

Mixture distribution

In probability and statistics, a **mixture distribution** is the probability distribution of a random variable that is derived from a collection of other random variables as follows: first, a random variable is selected by chance from the collection according to given probabilities of selection, and then the value of the selected random variable is realized. The underlying random variables may be random real numbers, or they may be random vectors (each having the same dimension), in which case the mixture distribution is a multivariate distribution.

In cases where each of the underlying random variables is continuous, the outcome variable will also be continuous and its probability density function is sometimes referred to as a **mixture density**. The cumulative distribution function (and the probability density function if it exists) can be expressed as a convex combination (i.e. a weighted sum, with non-negative weights that sum to 1) of other distribution functions and density functions. The individual distributions that are combined to form the mixture distribution are called the **mixture components**, and the probabilities (or weights) associated with each component are called the **mixture weights**. The number of components in a mixture distribution is often restricted to being finite, although in some cases the components may be countably infinite in number. More general cases (i.e. an uncountable set of component distributions), as well as the countable case, are treated under the title of compound distributions.

A distinction needs to be made between a random variable whose distribution function or density is the sum of a set of components (i.e. a mixture distribution) and a random variable whose value is the sum of the values of two or more underlying random variables, in which case the distribution is given by the convolution operator. As an example, the sum of two jointly normally distributed random variables, each with different means, will still have a normal distribution. On the other hand, a mixture density created as a mixture of two normal distributions with different means will have two peaks provided that the two means are far enough apart, showing that this distribution is radically different from a normal distribution.

Mixture distributions arise in many contexts in the literature and arise naturally where a statistical population contains two or more subpopulations. They are also sometimes used as a means of representing non-normal distributions. Data analysis concerning statistical models involving mixture distributions is discussed under the title of mixture models, while the present article concentrates on simple probabilistic and statistical properties of mixture distributions and how these relate to properties of the underlying distributions.

Contents

Finite and countable mixtures

Uncountable mixtures

Mixtures within a parametric family

Properties

Convexity

Moments

Modes

Examples

Two normal distributions

A normal and a Cauchy distribution

Applications**See also**[Mixture](#)[Hierarchical models](#)**Notes****References**

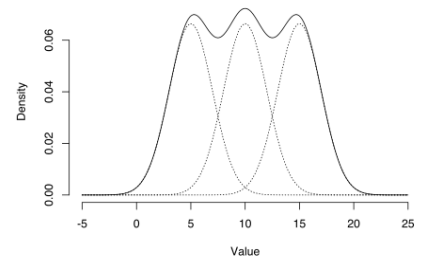
Finite and countable mixtures

Given a finite set of probability density functions $p_1(x)$, ..., $p_n(x)$, or corresponding cumulative distribution functions $P_1(x)$, ..., $P_n(x)$ and **weights** w_1 , ..., w_n such that $w_i \geq 0$ and $\sum w_i = 1$, the mixture distribution can be represented by writing either the density, f , or the distribution function, F , as a sum (which in both cases is a convex combination):

$$F(x) = \sum_{i=1}^n w_i P_i(x),$$

$$f(x) = \sum_{i=1}^n w_i p_i(x).$$

This type of mixture, being a finite sum, is called a **finite mixture**, and in applications, an unqualified reference to a "mixture density" usually means a finite mixture. The case of a countably infinite set of components is covered formally by allowing $n = \infty$.



Density of a mixture of three normal distributions ($\mu = 5, 10, 15$, $\sigma = 2$) with equal weights. Each component is shown as a weighted density (each integrating to 1/3)

Uncountable mixtures

Where the set of component distributions is uncountable, the result is often called a compound probability distribution. The construction of such distributions has a formal similarity to that of mixture distributions, with either infinite summations or integrals replacing the finite summations used for finite mixtures.

Consider a probability density function $p(x;a)$ for a variable x , parameterized by a . That is, for each value of a in some set A , $p(x;a)$ is a probability density function with respect to x . Given a probability density function w (meaning that w is nonnegative and integrates to 1), the function

$$f(x) = \int_A w(a) p(x; a) da$$

is again a probability density function for x . A similar integral can be written for the cumulative distribution function. Note that the formulae here reduce to the case of a finite or infinite mixture if the density w is allowed to be a generalized function representing the "derivative" of the cumulative distribution function of a discrete distribution.

Mixtures within a parametric family

The mixture components are often not arbitrary probability distributions, but instead are members of a parametric family (such as normal distributions), with different values for a parameter or parameters. In such cases, assuming that it exists, the density can be written in the form of a sum as:

$$f(x; a_1, \dots, a_n) = \sum_{i=1}^n w_i p(x; a_i)$$

for one parameter, or

$$f(x; a_1, \dots, a_n, b_1, \dots, b_n) = \sum_{i=1}^n w_i p(x; a_i, b_i)$$

for two parameters, and so forth.

Properties

Convexity

A general linear combination of probability density functions is not necessarily a probability density, since it may be negative or it may integrate to something other than 1. However, a convex combination of probability density functions preserves both of these properties (non-negativity and integrating to 1), and thus mixture densities are themselves probability density functions.

Moments

Let X_1, \dots, X_n denote random variables from the n component distributions, and let X denote a random variable from the mixture distribution. Then, for any function $H(\cdot)$ for which $\mathbf{E}[H(X_i)]$ exists, and assuming that the component densities $p_i(x)$ exist,

$$\begin{aligned} \mathbf{E}[H(X)] &= \int_{-\infty}^{\infty} H(x) \sum_{i=1}^n w_i p_i(x) dx \\ &= \sum_{i=1}^n w_i \int_{-\infty}^{\infty} p_i(x) H(x) dx = \sum_{i=1}^n w_i \mathbf{E}[H(X_i)]. \end{aligned}$$

The j th moment about zero (i.e. choosing $H(x) = x^j$) is simply a weighted average of the j th moments of the components. Moments about the mean $H(x) = (x - \mu)^j$ involve a binomial expansion:^[1]

$$\begin{aligned} \mathbf{E}[(X - \mu)^j] &= \sum_{i=1}^n w_i \mathbf{E}[(X - \mu_i + \mu_i - \mu)^j] \\ &= \sum_{i=1}^n w_i \sum_{k=0}^j \binom{j}{k} (\mu_i - \mu)^{j-k} \mathbf{E}[(X - \mu_i)^k], \end{aligned}$$

where μ_i denotes the mean of the i th component.

In the case of a mixture of one-dimensional distributions with weights w_i , means μ_i and variances σ_i^2 , the total mean and variance will be:

$$\mathbf{E}[X] = \mu = \sum_{i=1}^n w_i \mu_i,$$

$$\begin{aligned}
\mathbf{E}[(X - \mu)^2] &= \sigma^2 \\
&= \mathbf{E}[X^2] - \mu^2 && \text{(standard variance reformulation)} \\
&= \left(\sum_{i=1}^n w_i (\mathbf{E}[X_i^2]) \right) - \mu^2 \\
&= \sum_{i=1}^n w_i (\sigma_i^2 + \mu_i^2) - \mu^2 && \text{(from } \sigma_i^2 = \mathbf{E}[X_i^2] - \mu_i^2, \text{ therefore } \mathbf{E}[X_i^2] = \sigma_i^2 + \mu_i^2.)
\end{aligned}$$

These relations highlight the potential of mixture distributions to display non-trivial higher-order moments such as skewness and kurtosis (fat tails) and multi-modality, even in the absence of such features within the components themselves. Marron and Wand (1992) give an illustrative account of the flexibility of this framework.^[2]

Modes

The question of multimodality is simple for some cases, such as mixtures of exponential distributions: all such mixtures are unimodal.^[3] However, for the case of mixtures of normal distributions, it is a complex one. Conditions for the number of modes in a multivariate normal mixture are explored by Ray & Lindsay^[4] extending earlier work on univariate^{[5][6]} and multivariate^[7] distributions.

Here the problem of evaluation of the modes of an n component mixture in a D dimensional space is reduced to identification of critical points (local minima, maxima and saddle points) on a manifold referred to as the ridgeline surface, which is the image of the ridgeline function

$$x^*(\alpha) = \left[\sum_{i=1}^n \alpha_i \Sigma_i^{-1} \right]^{-1} \times \left[\sum_{i=1}^n \alpha_i \Sigma_i^{-1} \mu_i \right],$$

where α belongs to the $(n - 1)$ -dimensional standard simplex: $\mathcal{S}_n = \{\alpha \in \mathbb{R}^n : \alpha_i \in [0, 1], \sum_{i=1}^n \alpha_i = 1\}$

and $\Sigma_i \in \mathbb{R}^{D \times D}$, $\mu_i \in \mathbb{R}^D$ correspond to the covariance and mean of the i^{th} component. Ray & Lindsay^[4] consider the case in which $n - 1 < D$ showing a one-to-one correspondence of modes of the mixture and those on the **ridge elevation function** $h(\alpha) = q(x^*(\alpha))$ thus one may identify the modes by solving

$$\frac{dh(\alpha)}{d\alpha} = 0 \text{ with respect to } \alpha \text{ and determining the value } x^*(\alpha).$$

Using graphical tools, the potential multi-modality of mixtures with number of components $n \in \{2, 3\}$ is demonstrated; in particular it is shown that the number of modes may exceed n and that the modes may not be coincident with the component means. For two components they develop a graphical tool for analysis by instead solving the aforementioned differential with respect to the first mixing weight w_1 (which also determines the second mixing weight through $w_2 = 1 - w_1$) and expressing the solutions as a function $\Pi(\alpha)$, $\alpha \in [0, 1]$ so that the number and location of modes for a given value of w_1 corresponds to the number of intersections of the graph on the line $\Pi(\alpha) = w_1$. This in turn can be related to the number of oscillations of the graph and therefore to solutions of $\frac{d\Pi(\alpha)}{d\alpha} = 0$ leading to an explicit solution for the case of a two component mixture with $\Sigma_1 = \Sigma_2 = \Sigma$ (sometimes called a homoscedastic mixture) given by

$$1 - \alpha(1 - \alpha)d_M(\mu_1, \mu_2, \Sigma)^2$$

where $d_M(\mu_1, \mu_2, \Sigma) = \sqrt{(\mu_2 - \mu_1)^T \Sigma^{-1} (\mu_2 - \mu_1)}$ is the Mahalanobis distance between μ_1 and μ_2 .

Since the above is quadratic it follows that in this instance there are at most two modes irrespective of the dimension or the weights.

For normal mixtures with general $n > 2$ and $D > 1$, a lower bound for the maximum number of possible modes, and – conditionally on the assumption that the maximum number is finite – an upper bound are known. For those combinations of n and D for which the maximum number is known, it matches the lower bound.^[8]

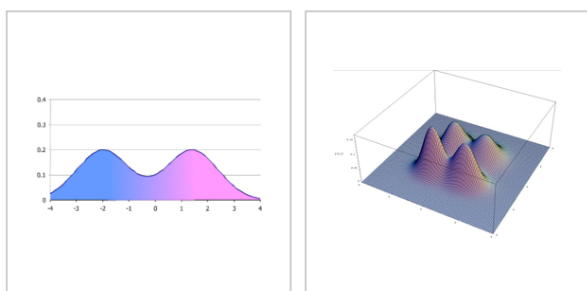
Examples

Two normal distributions

Simple examples can be given by a mixture of two normal distributions. (See Multimodal distribution#Mixture of two normal distributions for more details.)

Given an equal (50/50) mixture of two normal distributions with the same standard deviation and different means (homoscedastic), the overall distribution will exhibit low kurtosis relative to a single normal distribution – the means of the subpopulations fall on the shoulders of the overall distribution. If sufficiently separated, namely by twice the (common) standard deviation, so $|\mu_1 - \mu_2| > 2\sigma$, these form a bimodal distribution, otherwise it simply has a wide peak.^[9] The variation of the overall population will also be greater than the variation of the two subpopulations (due to spread from different means), and thus exhibits overdispersion relative to a normal distribution with fixed variation σ , though it will not be overdispersed relative to a normal distribution with variation equal to variation of the overall population.

Alternatively, given two subpopulations with the same mean and different standard deviations, the overall population will exhibit high kurtosis, with a sharper peak and heavier tails (and correspondingly shallower shoulders) than a single distribution.



Univariate mixture distribution, showing bimodal distribution Multivariate mixture distribution, showing four modes

A normal and a Cauchy distribution

The following example is adapted from Hampel,^[10] who credits John Tukey.

Consider the mixture distribution defined by

$$F(x) = (1 - 10^{-10}) (\text{standard normal}) + 10^{-10} (\text{standard Cauchy}).$$

The mean of i.i.d. observations from $F(x)$ behaves "normally" except for exorbitantly large samples, although the mean of $\bar{F}(x)$ does not even exist.

Applications

Mixture densities are complicated densities expressible in terms of simpler densities (the mixture components), and are used both because they provide a good model for certain data sets (where different subsets of the data exhibit different characteristics and can best be modeled separately), and because they can be more mathematically tractable, because the individual mixture components can be more easily studied than the overall mixture density.

Mixture densities can be used to model a statistical population with subpopulations, where the mixture components are the densities on the subpopulations, and the weights are the proportions of each subpopulation in the overall population.

Mixture densities can also be used to model experimental error or contamination – one assumes that most of the samples measure the desired phenomenon, with some samples from a different, erroneous distribution.

Parametric statistics that assume no error often fail on such mixture densities – for example, statistics that assume normality often fail disastrously in the presence of even a few outliers – and instead one uses robust statistics.

In meta-analysis of separate studies, study heterogeneity causes distribution of results to be a mixture distribution, and leads to overdispersion of results relative to predicted error. For example, in a statistical survey, the margin of error (determined by sample size) predicts the sampling error and hence dispersion of results on repeated surveys. The presence of study heterogeneity (studies have different sampling bias) increases the dispersion relative to the margin of error.

See also

- Compound distribution
- Contaminated normal distribution
- Convex combination
- Expectation-maximization (EM) algorithm
- Not to be confused with: list of convolutions of probability distributions
- Product distribution

Mixture

- Mixture (probability)
- Mixture model

Hierarchical models

- Graphical model
- Hierarchical Bayes model

Notes

1. Frühwirth-Schnatter (2006, Ch.1.2.4)

2. Marron, J. S.; Wand, M. P. (1992). "Exact Mean Integrated Squared Error" (<https://doi.org/10.1214%2Faos%2F1176348653>). *The Annals of Statistics*. **20** (2): 712–736. doi:10.1214/aos/1176348653 (<https://doi.org/10.1214%2Faos%2F1176348653>)., <http://projecteuclid.org/euclid.aos/1176348653>
3. Frühwirth-Schnatter (2006, Ch.1)
4. Ray, R.; Lindsay, B. (2005), "The topography of multivariate normal mixtures", *The Annals of Statistics*, **33** (5): 2042–2065, arXiv:math/0602238 (<https://arxiv.org/abs/math/0602238>), doi:10.1214/009053605000000417 (<https://doi.org/10.1214%2F009053605000000417>)
5. Robertson CA, Fryer JG (1969) Some descriptive properties of normal mixtures. *Skand Aktuarietidskr* 137–146
6. Behboodian, J (1970). "On the modes of a mixture of two normal distributions". *Technometrics*. **12**: 131–139. doi:10.2307/1267357 (<https://doi.org/10.2307%2F1267357>). JSTOR 1267357 (<https://www.jstor.org/stable/1267357>).
7. Carreira-Perpiñán, M Á; Williams, C (2003). *On the modes of a Gaussian mixture* (<http://faculty2.ucmerced.edu/mcarreira-perpinan/papers/EDI-INF-RR-0159.pdf>) (PDF). Published as: Lecture Notes in Computer Science 2695. Springer-Verlag. pp. 625–640. doi:10.1007/3-540-44935-3_44 (https://doi.org/10.1007%2F3-540-44935-3_44). ISSN 0302-9743 (<https://www.worldcat.org/issn/0302-9743>).
8. Améndola, C.; Engström, A.; Haase, C. (2020), "Maximum number of modes of Gaussian mixtures", *Information and Inference: A Journal of the IMA*, **9** (3): 587–600, arXiv:1702.05066 (<https://arxiv.org/abs/1702.05066>), doi:10.1093/imaia/iaz013 (<https://doi.org/10.1093%2Fimaia%2Fiaz013>)
9. Schilling, Mark F.; Watkins, Ann E.; Watkins, William (2002). "Is human height bimodal?". *The American Statistician*. **56** (3): 223–229. doi:10.1198/00031300265 (<https://doi.org/10.1198%2F00031300265>).
10. Hampel, Frank (1998), "Is statistics too difficult?", *Canadian Journal of Statistics*, **26**: 497–513, doi:10.2307/3315772 (<https://doi.org/10.2307%2F3315772>), hdl:20.500.11850/145503 (<https://hdl.handle.net/20.500.11850%2F145503>)

References

- Frühwirth-Schnatter, Sylvia (2006), *Finite Mixture and Markov Switching Models*, Springer, ISBN 978-1-4419-2194-9
- Lindsay, Bruce G. (1995), *Mixture models: theory, geometry and applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, vol. 5, Hayward, CA, USA: Institute of Mathematical Statistics, ISBN 0-940600-32-3, JSTOR 4153184 (<https://www.jstor.org/stable/4153184>)
- Seidel, Wilfried (2010), "Mixture models", in Lovric, M. (ed.), *International Encyclopedia of Statistical Science*, Heidelberg: Springer, pp. 827–829, arXiv:0909.0389 (<https://arxiv.org/abs/0909.0389>), doi:10.1007/978-3-642-04898-2 (<https://doi.org/10.1007%2F978-3-642-04898-2>), ISBN 978-3-642-04898-2

Retrieved from "https://en.wikipedia.org/w/index.php?title=Mixture_distribution&oldid=1068865381"

This page was last edited on 30 January 2022, at 15:01 (UTC).

Text is available under the Creative Commons Attribution-ShareAlike License 3.0; additional terms may apply. By using this site, you agree to the Terms of Use and Privacy Policy. Wikipedia® is a registered trademark of the Wikimedia Foundation, Inc., a non-profit organization.