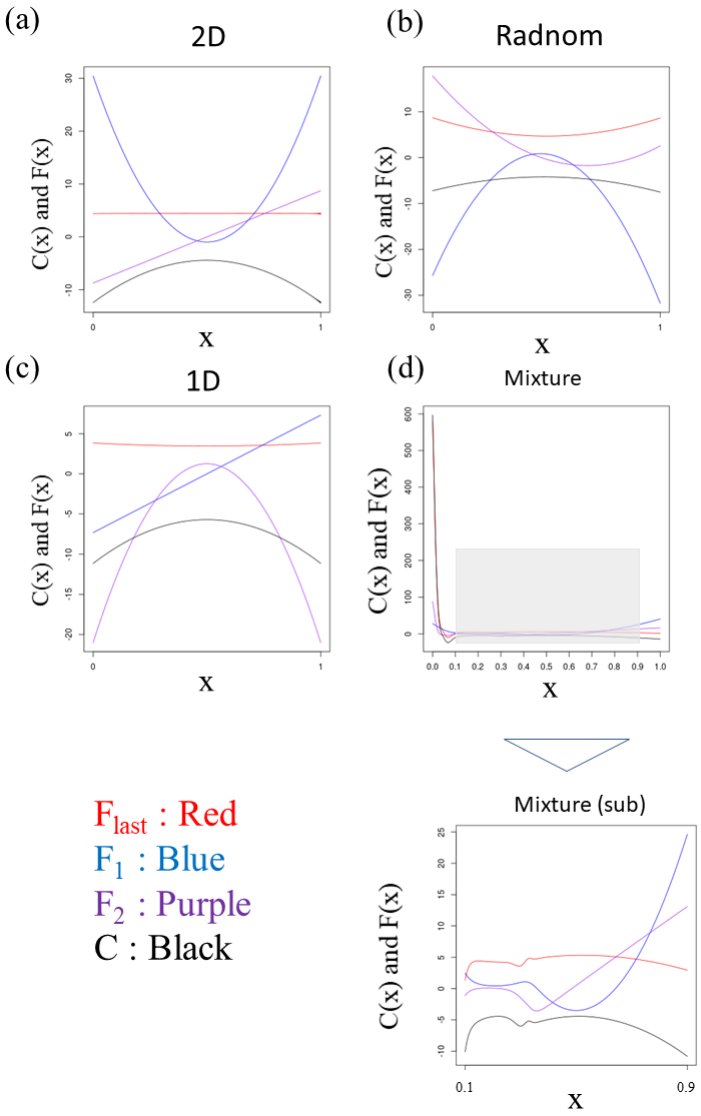


Fig C. Calculated $C(x)$, $F_{last}(x)$, $F_1(x)$, and $F_2(x)$ for distribution sets 2D, Random, 1D, and Mixture. For Mixture, extreme values tend to be estimated in the edge region. The range of discretization is scaled from 0 to 1. Part of the plot between 0.1 and 0.9 of the entire region (indicated by gray rectangle) was extracted from the left panel (Mixture(sub)). $C(x)$ represents the average pattern of the distribution set. $F_i(x)$ is a function that associates the corresponding θ_i with the original distribution.

evaluated performance. The reconstruction performance is defined as the difference between the original distribution and the reconstructed distribution (details are given in the main manuscript). With an increasing number of coordinates, the reconstructed distribution tends to approach the original distribution as a whole (fourth column in Fig A). Using all coordinates, the original distribution can be exactly reproduced. An example of the reconstruction of one distribution is shown in Fig D(a). Fig E shows the relationship between the position on the original parameter coordinates and performance. The performance improved as the number of coordinates increased to three. Instances located at the periphery of the distribution set tended to have worse reconstruction performance and required more θ coordinates to achieve performance similar to that of instances in the central area. The features shared by many instances were explained by a limited number of coordinates with relatively large eigenvalues, whereas those of instances at the periphery required more coordinates with relatively small eigenvalues. For example, Normal(-1, 4) is not sufficiently reproduced at the edge for K=3 or 4; K=5 is required (Fig D(b)). These performance features apply to all four distribution sets. These evaluations indicate that our method can identify the EEF expression of a set of distributions in the exponential family and can be applied to a set of mixture distributions that are not in the exponential family.

As another case of applying DEEF to simulation data, Fig F shows the case of an exponential distribution set. An exponential distribution was parameterized by one parameter and a distribution with non-symmetric shape that was unlike a normal distribution. We generated an exponential distribution set as another example of the application of DEEF. This set consisted of 900 instances whose lambda ranged from 1 to 5 at an interval of 0.0044. The number of grids was 10,000. The range for discretization was determined so that the section between the 0.5th percentile and the 99.5th percentile of all distributions was included. DEEF successfully extracted the parameter structure and reconstructed the distributions. Before applying DEEF, the first grid was removed for the calculation. Interestingly, the maximum eigenvalue was a positive eigenvalue, unlike the case for the normal distribution set. However, at least θ_1 and θ_{last} (K=2) are needed to obtain good performance in reconstruction.

2.2 Relationship between complexity and number of eigenvalues

The eigenvalue plots of distribution sets 2D, Random, 1D, and Mixture imply that the number of significant positive eigenvalues is the DoF of the set of distributions. As mentioned, the DoFs of the original parameter structures of 2D, Random, 1D, and Mixture were 2, 2, 1, and 2, respectively, which correspond to the numbers of significant positive eigenvalues. We thus quantitatively investigated whether the potential DoF of the distribution can be estimated using the DEEF method for other mixture normal distribution sets. The number of mixture components was changed from 2 to 10. All component normal distributions had the same sd (=1). The mean values of the component normal distributions

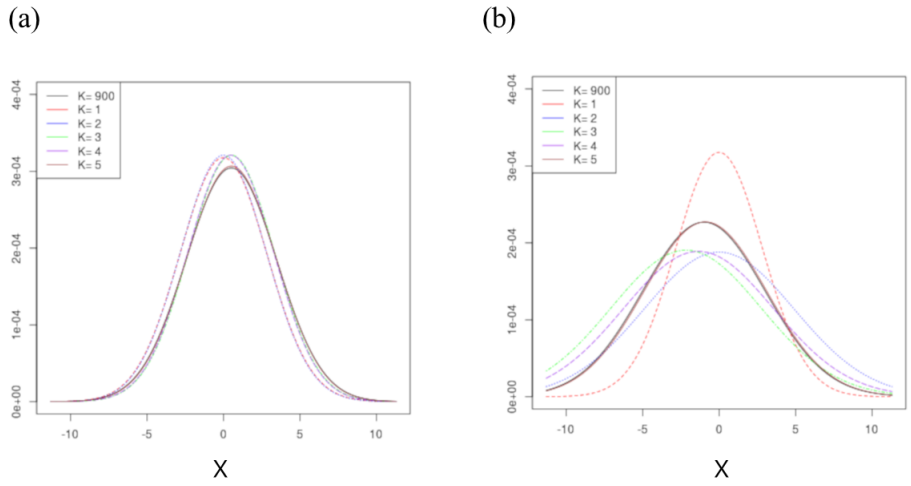


Fig D. Change in reconstructed distribution with K value (1, 2, 3, 4, 5, and all 900). (a) Normal(0.517, 2.97). The average and dispersion are roughly reproduced at K=3. The distribution reproduced with K=5 is almost equal to the original distribution (K=900). (b) Normal(-1, 4). The difference between this distribution and other distributions can be explained by θ_3 or θ_4, but not θ_1 or θ_2.

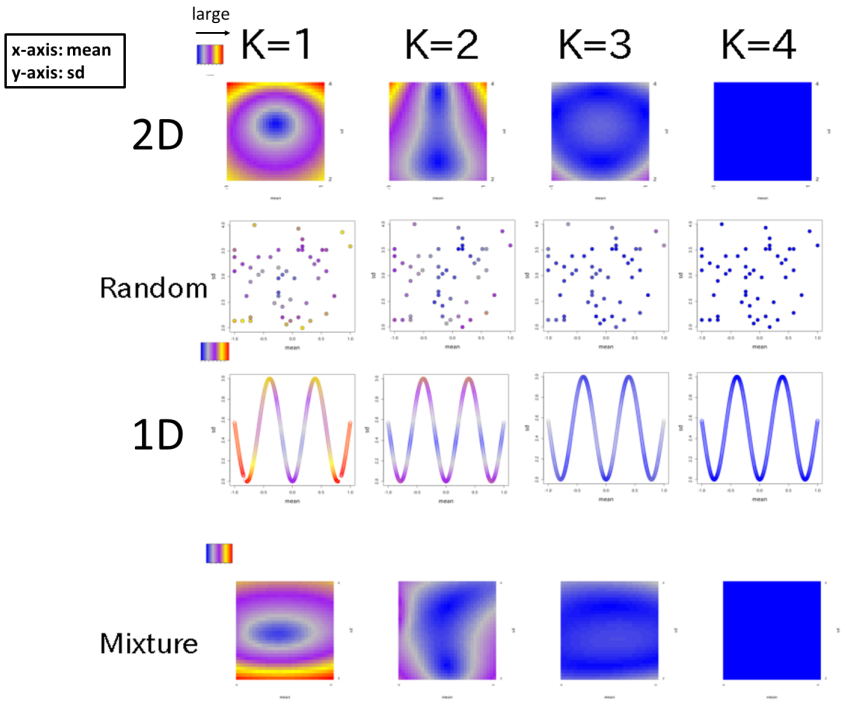


Fig E. Heat map of the performance of distributions reconstructed using only the top K coordinates for distribution sets 2D, 1D, Random, and Mixture. The performance, indicated by color, was evaluated in terms of the squared error between the true probability mass function and the probability mass function reconstructed using the top K θ coordinates. These panels show the relation between the location on the original parameter structure and the reconstruction performance.

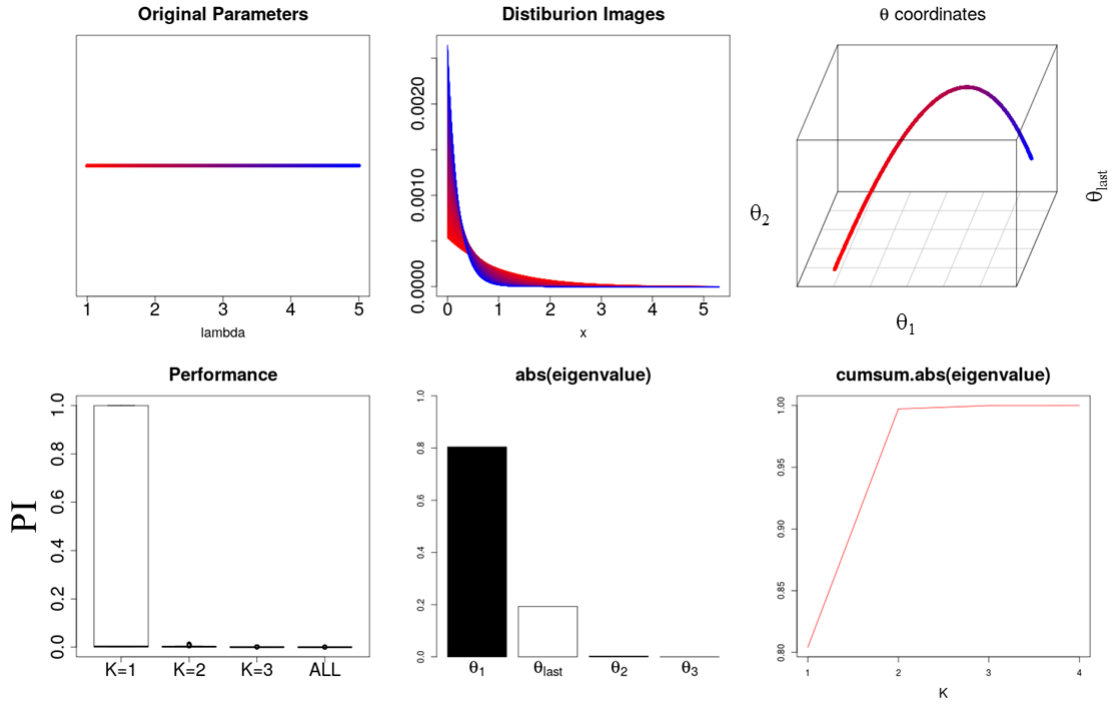


Fig F. Application of DEEF to an exponential distribution set. The meanings of the panels are the same as those for the normal distribution set.

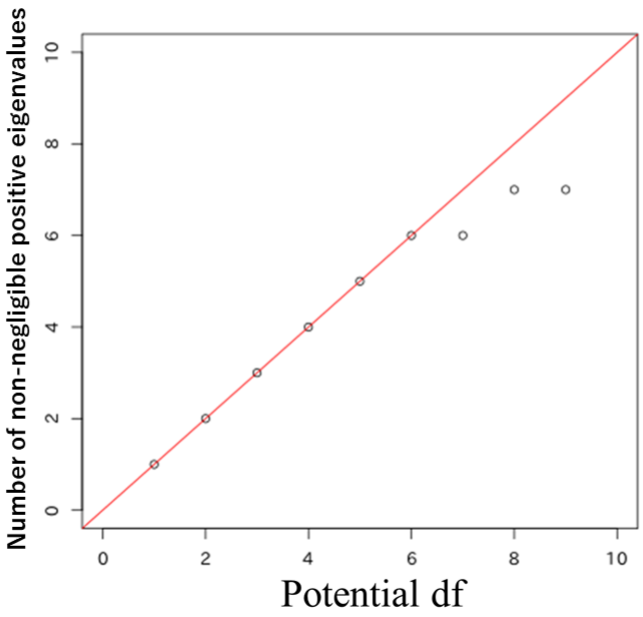


Fig G. Plot of potential degrees of freedom (DoFs) versus the number of non-negligible positive eigenvalues. The number of mixture components was varied from 2 to 10. The potential DoF can be defined as the number of mixture components - 1. The number of non-negligible positive eigenvalues is equal to the potential DoF of the distribution set; however, the relationship is not satisfied for large DoFs due to the limited resolution of the proposed method.

were evenly spaced between -10 and 10. The number of non-negligible positive eigenvalues was defined as the minimum number of eigenvalues whose sum exceeds 90% of the sum of all positive eigenvalues. A distribution set was composed of 100 distributions, among which only the mixture ratio was different. For this case, the potential DoF for the distribution set was defined as the number of components - 1 because only the mixture ratio varied.

Fig G shows a plot of the potential DoF and the number of non-negligible positive eigenvalues. This plot suggests that the potential DoF corresponds to the number of non-negligible positive eigenvalues when the number of components is small (red line). When the potential DoF was larger, the number of required eigenvalues decreased, which seemed to be due to the insufficient resolution of the decomposition by the proposed method because of the relatively small sample size compared to the complexity of the datasets. This result sug-

gests that our method can theoretically identify the potential DoF of datasets based on the quantity of meaningful information in the datasets.

2.3 Discussion in terms of information geometry

In information geometry, the geometric properties of the probability distribution space have been extensively researched. In particular, it is known that exponential families can be embedded into special manifolds equipped with two flat coordinate systems. The geometrical properties of these coordinate systems have been well investigated. The θ coordinate system of the DEEF method is quite similar to such systems. However, the potential function of an EEF is not convex and has imaginary axis (particularly when the inner product matrix is calculated from the probability mass function, a negative eigenvalue must be appear. (proof is shown in Appendix Theorem 3)). That makes the interpretation of the EEF space more complicated than that of a regular information geometry space. One specific feature of the EEF space is that it has subspaces where no distributions are assigned. One example of a space with an indeterminate inner product is the Minkowski space, which has a deep relationship with special relativity [1]. The properties of manifolds defined by these features should be further studied in the future. The investigation of these features in terms of theoretical geometry would further advance the understanding of statistical manifolds and probability distribution theory.

References

[1] Walter, Scott. "The non-Euclidean style of Minkowskian relativity." The Symbolic Universe, Editor J. Gray, Oxford University Press, Oxford (1999): 91-127.

Appendix

Theorem 1 If $P(x, \boldsymbol{\theta}^P)$ and $Q(x, \boldsymbol{\theta}^Q)$ are the members of an exponential family represented by Eq 1, then:

$$< P(x, \boldsymbol{\theta}^P), Q(x, \boldsymbol{\theta}^Q) >= \frac{e^{\psi(\boldsymbol{\theta}^P + \boldsymbol{\theta}^Q)}}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}}$$

Proof: The definition of an exponential family can be written as:

$$\begin{aligned} \log P(x, \boldsymbol{\theta}^P) &= \sum_{i=0} F_i(x) \boldsymbol{\theta}_i^P - \psi(\boldsymbol{\theta}^P) \\ \log Q(x, \boldsymbol{\theta}^Q) &= \sum_{i=0} F_i(x) \boldsymbol{\theta}_i^Q - \psi(\boldsymbol{\theta}^Q) \\ F_0(x) &= C(x), \theta_0 = Const \end{aligned}$$

Then, the inner product of the exponential distribution family is defined by the following procedure:

$$\begin{aligned} < P(x, \boldsymbol{\theta}^P), P(x, \boldsymbol{\theta}^Q) > &= \int P(x, \boldsymbol{\theta}^P) P(x, \boldsymbol{\theta}^Q) dx \\ &= \int e^{\sum_{i=0} F_i(x) \theta_i^P - \psi(\boldsymbol{\theta}^P)} e^{\sum_{i=0} F_i(x) \theta_i^Q - \psi(\boldsymbol{\theta}^Q)} dx \\ &= \int e^{\sum_{i=0} F_i(x) (\theta_i^P + \theta_i^Q) - (\psi(\boldsymbol{\theta}^P) + \psi(\boldsymbol{\theta}^Q))} dx \\ &= \frac{1}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}} \int e^{\sum_{i=0} F_i(x) (\theta_i^P + \theta_i^Q)} dx \end{aligned}$$

Let $\boldsymbol{\theta}^{P+Q}$ be $\boldsymbol{\theta}^P + \boldsymbol{\theta}^Q$. The above equation can then be rewritten as:

$$< P(x, \boldsymbol{\theta}^P), P(x, \boldsymbol{\theta}^Q) >= \frac{1}{e^{\psi(\boldsymbol{\theta}^P)} e^{\psi(\boldsymbol{\theta}^Q)}} \int e^{\sum_{i=0} F_i(x) \boldsymbol{\theta}^{P+Q}} dx$$

The following equation holds because: $\int P(x, \boldsymbol{\theta}^{P+Q}) dx = \int e^{\sum_{i=0} F_i(x) \theta_i^{P+Q} - \psi(\boldsymbol{\theta}^{P+Q})} dx = 1$

$$\int e^{\sum_{i=0} F_i(x) \theta_i^{P+Q}} dx = e^{\psi(\boldsymbol{\theta}^{P+Q})}$$

Then, the inner product between members of an exponential family is expressed as:

$$
\begin{aligned}
< P(x,\boldsymbol{\theta}^P), P(x,\boldsymbol{\theta}^Q) > &= \frac{1}{e^{\psi(\boldsymbol{\theta}^P)}e^{\psi(\boldsymbol{\theta}^Q)}} \int e^{\sum_{i=0} F_i(x)\boldsymbol{\theta}^{P+Q}} dx \\
&= \frac{1}{e^{\psi(\boldsymbol{\theta}^P)}e^{\psi(\boldsymbol{\theta}^Q)}} \int e^{\sum_{i=0} F_i(x)\theta_i^{P+Q} - \psi(\boldsymbol{\theta}^{P+Q})} e^{\psi(\boldsymbol{\theta}^{P+Q})} dx \\
&= \frac{e^{\psi(\boldsymbol{\theta}^{P+Q})}}{e^{\psi(\boldsymbol{\theta}^P)}e^{\psi(\boldsymbol{\theta}^Q)}} \int e^{\sum_{i=0} F_i(x)\theta_i^{P+Q} - \psi(\boldsymbol{\theta}^{P+Q})} dx \\
&= \frac{e^{\psi(\boldsymbol{\theta}^{P+Q})}}{e^{\psi(\boldsymbol{\theta}^P)}e^{\psi(\boldsymbol{\theta}^Q)}} \\
&= \frac{e^{\psi(\boldsymbol{\theta}^P + \boldsymbol{\theta}^Q)}}{e^{\psi(\boldsymbol{\theta}^P)}e^{\psi(\boldsymbol{\theta}^Q)}}
\end{aligned}
$$

Theorem 2 If $P(x,\boldsymbol{\theta}^P)$ and $Q(x,\boldsymbol{\theta}^Q)$ are EEFs as defined in Eq 2, then:

$$
\frac{1}{2}\log < P(x,\boldsymbol{\theta}^P), Q(x,\boldsymbol{\theta}^Q) > = \sum_{k=1} h_k \theta_k^P \theta_k^Q
$$

Proof: If P and Q are not members of the exponential family but are EEFs, the potential function $\psi(\theta)$ can be expressed as $\psi'(\theta) = \sum_{k=1} h_k \theta_k^2$. Theorem 1 is also satisfied if P and Q are EEFs. Then:

$$
\begin{aligned}
\frac{1}{2}\log < P(x,\boldsymbol{\theta}^P), Q(x,\boldsymbol{\theta}^Q) > &= \frac{1}{2}(\psi(\boldsymbol{\theta}^{P+Q}) - \psi(\boldsymbol{\theta}^P) - \psi(\boldsymbol{\theta}^Q)) \\
&= \frac{1}{2}(\sum_{k=1} h_k(\theta_k^P + \theta_k^Q)^2 - \sum_{k=1} h_k(\theta_k^P)^2 - \sum_{k=1} h_k(\theta_k^Q)^2) \\
&= \sum_{k=1} h_k \theta_k^P \theta_k^Q
\end{aligned}
$$

Theorem 3 Matrix $\mathbf{M}$ must have at least one negative eigenvalue.

Proof: Probability matrix $\mathbf{P}$ can be expressed as:

$$
\mathbf{P} = \mathbf{NA}
$$

where $\mathbf{A}$ has the same size as that of $\mathbf{P}$, $a_{i,j}$, the (i,j)-th element of $\mathbf{A}$, is non-negative, and $\mathbf{N}$ is a diagonal matrix used for the normalization of row sums whose i-th diagonal element is $N_i = \sum_{k=1}^m a_{i,k}$. Therefore, $\mathbf{Q} = \mathbf{PP}^T = \mathbf{NA}(\mathbf{NA})^\mathbf{T} = \mathbf{NAA}^T\mathbf{N}$. Then, denote $q_{i,i}$ as the i-th diagonal element of $\mathbf{Q}$. $q_{i,i} = \frac{\sum_{k=1}^m a_{i,k}^2}{(\sum_{k=1}^m a_{i,k})^2}$ because $a_{i,j} > 0$ and $q_{i,i} < 1$. $trace(M) = \sum_{i=1}^n m_{i,i} = \sum_{i=1}^n \log q_{i,i}$, which must be negative. Because $\mathbf{M}$ is a symmetric matrix, it has n real eigenvalues. The trace and the sum of the eigenvalues must match. From the above, $\mathbf{M}$ has at least one negative eigenvalue.