Chapter 1
Gaussian Mixture Models

Abstract In this chapter we first introduce the basic concepts of random variables and the associated distributions. These concepts are then applied to Gaussian random variables and mixture-of-Gaussian random variables. Both scalar and vector-valued cases are discussed and the probability density functions for these random variables are given with their parameters specified. This introduction leads to the Gaussian mixture model (GMM) when the distribution of mixture-of-Gaussian random variables is used to fit the real-world data such as speech features. The GMM as a statistical model for Fourier-spectrum-based speech features plays an important role in acoustic modeling of conventional speech recognition systems. We discuss some key advantages of GMMs in acoustic modeling, among which is the easy way of using them to fit the data of a wide range of speech features using the EM algorithm. We describe the principle of maximum likelihood and the related EM algorithm for parameter estimation of the GMM in some detail as it is still a widely used method in speech recognition. We finally discuss a serious weakness of using GMMs in acoustic modeling for speech recognition, motivating new models and methods that form the bulk part of this book.

1.1 Random Variables

The most basic concept in probability theory and in statistics is the random variable. A scalar random variable is a real-valued function or variable which takes its value based on the outcome of a random experiment. A vector-valued random variable is a set of scalar random variables, which may either be related to or be independent of each other. Since the experiment is random, the value assumed by the random variable is random as well. A random variable can be understood as a mapping from a random experiment to a variable. Depending on the nature of the experiment and of the design of the mapping, a random variable can take either discrete values, continuous values, or a mix of discrete and continuous values. We hence see the names of discrete random variable, continuous random variable, or hybrid random variable. All possible values which may be assumed by a random variable are sometimes called its domain. In this as well as a few other later chapters, we use the same notations to describe random variables and other concepts as those adopted in [16].

The fundamental characterization of a continuous-valued random variable,x, is its distribution or the probability density function (PDF), denoted generally by $p(x)$. The PDF for a continuous random variable at $x = a$ is defined by

$$p(a) \doteq \lim_{\Delta a \to 0} \frac{P(a - \Delta a < x \leq a)}{\Delta a} \geq 0 \tag{1.1}$$

where $P(\cdot)$ denotes the probability of the event.

The cumulative distribution function of a continuous random variable x evaluated at $x = a$ is defined by

$$P(a) \doteq P(x \leq a) = \int_{-\infty}^{a} p(x)dx. \tag{1.2}$$

A PDF has to satisfy the property of normalization:

$$P(x \leq \infty) = \int_{-\infty}^{\infty} p(x)dx = 1. \tag{1.3}$$

If the normalization property is not held, we sometime call the PDF an improper density or unnormalized distribution.

For a continuous random vector $\mathbf{x} = (x_1, x_2, \ldots, x_D)^{\mathrm{T}} \in \mathcal{R}^{\mathcal{D}}$, we can similarly define their joint PDF of $p(x_1, x_2, \ldots, x_D)$. Further, a marginal PDF for each of the random variable x_i in the random vector $\mathbf{x}$ is defined by

$$p(x_i) \doteq \int \int_{all\ x_j:\ x_j \neq x_i} \cdots \int p(x_1, \ldots, x_D) dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_D. \tag{1.4}$$

It has the same properties as the PDF for a scalar random variable.

1.2 Gaussian and Gaussian-Mixture Random Variables

A scalar continuous random variable x is normally or Gaussian distributed if its PDF is

$$p(x) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \doteq \mathcal{N}(x; \mu, \sigma^2),$$
$$(-\infty < x < \infty; \sigma > 0) \tag{1.5}$$

An equivalent notation for the above is

$$x \sim \mathcal{N}(\mu, \sigma^2),$$

denoting that random variable x obeys a normal distribution with meanμ and varianceσ^2. With the use of the precision parameter, a Gaussian PDF can also be written as

$$p(x) = \sqrt{\frac{r}{2\pi}} \exp\left[-\frac{r}{2}(x-\mu)^2\right]. \tag{1.6}$$

It is a simple exercise to show that for a Gaussian random variable x, $E(x) = \mu, var(x) = \sigma^2 = r^{-1}$.

The normal random vector $\mathbf{x} = (x_1, x_2, \ldots, x_D)^{\mathrm{T}}$, also called multivariate or vector-valued Gaussian random variable, is defined by the following joint PDF:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \doteq \mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{1.7}$$

An equivalent notation is$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu} \in \mathcal{R}^D, \boldsymbol{\Sigma} \in \mathcal{R}^{D\times D})$. It is also straightforward to show that for a multivariate Gaussian random variable, the expectation and covariance matrix are given by $E(\mathbf{x}) = \boldsymbol{\mu}$; $E[(\mathbf{x}-\overline{\mathbf{x}})(\mathbf{x}-\overline{\mathbf{x}})^{\mathrm{T}}] = \boldsymbol{\Sigma}$.

The Gaussian distribution is commonly used in many engineering and science disciplines including speech recognition. The popularity arises not only from its highly desirable computational properties, but also from its ability to approximate many naturally occurring real-world data thanks to the law of large numbers.

Let us now move to discuss the Gaussian-mixture random variable with the distribution called mixture of Gaussians. A scalar continuous random variable x has a Gaussian-mixture distribution if its PDF is specified by

$$\begin{aligned} p(x) &= \sum_{m=1}^{M} \frac{c_m}{(2\pi)^{1/2}\sigma_m} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_m}{\sigma_m}\right)^2\right] \qquad (1.8)\\ &= \sum_{m=1}^{M} c_m \mathcal{N}(x;\mu_m,\sigma_m^2) \qquad (-\infty < x < \infty; \sigma_m > 0;\ c_m > 0) \end{aligned}$$

where the positive mixture weights sum to unity: $\sum_{m=1}^{M} c_m = 1$.

The most obvious property of Gaussian mixture distribution is its multimodal one ($M > 1$ in Eq. 1.8), in contrast to the uni-modal property of the Gaussian distribution where $M = 1$. This makes it possible for a mixture Gaussian distribution to adequately describe many types of physical data (including speech data) exhibiting multi-modality poorly suited for a single Gaussian distribution. The multi-modality in data may come from multiple underlying causes each being responsible for one particular mixture component in the distribution. If such causes are identified, then the mixture distribution can be decomposed into a set of cause-dependent or context-dependent component distributions.

It is easy to show that the expectation of a random variable x with the mixture Gaussian PDF of Eq. 1.8 is $E(x) = \sum_{m=1}^{M} c_m \mu_m$. But unlike a (uni-modal) Gaussian distribution, this simple summary statistic is not very informative unless all the component means, $\mu_m, m = 1, ..., M$, in the Gaussian-mixture distribution are close to each other.

The multivariate generalization of the mixture Gaussian distribution has the joint PDF of

$$\begin{aligned} p(\mathbf{x}) &= \sum_{m=1}^{M} \frac{c_m}{(2\pi)^{D/2} |\mathbf{\Sigma}_m|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_m)^{\mathrm{T}} \boldsymbol{\Sigma}_m^{-1} (\mathbf{x} - \boldsymbol{\mu}_m) \right] \\ &= \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m), \qquad (c_m > 0). \end{aligned} \tag{1.9}$$

The use of this multivariate mixture Gaussian distribution has been one key factor contributing to improved performance of many speech recognition systems (prior to the rise of deep learning); e.g., [27, 14, 23, 24]. In most applications, the number of mixture components,M, is chosen a priori according to the nature of the problem, although attempts have been made to sidestep such an often difficult problem of finding the “right” number; e.g., [31].

In using the multivariate mixture Gaussian distribution of Eq. 1.8, if the variable x’s dimensionality, D, is large (say, 40, for speech recognition problems), then the use of full (non-diagonal) covariance matrices ($\boldsymbol{\Sigma}_m$) would involve a large number of parameters (on the order of $M \times D^2$). To reduce such a number, one can opt to use diagonal covariance matrices for$\boldsymbol{\Sigma}_m$. (When M is large, one can also constrain all covariance matrices to be the same; i.e., “tying” $\boldsymbol{\Sigma}_m$ for all mixture components,m.) An additional advantage of using diagonal covariance matrices is significant simplification of computations needed for the applications of the Gaussian-mixture distributions. Reducing full covariance matrices to diagonal ones may have seemed to impose uncorrelatedness among data vector components. This has been misleading, however, since a mixture of Gaussians each with a diagonal covariance matrix can at least effectively describe the correlations modeled by one Gaussian with a full covariance matrix.

1.3 Parameter Estimation

The Gaussian-mixture distributions we just discussed contain a set of parameters. In the multivariate case of Eq. 1.8, the parameter set consists of $\boldsymbol{\Theta} = \left\{ c_m, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m \right\}$. The parameter estimation problem, also called learning, is to determine the values of these parameters from a set of data typically assumed to be drawn from the Gaussian-mixture distribution.

It is common to think of Gaussian mixture modeling and the related parameter estimation as a missing data problem. To understand this, let us assume that the data points under consideration have "membership," or the component of the mixture, in one of the individual Gaussian distributions we are using to model the data. At the start, this membership is unknown, or missing. The task of parameter estimation is to *learn* appropriate parameters for the distribution, with the connection to the data points being represented as their membership in the individual Gaussian distributions.

Here we focus on maximum likelihood methods for parameter estimation of the Gaussian-mixture distribution, and the expectation maximization (EM) algorithm in particular. The EM algorithm is the most popular technique used to estimate the parameters of a mixture given a fixed number of mixture components, and it can be used to compute the parameters of any parametric mixture distribution. It is an iterative algorithm with two steps: an expectation or E-step and a maximization or M-step. We will cover the general statistical formulation of the EM algorithm, based on [5], in more detail in Chapter **??** on hidden Markov models and here we only discuss its practical use for the parameter estimation problem related to the Gaussian mixture distribution.

The EM algorithm is of particular appeal for the Gaussian mixture distribution as the main topic of this chapter, where closed-form expressions in the M-step are available as expressed in the following iterative fashion:[1]

$$c_m^{(j+1)} = \frac{1}{N} \sum_{t=1}^{N} h_m^{(j)}(t), \tag{1.10}$$

$$\boldsymbol{\mu}_m^{(j+1)} = \frac{\sum_{t=1}^{N} h_m^{(j)}(t) \boldsymbol{x}^{(t)}}{\sum_{t=1}^{N} h_m^{(j)}(t)}, \tag{1.11}$$

$$\boldsymbol{\Sigma}_m^{(j+1)} = \frac{\sum_{t=1}^{N} h_m^{(j)}(t) [\mathbf{x}^{(t)} - \boldsymbol{\mu}_m^{(j)}][\boldsymbol{x}^{(t)} - \boldsymbol{\mu}_m^{(j)}]^T}{\sum_{t=1}^{N} h_m^{(j)}(t)}, \tag{1.12}$$

where the posterior probabilities (also called the membership responsibilities) computed from the E-step are given by

[1] (Detailed derivation of these formuli can be found in [1], which we omit here. Related derivations for similar but more general models can be found in [6, 15, 18, 2, 3].)

$$h_m^{(j)}(t) = \frac{c_m^{(j)} \mathcal{N}(\boldsymbol{x}^{(t)}; \boldsymbol{\mu}_m^{(j)}, \boldsymbol{\Sigma}_m^{(j)})}{\sum_{i=1}^{n} c_i^{(j)} \mathcal{N}(\mathbf{x}^{(t)}; \boldsymbol{\mu}_i^{(j)}, \boldsymbol{\Sigma}_i^{(j)})}. \tag{1.13}$$

That is, on the basis of the current (denoted by superscript j above) estimate for the parameters, the conditional probability for a given observation$\boldsymbol{x}^{(t)}$being generated from mixture component$m$ is determined for each data sample point at $t = 1, \ldots, N$, where N is the sample size. The parameters are then updated such that the new component weights correspond to the average conditional probability and each component mean and covariance is the component specific weighted average of the mean and covariance of the entire sample.

It has been well established that each successive EM iteration will not decrease the likelihood, a property not shared by most other gradient based maximization techniques. Further, the EM algorithm naturally embeds within it constraints on the probability vector, and for sufficiently large sample sizes positive definiteness of the covariance iterates. This is a key advantage since explicitly constrained methods incur extra computational costs to check and maintain appropriate values. Theoretically, the EM algorithm is a first-order one and as such converges slowly to a fixed-point solution. However, convergence in likelihood is rapid even if convergence in the parameter values themselves is not. Another disadvantage of the EM algorithm is its propensity to spuriously identify local maxima and its sensitivity to initial values. These problems can be addressed by evaluating EM at several initial points in the parameter space although this may become computationally costly. Another popular approach to address these issues is to start with one Gaussian component and split the Gaussian components after each epoch.

In addition to the EM algorithm discussed above for parameter estimation that is rested on maximum likelihood or data fitting, other types of estimation aimed to perform discriminative estimation or learning have been developed for Gaussian or Gaussian mixtures, as special cases of the related but more general statistical models such as the Gaussian HMM and and its Gaussian-mixture counterpart; e.g., [26, 25, 33, 22].

1.4 Mixture of Gaussians as a Model for the Distribution of Speech Features

When speech waveforms are processed into compressed (e.g., by taking logarithm of) short-time Fourier transform magnitudes or related cepstra, the Gaussian-mixture distribution discussed above is shown to be quite appropriate to fit such speech features when the information about the temporal order is discarded. That is, one can use the Gaussian-mixture distribution as a *model* to represent frame-based speech features. We use the Gaussian mixture model (GMM) to refer to the use of the Gaussian-mixture distribution for representing the data distribution. In this case and in the remainder of this book, we generally use *model* or computational model to refer to a form of mathematical abstraction of aspects of some realistic physical process (such as the human speech process), following the guiding principles detailed in [9]. Such models are established often with necessary simplification and approximation aimed at mathematical or computational tractability. The tractability is crucial in making the mathematical abstraction amenable to computer or algorithmic implementation for practical engineering applications (such as speech analysis and recognition).

Both inside and outside the speech recognition domain, the GMM is commonly used for modeling the data and for statistical classification. GMMs are well known for their ability to represent arbitrarily complex distributions with multiple modes. GMM-based classifiers are highly effective with widespread use in speech research, primarily for speaker recognition, denoising speech features, and speech recognition. For speaker recognition, the GMM is directly used as a universal background model (UBM) for the speech feature distribution pooled from all speakers [32, 28, 34, 4]. In speech feature denoising or noise tracking applications, the GMM is used in a similar way and as a prior distribution [21, 19, 11, 12, 13, 10]. In speech recognition applications, the GMM is integrated into the doubly stochastic model of HMM as its output distribution conditioned on a state, a topic which will be discussed in a great detail in Chapter **??**.

When speech sequence information is taken into account, the GMM is no longer a good model as it contains no sequence information. A class of more general models, called the hidden Markov models (HMM) to be discussed in Chapter **??**, captures the sequence information. Given a fixed state of the HMM, the GMM remains a reasonably good model for the PDF of speech feature vectors allocated to the state.

GMMs have several distinct advantages that make them suitable for modeling the PDFs over speech feature vectors associated with each state of an HMM. With enough components, they can model PDFs to any required level of accuracy, and they are easy to fit to data using the EM algorithm described in Section 1.3. A huge amount of research has gone into finding ways of constraining GMMs to increase their evaluation speed and to optimize the tradeoff between their flexibility and the amount of training data required to avoid overfitting. This includes the development of parameter-tied or semi-tied GMMs and subspace GMMs.

Beyond the use of the EM algorithm for parameter estimation of the GMM parameters, the speech recognition accuracy obtained by a GMM-based system (which is interfaced with the HMM) has been drastically improved if the GMM parameters are discriminatively learned after they have been generatively trained by EM to maximize its probability of generating the observed speech features in the training data. This is especially true if the discriminative objective function used for training is closely related to the error rate on phones, words, or sentences. The accuracy can also be improved by augmenting (or concatenating) the input speech features with tandem or bottleneck features generated using neural networks, which we will discuss in a later chapter. GMMs had been very successful in modeling speech features and in acoustic modeling for speech recognition for many years (until around year 2010-2011 when deep neural networks were shown to outperform the GMMs).

Despite all their advantages, GMMs have a serious shortcoming. That is, GMMs are statistically inefficient for modeling data that lie on or near a non-linear manifold in the data space. For example, modeling the set of points that lie very close to the surface of a sphere only requires a few parameters using an appropriate model class, but it requires a very large number of diagonal Gaussians or a fairly large number of full-covariance Gaussians. It is well known that speech is produced by modulating a relatively small number of parameters of a dynamical system [20, 29, 30, 8, 17, 7]. This suggests that the true underlying structure of speech is of a much lower dimension than is immediately apparent in a window that contains hundreds of coefficients. Therefore, other types of model which can capture better properties of speech features are expected to work better than GMMs for acoustic modeling of speech. In particular, the new models should more effectively exploit information embedded in a large window of frames of speech features than GMMs. We will return to this important problem of characterizing speech features after discussing a model, the HMM, for characterizing temporal properties of speech in the next chapter.

References

1. Bilmes, J.: A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Tech. Rep. TR-97-021, ICSI (1997)
2. Bilmes, J.: What HMMs can do. IEICE Trans. Information and Systems **E89-D**(3), 869–891 (2006)
3. Bishop, C.: Pattern Recognition and Machine Learning. Springer (2006)
4. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech and Language Processing **19**(4), 788–798 (2011)
5. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum-likelihood from incomplete data via the EM algorithm. J. Royal Statist. Soc. Ser. B. **39** (1977)
6. Deng, L.: A generalized hidden markov model with state-conditioned trend functions of time for the speech signal. Signal Processing **27**(1), 65–78 (1992)
7. Deng, L.: Computational models for speech production. In: Computational Models of Speech Pattern Processing, pp. 199–213. Springer-Verlag, New York (1999)
8. Deng, L.: Switching dynamic system models for speech articulation and acoustics. In: Mathematical Foundations of Speech and Language Processing, pp. 115–134. Springer-Verlag, New York (2003)
9. Deng, L.: DYNAMIC SPEECH MODELS — Theory, Algorithm, and Applications. Morgan and Claypool (2006)
10. Deng, L., Acero, A., Plumpe, M., Huang, X.: Large vocabulary speech recognition under adverse acoustic environment. In: Proc. International Conference on Spoken Language Processing (ICSLP), pp. 806–809 (2000)
11. Deng, L., Droppo, J., A.Acero: Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. IEEE Transactions on Speech and Audio Processing **11**, 568–580 (2003)
12. Deng, L., Droppo, J., Acero, A.: A Bayesian approach to speech feature enhancement using the dynamic cepstral prior. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 1, pp. I–829 –I–832 (2002)
13. Deng, L., Droppo, J., Acero, A.: Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. IEEE Transactions on Speech and Audio Processing **12**(2), 133 – 143 (2004)
14. Deng, L., Kenny, P., Lennig, M., Gupta, V., Seitz, F., Mermelsten, P.: Phonemic hidden markov models with continuous mixture output densities for large vocabulary word recognition. IEEE Transactions on Acoustics, Speech and Signal Processing **39**(7), 1677–1681 (1991)

15. Deng, L., Mark, J.: Parameter estimation for markov modulated poisson processes via the em algorithm with time discretization. In: Telecommunication Systems (1993)
16. Deng, L., O'Shaughnessy, D.: SPEECH PROCESSING — A Dynamic and Optimization-Oriented Approach. Marcel Dekker Inc, NY (2003)
17. Deng, L., Ramsay, G., Sun, D.: Production models as a structural basis for automatic speech recognition. Speech Communication **33**(2-3), 93–111 (1997)
18. Deng, L., Rathinavelu, C.: A Markov model containing state-conditioned second-order non-stationarity: application to speech recognition. Computer Speech and Language **9**(1), 63–86 (1995)
19. Deng, L., Wang, K., Acero, A., Hon, H., Droppo, J., C. Boulis, Y.W., Jacoby, D., Mahajan, M., Chelba, C., Huang, X.: Distributed speech processing in mipad's multimodal user interface. IEEE Transactions on Audio, Speech and Language Processing **20**(9), 2409 –2419 (2012)
20. Divenyi, P., Greenberg, S., Meyer, G.: Dynamics of Speech Production and Perception. IOS Press (2006)
21. Frey, B., Deng, L., Acero, A., Kristjansson, T.: Algonquin: Iterating laplaces method to remove multiple types of acoustic distortion for robust speech recognition. In: Proc. European Conference on Speech Communication and Technology (EUROSPEECH) (2001)
22. He, X., Deng, L.: DISCRIMINATIVE LEARNING FOR SPEECH RECOGNITION: Theory and Practice. Morgan and Claypool (2008)
23. Huang, X., Acero, A., Hon, H.W., et al.: Spoken language processing, vol. 18. Prentice Hall Englewood Cliffs (2001)
24. Huang, X., Deng, L.: An overview of modern speech recognition. In: N. Indurkhya, F.J. Damerau (eds.) Handbook of Natural Language Processing, Second Edition. CRC Press, Taylor and Francis Group, Boca Raton, FL (2010). ISBN 978-1420085921
25. Jiang, H., Li, X.: Discriminative learning in sequential pattern recognition — a unifying review for optimization-oriented speech recognition. IEEE Signal Processing Magazine **27**(3), 115–127 (2010)
26. Jiang, H., Li, X., Liu, C.: Large margin hidden markov models for speech recognition. IEEE Transactions on Audio, Speech and Language Processing **14**(5), 1584–1595 (2006)
27. Juang, B.H., Levinson, S.E., Sondhi, M.M.: Maximum likelihood estimation for mixture multivariate stochastic observations of markov chains. IEEE International Symposium on Information Theory **32**(2), 307–309 (1986)
28. Kenny, P.: Joint factor analysis of speaker and session variability: Theory and algorithms. CRIM, Montreal,(Report) CRIM-06/08-13 (2005)
29. King, S., J., F., K., L., E., M., K., R., M., W.: Speech production knowledge in automatic speech recognition. Journal Acoustical Society of America **121**, 723–742 (2007)
30. Lee, L.J., Fieguth, P., Deng, L.: A functional articulatory dynamic model for speech production. In: Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 2, pp. 797–800. Salt Lake City (2001)
31. Rasmussen, C.E.: The infinite gaussian mixture model. In: Proc. Neural Information Processing Systems (NIPS) (1999)
32. Reynolds, D., Rose, R.: Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Transactions on Speech and Audio Processing **3**(1), 72–83 (1995)
33. Xiao, L., Deng, L.: A geometric perspective of large-margin training of Gaussian models. IEEE Signal Processing Magazine **27**, 118–123 (2010)
34. Yin, S.C., Rose, R., Kenny, P.: A joint factor analysis approach to progressive model adaptation in text-independent speaker verification. IEEE Transactions on Audio, Speech, and Language Processing **15**(7), 1999–2010 (2007)