Statistical Archetypal Analysis

Chenyue Wu* Esteban G. Tabak†

February 1, 2017

Abstract

Statistical Archetypal Analysis (SAA) is introduced for the dimensional reduction of a collection of probability distributions known via samples. Applications include medical diagnosis from clinical data in the form of distributions (such as distributions of blood pressure or heart rates from different patients), the analysis of climate data such as temperature or wind speed at different locations, and the study of bifurcations in stochastic dynamical systems. Distributions can be embedded into a Hilbert space with a suitable metric, and then analyzed similarly to feature vectors in Euclidean space. However, most dimensional reduction techniques –such as Principal Component Analysis– are not interpretable for distributions, as neither the components nor the reconstruction of input data by components are themselves distributions. To obtain an interpretable result, Archetypal Analysis (AA) is extended to distributions, requiring the components to be mixtures of the input distributions and approximating the input distributions by mixtures of components.

Keywords. archetypal analysis, dimension reduction, energy distance, kernel embedding

*Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA, chenyue@cims.nyu.edu

†Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, NY 10012, USA, tabak@cims.nyu.edu

1 Introduction

Finite collections of probability distributions appear naturally in a variety of settings, often as conditional distributions $\rho(x|z)$ where z adopts a discrete set of values. For instance, x may represent a collection of clinical variables such as body temperature, blood pressure and cholesterol level, and z may stand for covariates such as sex, age group or medical treatment. In an example that this paper analyzes in some detail, x is the atmospheric temperature measured at ground level and z stands for the station where the measurements are performed.

It is therefore a natural extension of data analysis to use as either labels or features, probability distributions instead of the more conventional discrete-valued variables, continuum scalars or vectors. Thus one might want to predict not the temperature at a particular location and time but its probability distribution, or cluster populations for medical purposes according to the probability distributions of a group of clinical variables.

A basic quantity that permeates data analysis is the distance between data points. There are several statistical distances in the literature that measure the dissimilarity between two probability distributions. Some are based on analogues of the Euclidean distance, some on information theory, some on optimal transport. Typically, each sheds a different light on what makes two distributions different. In this article, we use the energy distance as a measure of dissimilarity among distributions, as it is easy to evaluate efficiently from sample points and can be derived from an inner product, thus rendering accessible many data analysis tools.

We study the problem of dimensional reduction of sets of distributions. After being equipped with a metric and embedded into a Hilbert space, distributions can be analyzed similarly to conventional feature vectors. However, there is a gap between the dimensional reduction of distributions and vectors: interpretability. Traditional dimension reduction techniques, such as principal components analysis, lack interpretability when applied to probability distributions, as the projection of each distribution onto the low dimensional subspace found is almost surely not a probability distribution: even though probability distributions can be embedded into a Hilbert space, almost all elements in this space are not probability distributions, since these are constrained by positivity and normalization.

To overcome this difficulty in interpretation, we use the tools of archetypal analysis. Archetypal analysis finds a small number of "archetypes" that

are convex combinations of the original data points, and approximates the original data points again via convex combinations of these archetypes. A convex combination can be interpreted as a mixture of probability distributions, so the archetypes found by archetypal analysis are mixtures of the original distributions and the original distributions are approximated within the family of mixtures of the archetypes.

This paper is arranged as follows: Section 2 gives a review of archetypal analysis, of the algorithms for archetypal analysis in the general case and specifically for energy distance. Section 3 reviews reproducing kernel Hilbert space, energy distance, describes how distributions equipped with the energy distance can be embedded into a Hilbert space, and describes algorithms to evaluate the energy distance from samples. Section 4 introduces statistical archetypal analysis for the dimensional reduction of probability distributions and includes applications with numerical experiment.

2 Archetypal Analysis

Archetypal analysis approximates data points by convex combination of prototypes, where these prototypes, denoted "archetypes", are themselves convex combinations of the data points.

Archetypal analysis was introduced in Cutler and Breiman [1994] –see also Friedman et al. [2001]– as a dimensional reduction method alternative to principal components analysis (PCA), yielding more interpretable results. It originated in the study of a dataset consisting of 6 head dimensions for 200 soldiers, with the goal of designing face masks for the Swiss Army. For this dataset, PCA found principal components that did not resemble a head shape. To have patterns resembling "pure types" in the data, each entry in the dataset was approximated by a mixture of the patterns. To make patterns resemble the data, each pattern itself was a mixture of the data points.

For a data matrix $X = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ representing n observations, each of dimension m, Archetypal Analysis seeks $k \ll n$ m-dimensional archetypes $Z = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_k)$, such that each $\mathbf{x}_i$ can be approximated by a convex combination of the z_k:

$$x_i \approx a_{1i}\mathbf{z}_1 + a_{2i}\mathbf{z}_2 + \cdots + a_{ki}\mathbf{z}_k, \quad a_{ji} \geq 0, \quad \sum_j a_{ji} = 1,$$

where the $\mathbf{z}_j$ themselves are convex combinations of the data:

$$\mathbf{z}_j = b_{1j}\mathbf{x}_1 + b_{2j}\mathbf{x}_2 + \cdots + b_{nj}\mathbf{x}_n, \quad b_{ij} \geq 0, \quad \sum_i b_{ij} = 1.$$

After setting a number of archetypes k, the coefficients a and b arise from the optimization problem

$$\min_{a_{ji}, b_{lj}} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{j=1}^{k} a_{ji} \sum_{l=1}^{n} b_{lj}\mathbf{x}_l \right\|^2, \tag{1}$$

with constraints

$$a_{ji} \geq 0, \quad \sum_j a_{ji} = 1, \quad b_{ij} \geq 0, \quad \sum_i b_{ij} = 1,$$

or, in terms of the matrices $A = (a_{ji})_{k\times n}$ and $B = (b_{lj})_{n\times k}$,

$$\min_{A,B} \|X - XBA\|_F^2 \tag{2}$$

under the same constraints, with $\|\cdot\|_F$ denoting the Frobenius norm $\|M\|_F = \left(\sum_{i=1}^{p} \sum_{j=1}^{q} |m_{ij}|^2 \right)^{\frac{1}{2}}$. Alternatively, this can be written as:

$$A, B = \operatorname{argmin} \operatorname{tr}\left[(I_n - BA)^\intercal G (I_n - BA) \right], \tag{3}$$

where $G = X^\intercal X$ is the Gram matrix of data. This restatement is particularly convenient, as it will allow us to formulate the problem in terms of inner products among the data points instead of the points themselves, which in our problem are distributions. Thus we need a norm for distributions that derive from an inner product, for which we will adopt the energy distance.

3 Energy Distance

The energy distance is a metric defined on probability measures (Rizzo and Székely [2016], Székely and Rizzo [2013]), which we will use to measure dissimilarity among probability distributions.

Definition 1 (Energy Distance)**.** For probability measures μ, ν on $\mathbb{R}^d$, random vectors $X, X' \sim \mu(x), Y, Y' \sim \nu(y)$, $\mathbb{E}\|X\| < \infty$, $\mathbb{E}\|Y\| < \infty$, the energy distance between μ and ν, $D(\mu, \nu)$, is defined by

$$D^2(\mu, \nu) = 2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|, \tag{4}$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$, and X, X', Y and Y' are pairwise independent.

The energy distance as defined above is a metric on distributions (Klebanov [2002], Székely and Rizzo [2005]). It can be viewed as the metric induced by kernel embedding (Rachev et al. [2013]) with kernel

$$k(x, y) = \|x - x_0\| + \|y - y_0\| - \|x - y\|, \tag{5}$$

where x_0 is a fixed value in $\mathbb{R}^d$, whose choice does not affect the induced metric. The kernel induces an inner product between distributions P and Q:

$$\langle P, Q \rangle = \mathbb{E}_{X,Y} k(X, Y) \tag{6}$$

where $X \sim P$, $Y \sim Q$, with X and Y independent. The corresponding square-distance is given by

$$\begin{aligned} \gamma_k^2(P, Q) &= \langle P, P \rangle + \langle Q, Q \rangle - 2\langle P, Q \rangle \\ &= \mathbb{E}_{XX'} k(X, X') + \mathbb{E}_{YY'} k(Y, Y') - 2\mathbb{E}_{XY} k(X, Y), \end{aligned} \tag{7}$$

where the random vectors $X, X' \sim P(x), Y, Y' \sim Q(y)$ are pairwise independent (conditions for kernels to yield a metric can be found in Klebanov [2002], Sriperumbudur et al. [2010]). In terms of the kernel in (5),

$$\begin{aligned} \gamma_k^2(\mu, \nu) = 2\mathbb{E}\|X - x_0\| - \mathbb{E}\|X - X'\| + 2\mathbb{E}\|Y - x_0\| - \mathbb{E}\|Y - Y'\| \\ -2\mathbb{E}\|X - x_0\| - 2\mathbb{E}\|Y - x_0\| + 2\mathbb{E}\|X - Y\| = D^2(\mu, \nu). \end{aligned}$$

A number of distances for distributions is available in the literature of statistics, probability and information theory, such as the Kullback-Leibler divergence (Bishop [2006], Kullback [1968]) and the p-Wasserstein metric between two probability measures $\mu(x)$ and $\nu(x)$ on a metric space (M, d) (Givens et al. [1984]). We chose the energy distance because it can be estimated efficiently from samples and it embeds the probability measures into a Hilbert space, which facilitates further analysis.

Energy distance has a kernel interpretation, which induces an inner product.

3.1 Estimating the Energy Distance from Data

In calculating the energy distance between two distributions μ and ν given independent random vectors $X \sim \mu$, $Y \sim \nu$ and their i.i.d. copies X', Y',

$$D(\mu, \nu) = \sqrt{2\mathbb{E}\|X - Y\| - \mathbb{E}\|X - X'\| - \mathbb{E}\|Y - Y'\|},$$

one needs to evaluate three expectations: $\mathbb{E}\|X - Y\|$, $\mathbb{E}\|X - X'\|$ and $\mathbb{E}\|Y - Y'\|$. If we only have samples of μ and ν, these expectation can be approximated by their empirical means.

Specifically, when we have samples $\{x_i\}_{i=1}^{n_X}$ of μ and $\{y_j\}_{j=1}^{n_Y}$ of ν, we can estimate the energy distance between μ and ν by the energy distance between their corresponding empirical distributions $\hat{\mu}$ and $\hat{\nu}$:

$$D(\hat{\mu}, \hat{\nu}) = \sqrt{2\mathbb{E}\|\hat{X} - \hat{Y}\| - \mathbb{E}\|\hat{X} - \hat{X}'\| - \mathbb{E}\|\hat{Y} - \hat{Y}'\|}. \tag{8}$$

In the equations above,

$$\mathbb{E}\|\hat{X} - \hat{Y}\| = \frac{1}{n_X n_Y} \sum_{i,j=1}^{i=n_X, j=n_Y} \|x_i - y_j\| \tag{9}$$

is the empirical mean of $\mathbb{E}\|X - Y\|$. For $\hat{X}'$, we use the same samples available for X,

$$\mathbb{E}\|\hat{X} - \hat{X}'\| = \frac{1}{n_X n_X} \sum_{i,i'=1}^{i=n_X, i'=n_X} \|x_i - x_{i'}\|. \tag{10}$$

Similarly,

$$\mathbb{E}\|\hat{Y} - \hat{Y}'\| = \frac{1}{n_Y n_Y} \sum_{j,j'=1}^{j=n_Y, j=n_Y} \|y_j - y_{j'}\|. \tag{11}$$

According to the formulations above for estimating energy distance from samples, if we have n_X sample points for μ and n_Y sample points for ν, the time complexity of estimating their energy distance is $O(n_X n_Y + n_X^2 + n_Y^2)$.

The corresponding inner product between distributions μ and ν, given independent random vectors $X \sim \mu$, $Y \sim \nu$ and X', Y', is

$$\langle \mu, \nu \rangle = \mathbb{E}\|X - x_0\| + \mathbb{E}\|Y - x_0\| - \mathbb{E}\|X - Y\|, \tag{12}$$

where x_0 is a fixed point. Similarly, calculation of this inner product involves three expectations: $\mathbb{E}\|X - x_0\|$, $\mathbb{E}\|Y - x_0\|$, $\mathbb{E}\|X - Y\|$. When μ and ν are

known via their samples $\{x_i\}_{i=1}^{n_X}$ and $\{y_j\}_{j=1}^{n_Y}$, their inner product (μ, ν) is estimated by

$$\langle \hat{\mu}, \hat{\nu} \rangle = \mathbb{E}\|\hat{X} - x_0\| + \mathbb{E}\|\hat{Y} - x_0\| - \mathbb{E}\|\hat{X} - \hat{Y}\|, \tag{13}$$

where

$$\mathbb{E}\|\hat{X} - x_0\| = \frac{1}{n_X} \sum_{i=1}^{n_X} \|x_i - x_0\|, \tag{14}$$

$$\mathbb{E}\|\hat{Y} - x_0\| = \frac{1}{n_Y} \sum_{j=1}^{n_Y} \|y_j - x_0\|, \tag{15}$$

$$\mathbb{E}\|\hat{X} - \hat{Y}\| = \frac{1}{n_X n_Y} \sum_{i,j=1}^{i=n_X, j=n_Y} \|x_i - y_j\|, \tag{16}$$

and the time complexity of estimating this inner product is $O(n_X n_Y)$. If we have n sample points for both μ and ν, the time complexity is $O(n^2)$.

Notice that estimating the energy distance and the corresponding inner product from n sample points of both distributions have time complexities $O(n^2)$, which becomes computationally expensive when using a large number of sample points. In the following section, a fast algorithm for energy distance between one-dimensional distributions is introduced, making the application of energy distance much more efficient.

3.2 Fast Algorithm in One Dimension

According to eqs. (9) to (11) and (14) to (16), both the data-based computations of energy distance (8) and corresponding inner product (13) have the same complexity of evaluating

$$\mathbb{E}\|\hat{X} - \hat{Y}\| = \frac{1}{n_X n_Y} \sum_{i=1}^{n_X} \sum_{j=1}^{n_Y} \|x_i - y_j\|, \tag{17}$$

where the (x_i, y_j) are (n_X, n_Y) samples of (X, Y). Generally, the time complexity of evaluating (17) is $O(n^2)$ via Algorithm 1, which simply takes the arithmetic mean of $\|x_i - y_j\|$.

naive computation $O(n^2)$

Algorithm 1 Generic algorithm for estimating the energy distance

Input: Samples x_i, y_j of X, Y respectively.
Output: Empirical estimation of $\mathbb{E}\|X - Y\|$.

```
procedure ENERGY({x_i}, {y_j})
    sum = 0
    for all x_i do
        for all y_j do
            sum = sum + ||x_i − y_j||
        end for
    end for
    return sum / (n_X n_Y)
end procedure
```

In one-dimensional space, however, the fact that $\|\cdot\| = |\cdot|$ enables us to use the identity $|x-y| = \mathbf{1}_{x-y>0}(x-y) - \mathbf{1}_{x-y\leq 0}(x-y)$ to obtain

(*)

$$
\begin{aligned}
&\mathbb{E}\|\hat{X} - \hat{Y}\| \\
=&\frac{1}{n_X n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y}|x_i - y_j| \\
=&\frac{1}{n_X n_Y}\sum_{i=1}^{n_X}\sum_{j=1}^{n_Y}\mathbf{1}_{\{x_i-y_j>0\}}(x_i-y_j) - \mathbf{1}_{\{x_i-y_j\leq 0\}}(x_i-y_j) \\
=&\frac{1}{n_X}\sum_{i=1}^{n_X}\frac{\#\{j|y_j<x_i\} - \#\{j|y_j\geq x_i\}}{n_Y}x_i + \frac{1}{n_Y}\sum_{j=1}^{n_Y}\frac{\#\{i|x_i\leq y_j\} - \#\{i|x_i>y_j\}}{n_X}y_j,
\end{aligned}
$$

nice $\rightarrow O(n)$

where $\#\{\cdots\}$ denote the number of elements in a set.

If $\{x_i\}_{i=1}^{n_X}$ and $\{y_j\}_{j=1}^{n_Y}$ are sorted arrays, the latter expression can be calculated in the linear time $O(n_X+n_Y)$, since each of $\#\{j|y_j<x_i\}$, $\#\{j|y_j\geq x_i\}$, $\#\{i|x_i\leq y_j\}$ and $\#\{i|x_i>y_j\}$ can be calculated in linear time by merging $\{x_i\}_{i=1}^{n_X}$ and $\{y_j\}_{j=1}^{n_Y}$ into one sorted array (Algorithm 2.)

If given unsorted samples, we need to sort them before applying Algorithm 2. Feasible sorting algorithms are quick sort, which has an $O(n\log n)$ average complexity and an $O(n^2)$ worst case complexity, heap sort and merge sort, which have an $O(n\log n)$ worst case complexity. Therefore even for unsorted samples, the complexity of estimating the energy distance can be bounded by $O(n\log n)$.

message: In 1D, energy distance can be computed in $O(n\log n)$ instead of $O(n^2)$!
However, this should be true for other euclidean based metrics as well, since that it only uses (*).

Algorithm 2 Fast algorithm for estimating the energy distance in 1D

Input: Sorted samples x_i, y_j of 1D random variable X, Y respectively

Output: Empirical estimation of $\mathbb{E}\|X - Y\|$

procedure FastEnergy($\{x_i\}, \{y_j\}$)
  $\text{sum}_X = 0, \text{sum}_Y = 0, i = 1, j = 1$
  while $i \leq n_X$ **and** $j \leq n_Y$ **do**
    if $x_i \leq y_j$ **then**
      $\text{sum}_X = \text{sum}_X + \frac{(j-1)-[n_Y-(j-1)]}{n_Y} x_i$
      $i = i + 1$
    else
      $\text{sum}_Y = \text{sum}_Y + \frac{(i-1)-[n_X-(i-1)]}{n_X} y_j$
      $j = j + 1$
    end if
  end while
  if $i > n_X$ **then**
    $\text{sum}_Y = \text{sum}_Y + \sum_{k=j}^{n_Y} y_k$
  else
    $\text{sum}_X = \text{sum}_X + \sum_{k=i}^{n_X} x_k$
  end if
  return $\text{sum}_X/n_X + \text{sum}_Y/n_Y$
end procedure

4 Statistical Archetypal Analysis

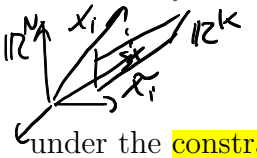
4.1 Dimensional Reduction

In this section, we study the dimensional reduction of probability distributions, mapping a collection of distributions to a low-dimensional space with minimal loss of information. Probability distributions have infinite dimension; when they are known via samples, they can be said to have a dimensionality of the order of the number of samples points. Our dimensional reduction on this high-dimensional dataset consists of two steps: we embed the distributions into an Euclidean space, and then use dimensional reduction methods developed for Euclidean spaces.

Probability distributions equipped with the energy distance form a convex subset of a Hilbert space. Therefore a collection of N distributions μ_i can

be naturally embedded into an N-dimensional Euclidean space, since every finite dimension subspace of a Hilbert space is isometric to an Euclidean space.

Assume that $x_i \in \mathbb{R}^N$, $i \in [1, 2, \cdots, N]$, are points in $\mathbb{R}^N$ such that $\|x_i - x_j\| = D(\mu_i, \mu_j)$ where $D(\cdot)$ is the energy distance (In other words, x_i is the image of μ_i under the embedding into an Euclidean space.) Principal Components Analysis (PCA) solves the following optimization problem for centered x_i:



$$\min_{z_j, a_{ji}} \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{K} a_{ji} z_j \right\|^2 \tag{18}$$

under the constraints that the z_j are orthonormal vectors. Thus PCA maps each data point to the closest point in the vector space spanned by the z_j. Here K is the dimension of the low dimensional space sought, and $\sum_{j=1}^{K} a_{ji} z_j$ is the image of x_i under this dimensional reduction.

PCA and other mainstream dimensional reduction techniques are not appropriate for probability distributions from two perspectives: 1) the z_j in eq. (18) is generally not a probability distribution, neither are almost all points in the space spanned by the z_j. 2) the coefficients a_{ji} for each x_i in eq. (18) may be negative, so they cannot clearly express how each z_j contributes to the representation of x_i. We will use instead archetypal analysis for the dimensional reduction of distributions, which does not suffer from this lack of interpretability.

4.2 Statistical Archetypal Analysis

As seen in section 2, archetypal analysis has a formulation similar to eq. (18), except that it requires the optimization of

$$\min_{z_j, a_{ji}} \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{K} a_{ji} z_j \right\|^2 \tag{19}$$

under the constraints that $a_{ji} \geqslant 0$, $\sum_{j=1}^{K} a_{ji} = 1$, and each z_j is a convex combination of the x_i, i.e. $z_j = \sum_{l=1}^{N} b_{lj} x_l$, with $b_{lj} \geqslant 0$, $\sum_{l=1}^{N} b_{lj} = 1$.

$x_i \to \mu_i$
$z_j \to \nu_j$
$\|\ \|^2 \to D(\cdot\,,\cdot)$

Switching from vectors x_i, z_j to distributions μ_i, ν_j and using energy distance instead of Euclidean distance, statistical archetypal analysis adopts the form

$$\min_{\nu_j, a_{ji}} \sum_{i=1}^{N} \left\| \mu_i - \sum_{j=1}^{K} a_{ji} \nu_j \right\|^2, \quad \nu_j = \sum_{l=1}^{N} b_{lj} \mu_l, \tag{20}$$

with the same constraints over the a and b, which now adopt the natural interpretation that the ν_j are mixtures of the μ_i and the latter are well-approximated by mixtures of the ν_j. $\|\cdot\|$ in (20) is the energy distance, but can naturally be extended to any metric induced by kernel embedding as discussed in Section 3.

Since the energy distance, that we shall use for the norm in (20), derives from an inner product, statistical archetypal analysis can be rewritten as in (3):

$$\operatorname*{argmin}_{A,B} \operatorname{tr}\left[(I_n - BA)^\intercal G (I_n - BA)\right], \tag{21}$$

where each column of A represents one archetype as a convex combination of the original distributions, and each column of B contains the coefficients for the approximate reconstruction of each original distribution from the archetypes. G is the Gram matrix of pairwise inner products among the distributions,

$$G_{ij} = \mathbb{E}k(\mathbf{X}_i, \mathbf{X}_j)$$

for independent $\mathbf{X}_i \sim \mu_i$ and $\mathbf{X}_j \sim \mu_j$ and kernel k.

When each μ_i is known via samples $\{y_m^{(i)}\}_{m=1}^{M_i}$ of size M_i, we can replace μ_i by its empirical distribution at data points $y_m^{(i)}$ with weights $\frac{1}{M_i}$. In this setting, ν_j becomes an empirical distribution concentrated at the union of the $y_m^{(l)}$ over $l = 1, 2, \cdots, N$, with weights $\frac{b_{lj}}{M_l}$ for all m. The resulting number of samples of ν_j appears large, since it contains the support of every empirical distribution μ_i. However, since the solution of eq. (19) is sparse, most entries in b_{lj} are zero, so we only need to keep those dataponts $y_m^{(l)}$ for ν_j where b_{lj} is non-zero.

Statistical archetypal analysis overcomes the two difficulties in interpretation when applying dimension reduction on probability distribution. Archetypes $\{\nu_i\}_{i=1}^{k}$ found by archetypal analysis, which are mixtures of the $\{\mu_i\}_{i=1}^{n}$, are all probability distributions. The low-dimensional space used to capture information of the dataset of distributions in this case is the convex hull of all archetypes, i.e. the family of mixtures of all archetypes. Each

coefficient a_{ji} in eq. (19) stands for the contribution of the j^{th} archetype ν_j to μ_i.

4.3 Numerical Examples

4.3.1 Synthetic Data

In our first example, we simulate 100 probability distributions $\{\mu_i\}_{i=1}^{100}$, each a Gaussian mixture $\mu_i = \lambda_i \mathcal{N}(-6, 2) + (1 - \lambda_i)\mathcal{N}(6, 1)$, where each λ_i is drawn independently from the uniform distribution in $[0, 1]$.

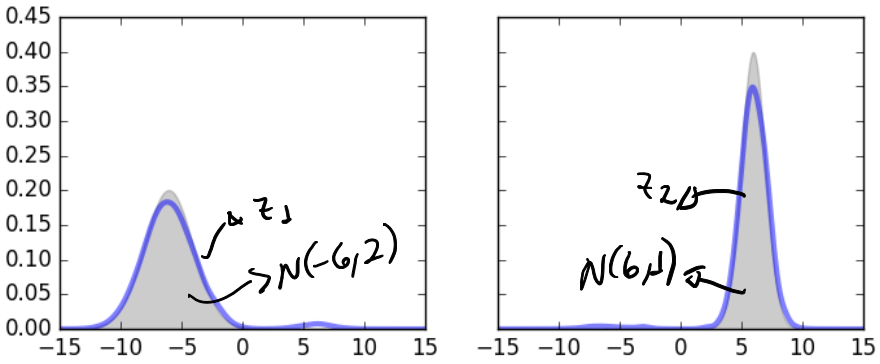


Figure 1: Archetypes of synthetic data for k=2. The curves are the found archetypes and the shadows are the two components $\mathcal{N}(-6, 2)$ and $\mathcal{N}(6, 1)$ in the mixture family respectively.

We set number of archetypes k to 2 and perform archetypal analysis on the synthetic data. The two archetypes found are shown and compared to $\mathcal{N}(-6, 2)$ and $\mathcal{N}(6, 1)$, the two components in the mixture family, in Figure 1. Both of them are close to the components except at the center and tail part. This is due to the definition of archetypes, which is a mixture of input distributions. Unless we have exactly these two components as input, the archetypes will always have a heavier tail.

4.3.2 Temperature Data

We work with ground temperature data, measured hourly in 43 cities across the United States and publicly available at the website `http://www1.`

ncdc.noaa.gov/pub/data/uscrn/products/hourly02. We operate on data from which the diurnal and seasonal signal has been removed using the optimal-transport based methodology in Tabak and Trigila [2016]. In addition, this dataset has missing values, which are filled using a low rank approximation to the data matrix. Figure 2 shows the 43 cities on the map with Table 1 a complete list of the cities.

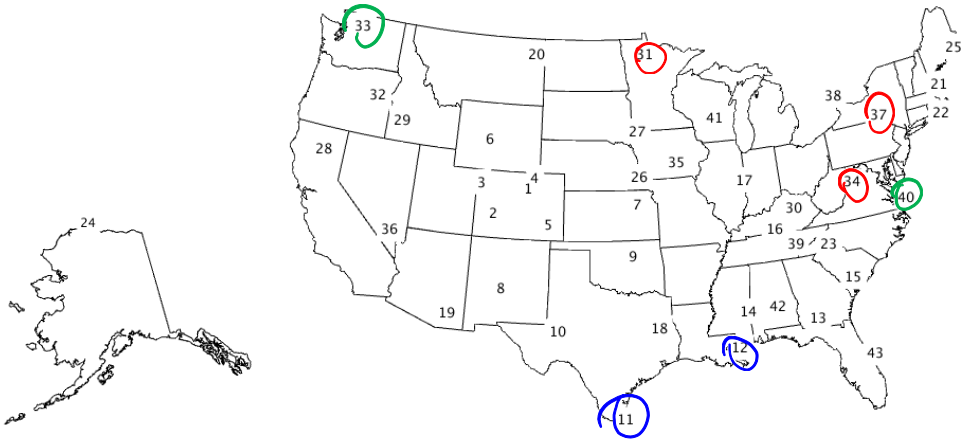


Figure 2: Locations on the map where the data were collected.

Table 1: Locations on the map where the data were collected

Index	City	Index	City	Index	City
1	Boulder	16	KY-Bowling Green	31	MN-Goodridge
2	Montrose	17	IL-Champaing	32	OR-John Day
3	Dinosaur	18	TX-Palestine	33	WA-Darrington
4	Nunn	19	AZ-Tucson	34	WV-Elkins
5	LaJunta	20	MT-Wolf Point	35	IA-Des Moines
6	Lander	21	NH-Durham	36	NV-Mercury
7	ManhattanKs	22	RI-Kingston	37	NY-Ithaca
8	Socorro	23	NC-Asheville	38	ON-Egbert
9	Stillwater	24	AK-Barrow	39	TN-Crossville
10	Monahans	25	ME-Old Town	40	VA-Cape Charles
11	Edinburg	26	NE-Lincoln	41	WI-Necedah
12	Lafayette	27	SD-Sioux Falls	42	AL-Selma
13	Newton	28	CA-Redding	43	FL-Titusville
14	MsNewton	29	ID-Murphy		
15	SC-Blackville	30	KY-Versailles		

We choose alternatively $K = 3, 5$ as the number of archetypes. For $K = 3$, the resulting archetypes are shown in Figure 3; the corresponding mixtures are as follows:

$$\begin{aligned}
\text{Archetype 1:}\quad & 0.66875 \times \text{MN-Goodridge} \\
& + 0.01916 \times \text{NY-Ithaca} + 0.31209 \times \text{WV-Elkins}, \\
\text{Archetype 2:}\quad & 0.22233 \times \text{Edinburg} + 0.77767 \times \text{Lafayette}, \\
\text{Archetype 3:}\quad & 0.01292 \times \text{VA-Cape Charles} + 0.98707 \times \text{WA-Darrington}.
\end{aligned}$$

The main difference between these three archetypes is how much they are spread. The first archetype has the heaviest tail among the three while the last archetype has the largest peak at center. The second archetype also has a marked asymmetry.

Figure 4 shows the plane spanned by these three archetypes. The bottom left cross is the first archetype, the bottom right cross is the second archetype and the top cross is the third archetype, which consists almost exclusively of the distribution at WA-Darrington. Each point represents the best approximation within the convex hull to its corresponding distribution for one city.

The approximation of distributions at each station by mixtures of archetypes are shown in Figures 5–9. We can see that, except for the distribution at Titusville, FL, the distributions at all 43 stations can be well approximated by mixtures of just three archetypes. These results indicate strongly that there is a low dimension structure underlying this dataset.

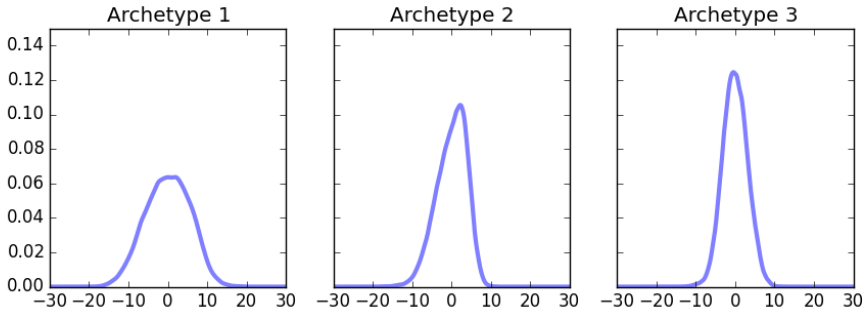


Figure 3: Archetypes of temperature data for k=3.

For K = 5, the archetypes are shown in Figure 10; the corresponding mixtures are:

Archetype 1: $0.07329 \times$ AK-Barrow $+ 0.71803 \times$ NH-Durham $+ 0.20867 \times$ RI-Kingston,

Archetype 2: $0.68181 \times$ Dinosaur $+ 0.31819 \times$ Lafayette,

Archetype 3: $0.09892 \times$ Edinburg $+ 0.90108 \times$ FL-Titusville,

Archetype 4: $0.44146 \times$ MN-Goodridge $+ 0.41827 \times$ MT-Wolf Point $+ 0.14027 \times$ ManhattanKs,

Archetype 5: $0.01306 \times$ VA-Cape Charles $+ 0.98694 \times$ WA-Darrington.

When the number of archetypes K is increased from 3 to 5, the archetypes found for $K = 3$ are not the same as for $K = 5$: only the last archetypes for $K = 3$ and $K = 5$ are close. This is due to the fact that in archetypal analysis, when the number of archetypes is increased, the shape of convex hull of archetypes changes so as to be as close to the data points as possible.

The approximation to the original distributions by mixtures of archetypes are shown in Figures 11–15. In this example, we find that when the number of archetypes is increased to 5, the mixtures of archetypes offer an almost perfect approximation to the distributions for all the 43 cities.

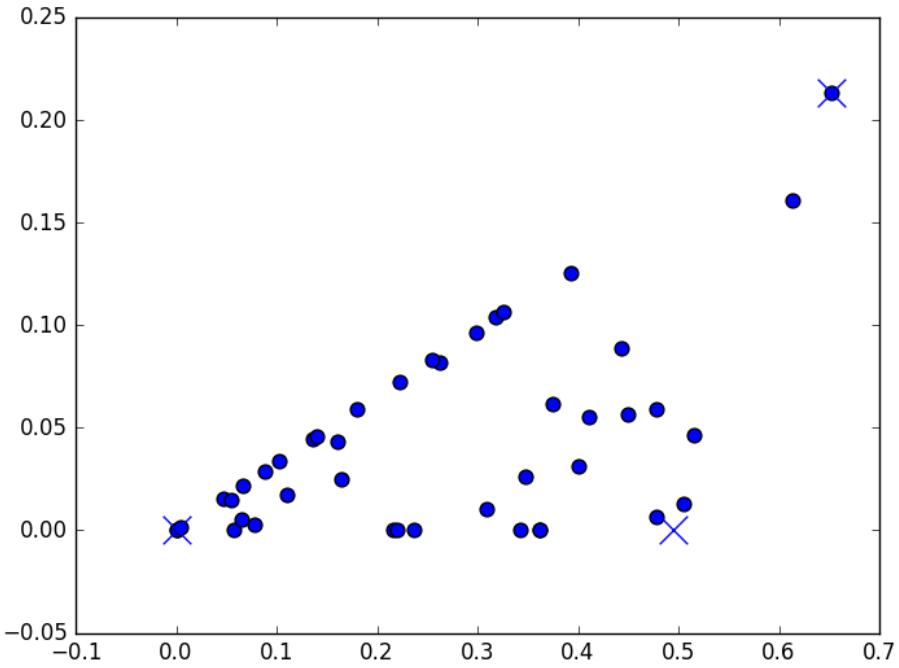


Figure 4: Convex hull spanned by archetypes of temperature data for k=3. A cross stands for one archetype and a point for the distribution at each city.

5 Conclusions

This article develops statistical archetypal analysis for dimension reduction of probability distributions. Archetypal analysis constrains the archetypes –analogues of principal components– to convex combinations of the data, and approximates the data as convex combinations of these archetypes, hence providing an interpretable fit for distributions, with patterns that can be interpreted as mixtures of distributions.

In order to perform archetypal analysis on distributions, one needs a metric and a linear structure. A natural way to introduce these is through an embedding of the distributions into a Hilbert space, for which we have used the energy distance (one of the many choices provided by the theory of reproducing kernel Hilbert spaces for distributions.)

As a proof of concept, statistical archetypal analysis was applied to both synthetic and temperature data. Statistical archetypal analysis recovers the components of a mixture family used to generate synthetic data, and reveals a low dimensional structure in the distributions of temperature data across the United States.

References

Bishop, C. M. (2006). Pattern recognition. *Machine Learning*.

Cutler, A. and Breiman, L. (1994). Archetypal analysis. *Technometrics*, 36(4):338–347.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Givens, C. R., Shortt, R. M., et al. (1984). A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, 31(2):231–240.

Klebanov, L. B. (2002). A class of probability metrics and its statistical applications. In *Statistical Data Analysis Based on the L1-Norm and Related Methods*, pages 241–252. Springer.

Kullback, S. (1968). *Information theory and statistics*. Courier Corporation.

Rachev, S. T., Klebanov, L., Stoyanov, S. V., and Fabozzi, F. (2013). *The methods of distances in the theory of probability and statistics*. Springer Science & Business Media.

Rizzo, M. L. and Székely, G. J. (2016). Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38.

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561.

Székely, G. J. and Rizzo, M. L. (2005). A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80.

Székely, G. J. and Rizzo, M. L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272.

Tabak, E. G. and Trigila, G. (2016). Explanation of variability and removal of confounding factors from data through optimal transport. In preparation.

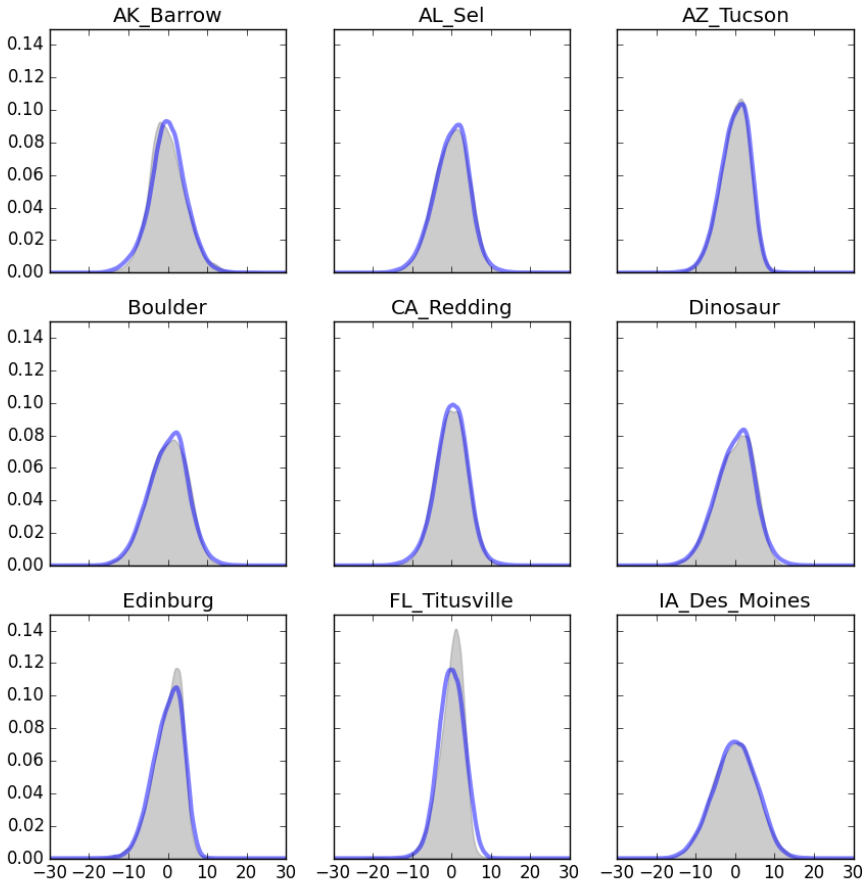


Figure 5: Reconstruction of distribution at each city by 3 archetypes. The original distributions are depicted as shadows; their approximation by mixtures of archetypes as solid curves.

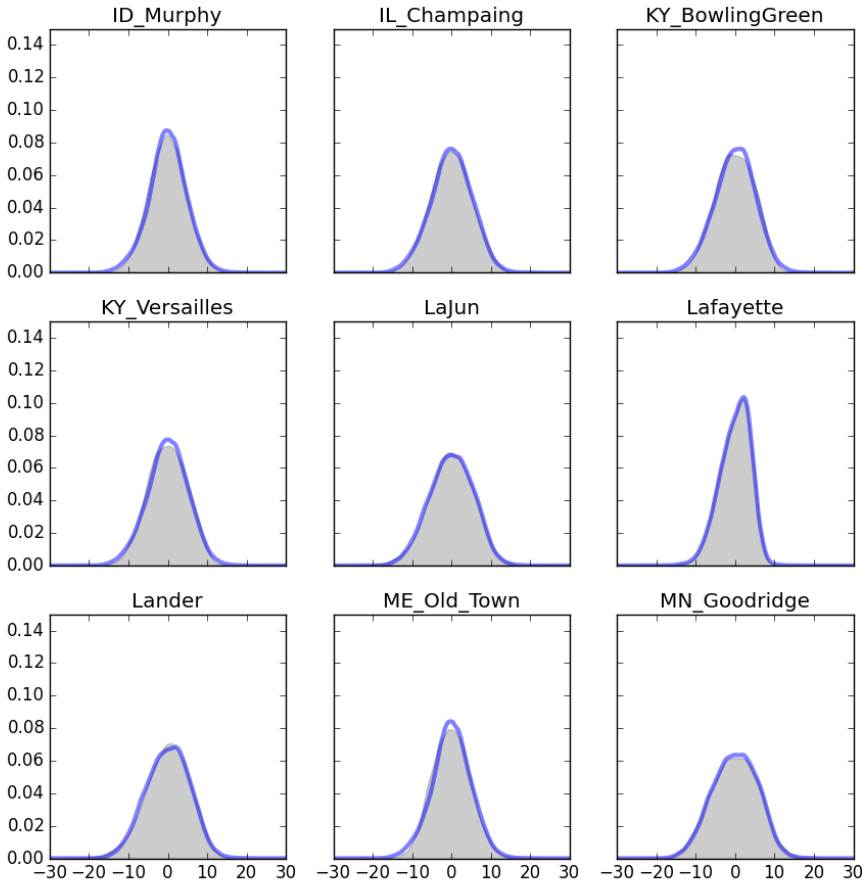


Figure 6: Reconstruction of distribution at each city by 3 archetypes. The original distributions are depicted as shadows; their approximation by mixtures of archetypes as solid curves.

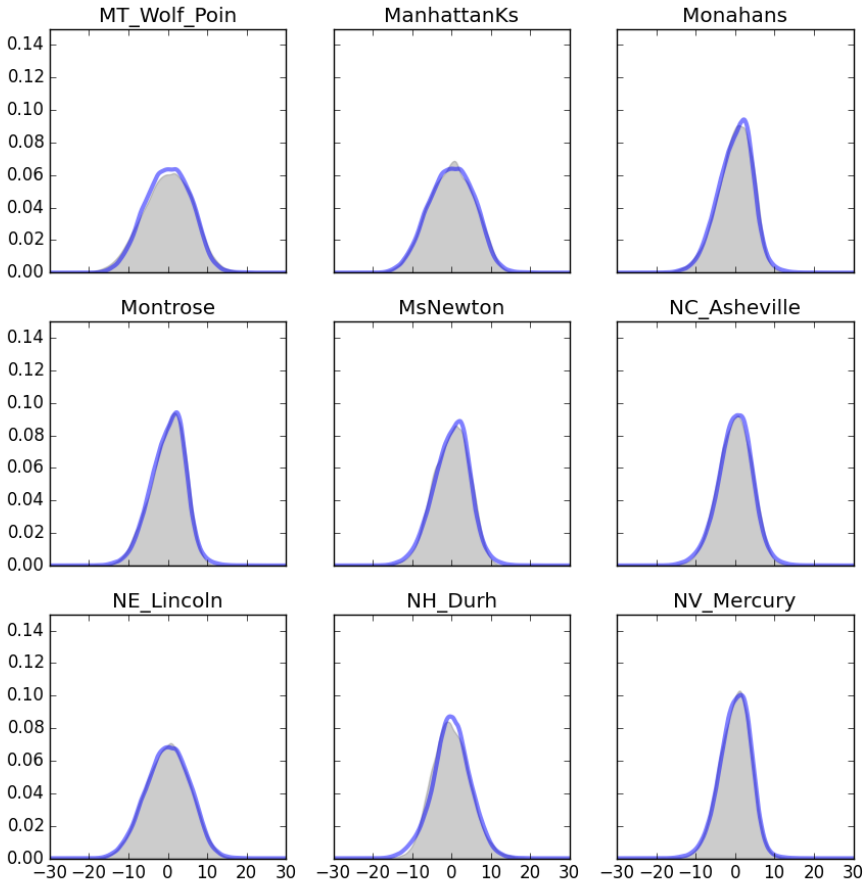


Figure 7: Reconstruction of distribution at each city by 3 archetypes. The original distributions are depicted as shadows; their approximation by mixtures of archetypes as solid curves.

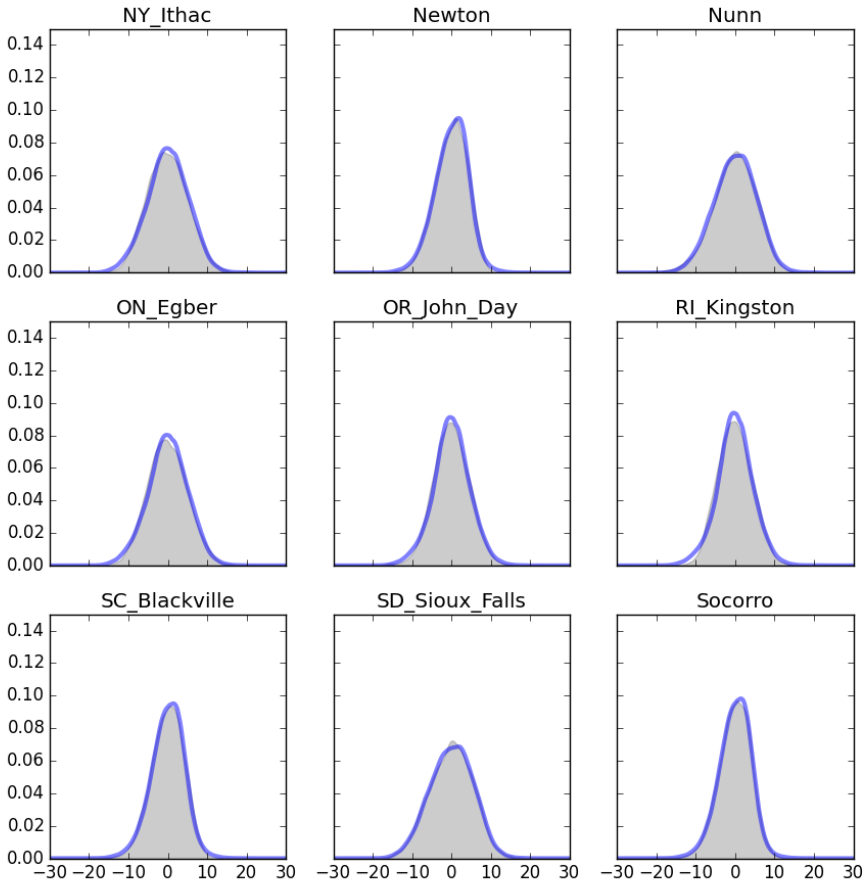


Figure 8: Reconstruction of distribution at each city by 3 archetypes. The original distributions are depicted as shadows; their approximation by mixtures of archetypes as solid curves.

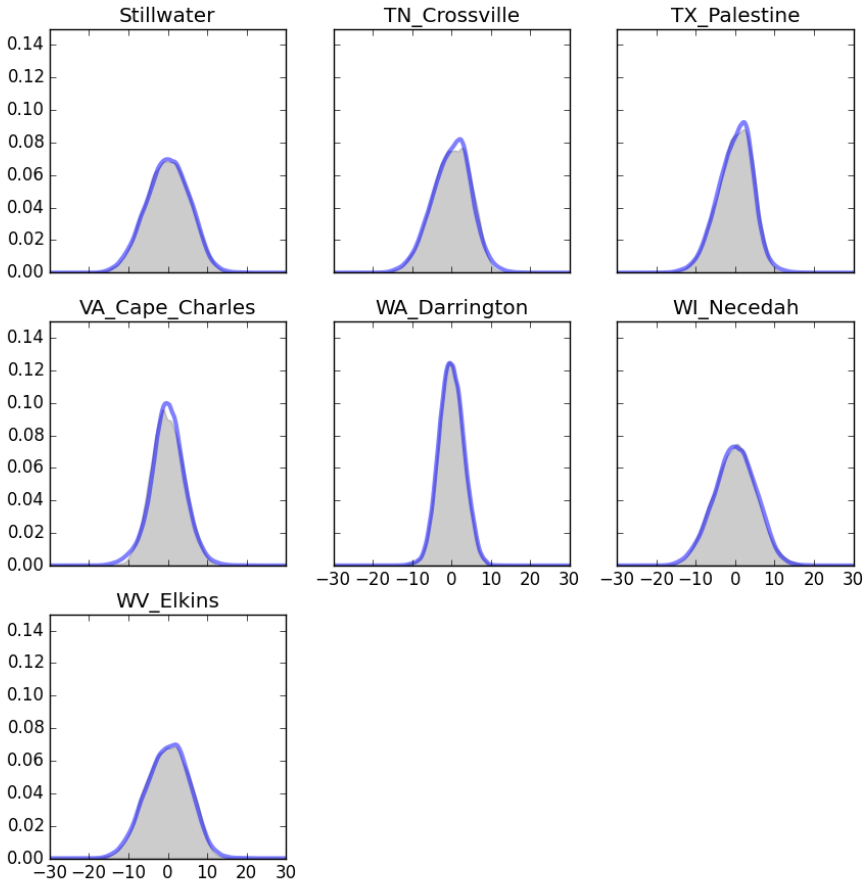


Figure 9: Reconstruction of distribution at each city by 3 archetypes. The original distributions are depicted as shadows; their approximation by mixtures of archetypes as solid curves.

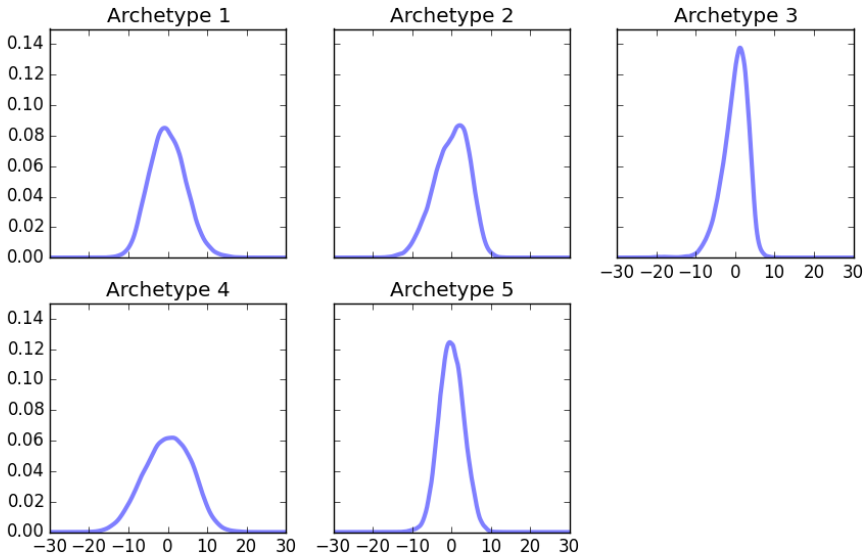


Figure 10: Archetypes of temperature data for k=5.

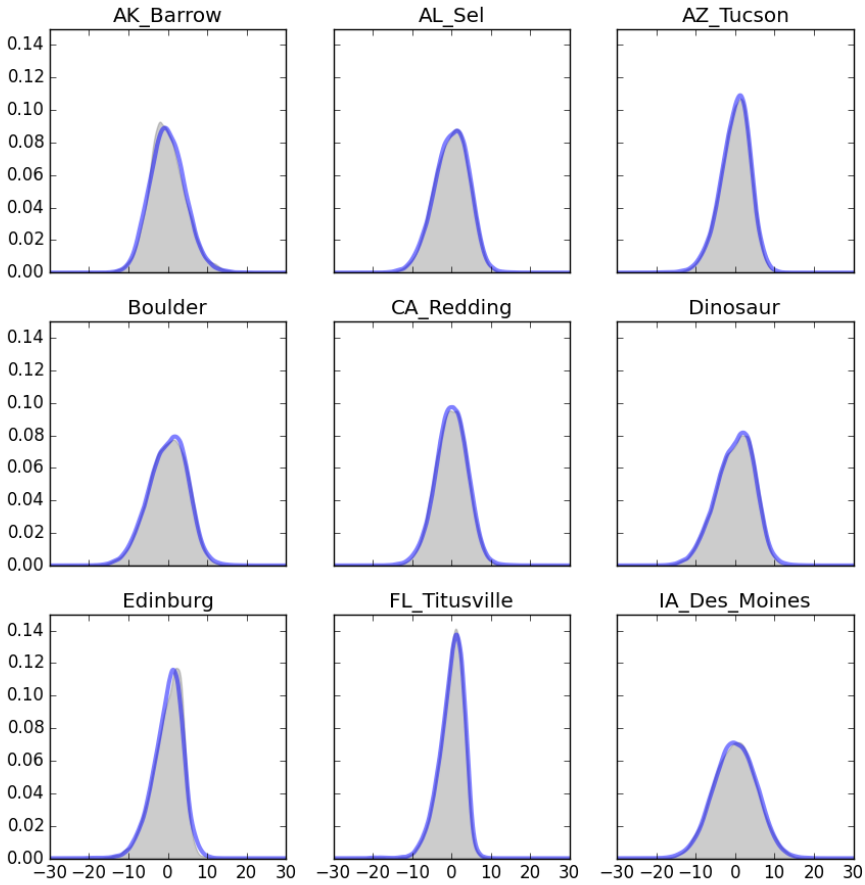


Figure 11: Reconstruction of distribution at each city by 5 archetypes. The original distributions are depicted as shadows; their approximation by mixtures of archetypes as solid curves.

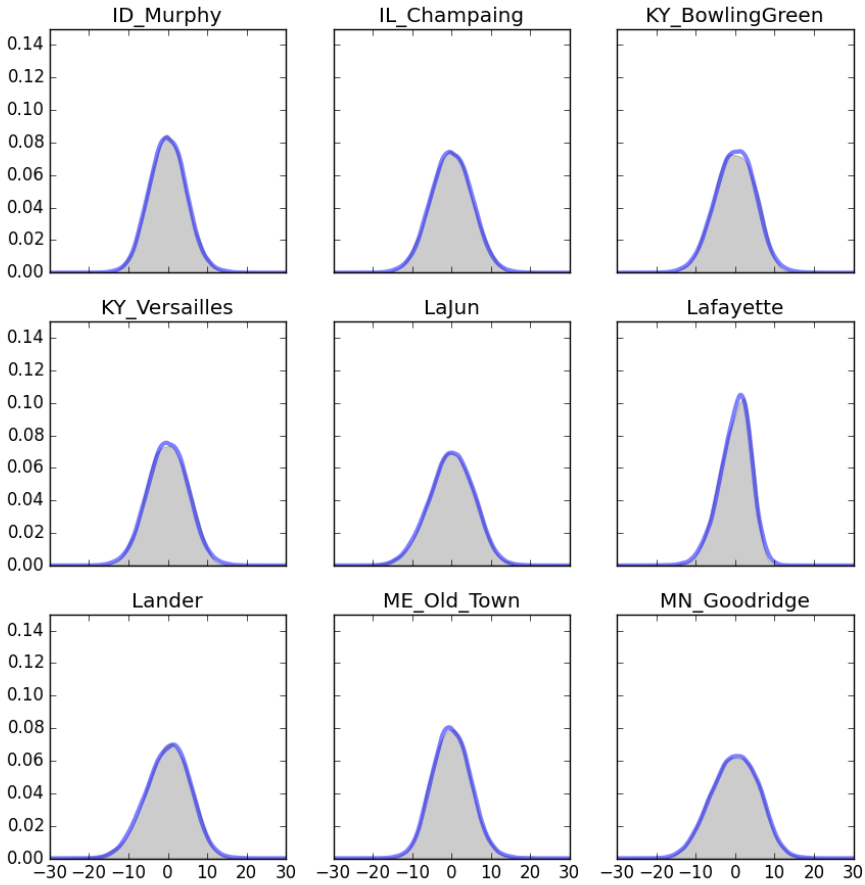


Figure 12: Reconstruction of distribution at each city by 5 archetypes. The original distributions are depicted as shadows; their approximation by mixtures of archetypes as solid curves.

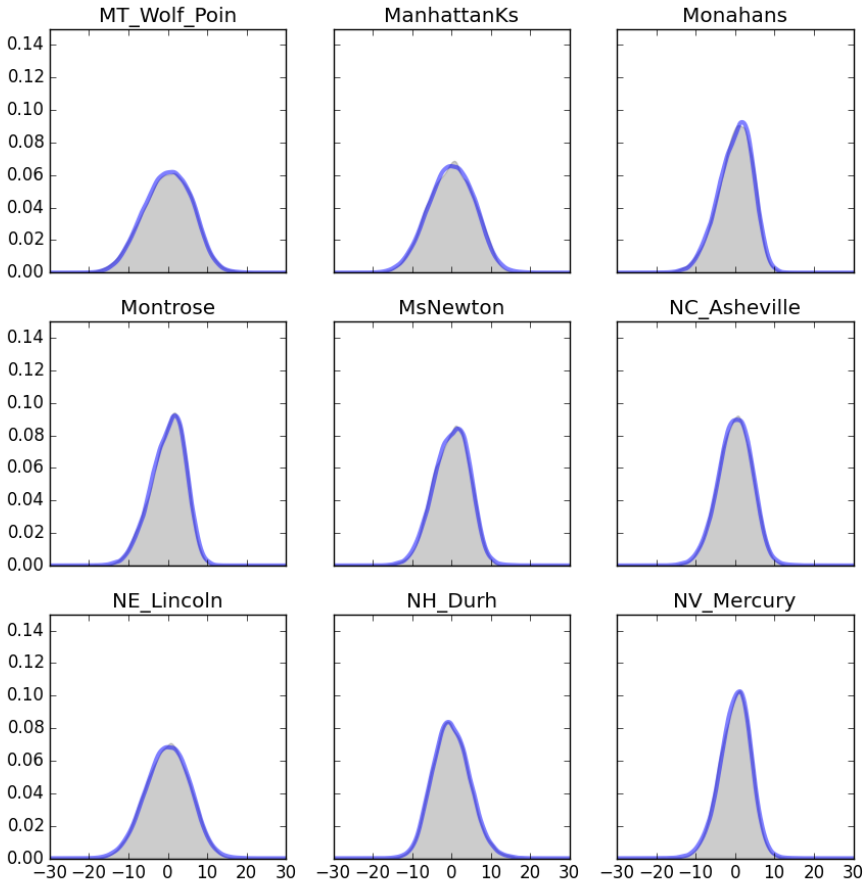


Figure 13: Reconstruction of distribution at each city by 5 archetypes. The original distributions are depicted as shadows; their approximation by mixtures of archetypes as solid curves.

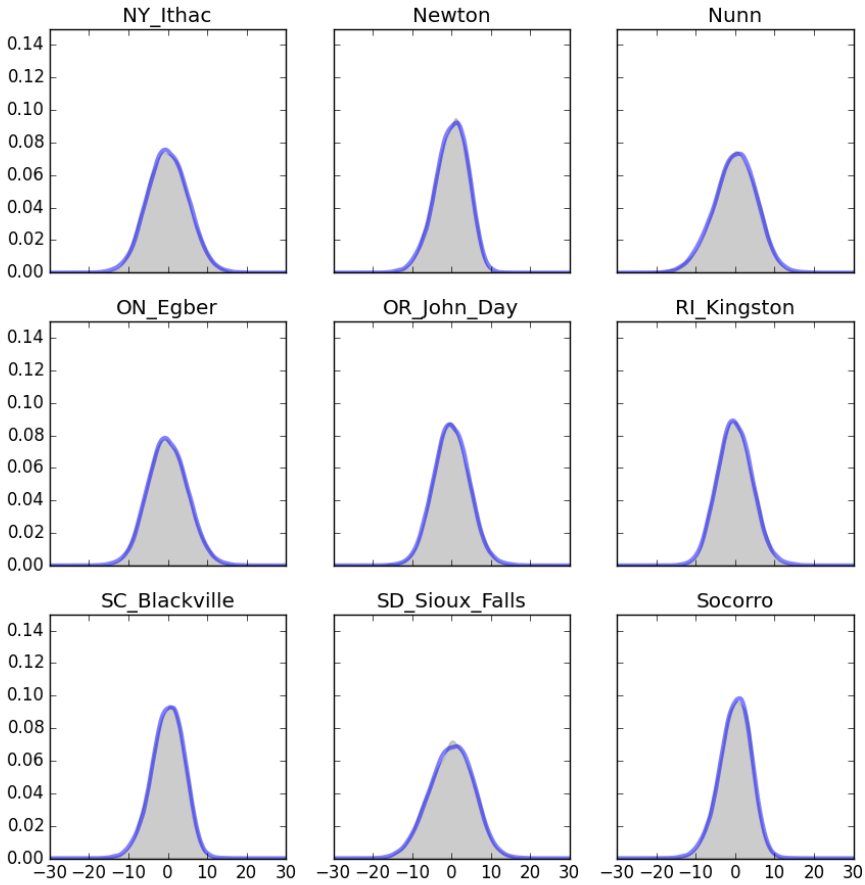


Figure 14: Reconstruction of distribution at each city by 5 archetypes. The original distributions are depicted as shadows; their approximation by mixtures of archetypes as solid curves.

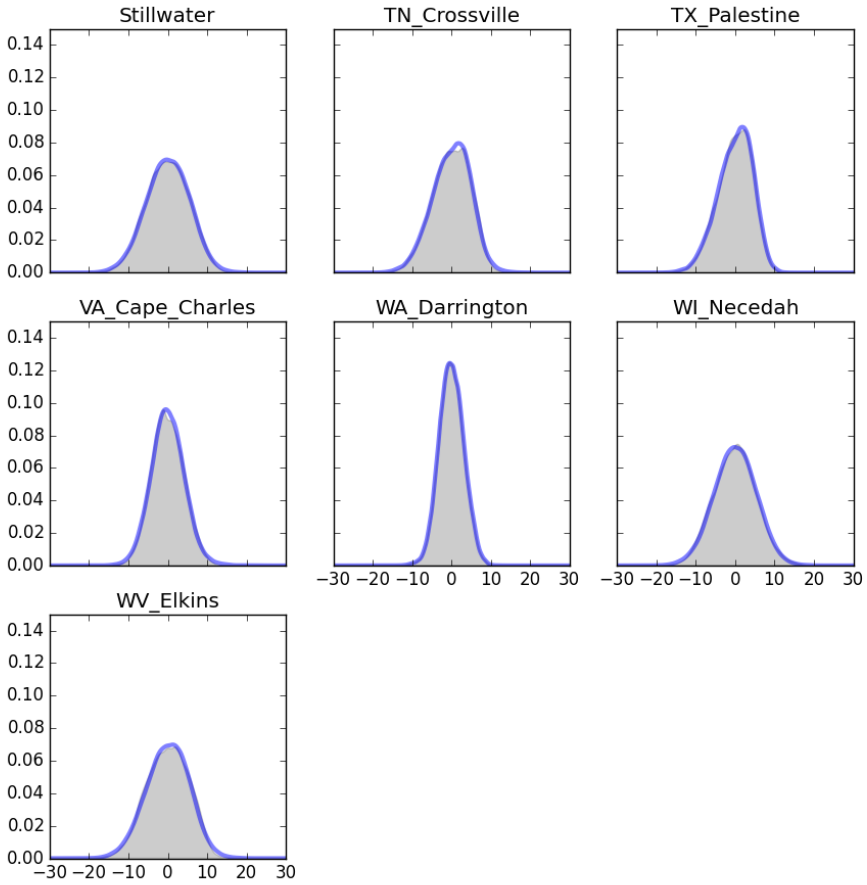


Figure 15: Reconstruction of distribution at each city by 5 archetypes. The original distributions are depicted as shadows; their approximation by mixtures of archetypes as solid curves.