1 Normal distribution

Univariate: A random variable ξ is said to have a normal distribution with mean μ and variance σ^2 if its probability density function (pdf) is given by

$$f_\xi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \tag{1}$$

where μ is the mean and σ^2 is the variance of the distribution. More compactly, it can be written as

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

Uncorrelated multivariate: A random vector $\boldsymbol{\xi} = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_k \end{pmatrix}$ is said to have an uncorrelated multivariate normal distribution with mean $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix}$ and variances $\sigma_1^2, ..., \sigma_k^2$ if the pdf of every random component of $\boldsymbol{\xi}$ is given by

$$f_{\xi_j}(x) = \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_j}{\sigma_j}\right)^2\right\} \tag{3}$$

where μ_j is the mean and σ_j^2 is the variance of the j-th component.

All components of $\boldsymbol{\xi}$ are assumed to be independent, so the joint pdf of $\boldsymbol{\xi}$ is the product of the pdfs of its components:

$$\begin{aligned} f_{\boldsymbol{\xi}}(x_1, ..., x_k) &= \prod_{i=1}^{k} f_{\xi_i}(x_i) \\ &= \prod_{i=1}^{k} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right\} \end{aligned} \tag{4}$$

The uncorrelated multivariate normal distribution is a special case of the general multivariate normal. When components are uncorrelated, the covariance matrix is diagonal, which simplifies many calculations.

Covariance matrix: All variance parameters $\sigma_1^2, ..., \sigma_k^2$ can be combined into a covariance matrix Σ. The covariance matrix is a symmetric positive definite matrix that describes the covariance between the components of $\boldsymbol{\xi}$.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_k^2 \end{pmatrix} \tag{5}$$

Here, the covariance matrix is diagonal (all off-diagonal elements are zero), because we assumed that the components of $\boldsymbol{\xi}$ are uncorrelated, *i.e.*, $\operatorname{Cov}[\xi_i, \xi_j] = 0$ for all $i \neq j$.

The pdf of the multivariate normal distribution can be written in terms of the covariance matrix:

$$f_{\boldsymbol{\xi}}(x_1, ..., x_k) = \frac{\exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\, \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}}{\sqrt{(2\pi)^k \det \Sigma}} \tag{6}$$

The covariance matrix Σ above is a diagonal matrix, but in general, it's a symmetric positive definite matrix that describes the covariance between the components of $\boldsymbol{\xi}$:

$$\Sigma \coloneqq \begin{pmatrix} \operatorname{Cov}[\xi_1, \xi_1] & ... & \operatorname{Cov}[\xi_1, \xi_k] \\ \vdots & \ddots & \vdots \\ \operatorname{Cov}[\xi_k, \xi_1] & ... & \operatorname{Cov}[\xi_k, \xi_k] \end{pmatrix}. \tag{7}$$

If we substitute the non-diagonal covariance matrix Σ into the pdf, we get the general form of the multivariate normal distribution.

Technically, each component of Σ is the covariance between the corresponding components

$$\Sigma_{i,j} \coloneqq \operatorname{Cov}[\xi_i, \xi_j] = \mathbb{E}[(\xi_i - \mu_i)(\xi_j - \mu_j)]. \tag{8}$$

The term $\det \Sigma$ is the generalized variance.

For a sample $X = \{x_1, ..., x_\ell\} \subset \mathbb{R}$, the variance is the average of the squared differences from the mean:

$$\mathbb{D}[X] \coloneqq \frac{1}{\ell}\sum_{i=1}^{\ell}(x_i - \bar{x})^2.$$

Given another sample $Y = \{y_1, ..., y_\ell\} \subset \mathbb{R}$, the *co*-variance between two samples is characterized by how much they vary together:

$$\operatorname{Cov}[X, Y] \coloneqq \frac{1}{\ell}\sum_{i=1}^{\ell}(x_i - \bar{x}) \cdot (y_i - \bar{y}).$$

Both per sample variance and two samples covariance can be combined into a covariance matrix.

$$\Sigma = \begin{pmatrix} \operatorname{Cov}[x,x] & \operatorname{Cov}[x,y] \\ \operatorname{Cov}[y,x] & \operatorname{Cov}[y,y] \end{pmatrix} = \begin{pmatrix} \mathbb{D}[x] & \operatorname{Cov}[x,y] \\ \operatorname{Cov}[y,x] & \mathbb{D}[y] \end{pmatrix}$$

Mahalanobis distance: The distance between a point $\boldsymbol{x}$ and the distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ can be measured using the Mahalanobis distance.

The premise is that the covariance matrix Σ captures the correlations between the components of $\boldsymbol{\xi}$. The Mahalanobis distance is a measure of how many standard deviations away a point $\boldsymbol{x}$ is from the mean $\boldsymbol{\mu}$, taking into account the correlations between the components of $\boldsymbol{\xi}$.

We can define a quadratic form

$$\begin{aligned} Q(\boldsymbol{x}) &:= (\boldsymbol{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \\ &= \sum_{i,j} (x_i - \mu_i)(\Sigma^{-1})_{i,j}(x_j - \mu_j). \end{aligned} \tag{9}$$

The square root of this quadratic form $\sqrt{Q(\boldsymbol{x})}$ is the Mahalanobis distance between a point $\boldsymbol{x}$ and the distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

Quadratic form $Q(\boldsymbol{x})$ is a scalar function of a vector $\boldsymbol{x}$ that can be expressed as a weighted sum of the squares of the components of $\boldsymbol{x}$:

$$Q(\boldsymbol{x}) = \sum_{i,j} w_{i,j} x_i x_j.$$

These weights can be gathered into a matrix W, and the quadratic form can be written as a matrix product:

$$Q(\boldsymbol{x}) = \boldsymbol{x}^\top W \boldsymbol{x}.$$

2 Multivariate Normal Distribution

General form: The probability density function of the multivariate normal distribution is given by:

$$f_{\boldsymbol{\xi}}(x_1, ..., x_k) := \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\,\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}}{\sqrt{(2\pi)^k \det\Sigma}} \tag{10}$$

where:

- $\boldsymbol{\xi} = \begin{pmatrix}\xi_1\\ \vdots \\ \xi_k\end{pmatrix}$ is the random vector
- $\boldsymbol{\mu} := \mathbb{E}\boldsymbol{\xi} = \begin{pmatrix}\mathbb{E}\xi_1\\ \vdots \\ \mathbb{E}\xi_k\end{pmatrix}$ is the mean vector
- $\Sigma_{i,j} := \mathrm{Cov}\left[\xi_i, \xi_j\right] = \mathbb{E}\left[(\xi_i - \mu_i)(\xi_j - \mu_j)\right]$ is the covariance matrix (symmetric positive definite)
- $\det\Sigma$ is the generalized variance
- $Q(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\,\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})$ is a quadratic form
- $\sqrt{Q(\boldsymbol{x})}$ is the Mahalanobis distance between a point $\boldsymbol{x}$ and the distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

In compact notation, this is written as: $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Uncorrelated components: When the components of the distribution are uncorrelated: $\mathrm{Cor}\left[\xi_i, \xi_j\right] = 0 \quad \forall i \neq j$

This has geometric implications - the axes of the probability density ellipsoid are parallel to the coordinate axes.

The covariance matrix simplifies to a diagonal matrix: $\Sigma = \mathrm{diag}(\sigma_1^2, ..., \sigma_k^2)$

For uncorrelated components, the multivariate normal pdf can be expressed as a product of univariate normal pdfs:

$$\begin{aligned} f_{\boldsymbol{\xi}}(x_1, ..., x_k) &:= \frac{\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\,\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}}{\sqrt{(2\pi)^k \det\Sigma}} \\ &= \prod_{i=1}^{k} \frac{\exp\left\{-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right\}}{\sigma_i\sqrt{2\pi}} \end{aligned} \tag{11}$$

This factorization enables simpler parameter estimation methods.

Decorrelation transformation:

Decorrelation is a critical preprocessing step in many machine learning applications. It simplifies the data by removing linear dependencies between features.

For correlated components, we need to find a transformation that makes the components uncorrelated. When components of a multivariate normal distribution are correlated, the covariance matrix is not diagonal.

To decorrelate the components:

1. Apply spectral decomposition to the covariance matrix (a special case of SVD for symmetric matrices):

$\Sigma = VSV^\mathsf{T}, \quad S = \mathrm{diag}(\lambda_1^2, ..., \lambda_k^2)$

2. The quadratic form can be rewritten as:

$$\begin{aligned} Q = (\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\,\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= (\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\,(VSV^\mathsf{T})^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= (\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\,V^{-T}S^{-1}V^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= (\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\,VS^{-1}V^\mathsf{T}\,(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= \left(V^\mathsf{T}\,(\boldsymbol{x}-\boldsymbol{\mu})\right)^\mathsf{T}\,S^{-1}\left(V^\mathsf{T}\,(\boldsymbol{x}-\boldsymbol{\mu})\right) \end{aligned} \tag{12}$$

3. Define the decorrelation transformation:

$\boldsymbol{x}' := V^\mathsf{T}\,\boldsymbol{x}$

4. The transformed parameters are:

$$\boldsymbol{\mu}' = V^\top \boldsymbol{\mu}$$
$$\Sigma' = S \tag{13}$$

5. To transform parameters back to the original space:

$$\boldsymbol{\mu} = V\boldsymbol{\mu}'$$
$$\Sigma = VSV^\top \tag{14}$$

For orthogonal (rotation) matrices: $V^{-1} = V^\top$, $V^{-T} = V$

3 Parameter Estimation for Normal Distribution

Maximum likelihood estimation: When a sample is generated from a Gaussian distribution, we can estimate its parameters using maximum likelihood estimation:

$$\begin{aligned} \frac{\partial}{\partial \mu} \ln L\big(\boldsymbol{\mu}, \Sigma \mid X^\ell\big) = 0 \Rightarrow \hat{\boldsymbol{\mu}} &= \frac{1}{\ell} \sum_{\boldsymbol{x} \in X^\ell} \boldsymbol{x} \\ \frac{\partial}{\partial \Sigma} \ln L\big(\boldsymbol{\mu}, \Sigma \mid X^\ell\big) = 0 \Rightarrow \hat{\Sigma} &= \frac{1}{\ell} \sum_{\boldsymbol{x} \in X^\ell} (\boldsymbol{x} - \hat{\boldsymbol{\mu}})(\boldsymbol{x} - \hat{\boldsymbol{\mu}})^\mathsf{T} \end{aligned} \tag{15}$$

The MLE estimates are asymptotically unbiased and efficient, making them optimal for large samples. For smaller samples, however, the covariance estimate is biased.

Multicollinearity issues:

When estimating parameters from data, multicollinearity can cause problems:

- The sample covariance matrix $\hat{\Sigma}$ is constructed from ℓ rank-1 matrices, so $\operatorname{rank} \hat{\Sigma} \leq \ell$
- If the number of features exceeds the sample size, $\hat{\Sigma}$ will be singular ($\det = 0$)
- A singular covariance matrix cannot be inverted, making it impossible to evaluate the density function

Multicollinearity is a statistical phenomenon where two or more predictor variables in a model are highly correlated. This creates redundant information that doesn't contribute uniquely to the model's explanatory power.

Solutions to multicollinearity include:

1. Reducing the number of features through feature selection methods (PCA, SFS, *etc.*)
2. Increasing the sample size
3. Adding regularization to the covariance matrix:

$\hat{\Sigma}' \leftarrow \hat{\Sigma} + \tau I$

4. Assuming uncorrelated features (diagonal covariance matrix)

4 Non-parametric Density Estimation

Parametric vs non-parametric approaches:

Parametric methods for density estimation assume a specific distribution shape (*e.g.*, normal), which simplifies calculations and requires less data. However, they can be inaccurate if the actual data distribution differs significantly from the assumed form and are limited in capturing complex structures.

Non-parametric methods offer greater flexibility and can provide more accurate estimates for complex and multimodal distributions without assuming a specific form. However, they typically require larger data sets, are more computationally intensive, and their results are harder to interpret due to the lack of explicit parameters.

The choice between parametric and non-parametric methods involves a bias-variance tradeoff. Parametric methods generally have higher bias but lower variance, while non-parametric methods have lower bias but higher variance.

Empirical density estimator: The simplest non-parametric estimator for a probability density function is:

$$\hat{f}(\boldsymbol{x}) \coloneqq \frac{1}{\ell} \sum_{\boldsymbol{x}' \in X^\ell} [\![\boldsymbol{x} = \boldsymbol{x}']\!], \tag{16}$$

and for the cumulative distribution function:

$$\hat{F}(\boldsymbol{x}) \coloneqq \frac{1}{\ell} \sum_{\boldsymbol{x}' \in X^\ell} [\![x'_1 \le x_1]\!] \cdots [\![x'_k \le x_k]\!] \tag{17}$$

where:

- f is a pdf/pmf
- F is a cdf
- $\boldsymbol{x}' = (x'_1, ..., x'_k)$ is a point from the sample X^ℓ
- $\boldsymbol{x} = (x_1, ..., x_k)$ is a new point

Histogram density estimator: A more practical approach divides the feature space into bins:

$$\hat{f}(\boldsymbol{x}) \coloneqq \frac{1}{\ell} \cdot \#\big(B(\boldsymbol{x}) \cap X^\ell\big), \tag{18}$$

and for the CDF:

$$\hat{F}(\boldsymbol{x}) \coloneqq \frac{1}{\ell} \sum_{B} [\![B \le B(\boldsymbol{x})]\!] \cdot \#B \tag{19}$$

where:

- $X^\ell \subseteq \bigsqcup_B B$ is the partition into bins
- $\#B$ is the bin size
- $B(\boldsymbol{x})$ is the specific bin containing $\boldsymbol{x}$
- $B \le B'$ iff $\bigwedge_{j=1}^{k} \sup B_j \le \sup B'_j$ allows ordering of bins
- n_j is the number of bins for the jth feature
- $h_j \coloneqq \frac{f_j^{\max} - f_j^{\min}}{n}$ is the corresponding bin width

Window averaging:

Instead of bins, we can use a window function centered at each data point:

$$\hat{f}(\boldsymbol{x}) \coloneqq \frac{1}{\ell} \cdot \frac{1}{2h} \sum_{\boldsymbol{x}' \in X^\ell} \left[\!\!\left[\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|}{h} < 1 \right]\!\!\right], \tag{20}$$

The choice of window width h is critical in kernel density estimation and presents the classic bias-variance tradeoff.

and for the CDF:

$$\hat{F}(\boldsymbol{x}) \coloneqq \frac{1}{\ell} \sum_{\boldsymbol{x}' \in X^\ell} [\![\boldsymbol{x}' \le \boldsymbol{x} \oplus h]\!] \tag{21}$$

where:

- h is the window width (radius)
- $\boldsymbol{x}' \leq \boldsymbol{x}$ means all components are less or equal
- $\oplus$ means componentwise addition
- $K(r) = \frac{1}{2}[\![|r| < 1]\!]$ is a kernel function

5 Kernel Density Estimation

Parzen-Rosenblatt window method:

There are two main approaches for multivariate density estimation:

Kernel density estimation is often called the Parzen-Rosenblatt window method after its developers Emanuel Parzen and Murray Rosenblatt.

1. Product kernel approach (assumes local independence):

$$\hat{f}(\boldsymbol{x}) := \frac{1}{\ell} \sum_{\boldsymbol{x}' \in X^\ell} \prod_{j=1}^{k} \frac{1}{h_j} \cdot K\left(\frac{\boldsymbol{x}^j - \boldsymbol{x}'^j}{h_j}\right) \tag{22}$$

2. Multivariate kernel approach:

$$\hat{f}(\boldsymbol{x}) := \frac{1}{\ell} \cdot \frac{1}{V(h)} \sum_{\boldsymbol{x}' \in X^\ell} K\left(\frac{\rho(\boldsymbol{x}, \boldsymbol{x}')}{h}\right) \tag{23}$$

where:

- $K(r)$ is a kernel function satisfying:
  - $\int K(r)\,\mathrm{d}r = 1$
  - $K(r) > 0$
  - For all $r > 0$: $K(r) \searrow$ (non-increasing)
- $V(h) := \int_{-\infty}^{+\infty} K\left(\frac{\rho}{h}\right) \mathrm{d}\rho$

Kernel functions: Several standard kernel functions are used in practice:

- Epanechnikov kernel: $E(r) = \frac{3}{4}(1 - r^2)_+$
- Quartic (biweight) kernel: $Q(r) = \frac{15}{16}(1 - r^2)^2 \cdot [\![|r| > 01]\!]$
- Triangle kernel: $T(r) = (1 - |r|)_+$
- Gaussian kernel: $G(r) = \frac{1}{\sqrt{2\pi}} e^{-\frac{r^2}{2}}$
- Uniform kernel: $\Pi(r) = \frac{1}{2} \cdot [\![|r| \geq 1]\!]$

The Epanechnikov kernel is theoretically optimal in terms of mean integrated squared error, but in practice, the choice of kernel usually has less impact than the bandwidth selection.

Bandwidth selection: The optimal bandwidth can be found by minimizing the cross-validation criterion:

$$Q(h) = -\sum_{\boldsymbol{x} \in X^\ell} \ln \hat{f}\left(\boldsymbol{x} \mid X^\ell \setminus \boldsymbol{x}, h\right) \to \min_h \tag{24}$$

where:

- h is the window width (radius)
- $\hat{f}$ is the estimated density function
- The notation indicates leave-one-out cross-validation

This approach:

1. Excludes each point from the training set
2. Estimates the density at that point using the remaining points
3. Minimizes the negative log-likelihood of these estimates

Cross-validation provides a data-driven approach to bandwidth selection, avoiding both oversmoothing and undersmoothing.

Relationship to other methods:

Kernel methods form a common framework that includes:

1. Density estimation: $a_1(\boldsymbol{x}) = \frac{1}{\ell \cdot V(h)} \sum_{\boldsymbol{x}' \in X^\ell} K\left(\frac{\rho(\boldsymbol{x}, \boldsymbol{x}')}{h}\right)$
2. Classification (Parzen window): $a_2(\boldsymbol{x}) = \arg\max_{y \in Y} \sum_{\boldsymbol{x}' \in X^\ell} [\![y(\boldsymbol{x}') = y]\!] \cdot K\left(\frac{\rho(\boldsymbol{x}, \boldsymbol{x}')}{h}\right)$
3. Regression (Nadaraya-Watson): $a_3(\boldsymbol{x}) = \frac{\sum_{\boldsymbol{x}' \in X^\ell} y(\boldsymbol{x}') \cdot K\left(\frac{\rho(\boldsymbol{x}, \boldsymbol{x}')}{h}\right)}{\sum_{\boldsymbol{x}'} K\left(\frac{\rho(\boldsymbol{x}, \boldsymbol{x}')}{h}\right)}$

The connection between these methods highlights the unified theoretical foundation underlying non-parametric approaches.

In these methods:

- ρ is a distance function
- $K(\rho)$ is a similarity function (larger distance means smaller similarity)

6 Mixture Models

Definition:

A mixture model combines multiple probability distributions:

$$f(\boldsymbol{x}) \coloneqq \sum_{n=1}^{N} w_n \cdot f_n(\boldsymbol{x} \mid \boldsymbol{\theta}_n) \tag{25}$$

Mixture models provide a flexible framework that can approximate any continuous distribution with arbitrary precision.

where:

- $\boldsymbol{x} \sim \mathcal{D}_1, ..., \mathcal{D}_N$ indicates data generated by N different distributions
- $w_n \coloneqq \mathbb{P}(\boldsymbol{x} \sim \mathcal{D}_n)$ is the probability of being generated by the nth distribution
- $\sum_{n=1}^{N} w_n = 1, \quad w_n \geq 0$
- $f_n(\boldsymbol{x} \mid \boldsymbol{\theta}_n)$ is the pdf/pmf of the nth distribution

Parameter estimation: The log-likelihood function for a mixture model is:

$$l(\boldsymbol{w}, \boldsymbol{\theta}) = \sum_{\boldsymbol{x} \in X^{\ell}} \ln \sum_{n=1}^{N} w_n \cdot f_n(\boldsymbol{x} \mid \boldsymbol{\theta}_n) \to \max_{\boldsymbol{w}, \boldsymbol{\theta}} \tag{26}$$

where:

- $\sum_n w_n = 1, \quad w_n \geq 0$ are constraints
- $f_n(\boldsymbol{x} \mid \boldsymbol{\theta}_n)$ is the pdf/pmf of the nth distribution

Direct optimization is challenging because the logarithm of a sum doesn't simplify easily. The EM algorithm provides an iterative solution.

Fixed point method:

Before introducing EM, it's helpful to understand the fixed point iteration method:

The fixed point method underlies many iterative algorithms in machine learning, including the EM algorithm.

$x_{n+1} = f(x_n)$

This method converges if: $|f'(x^*)| < 1$

where x^* is the fixed point such that $f(x^*) = x^*$.

7 Expectation-Maximization Algorithm

EM algorithm:

The EM algorithm was formalized by Dempster, Laird, and Rubin in 1977, though similar approaches had been used earlier.

The EM algorithm iteratively optimizes parameters w_n and $\boldsymbol{\theta}_n$ for mixture components. Each iteration consists of two steps:

1. Expectation step (E-step): Calculate the posterior probability for each data point:

$$w'_n(\boldsymbol{x}) := \mathbb{P}[\boldsymbol{x} \sim f_n \mid \boldsymbol{x}] = \frac{w_n \cdot f_n(\boldsymbol{x} \mid \boldsymbol{\theta}_n)}{\sum_{m=1}^{N} w_m f_{m(\boldsymbol{x},\boldsymbol{\theta}_m)}} \tag{27}$$

2. Maximization step (M-step): Update parameters:

$$\boldsymbol{\theta}_n \leftarrow \arg\max_{\boldsymbol{\theta}} \sum_{\boldsymbol{x} \in X^\ell} w'_n(\boldsymbol{x}) \cdot \ln f(\boldsymbol{x} \mid \boldsymbol{\theta}_n) \tag{28}$$

$$w_n \leftarrow \frac{1}{\ell} \sum_{\boldsymbol{x} \in X^\ell} w'_n(\boldsymbol{x}) \tag{29}$$

Theoretical foundation:

The EM algorithm can be derived using the Lagrangian function and Karush-Kuhn-Tucker conditions.

The EM algorithm optimizes a Lagrangian function:

$$Q(\boldsymbol{w}, \Theta) = \sum_{\boldsymbol{x} \in X^\ell} \ln \left(\sum_{n=1}^{N} w_n \cdot f_n(\boldsymbol{x} \mid \boldsymbol{\theta}_n) \right) - \lambda \cdot \left(\sum_{n=1}^{N} w_n - 1 \right) \to \max_{\boldsymbol{w}, \Theta} \tag{30}$$

where:

- $\Theta = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N]$ is the parameters matrix
- λ is the Lagrange multiplier
- f_n is the pdf/pmf of the nth distribution
- $w_n := \mathbb{P}[\boldsymbol{x} \sim f_n]$ is the probability of coming from the nth distribution

Variants of EM:

Generalized EM (GEM) relaxes the maximization requirement:

- Standard EM: $\boldsymbol{\theta}^{(t+1)} \leftarrow \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$
- GEM: $\boldsymbol{\theta}^{(t+1)} \leftarrow \boldsymbol{\theta}^* : \ell(\boldsymbol{\theta}^*) > \ell\left(\boldsymbol{\theta}^{(t)}\right)$

Stochastic EM (SEM) optimizes parameters using sampled subsets:

- Generate samples from the estimated distributions
- Optimize parameters independently for each component
- This often accelerates convergence and can use standard maximum likelihood methods

Determining the number of components: Several approaches can determine the optimal number of distributions N in a mixture:

1. Greedy addition: Start with fewer components; add new components if the likelihood for some data points is below a threshold
2. Greedy deletion: Start with more components; remove components with small weights
3. AddDel: Combination of both approaches
4. Regularization: Use cross-entropy regularization to encourage sparsity in component weights

Information criteria like AIC or BIC can also be used to select the optimal number of components, balancing model complexity with goodness-of-fit.

Hierarchical EM: Hierarchical EM extends the standard algorithm to restore hierarchical relationships in the data. It operates by greedily adding components and splitting components with low likelihood. This approach is useful for clustering and enhancing data understanding by revealing structure where single clusters may have multiple subclusters.

8 Comparing Density Estimation Methods

Framework comparison: Different approaches to density estimation include:

1. Parametric: Assumes a specific functional form

$$\hat{f}(\boldsymbol{x}) = f(\boldsymbol{x} \mid \boldsymbol{\theta})$$

2. Non-parametric kernel: Based on local estimations around each training point

$$f(\boldsymbol{x}) = \frac{1}{\ell} \sum_{\boldsymbol{x}' \in X^{\ell}} \frac{1}{V(h)} K\left(\frac{\rho(\boldsymbol{x}, \boldsymbol{x}')}{h}\right)$$

3. Mixture models: Combines multiple distributions

$$\hat{f}(\boldsymbol{x}) = \sum_{n=1}^{N} w_n \cdot f_n(\boldsymbol{x} \mid \boldsymbol{\theta}_n)$$

These approaches represent a spectrum: mixture models generalize both parametric methods (when $N = 1$) and non-parametric approaches (when N equals the sample size).

Mixture models provide a unified framework - when $N = 1$, they reduce to parametric methods, and as N approaches the sample size, they approximate non-parametric methods.