Analysis of Big Data

Billy Fang
Instructor: Prof. Han Liu

Spring 2015

The following are notes for a course taught by Prof. Han Liu. Any errors are my own.

Contents

Introduction

We would like to analyze data that is massive, high-dimensional, and/or complex.

Perspective	Machine learning	Statistics
Foundation	**Concentration principle** (LLN)	**Likelihood principle**
Approach	risk minimization (model-free)	maximum likelihood estimation (model-based)
Goal	prediction, generalization	consistency, model selection, understanding/explaining

Definition 0.1 (General principles)**.** The following are rough descriptions of three important principles.

1. **Likelihood principle.** Everything is model-based. This allows us to derive theory, such as asymptotic theory. It also provides the sufficiency principle (data reduction) which helps cope with massive data.

2. **Concentration principle.** We assume that data are noisy signals, and we want to recover the signal given the data. This inverse problem can be approached due to the concentration phenomena, which, loosely speaking, states that averaging over many samples gets rid of the noise. See the law of large numbers (Theorem 1.6).

3. **Regularization/parsimony principle.** If two explanations are equally good at explaining a phenomenon, we prefer the simpler one. We always seek dramatically simplified models to analyze complex data.

Warning: Many techniques that are good for "big" data may not be good for "small" data, e.g. the Naive Bayes classifier.

Warning: All the simple models that we use are wrong; true models are complex. However, simple models may still be useful for inference, prediction, etc.

1 Theoretical foundation

1.1 Statistical models and parameter spaces

Definition 1.1. A **statistical model** $\mathcal{P}$ is a set of probability distributions indexed by a parameter space Θ.

$$\mathcal{P} := \{p_\theta : \theta \in \Theta\}.$$

A statistical model is called a **parametric model** if it can be indexed by a finite-dimensional parameter space Θ. If no finite-dimensional parameter space can index the model, then it is called a **nonparametric model**.

Example 1.2 (Gaussian model)**.**

$$\mathcal{P} := \left\{ p_{\mu,\sigma^2}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \mu \in \mathbb{R}, \sigma^2 > 0 \right\}.$$

Here, $\theta = (\mu, \sigma^2)^\top \in \Theta := \mathbb{R} \times \mathbb{R}_+$, so this is a parametric model.

Example 1.3 (Sobolev space)**.**

$$\mathcal{P} := \left\{ p(x) \text{ continuous density and } \int p''(t)^2 \, dt < \infty \right\}.$$

This is a nonparametric model.

1.2 Limit theorems

Definition 1.4. A sequence of random variables $(X_n)_n$ is said to **converge in probability** to a random variable X if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

We denote this by $X_n \xrightarrow{P} X$.

Definition 1.5. A sequence of random variables $(X_n)_n$ (with corresponding cumulative distribution functions F_{X_n}) is said to **converge in distribution** to a random variable X (with cdf F_X) if for every x at which F_X is continuous,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

We denote this $X_n \xrightarrow{D} X$.

Theorem 1.6 ([Weak] Law of Large Numbers)**.** *If $X_1, \ldots, X_n$ are i.i.d. random variables with expectation μ, then*

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \xrightarrow{P} \mathbb{E}[X_i] =: \mu \quad \text{as } n \to \infty.$$

Theorem 1.7 (Central Limit Theorem)**.** *If $X_1, \ldots, X_n$ are i.i.d. random variables with expectation μ and variance σ^2, then*

$$\sqrt{n}\left(\frac{\overline{X}_n - \mu}{\sigma} \right) \xrightarrow{D} \mathcal{N}(0,1) \quad \text{as } n \to \infty.$$

1.3 Estimation theory

Definition 1.8. In **point estimation** of a parameter, we are given i.i.d. random variables $X_1, \ldots, X_n$ that follow a distribution p_θ, and want to find a single best guess (**estimate**) for θ. An **estimator** is a rule for computing an estimate, given random samples $X_1, \ldots, X_n$. We treat the estimator as a function of $X_1, \ldots, X_n$ (so it is a random variable), and denote it $\widehat{\theta}_n$. We sometimes also let "estimator" refer to the sequence $(\widehat{\theta}_n)_n$ as well.

Definition 1.9.

- An estimator $(\widehat{\theta}_n)_n$ is **consistent** if $\widehat{\theta}_n \xrightarrow{P} \theta$.
- An estimator $(\widehat{\theta}_n)_n$ is **unbiased** if $\mathbb{E}[\widehat{\theta}_n] = \theta$ for all n. Otherwise, the **bias** of an estimator is $\mathbb{E}[\widehat{\theta}_n] - \theta$.

Proposition 1.10. *Neither consistency nor unbiasedness imply each other.*

Proof. Let $X_1, \dots, X_n$ be i.i.d. following the distribution $\mathcal{N}(\mu, 1)$, and consider estimation of μ. The estimator $\widehat{\mu}_n := X_1$ is unbiased, but not consistent. On the other hand, the estimator $\widehat{\mu}_n := \frac{1}{n+1}\sum_{i=1}^n X_i$ is consistent, but not unbiased.

$$\begin{aligned}
&\mathbb{P}\left(\left|\mu - \frac{1}{n+1}\sum_{i=1}^n X_i\right| > \varepsilon\right) \\
&\leq \mathbb{P}\left(\left|\frac{n+1}{n}\mu - \frac{1}{n}\sum_{i=1}^n X_i\right| > \varepsilon\right) \\
&\leq \mathbb{P}\left(\left|\mu - \frac{1}{n}\sum_{i=1}^n X_i\right| > \varepsilon/2\right) && \text{for large } n \text{ such that } |\mu|/n < \varepsilon/2 \\
&\to 0. && \text{as } n \to \infty
\end{aligned}$$

□

Proposition 1.11. *If an estimator $(\widehat{\theta}_n)_n$ is consistent*[1]*, then $\lim_{n\to\infty} \mathbb{E}[\widehat{\theta}_n] = \theta$. For this reason, we sometimes say consistent estimators are **asymptotically unbiased**.*

We will see throughout the rest of the course that unbiasedness does not necessarily make an estimator good.

1.4 Likelihood-based estimation

Definition 1.12. The **likelihood** function of θ with respect to the random sample X_i is

$$\mathcal{L}(X_i, \theta) := p_\theta(X_i).$$

Although $\mathcal{L}$ is a function of X_i and θ, we typically keep X_i fixed and think of it as a function of θ. Nonetheless, it is still a random quantity because X_i is a random variable.

Definition 1.13. The **joint likelihood function** of θ with respect to the entire set of random samples $X_1, \dots, X_n$ is

$$\mathcal{L}_n(\theta) := p_\theta(X_1, \dots, X_n).$$

Note that this definition involves a general joint distribution of the random samples. In the special case where the samples are i.i.d. following distribution p_θ, then we have

$$\mathcal{L}_n(\theta) = \prod_{i=1}^n p_\theta(X_i).$$

Definition 1.14. The **joint log-likelihood** of θ with respect to $X_1, \dots, X_n$ is

$$\ell(\theta) := \log(\mathcal{L}_n(\theta)).$$

Again, if the samples are i.i.d., then

$$\ell(\theta) = \sum_{i=1}^n \log(p_\theta(X_i)).$$

[1] Some other conditions are required. A uniform bound on variance of the $\widehat{\theta}_n$ suffices.

Because the logarithm is an increasing function,

$$\operatorname*{argmax}_{\theta\in\Theta} \mathcal{L}_n(\theta) = \operatorname*{argmax}_{\theta\in\Theta} \ell(\theta).$$

Definition 1.15. If an estimator $\widehat{\theta}$ satisfies $\mathcal{L}_n(\widehat{\theta}_n) \geq \mathcal{L}_n(\theta)$ for any $\theta \in \Theta$, then we call it a **maximum likelihood estimator (MLE)**. Note that the MLE may not be unique, but in most cases it will be. If the likelihood function attains a unique maximum (over $\theta \in \Theta$), we denote it by

$$\widehat{\theta}_n := \operatorname*{argmax}_{\theta\in\Theta} \mathcal{L}_n(\theta).$$

Example 1.16 (MLE of Gaussian distribution)**.** Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ are i.i.d. random variables, with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Again, $\theta = (\mu, \sigma^2)^\top$.

$$\mathcal{L}_n(\theta) = \prod_{i=1}^n p_\theta(X_i) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2\right)$$
$$\ell(\theta) = \sum_{i=1}^n \log(p_\theta(X_i)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\mu)^2.$$

We want to maximize $\ell(\theta)$. Because of the nature of this particular expression for $\ell(\theta)$, we may hold σ^2 fixed and maximize with respect to μ first. This reduces to minimizing $\sum_{i=1}^n (X_i-\mu)^2$. Taking the derivative with respect to μ and setting it equal to zero gives $0 = -2\sum_{i=1}^n (X_i - \mu)$. Solving gives

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i =: \overline{X},$$

which is the sample mean.

Knowing this, we can hold μ fixed at $\widehat{\mu}$ and maximize $\ell(\theta)$ with respect to $\sigma^2 > 0$. The derivative with respect to σ^2 is

$$-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (X_i - \overline{X})^2.$$

Setting this equal to zero and solving gives

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2.$$

Note that this is slightly different than the sample variance $s^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$. The sample variance s^2 is unbiased[2], so $\widehat{\sigma}^2$ is biased. This shows that the MLE is not necessarily unbiased.

We are interested in the MLE because it gives a "unified" treatment to construct estimators that are "good" in some sense. It is not necessarily the "best" estimator, but in the "big data" regime (large n), it is not bad.

[2] If μ and σ^2 are the true parameters, then

$$\begin{aligned}
\mathbb{E}[s^2] &= \frac{1}{n-1}\sum_{i=1}^n \mathbb{E}[(X_i - \overline{X})^2] \\
&= \frac{1}{n-1}\sum_{i=1}^n (\mathbb{E}[X_i^2] - 2\mathbb{E}[X_i\overline{X}] + \mathbb{E}[\overline{X}^2]) \\
&= \frac{n}{n-1}\mathbb{E}[X_1^2] - \frac{2}{n-1}(\mathbb{E}[X_1^2] + (n-1)\mu^2) + \frac{1}{n(n-1)}(n\mathbb{E}[X_1^2] + n(n-1)\mu^2) && \text{i.i.d.} \\
&= \mathbb{E}[X_1^2] - \mu^2 \\
&= \sigma^2.
\end{aligned}$$

Definition 1.17. Given a statistical model $\{p_\theta : \theta \in \Theta\}$ indexed by θ such that $\log p_\theta(x)$ is twice differentiable with respect to θ, the **Fisher information** is defined by

$$I(\theta) := -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\right] = -\int\left(\frac{\partial^2}{\partial\theta^2}\log p_\theta(x)\right)p_\theta(x)\,dx\,.$$

One can think of the Fisher information as the "expected curvature" of $\log p_\theta(X)$. As we can see below, a higher curvature corresponds with higher confidence that we have maximized $\log p_\theta(X)$, which results in a lower variance in the limiting distribution.

Theorem 1.18 (Asymptotic normality of MLE)**.** *Let θ be the true parameter. Under certain conditions,*[3] *the MLE is asymptotically normal, i.e.,*

$$\sqrt{n}(\widehat{\theta}_n^{MLE} - \theta) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{I(\theta)}\right) \quad \text{as } n \to \infty,$$

where $1/I(\theta)$ denotes the inverse of the matrix $I(\theta)$.

In addition, the variance of any unbiased estimator is at least as high as that of the MLE. That is, if $\widetilde{\theta}$ is an unbiased estimator, then in general we will have $\sqrt{n}(\widetilde{\theta}_n - \theta) \xrightarrow{D} \mathcal{N}(0, \Gamma)$ as $n \to \infty$ for some Γ; then $\Gamma \geq 1/I(\theta)$.

When applying this theorem, we often use $I(\widehat{\theta})$ or $\widehat{I}(\widehat{\theta})$ in place of $I(\theta)$, since θ is unknown.

$$I(\widehat{\theta}_n) = -\,\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)\right]\Bigg|_{\theta=\widehat{\theta}_n}$$
$$\widehat{I}(\widehat{\theta}) = -\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial^2}{\partial\theta^2}\log p_\theta(X_i)\right]\Bigg|_{\theta=\widehat{\theta}_n}$$

1.5 Likelihood-based model selection

Suppose we have i.i.d. random variables $X_1, \ldots, X_n$ that follow a completely unknown distribution. Assume we have K candidate models

$$\begin{aligned}\mathcal{M}_1 &:= \{p_{\theta_1}^{(1)}(x)\}\\ &\vdots\\ \mathcal{M}_K &:= \{p_{\theta_K}^{(K)}(x)\}.\end{aligned}$$

For example, $\mathcal{M}_1$ could be a family of Gaussian distributions, $\mathcal{M}_2$ could be a family of Poisson distributions, and so on. Our goal is to choose the model that "best" fits the data (even though it is possible that none of the models are "correct").

Definition 1.19 (Akaike Information Criterion (AIC))**.** The **AIC score** for model $\mathcal{M}_k$ is

$$\mathrm{AIC}(k) = -2\log p_{\widehat{\theta}_k}^{(k)}(X_1, \ldots, X_n) + 2d_k,$$

where $\widehat{\theta}_k$ is the MLE under model $\mathcal{M}_k$, and d_k is the number of "free parameters" in $\mathcal{M}_k$. The AIC criterion selects the model with the lowest AIC score.

The first term in the AIC score rewards the fitness of the model, while the second term penalizes the complexity of the model. It turns out that the AIC criterion chooses the model that minimizes the Kullback-Leibler divergence with respect to the true joint density.

[3] The certain conditions are that $I(\theta) > 0$ for all θ, and that the first derivative $I'(\theta)$ is continuous.

Definition 1.20. If f and g are densities, the **Kullback-Leibler divergence (KL divergence)** of g from f is defined to be

$$D(f\|g) := \int f(x) \log \frac{f(x)}{g(x)}\, dx\,.$$

We also have $D(f\|g) \geq 0$ (Gibb's inequality, follows from Jensen's inequality), with equality if and only if $f = g$. However, the KL divergence is generally not symmetric.

Proposition 1.21. *The AIC criterion chooses the model that minimizes $D(f^*\|p^{(k)}_{\widehat{\theta}_k})$, where f^* is the true density.*

Proof sketch.

$$\begin{aligned} D(f^*\|p^{(k)}_{\widehat{\theta}_k}) &= \int f^*(x) \log f^*(x)\, dx - \int f^*(x) \log p^{(k)}_{\widehat{\theta}_k}(x)\, dx \\ \operatorname*{argmin}_k D(f^*\|p^{(k)}_{\widehat{\theta}_k}) &= \operatorname*{argmin}_k \left[- \int f^*(x) \log p^{(k)(x)}_{\widehat{\theta}_k}\, dx \right] \\ &= \operatorname*{argmax}_k \underbrace{\mathbb{E}_{f^*}[\log p^{(k)}_{\widehat{\theta}_k}(X)]}_{=:J(k)} \end{aligned}$$

Let $\widehat{J}(k) := \frac{1}{n}\sum_{i=1}^n \log p^{(k)}_{\widehat{\theta}_k}(X_i)$. By the law of large numbers (Theorem 1.6), $\widehat{J}(k) \xrightarrow{P} J(k)$ as $n \to \infty$. However, $\widehat{J}(k)$ is highly biased because we use the data once to produce the MLE $\widehat{\theta}_k$, and a second time to compute $\widehat{J}$. Akaike proved that the "bias" is approximately d_k/n. To correct this, we define

$$\widetilde{J}(k) := \frac{1}{n}\sum_{i=1}^n \log p^{(k)}_{\widehat{\theta}_k}(X_i) - \frac{d_k}{n} = -\frac{\mathrm{AIC}(k)}{2n}.$$

□

Unfortunately, AIC requires many assumptions (in Akaike's proof) and works only for large n.

One way to avoid this issue is **data splitting**, in which we partition the data into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$, get MLEs $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ based only on $\mathcal{D}_1$, and pick the model that minimizes

$$\mathrm{DS}(k) := -\frac{1}{|\mathcal{D}_2|}\sum_{i\in\mathcal{D}_2} \log p^{(k)}_{\widehat{\theta}_k}(X_i).$$

This setup is unbiased because we do not use the data twice as in the AIC criterion.

Going one step farther, we have **cross-validation**, in which we partition the data into subsets $\mathcal{D}_1, \ldots, \mathcal{D}_J$. We define

$$\mathrm{CV}_j(k) := -\frac{1}{|\mathcal{D}_j|}\sum_{i\in\mathcal{D}_j} \log p^{(k)}_{\widehat{\theta}_k}(X_i),$$

where $\widehat{\theta}_k$ is the MLE based on $\mathcal{D} \setminus \mathcal{D}_j$. We then choose the model that minimizes

$$\mathrm{CV}(k) := \frac{1}{J}\sum_{i=1}^J \mathrm{CV}_j(k).$$

Definition 1.22 (Bayesian Information Criterion (BIC))**.** The **BIC score** for model $\mathcal{M}_k$ is

$$\mathrm{BIC}(k) := -2 \log p^{(k)}_{\widehat{\theta}_k}(X_1, \ldots, X_n) + (\log n) d_k.$$

The BIC criterion selects the model with the smallest BIC score.

Note that the BIC score is simply the AIC score but with the second factor of 2 replaced by $\log n$. This is a harsher penalty on the complexity of a model; in general BIC selects simpler models than AIC does.

The intuition for the BIC comes from the Bayesian approach.

$$\mathbb{P}(\mathcal{M}_j \mid X_{1:n}) = \frac{\mathbb{P}(X_{1:n} \mid \mathcal{M}_j)\mathbb{P}(\mathcal{M}_j)}{\mathbb{P}(X_{1:n})}. \qquad \text{Bayes's formula}$$

One example of a prior distribution is the uniform prior $\mathbb{P}(\mathcal{M}_1) = \cdots = \mathbb{P}(\mathcal{M}_K) = \frac{1}{K}$. In this case, the only relevant term is $\mathbb{P}(X_{1:n} \mid \mathcal{M}_j)$.[4] It turns out that the BIC score satisfies

$$2 \log \frac{\mathbb{P}(\mathcal{M}_j \mid X_{1:n})}{\mathbb{P}(\mathcal{M}_k \mid X_{1:n})} \approx \mathrm{BIC}(k) - \mathrm{BIC}(j).$$

In applications, we use AIC and cross-validation if we care more about prediction, and we use BIC if we care more about explanation or finding the "true" model. Adding junk features to the model may give AIC more predictive power even though in reality they may not have any real influence.

1.6 Sufficient statistics

Sufficient statistics are an effect approach to deal with massive data.

Definition 1.23. Data reduction is the process of minimizing the amount of data needed to be stored to do inference. There are two types of data reduction.

a) lossless (sufficient statistics, used for large n)

b) lossy (dimensionality reduction, used for large d)

Example 1.24. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d., and suppose we want to estimate (μ, σ^2). To find the MLE, we only need the first and second moments $\overline{X} := \frac{1}{n}\sum_i X_i$ and $\frac{1}{n}\sum_i X_i^2$, since

$$\widehat{\mu} = \overline{X},$$
$$\widehat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \overline{X})^2 = \frac{1}{n}\sum_i X_i^2 - \frac{2}{n}\overline{X}\sum_i X_i + \overline{X}^2 = \left(\frac{1}{n}\sum_i X_i^2\right) - \overline{X}^2.$$

Definition 1.25. A **statistic** $T(X_{1:n})$ is a function of the random samples $X_{1:n}$. A statistic is a **sufficient statistic** for the parameter θ if the conditional distribution $X_{1:n} \mid T(X_{1:n})$ does not depend on θ.

In some sense, a sufficient statistic $T(X_{1:n})$ contains all the information about θ that $X_{1:n}$ has. Note that the trivial statistic $T(X_{1:n}) := X_{1:n}$ is always sufficient.

Note that in the continuous case, we need Radon-Nikodym derivatives to justify using the density functions.

Example 1.26. If $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ be i.i.d., we claim that $T(X_{1:n}) := \overline{X}$ is sufficient.

[4] This term is not quite the likelihood; it is actually the expected likelihood.

$$\mathbb{P}(X_{1:n} \mid \mathcal{M}_j) = \int \mathbb{P}(X_{1:n} \mid \theta_j, \mathcal{M}_j)\mathbb{P}(\theta_j \mid \mathcal{M}_j)\, dP_{\theta_j}.$$

Note that $\overline{X} \sim \mathcal{N}(\mu, 1/n)$.

$$\begin{aligned} p(X_{1:n} = x_{1:n} \mid T(X_{1:n}) = T(x_{1:n})) &= \frac{p(X_{1:n} = x_{1:n}, T(X_{1:n}) = T(x_{1:n}))}{p(T(X_{1:n}) = T(x_{1:n}))} \\ &= \frac{p(X_{1:n} = x_{1:n})}{p(T(X_{1:n}) = T(x_{1:n}))} \\ &= \frac{\prod_{i=1}^{n} p(X_i = x_i)}{p(\overline{X} = \overline{x})} \\ &= \frac{(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_i (x_i - \mu)^2\right)}{\sqrt{\frac{n}{2\pi}} \exp\left(-\frac{n}{2}(\overline{x} - \mu)^2\right)} \\ &= \frac{(2\pi)^{-n/2}}{\sqrt{\frac{n}{2\pi}}} \exp\left(-\frac{1}{2}\sum_i (x_i - \overline{x})^2\right), \end{aligned}$$

where the last equality is due to

$$\begin{aligned} \sum_i (x_i - \mu)^2 &= \sum_i (x_i - \overline{x} + \overline{x} - \mu)^2 \\ &= \sum_i (x_i - \overline{x})^2 + n(\overline{x} - \mu)^2 + 2(\overline{x} - \mu)\sum_i (x_i - \overline{x}) \\ &= \sum_i (x_i - \overline{x})^2 + n(\overline{x} - \mu)^2. \end{aligned}$$

The density function for the conditional distribution does not depend on μ, so $\overline{X}$ is indeed a sufficient statistic.

Although the definition gives us a way to verify if something is a sufficient statistic, it does not provide a method to find sufficient statistics. The following theorem gives an equivalent characterization of sufficient statistics.

Theorem 1.27 (Fisher-Neyman Factorization Theorem)**.** *Let $p_\theta(X_{1:n})$ be the joint density/mass function of random samples $X_1, \ldots, X_n$. A statistic $T(X_{1:n})$ is sufficient for parameter θ if and only if there exist functions g_θ (may depend on θ) and h (free of θ) such that for all empirical realizations $x_{1:n}$ and all θ, we have*

$$p_\theta(x_{1:n}) = g_\theta(T(x_{1:n})) \cdot h(x_{1:n}).$$

Proof. We will only prove one direction. Suppose $T(X_{1:n})$ is sufficient. We define

$$\begin{aligned} g_\theta(t) &:= p_\theta(T(X_{1:n}) = t), \\ h(t) &:= p(X_{1:n} = x_{1:n} \mid T(X_{1:n}) = T(x_{1:n})). \end{aligned}$$

Note that $h(t)$ is free of θ because $T(X_{1:n})$ is a sufficient statistic. Then,

$$p_\theta(X_{1:n} = x_{1:n}) = h(x_{1:n}) \cdot g_\theta(T(x_{1:n})).$$

□

Example 1.28. Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ be i.i.d., and suppose we are doing inference on μ.

$$p_\mu(x_{1:n}) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_i (x_i - \mu)^2\right) = \underbrace{(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_i (x_i - \overline{x})^2\right)}_{h(x_{1:n})} \underbrace{\exp\left(-\frac{1}{2} n(\overline{x} - \mu)^2\right)}_{g_\mu(\overline{x})}.$$

Definition 1.29. A sufficient statistic is **minimal** if it can be represented as a function of any other sufficient statistic.

Definition 1.30. The **sufficiency principle** states that a sufficient statistic contains all information from the data relevant to inference about θ.

Definition 1.31. The **likelihood principle** states that the likelihood function contains all information from the data relevant to inference about θ.

Note that the Fisher-Neyman factorization principle shows that if we assume the likelihood principle, then the sufficiency principle follows.

2 Predictive analysis (supervised learning)

Predictive analysis techniques analyze current and past data in order to make predictions about the future. A rough overview of a typical predictive analysis process is as follows. Given training data $(Y_1, X_1), \ldots, (Y_n, X_n)$, build a prediction function $\widehat{f}$, then given a new observation x, predict $\widehat{y} := \widehat{f}(x)$.

Two learning tasks are **prediction** (given new x predict y) and **variable selection** (find a small subset of predictors that keep the most predictive power).

2.1 Regression

Regression analysis is the art of summarizing the relationship between two variables X and Y. Given observed data $(Y_1, X_1), \ldots, (Y_n, X_n) \sim P_{Y,X}$, we want a function f such that $f(X)$ is "close" to Y.

First, we need to identify the notion of closeness. Some examples of **loss functions** are L^1 loss

$$L(f(X), Y) := |f(X) - Y|,$$

and L^2 loss

$$L(f(X), Y) := |f(X) - Y|^2.$$

We will primarily study L^2 loss here because it is mathematically simple and statistically justifiable (see below).

Second, note that the loss is still a random quantity. We define the **risk** function by

$$R(f) := \mathbb{E}_{P_{Y,X}}[L(f(X), Y)] = \mathbb{E}[|Y - f(X)|^2].$$

We would like to find $f^* := \operatorname{argmin}_f R(f)$.

Theorem 2.1. *The function that minimizes the L^2 risk is the **mean function** (a.k.a. **regression function**)*

$$f^*(x) = \mathbb{E}[Y \mid X = x].$$

Proof. Define $\overline{f} := \mathbb{E}[Y \mid X = x]$. We want to show $f^* = \overline{f}$.

$$\begin{aligned} R(f) &= \mathbb{E}[|Y - f(X)|^2] \\ &= \mathbb{E}[|Y - \overline{f}(X) + \overline{f}(X) - f(X)|^2] \\ &= \mathbb{E}[|Y - \overline{f}(X)|^2] + \mathbb{E}[|\overline{f}(X) - f(X)|^2] + 2\underbrace{\mathbb{E}[(Y - \overline{f}(X))(\overline{f}(X) - f(X))]}_{=0} \end{aligned}$$

$$\implies \operatorname*{argmin}_f R(f) = \operatorname*{argmin}_f \mathbb{E}|\overline{f}(X) - f(X)|^2 = \overline{f}.$$

To show $\mathbb{E}[(Y - \overline{f}(X))(\overline{f}(X) - f(X))] = 0$, note that

$$\begin{aligned} \mathbb{E}[(Y - \overline{f}(X))(\overline{f}(X) - f(X))] &= \mathbb{E}_X[\mathbb{E}[(Y - \overline{f}(X))(\overline{f}(X) - f(X)) \mid X]] \\ &= \mathbb{E}_X[(\overline{f}(X) - f(X))\underbrace{\mathbb{E}[(Y - \overline{f}(X)) \mid X]}_{=0}] \\ &= 0 \end{aligned}$$

□

So, $\mathbb{E}[Y \mid X = x]$ minimizes L_2-loss. To minimize $R(f)$, the expectation is with respect to the true distribution $P_{Y,X}$. Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, we use the concentration principle to approximate the **population (true) risk** $R(f)$ by the **empirical risk**

$$\widehat{R}(f) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2.$$

By the law of large numbers, $\widehat{R}(f) \xrightarrow{P} R(f)$ as $n \to \infty$.

However, minimizing $\widehat{R}(f)$ is without any further assumptions on f is problematic, as *any* function f satisfying $f(X_i) = Y_i$ for each i will minimize $\widehat{R}$, regardless of how it acts outside of $X_1, \ldots, X_n$.

Definition 2.2. Overfitting is a phenomenon that occurs when a statistical model has too many parameters or degrees of freedom, so that the model not only fits the signal, but also the noise.

One solution is **regularization**, where we introduce additional constraints to control the degrees of freedom of a statistical model.

Example 2.3. We can consider functions $f(x) := \mathbb{E}[Y \mid X = x]$ that satisfy one of the following.

- $f(x) = \beta^\top x$, where $\beta \in \mathbb{R}^d$ (linear model)
- $f(x) = \text{Poly}(x)$ (polynomial model)
- f satisfies $\int (f''(x))^2 \, dx < \infty$ (nonparametric model)
- $f(x) = \beta^\top x$ where many components of β are zero (sparse linear model)

2.1.1 Ordinary least squares (OLS) regression

Let $X_i \in \{1\} \times \mathbb{R}^{d-1}$ (the first component is the bias term) and $Y_i \in \mathbb{R}$. We define

$$\widehat{\beta}^{\text{OLS}} := \operatorname*{argmin}_\beta \frac{1}{n} \sum_{i=1}^n (Y_i - \beta^\top X_i)^2.$$

We can rewrite this in vector/matrix notation. Let

$$Y := (Y_1, \ldots, Y_n)^\top,$$
$$X := [X_{i,j}] \in \mathbb{R}^{n \times d},$$

where the first column of X is $(1, \ldots, 1)^\top$. If we define $\|\beta\|_2 = \sqrt{\beta^\top \beta}$, we have

$$\widehat{\beta}^{\text{OLS}} = \operatorname*{argmin}_\beta \|Y - X\beta\|_2^2.$$

Let

$$F(\beta) := \|Y - X\beta\|_2^2 = Y^\top Y + \beta^\top X^\top X \beta - 2Y^\top X \beta.$$

Then the gradient of F is

$$\frac{\partial F(\beta)}{\partial \beta} = 2X^\top X\beta - 2X^\top Y = 0.$$

$$\widehat{\beta}^{\text{OLS}} = (X^\top X)^{-1} X^\top Y.$$

We will assume $d < n$ and that $X^\top X \in \mathbb{R}^{d \times d}$ is invertible.

We have defined $\widehat{\beta}^{\text{OLS}}$ to be the minimizer of the empirical risk. It turns out that $\widehat{\beta}^{\text{OLS}}$ also naturally appears as the MLE of the **Gaussian noise model**

$$Y = \beta^\top X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

In other words, we assume

$$P(Y, X) = P(Y \mid X) P(X)$$

where

$$Y \mid X \sim \mathcal{N}(\beta^\top X, \sigma^2),$$
$$X \sim P_X,$$

where P_X is an arbitrary distribution.

The log-likelihood is

$$\begin{aligned}\ell(\beta, \sigma^2) &= \sum_{i=1}^{n} \log p_{\beta,\sigma^2}(Y_i, X_i) \\ &= \sum_{i=1}^{n} \log p_{\beta,\sigma^2}(Y_i \mid X_i) + \sum_{i=1}^{n} \log p(X_i),\end{aligned}$$

and the MLE is

$$\begin{aligned}\widehat{\beta}^{\mathrm{MLE}} &= \operatorname*{argmax}_{\beta,\sigma^2} \ell(\beta, \sigma^2) \\ &= \operatorname*{argmax}_{\beta,\sigma^2} \sum_{i=1}^{n} \log p_{\beta,\sigma^2}(Y_i \mid X_i) \\ &= \operatorname*{argmax}_{\beta,\sigma^2} \left[-\frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} (Y_i - \beta^\top X_i)^2 \right) - n \log \sqrt{2\pi\sigma^2} \right] \\ &= \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} (Y_i - \beta^\top X_i)^2 \qquad \text{only concerned with } \beta\end{aligned}$$

This is the objective function in the definition of $\widehat{\beta}^{\mathrm{OLS}}$.

2.2 High-dimensional data analysis

High-dimensional data involves data with many features ($d > n$). When $d > n$, what happens to the OLS? The system $Y = X\beta$ becomes underdetermined, so there are infinitely many β that perfectly determine $Y = X\beta$. Also, since $\operatorname{rank}(X^\top X) = \operatorname{rank}(X) \le \min\{n, d\} \le n < d$ and $X^\top X \in \mathbb{R}^{d \times d}$, the matrix $X^\top X$ is not invertible.

We have various regularization techniques to cope with this issue.

2.2.1 Ridge estimator

The **ridge estimator** with parameter $\lambda > 0$ is defined as

$$\widehat{\beta}^{\mathrm{Ridge},\lambda} := (X^\top X + \lambda I_d)^{-1} X^\top Y.$$

The matrix $X^\top X + \lambda I_d$ is invertible because $X^\top X$ is positive semidefinite, so adding λI_d makes it positive definite and thus invertible.

Note that the ridge estimator satisfies

$$\widehat{\beta}^{\mathrm{Ridge},\lambda} = \operatorname*{argmin}_{\beta} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2.$$

Indeed, letting $F(\beta) := \|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$, we have

$$\begin{aligned}0 = \frac{\partial F(\beta)}{\partial \beta} &= -2X^\top (Y - X\beta) + 2\lambda\beta \\ (X^\top X + \lambda I_d)\beta &= X^\top Y \\ \beta &= (X^\top X + \lambda I_d)^{-1} X^\top Y.\end{aligned}$$

We can interpret the ridge estimator in yet another way by using Lagrangian duality.

Lemma 2.4. *For each $\lambda > 0$, there exists unique t such that*

$$\widehat{\beta}^{Ridge,\lambda} = \operatorname*{argmin}_{\beta} \|Y - X\beta\|_2^2 \ \textit{s.t.}\ \|\beta\|_2^2 \le t.$$

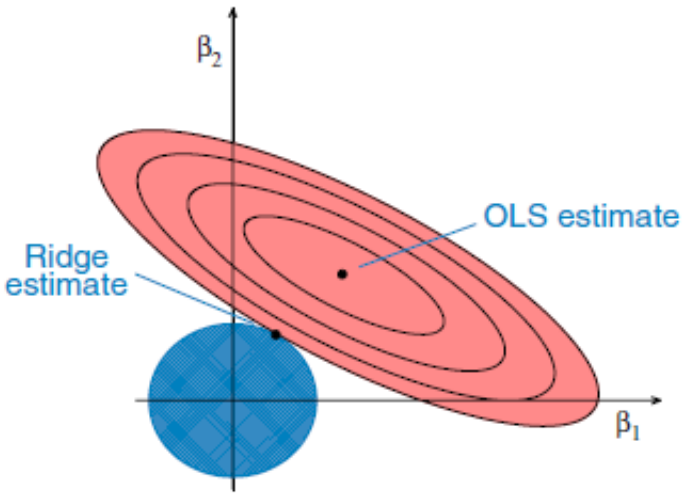

Figure 1: Ridge regression for $d = 2$. The countour lines for $\|Y - X\beta\|_2^2$ are ellipses because the objective function is quadratic in β. The minimum of the objective function is the OLS estimate, but we are restricted to the constraint region $\|\beta\|_2^2 \leq t$.

This is a "regularized" version of OLS because there is an additional constraint on the parameter β. This interpretation has a simple geometric interpretation, for example see Figure 1.

We typically use cross validation to choose the tuning parameter λ: if we select a set of candidates for λ (about thirty or so), each defines a model, so we may perform cross validation to select a model (and thus a λ).

2.2.2 Bridge estimator

The **bridge estimator** with parameters $p \in (0, \infty)$ and $\lambda > 0$ is defined as

$$\widehat{\beta}^{\text{Bridge},\lambda} := \underset{\beta}{\text{argmin}} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_p^p,$$

where $\|\beta\|_p^p := \sum_i |\beta_i|^p$. Note that the ridge estimator is the bridge estimator with $p = 2$.

We remark that if $1 \leq p < \infty$, then $\|\cdot\|_p$ is a norm, but if $0 < p < 1$, is not a norm (the triangle inequality fails).

2.2.3 Lasso estimator

The **lasso estimator** (Least Absolute Shrinkage and Selection Operator) is the bridge estimator with $p = 1$, defined as

$$\widehat{\beta}^{\text{Lasso},\lambda} := \underset{\beta}{\text{argmin}} \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1.$$

Similar to the case of the ridge estimator, Lagrangian duality implies that for each $\lambda > 0$, there exists a unique t such that

$$\widehat{\beta}^{\text{Lasso},\lambda} = \underset{\beta}{\text{argmin}} \|Y - X\beta\|_2^2 \text{ s.t. } \|\beta\|_1 \leq t.$$

See Figure 2 for a geometric interpretation when $d = 2$.

When $d = 2$, the lasso estimate tends to be a corner of constraint region, which makes one component equal to zero. In higher dimensions, the lasso estimate also tends to have several components equal to zero; this property is called **sparsity**. Sparsity helps with the task of **variable selection** and also is aligned with the parsimony principle.

Consider the different shapes of the constraint regions when p varies (Figure 3).

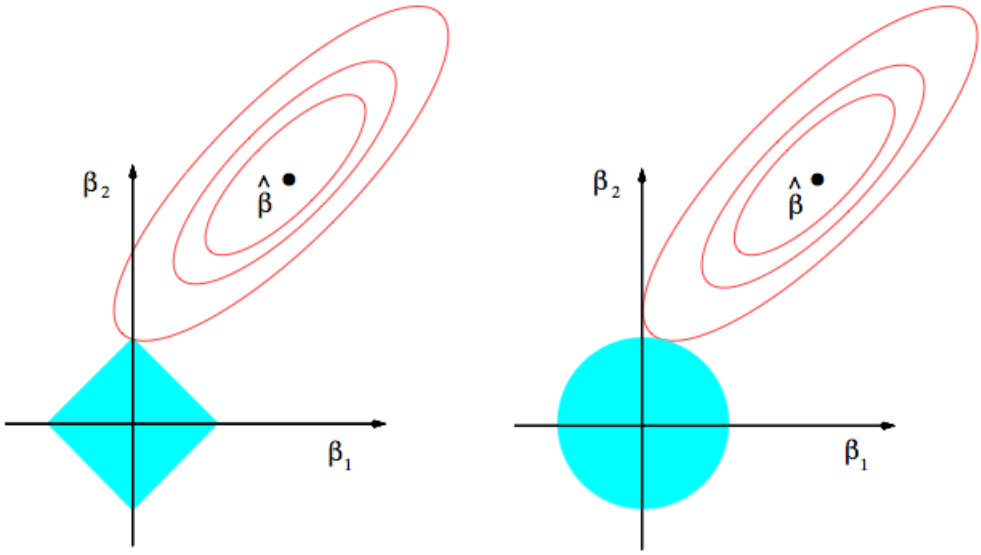

Figure 2: Lasso (left) and ridge (right) regression for $d = 2$. The countour lines for $\|Y - X\beta\|_2^2$ are ellipses because the objective function is quadratic in β. The minimum of the objective function is the OLS estimate, but we are restricted to the constraint regions $\|\beta\|_1 \leq t$ and $\|\beta\|_2^2 \leq t$.

Figure 3: The constraint region $\|\beta\|_p^p \leq t$ when $d = 3$.

We see that with $0 < p \leq 1$, we have sparsity because the estimator will tend to be on corners or edges where several components are zero. However, only for $1 \leq p < \infty$ is the objective function is convex, in which case we may use convex optimization techniques. The lasso estimator satisfies both these properties, which is why it is a particularly notable case of the bridge estimator.

Sparsity can help with prediction and cope with noise accumulation. Note that if we have an estimate $\widehat{\beta}$, our prediction function is

$$\widehat{f}(x) = \sum_{i=1}^{d} \widehat{\beta}_i x_i.$$

Each $\widehat{\beta}_i$ contributes a little bit of noise/error, but when d is large, this accumulates significantly. Sparsity limits the number of dimensions and helps avoid this issue. One might raise the question that it is bad to make the assumption that the underlying model is sparse, but even so, the estimator works well for prediction, even if the underlying model is not truly sparse.

Ridge		Lasso
not sparse	$<$	sparse, good for variable selection
closed form solution: $(X^\top X + \lambda I)^{-1} X^\top Y$	$>$	optimization: $\text{argmin}_\beta \|Y - X\beta\|_2^2 + \lambda\|\beta\|_1$
computationally difficult	$<$	computationally easy (sparsity makes optimization easier)
can handle multicollinearity	$>$	cannot handle multicollinearity

Multicollinearity occurs when predictor variables are highly correlated with each other, meaning that one can be linearly predicted from the others with reasonable accuracy. Handling multicollinearity is the

biggest advantage of ridge over lasso.

It is relatively easy to detect the existence of multicollinearity, but hard to determine the cause because there are too many possible combinations (e.g., two variables are correlated with linear combination of three other variables, etc.).

Consider $d = 2$.

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} \\ \vdots & \vdots \\ x_{n,1} & x_{n,2} \end{bmatrix}$$

If the two columns of X are highly correlated, the contour plot of $\|Y - X\beta\|_2^2$ is a very "flat" ellipse. If the contours are "parallel" to edge of the lasso constraint, the system is unstable (small change in λ result in big changes in β), and moreover the estimate may not be sparse. However, ridge regression maintains stability.

The following estimator combines the advantages of ridge and lasso.

Definition 2.5. The **elastic net** estimator is defined by

$$\widehat{\beta}^{\text{Elastic},\lambda,\alpha} := \underset{\beta}{\text{argmin}} \|Y - X\beta\|_2^2 + \lambda(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2).$$

Note that when $\alpha = 1$, we have the lasso estimator, and when $\alpha = 0$, we have the ridge estimator. By default, typically use $\alpha = 0.63$.

Regularization paths show the value of the components of the lasso, bridge, ridge, or elastic net estimator as function of λ. In lasso, variables are sent to zero one at a time, and result in a sparse estimate when λ is large. In ridge, components appraoch zero as λ increases, but are never zero. Note that when $\lambda = 0$, we have the OLS estimate. The elastic net regularization path resembles that of the lasso estimator.

Regularization paths can be used to detect multicollinearity.

1. Set $\alpha = 1$, fit lasso, visualize regularization path.
2. Set $\alpha = 0.6$, fit elastic net, visualize regularization path.
3. Compare these two plots, look for any "dramatic" change. If no, then there is probably no multicollinearity, so we use $\alpha = 1$ to take the full advantage of the lasso estimator. If yes, it is probably due to the instability of lasso under multicollinearity, so we use $\alpha = 0.6$ to cope.
4. Then, choose λ by cross validation.

The linear model is not as restrictive as it seems. We describe a few methods of moving from linear models to nonlinear models.

1. Inputs can be transformations of original features. For example, consider

   $$Y \sim \beta_1 f(X_1) + \cdots + \beta_d f(X_d),$$

   where f can be the logarithm, square root, square, etc.

2. Inputs can have interaction terms. For example, in addition to $X_1, \ldots, X_d$, we can also include

   $$X_1X_2, X_1X_3, \ldots, X_{d-1}X_d$$

   as variables. However, we pay the price of adding more variables (bivariate interaction adds $\sim d^2/2$, trivariate interaction adds $\sim d^3/6$). Including interaction terms naturally transforms the data into high-dimensional data even if they were not originally so.

3. Inputs can have basis expansions. For example, for each X_i, include $X_i^2, X_i^3, \ldots$. If the number of basis elements is allowed to increase, we enter the world of nonparametric models. For example, can use polynomial basis to approximate functions in the Sobolev space $\left\{f : \int (f'')^2 < \infty\right\}$. Instead of using $f(X) = \sum_{j=1}^d \beta_j X_j$, we can consider $f(X) = \sum_{j=1}^p \beta_j h_j(X)$. If $p \to \infty$, we have a nonparametric model.

Example 2.6 (Tree regression)**.** Let $d = 1$. Let $h_i(x) := \mathbf{1}\{x \le t_i\}$ for $1 \le i \le p$, where $t_1, \dots, t_p \in \mathbb{R}$ are given. Then $Y = \sum_{j=1}^{p} \beta_j h_j(X) + \varepsilon$ is tree model.

We can also handle situations where the input variable is categorical.

Definition 2.7. A **categorical variable** is a variable that takes on values from a finite [typically unordered] set.

Dummy coding Given a categorical random variable with K categories, encode it using $K - 1$ dummy variables.

$$\begin{aligned} 1 &= (0, 0, 0, \dots, 0) \\ 2 &= (1, 0, 0, \dots, 0) \\ 3 &= (0, 1, 0, \dots, 0) \\ &\vdots \\ K &= (0, 0, 0, \dots, 1) \end{aligned}$$

2.3 Classification and discriminant analysis

We quickly summarize our discussion of regression so far. The intuition behind regression is to minimize the risk

$$R(f) := \mathbb{E}[|Y - f(X)|^2],$$

and that this perspective is completely model-free (i.e., we assume nothing about the joint distribution $(X, Y) \sim P_{X,Y}$). We showed that the minimizing f is $f^*(x) := \mathbb{E}[Y \mid X = x]$. Because we cannot observe $R(f)$, we instead consider the empirical risk

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} |Y_i - f(X)|^2.$$

However, minimizing empirical risk leads to overfitting. To combat this issue, we regularize by considering linear models $f(X) = \beta^\top X$. This gives the OLS $\widehat{\beta} = (X^\top X)^{-1} X^\top Y$; we also showed that this solution can be derived from a model-based perspective (the Gaussian noise model). This approach fails with high-dimensional data, so we consider ridge, bridge, and lasso, as well as elastic-net. We discussed how to move from from linear to nonlinear models, parametric to nonparametric models, and numerical to categorical data.

Definition 2.8. Classification is regression with a categorical response variable $Y = \{1, -1\}$.

The goal of classification is still the same as in general regression; we want to find a mapping h such that Y and $h(X)$ are "close" to each other. Unlike general regression however, here we may assume that the range of h is $\{1, -1\}$.

We may use the L^2 loss as before, but in classification it is equivalent to **0-1 loss** up to a scalar multiple.

$$L(h) := \|Y - h(X)\|_2^2 = 4 \cdot \mathbf{1}\{Y \neq h(X)\}.$$

We will only use 0-1 loss for classification.

The **risk** function is again defined by

$$R(h) := \mathbb{E}[L(h)] = \mathbb{E}[\mathbf{1}\{Y \neq h(X)\}] = \mathbb{P}(Y \neq h(X)).$$

Definition 2.9. The **Bayes classification rule** is defined by

$$h^* = \operatorname*{argmin}_h R(h).$$

The **Bayes risk** is $R^* := R(h^*)$.

Theorem 2.10 (Bayes classification rule)**.** *The risk-minimizing function is*

$$h^*(x) = \begin{cases} 1 & \mathbb{P}(Y = 1 \mid X = x) > \frac{1}{2}, \\ -1 & \textit{otherwise.} \end{cases}$$

Proof.

$$\begin{aligned} R(h) &= \mathbb{P}(Y \neq h(X)) \\ &= 1 - \mathbb{P}(Y = h(X)) \\ &= 1 - \sum_{y \in \{1,-1\}} \mathbb{P}(Y = y, h(X) = y) \\ &= 1 - \sum_{y \in \{1,-1\}} \mathbb{E}[\mathbf{1}\{Y = y, h(X) = y\}] \\ &= 1 - \sum_{y \in \{1,-1\}} \mathbb{E}_X[\mathbb{E}[\mathbf{1}\{Y = y\} \cdot \mathbf{1}\{h(X) = y\} \mid X]] \\ &= 1 - \sum_{y \in \{1,-1\}} \mathbb{E}_X[\mathbf{1}\{h(X) = y\}\mathbb{E}[\mathbf{1}\{Y = y\} \mid X]] \\ &= 1 - \sum_{y \in \{1,-1\}} \mathbb{E}_X[\mathbf{1}\{h(X) = y\}\mathbb{P}(Y = y \mid X)] \\ &= 1 - \int (\mathbf{1}\{h(x) = 1\}\mathbb{P}(Y = 1 \mid X = x) + \mathbf{1}\{h(x) = -1\}\mathbb{P}(Y = -1 \mid X = x)) \cdot p(x)\, dx \end{aligned}$$

We want to maximize the integrand, so we want

$$h(x) = \begin{cases} 1 & \mathbb{P}(Y = 1 \mid X = x) > \mathbb{P}(Y = -1 \mid X = x), \\ -1 & \text{otherwise.} \end{cases}$$

Noting that $\mathbb{P}(Y = 1 \mid X = x) + \mathbb{P}(Y = -1 \mid X = x) = 1$ finishes the proof. □

Recall that the function that minimized the L^2 risk in general regression is $\mathbb{E}[Y \mid X = x]$ (Theorem 2.1). In the classification setting, this function takes the form

$$\begin{aligned} \mathbb{E}[Y \mid X = x] &= \mathbb{P}(Y = 1 \mid X = x) - \mathbb{P}(Y = -1 \mid X = x) \\ &= 2\mathbb{P}(Y = 1 \mid X = x) - 1. \end{aligned}$$

So,

$$\operatorname{sign} \mathbb{E}[Y \mid X = x] = h(x),$$

where we define

$$\operatorname{sign}(t) := \begin{cases} 1 & t > 0, \\ -1 & t \leq 0. \end{cases}$$

This result is intuitive. Without the restriction to the class of functions whose range is $\{-1, 1\}$, the risk-minimizing function would be $\mathbb{E}[Y \mid X = x]$. The "closest" function in the restricted class is the one that matches the sign of $\mathbb{E}[Y \mid X = x]$.

Definition 2.11. Let $r(x) := \mathbb{P}(Y = 1 \mid X = x)$. We define the **decision boundary** by

$$D(r) := \{x : r(x) = 1/2\}.$$

The **empirical risk** is defined by

$$\widehat{R}(h) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{Y_i \neq h(X_i)\}.$$

Again, minimizing empirical risk leads to overfitting, since any function that is correct on the observed data will minimize empirical risk, regardless of how it behaves on unobserved inputs.

As before, we turn to regularization to deal with this issue. We impose the restriction that h is of the form $h(X) = \text{sign}(\beta^\top X)$.

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} \sum_{i=1}^{n} \mathbf{1}\{Y_i \neq \text{sign}(\beta^\top X_i)\}.$$

Unfortunately, this optimization problem is hard to compute, since the indicator function is nonconvex! This is in contrast to the OLS from earlier where we were even able to have a closed-form solution.

We leave the risk-based approach for the moment and consider the model-based approach. How do we model $r(x) := \mathbb{P}(Y = 1 \mid X = x)$? We consider **logistic modeling**. The **logistic function** is defined by $g(t) := \frac{1}{1+e^{-t}}$. This is a "smooth" version of the sign function (see Figure 4). Given a function $f : \mathbb{R}^d \to \mathbb{R}$, we consider the model where

$$\mathbb{P}_f(Y = 1 \mid X = x) := \frac{1}{1 + e^{-f(x)}},$$

which consequently implies

$$\mathbb{P}_f(Y = -1 \mid X = x) = \frac{1}{1 + e^{f(x)}}.$$

We can combine these two expressions into the following form.

$$\mathbb{P}_f(Y = y \mid X = x) = \frac{1}{1 + e^{-yf(x)}}.$$

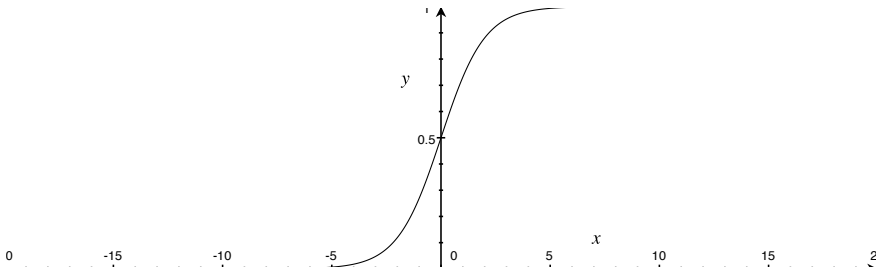

Figure 4: Plot of $\frac{1}{1+e^{-t}}$

Other possible models include inverse tangent, cdf of normal, etc. However, these examples still suffer from nonconvexity. The logistic model has "won" in popularity due to its convexity which leads to low computational complexity.

Restricting the type of functions f allows for different types of logistic models.

Example 2.12 (Linear logistic regression)**.**

$$f(x) := \beta_0 + \beta_1^\top x$$
$$P(Y = 1 \mid X = x) = \frac{1}{1 + e^{-\beta_0 + \beta_1^\top x}}$$

Example 2.13 (Nonparametric logistic regression)**.** Let $f(x)$ be "smooth", e.g. $\int (f'')^2 < \infty$.

However, note that choosing a class of functions f does not yet fully define the joint distribution of X and Y. The statistical model of logistic regression is

$$\{p(y, x) := \mathbb{P}_f(Y = y \mid X = x)p_X(x) : f, p_X\},$$

where p_X is the marginal density of X, and $\mathbb{P}_f(Y = y \mid X = x) := \frac{1}{1+e^{-yf(x)}}$. The p_X is called a "nuisance parameter" because it is needed to make the model valid, but we do not want to infer it at all. The f is the "parameter of interest" because it is the parameter we want to infer.

Given the random samples $\{(X_i, Y_i)\}_{i=1}^n$, we compute the MLE under this statistical model.

$$\begin{aligned}
\mathcal{L}(f) &:= \prod_{i=1}^n \mathbb{P}_f(Y_i \mid X_i) p_X(X_i) \\
\mathcal{L}(f) &= \prod_{i=1}^n \frac{1}{1 + e^{-Y_i f(X_i)}} p_X(X_i) \\
\ell(f) &= -\sum_{i=1}^n \log(1 + e^{-Y_i f(X_i)}) + \sum_{i=1}^n \log p_X(X_i) \\
\widehat{f} := \operatorname*{argmax}_f \ell(f) &= \operatorname*{argmin}_f \sum_{i=1}^n \log(1 + e^{-Y_i f(X_i)})
\end{aligned}$$

This motivates the definition of the following new loss function, which we call **logistic loss**, induced by the logistic model.

$$\ell^{\text{Logistic}}(y, f(x)) := \log(1 + e^{-yf(x)}).$$

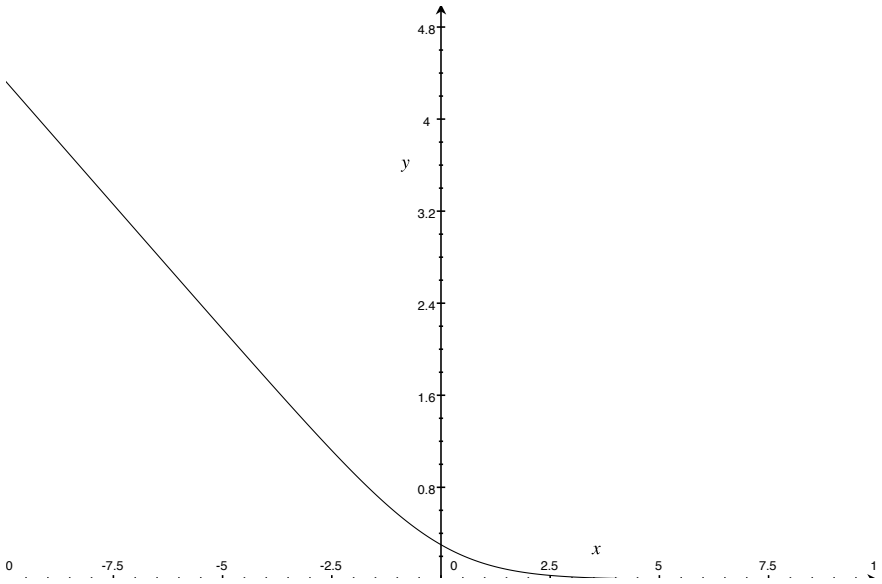

Figure 5: Plot of $\log(1 + e^{-yf(x)})$ vs. $y \cdot f(x)$

The quantity $y \cdot f(x)$ is called the **functional margin**, and we want it to be large to incur less loss, i.e., we encourage Y_i and $f(X_i)$ to have the same sign. Alternatively, consider

$$\begin{aligned}
\mathbb{P}(Y_i = 1 \mid X_i = x_i) &= \frac{1}{1 + e^{-f(X_i)}} \\
\implies \log \frac{\mathbb{P}(Y_i = 1 \mid X_i = x_i)}{\mathbb{P}(Y_i = -1 \mid X_i = x_i)} &= f(x_i).
\end{aligned}$$

This also captures the idea that we want $f(X_i)$ and Y_i to have the same sign. Note that the logistic loss not only encourages $Y_i f(X_i)$ to be positive, but also to be far away from zero.

To handle high-dimensional data, the ridge, bridge, lasso, and elastic net regressions are still applicable and effective. For instance, the ridge estimator for linear logistic regression is

$$\widehat{\beta}^{\text{Ridge},\lambda} := \operatorname*{argmin}_\beta \sum_{i=1}^n \log(1 + e^{-Y_i(\beta^\top X_i)}) + \lambda \|\beta\|_2^2.$$

[Note that β_0 still exists above, but we suppress it as an appended component to the X_i for ease of notation.] Even though $\sum_{i=1}^n \log(1 + e^{-Y_i(\beta^\top X_i)})$ is not quadratic, it is still convex, so the contours still look similar to those of $\|Y - X\beta\|_2^2$ from the previous section. The other approaches are analogous.

How does this compare to the OLS estimator, which is derived from using L^2-loss? Recall that $\widehat{\beta}^{\mathrm{OLS}} := \operatorname{argmin}_\beta \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^\top X_i)^2$. Because $(\widehat{\beta}^{\mathrm{OLS}})^\top x$ is a real number, our estimator is $\operatorname{sign}((\widehat{\beta}^{\mathrm{OLS}})^\top x)$. Note however that we can rewrite the definition of $\widehat{\beta}^{\mathrm{OLS}}$ as

$$\widehat{\beta}^{\mathrm{OLS}} := \operatorname*{argmin}_\beta \frac{1}{n}\sum_{i=1}^n (1 - Y_i\beta^\top X_i)^2$$

because $Y \in \{1, -1\}$. Then $Y_i\beta^\top X_i$ is the functional margin, and the summand corresponds to a quadratic loss function $\ell(u) = (1-u)^2$. This loss function is small when the functional margin is near 1, but when the functional margin $Y_i\beta^\top X_i$ is large, it still incurs a large loss despite being correct. This is a significant drawback of L^2 loss.

However, using L^2 loss not only encourages the functional margin to be away from the negative horizontal axis (misclassification), but also to be far from the ambiguous zone near zero on the horizontal axis; it constrains the functional margin to be in a small region, but it is in the correct region. Logistic loss encourages the functional margin to be large, but does a worse job of keeping it away from the ambiguous region near zero.

L^2 loss leads to linear discriminant analysis (LDA), which is powerful. Moreover, using logistic loss might require more data than L^2 loss. Neither L^2 loss or logistic loss is clearly better than the other.

We could also combine the two loss functions by considering considering the truncated quadratic loss $\ell(u) := (1-u)^2\mathbf{1}\{u \le 1\}$; this is sometimes better and sometimes worse than the previous two loss functions. In short, the question of whether a loss is good or not depends on the context and various other factors. In some applications, a certain loss function has been empirically shown to be better than others.

- Text mining: sparse logistic regression
- Image analysis: boosting (exponential loss)
- Genomics : LDA (quadratic loss)

The **hinge loss** is defined by $[x]_+ := \max(x, 0)$. Then the **SVM estimator** is defined by

$$\widehat{\beta}^{\mathrm{SVM}} = \operatorname*{argmin}_\beta \frac{1}{n}\sum_{i=1}^n [1 - Y_i\beta^\top X_i]_+ + \lambda\|\beta\|_2^2.$$

There are more involved interpretations of SVM that involve hyperplanes and margins, but from a statistical perspective, SVM is simply ridge logistic regression with hinge loss.

In **boosting** we consider the exponential loss $e^{-Y_i f(X_i)}$. Again, there are more involved interpretations of boosting, but from a statistical perspective it simply uses a different loss function.

Recall that in the statistical model of logistic regression, we model the distribution of $Y \mid X$, but not of X; this is a **discriminative model**, and we are unable to generate new data (Y, X) due to our lack of knowledge of the distribution of X. On the other hand, in **generative modeling** we model $X \mid Y$ and Y, and then use Bayes formula to model $Y \mid X$; note that we can generate new data (Y, X) because we model the joint distribution.

$$\begin{aligned}\mathbb{P}(Y = 1 \mid X = x) &= \frac{p(x \mid Y = 1)\mathbb{P}(Y = 1)}{p(x \mid Y = 1)\mathbb{P}(Y = 1) + p(x \mid Y = -1)\mathbb{P}(Y = -1)} \\ &=: \frac{p_+(x)\eta}{p_+(x)\eta + p_-(x)(1-\eta)},\end{aligned}$$

where we let $p_+(x) := p(x \mid Y = 1)$, $p_-(x) := p(x \mid Y = -1)$, and $\eta := P(Y = 1)$.

We need to model $\mathbb{P}(Y = \pm 1)$, $p(x \mid Y = 1)$, and $p(x \mid Y = -1)$. Clearly $Y \sim \mathrm{Ber}(\eta)$ for some η (note, we let Bernoulli to take on values ± 1 rather than 1 and 0); this is the only way we can model Y.

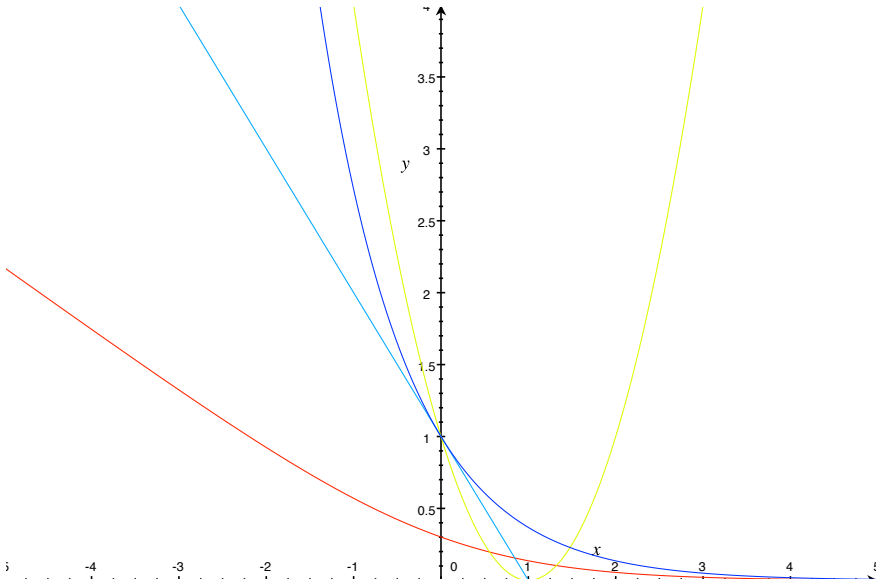

Figure 6: Plots of logistic loss $\log(1 + e^{-yf(x)})$, quadratic loss $(1 - yf(x))^2$, hinge loss $[1 - yf(x)]_+$, and exponential loss $e^{-yf(x)}$ vs. $y \cdot f(x)$

We have more freedom in modeling X. In **Gaussian discriminant analysis (GDA)** we have

$$X \mid (Y = 1) \sim \mathcal{N}(\mu_+, \Sigma_+)$$
$$X \mid (Y = -1) \sim \mathcal{N}(\mu_-, \Sigma_-),$$

i.e.,

$$p_+(x) = (2\pi)^{-d/2}|\Sigma_+|^{-1/2} \exp\left(-\frac{(x-\mu_+)^\top \Sigma_+^{-1}(x-\mu_+)}{2}\right)$$
$$p_-(x) = (2\pi)^{-d/2}|\Sigma_-|^{-1/2} \exp\left(-\frac{(x-\mu_-)^\top \Sigma_-^{-1}(x-\mu_-)}{2}\right)$$

Note again that we are not assuming the true conditional distributions are Gaussian; we are using a simplified model for the sake of regularization.

Logistic regression and other discriminative models care only about prediction, and thus only model $Y \mid X$ and disregard the marginal distribution of X. Generative modeling involves a belief/philosophy of how the data was generated. In this sense, discriminative modeling relies on fewer assumptions.

The Bayes Rule under GDA can be rewritten as follows.

$$\begin{aligned} & \mathbb{P}(Y = 1 \mid X = x) > \mathbb{P}(Y = -1 \mid X = x) \\ \iff & p_+(x)\eta > p_-(x)(1-\eta) \\ \iff & \log \frac{p_+(x)}{p_-(x)} + \log \frac{\eta}{1-\eta} > 0 \end{aligned}$$

The first term is

$$\frac{p_+(x)}{p_-(x)} = \frac{|\Sigma_-|^{1/2}}{|\Sigma_+|^{1/2}} \exp\left(-\frac{(x-\mu_+)^\top \Sigma_+^{-1}(x-\mu_+)}{2} + \frac{(x-\mu_-)^\top \Sigma_-^{-1}(x-\mu_-)}{2}\right)$$
$$\log \frac{p_+(x)}{p_-(x)} = \frac{1}{2} \log \frac{|\Sigma_-|^{1/2}}{|\Sigma_+|^{1/2}} - \frac{(x-\mu_+)^\top \Sigma_+^{-1}(x-\mu_+)}{2} + \frac{(x-\mu_-)^\top \Sigma_-^{-1}(x-\mu_-)}{2}$$

If we define the Mahalanobis distances

$$\gamma_+^2(x) := (x-\mu_+)^\top \Sigma_+^{-1}(x-\mu_+),$$
$$\gamma_-^2(x) := (x-\mu_-)^\top \Sigma_-^{-1}(x-\mu_-),$$

we have

$$h^*(x) = \begin{cases} 1 & \text{if } \frac{1}{2}\gamma_-^2(x) - \frac{1}{2}\gamma_+^2(x) + \frac{1}{2}\log\frac{|\Sigma_-|}{|\Sigma_+|} + \log\frac{\eta}{1-\eta} > 0 \\ -1 & \text{otherwise} \end{cases}$$

The condition is a quadratic form of x, i.e., it is of the form $x^\top Ax + b^\top x + c$. For this reason, GDA is sometimes called **quadratic discriminant analysis (QDA)**.

Given only the data, how to estimate μ_+, μ_-, Σ_+, Σ_-, η? MLE.

$$\begin{aligned}
n_+ &:= \sum_{i=1}^n \mathbf{1}\{Y_i = 1\} \\
n_- &:= n - n_+ \\
\widehat{\mu}_+ &:= \frac{1}{n_+}\sum_{i:Y_i=1} X_i \\
\widehat{\mu}_- &:= \frac{1}{n_-}\sum_{i:Y_i=-1} X_i \\
\widehat{\eta} &:= \frac{n_+}{n} \\
\widehat{\Sigma}_+ &:= \frac{1}{n_+}\sum_{i:Y_i=1} (X_i - \widehat{\mu}_+)(X_i - \widehat{\mu}_+)^\top \\
\widehat{\Sigma}_- &:= \frac{1}{n_-}\sum_{i:Y_i=-1} (X_i - \widehat{\mu}_-)(X_i - \widehat{\mu}_-)^\top
\end{aligned}$$

Linear discriminant analysis (LDA) is the special case of QDA with the extra condition of a common covariance matrix: $\Sigma_+ = \Sigma_- = \Sigma$.

$$\begin{aligned}
&(x-\mu_-)^\top\Sigma^{-1}(x-\mu_-) - (x-\mu_+)^\top\Sigma^{-1}(x-\mu_+) \\
&= x^\top\Sigma^\top x - 2\mu_-^\top\Sigma^{-1}x + \mu_-^\top\Sigma^{-1}\mu_- - x^\top\Sigma^\top x + 2\mu_+^\top\Sigma^{-1}x - \mu_+^\top\Sigma^{-1}\mu_+ \\
&= -2\mu_-^\top\Sigma^{-1}x + \mu_-^\top\Sigma^{-1}\mu_- + 2\mu_+^\top\Sigma^{-1}x - \mu_+^\top\Sigma^{-1}\mu_+
\end{aligned}$$

So, if we define

$$\beta := \Sigma^{-1}(\mu_+ - \mu_-) \tag{1}$$
$$\beta_0 := \frac{1}{2}\mu_-^\top\Sigma^{-1}\mu_- - \frac{1}{2}\mu_+^\top\Sigma^{-1}\mu_+ + \log\frac{\eta}{1-\eta}, \tag{2}$$

we have

$$h^*(x) = \begin{cases} 1 & \text{if } \beta^\top x + \beta_0 > 0 \\ -1 & \text{otherwise} \end{cases}$$

The decision boundary is linear in x, hence the name "linear discriminant analysis."

Let us return to the condition distribution $Y \mid X$ under the LDA framework.

$$\begin{aligned}\mathbb{P}(Y=1 \mid X=x) &= \frac{p(x \mid Y=1)\mathbb{P}(Y=1)}{p(x \mid Y=-1)\mathbb{P}(Y=-1)+p(x \mid Y=1)\mathbb{P}(Y=1)} \\ &= \frac{1}{1+\frac{p(x|Y=-1)(1-\eta)}{p(x|Y=1)\eta}} \\ &= \frac{1}{1+\exp(\log\frac{p(x|Y=1)}{p(x|Y=-1)}-\log\frac{\eta}{1-\eta})} \\ &= \frac{1}{1+e^{-\beta_0+\beta^\top x}}.\end{aligned}$$

This appears to be precisely the conditional distribution in the setting of linear logistic regression. However in LDA, the β_0 and β are constrained to take the particular form (see equations (1) and (2)) in terms of the parameters of the Gaussian distributions, whereas in logistic regression, β_0 and β are arbitrary. Thus, LDA is a special case of linear logistic regression; the assumption of the Gaussian conditional distributions makes it more regularized than linear logistic regression.

We already know the logistic model corresponds to logistic loss; it turns out that the LDA model corresponds to quadratic loss (difficult topic). If the true model does involve the extra assumption of conditional Gaussian, then both losses give the right parameters.

This leads us to the question of which technique we should use: logistic regression or LDA? In applications, we generally use logistic regression. In most applications, if LDA assumption is correct, then logistic regression works and is usually more efficient than LDA.

In high dimensions ($d > n$), we use **diagonal linear discriminant analysis (DLDA)**, which is LDA with the further assumption that the covariance matrix is diagonal.

$$\Sigma = \begin{bmatrix}\sigma_1^2 & & & \\ & \sigma_2^2 & & \\ & & \ddots & \\ & & & \sigma_d^2\end{bmatrix}$$

Note that a Gaussian distribution with a diagonal covariance matrix has contours that are ellipses/ellipsoids whose axes are parallel to the component axes. In particular, the components are independent of each other, so we may

$$\begin{aligned}Y &\sim \text{Ber}(\eta) \\ X_j \mid (Y=1) &\sim \mathcal{N}(\mu_{+,j}, \sigma_j^2) \\ X_j \mid (Y=-1) &\sim \mathcal{N}(\mu_{-,j}, \sigma_j^2)\end{aligned}$$

Even though independence of the components conditioned on the class Y (this is known as the "Naïve Bayes assumption") may not actually hold, using dramatically simplified (and wrong) models often still works well.

$$p(x \mid Y=1) = \prod_{j=1}^{d} p(x_j \mid Y=1) = \prod_{j=1}^{d} \frac{\exp\left(-\frac{(x_j-\mu_{+,j})^2}{2\sigma_j^2}\right)}{\sqrt{2\pi\sigma_j^2}}.$$

The MLEs $\widehat{\mu}_+$ and $\widehat{\mu}_-$ are the same as for LDA/QDA. For the covariance matrix, we have

$$\widehat{\sigma}_j^2 = \frac{n_+\widehat{S}_{+,j} + n_-\widehat{S}_{-,j}}{n},$$

where $n_+ := \sum_{i=1}^n \mathbf{1}\{Y_i = 1\}$ and $\widehat{S}_{+,j} := \frac{1}{n_+}\sum_{i:Y_i=1}(X_{i,j} - \widehat{\mu}_{+,j})^2$, with analogous definitions for n_- and $\widehat{S}_{-,j}$.

DLDA is a special case of LDA, but is also a special case of a Naïve Bayes classifier.

Definition 2.14. The **Naïve Bayes classifiers** form a family of generative classification methods that exploit the regularization "$X_1, \dots, X_d$ (components) are conditionally independent given Y," or more explicitly,

$$p(x_1, \dots, x_d \mid Y) = \prod_{j=1}^{d} p(x_j \mid Y).$$

We describe classification in general under the Naïve Bayes assumption. Recall the Bayes classification rule

$$h^*(x) = \begin{cases} 1 & f(x) > 0 \\ -1 & \text{otherwise} \end{cases}$$

where

$$f(x) := \log \frac{\mathbb{P}(Y = 1 \mid X = x)}{\mathbb{P}(Y = -1 \mid X = x)}.$$

Then,

$$\begin{aligned} f(x) &= \log \frac{\mathbb{P}(Y = 1 \mid X = x)}{\mathbb{P}(Y = -1 \mid X = x)} \\ &= \log \frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = -1)} + \log \frac{p(x \mid Y = 1)}{p(x \mid Y = -1)} \\ &= \log \frac{\eta}{1 - \eta} + \sum_{i=1}^{d} \log \frac{p(x_j \mid Y = 1)}{p(x_j \mid Y = -1)} && \text{Naïve Bayes assumption} \\ &=: \beta_0 + \sum_{j=1}^{d} f_j(x_j), \end{aligned}$$

where we have defined $f_j(x_j) := \log \frac{p(x_j|Y=1)}{p(x_j|Y=-1)}$.To compute the MLE, we can consider each component separately.

- In full QDA, we have μ_+, μ_-, Σ_+, Σ_-, and η, so total number of parameters is $d+d+\frac{d(d+1)}{2}+\frac{d(d+1)}{2}+1 = d(d+1) + 2d + 1$.
- In LDA, we have μ_+, μ_-, Σ, and η, so total number of parameters is $\frac{d(d+1)}{2} + 2d + 1$.
- In DLDA, same as LDA, but Σ is diagonal, so the number of parameters is $3d + 1$.
- In DQDA (two diagonal covariance matrices), the number of parameters is $4d + 1$.

3 Generalized linear models

Generalized linear models give a systematic view of regression and classification. We have seen a few types of regression already, like Gaussian linear regression and logistic regression.

These models can be decomposed into the following components

1. **Stochastic component.**
2. **Systematic component**
3. **Link function**

We first state the components for Gaussian linear regression and logistic regression in order to motivate the generalization.

	Gaussian linear regression	Logistic regression
Model	$Y_i = \beta^\top X_i + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$	$\mathbb{P}(Y_i = 1 \mid X_i) = \frac{1}{1+e^{-\beta^\top X_i}}$
Stochastic component	$Y_i \mid X_i \sim \mathcal{N}(\mu_i, \sigma^2)$ where $\mu_i := \beta^\top X_i$	$Y_i \mid X_i \sim \text{Ber}(\mu_i)$, where $\mu_i := \mathbb{E}[Y_i \mid X_i]$
Systematic component	$\eta_i := \beta^\top X_i$	$\eta_i := \beta^\top X_i$
Link function	$\eta_i = g(\mu_i)$, with g the identity	$\eta_i = g(\mu_i)$, with what g?

The link function g for logistic regression is derived as follows.

$$\begin{aligned}
\mu_i &= \mathbb{E}[Y_i \mid X_i] \\
\mu_i &= \mathbb{P}(Y_i = 1 \mid X_i) - \mathbb{P}(Y_i = -1 \mid X_i) \\
\mu_i &= \frac{1}{1+e^{-\eta_i}} - \frac{1}{1+e^{\eta_i}} \\
\mu_i &= \frac{e^{\eta_i} - 1}{1+e^{\eta_i}} \\
\mu_i + \mu_i e^{\eta_i} &= e^{\eta_i} - 1 \\
(1-\mu_i)e^{\eta_i} &= 1 + \mu_i \\
\eta_i &= \log \frac{1+\mu_i}{1-\mu_i} \\
g(t) &= \log \frac{1+t}{1-t}
\end{aligned}$$

This g is known as the **logit function**.

Remark: why not set $\eta_i := e^{-\beta^\top X_i}$ in logistic regression? We could model it this way, but when we try to use inference, we will run into computational issues (non-convexity causes intractability etc.)

Note that the link function links the systematic component (which is the input with possibly a transformation) to the mean of the stochastic component (the expected response of the input).

Generalized linear models (GLMs) are an extension of the three-component modeling scenario above.

1. We allow the stochastic components to follow any distribution, not just Gaussian or Bernoulli.
2. We allow the link functions to be more general, not just identity or logit.
3. We allow $\eta_i = f(x_i)$ where f is more general, not just $\beta^\top x_i$.

In practice,

1. Stochastic component: exponential dispersion family.
2. Link function: canonical link.
3. Systematic component: general.

Definition 3.1. The **exponential dispersion family (EDF)** contains any univariate distribution whose density takes the form

$$p_{\theta,\tau}(y) = h_\tau(y) \exp\left(\frac{\theta y - A(\theta)}{\tau}\right).$$

The parameter θ is the **canonical parameter**, τ is the **dispersion parameter**. The function $A(\theta)$ is the **normalization function** and may depend on τ. The function $h_\tau(y)$ is called the **base measure** and must be free of θ.

In the Gaussian distribution, τ is the variance, which is some sense the reasoning behind the term "dispersion."

The exponential dispersion family is more general than the exponential family.

The exponential dispersion family is in some sense the largest family of distributions for which any computation involving inference is tractable. In the past, people tried larger families of distributions, but ran into issues with non-convexity making problems intractable.

Example 3.2 (Gaussian distribution belongs to the EDF)**.**

$$\begin{aligned}
p(y) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2+\mu^2-2y\mu}{2\sigma^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}\right) \exp\left(\frac{\mu y - \frac{\mu^2}{2}}{\sigma^2}\right) \\
&= \left(\underbrace{\frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{y^2}{2\tau}\right)}_{h_\tau(y)}\right) \exp\left(\frac{\theta y - A(\theta)}{\tau}\right) \qquad \theta := \mu, A(\theta) := \theta^2/2, \tau = \sigma^2
\end{aligned}$$

Example 3.3 (Bernoulli distribution belongs to the EDF)**.** $\mathbb{P}(Y=1) = q$ and $\mathbb{P}(Y=-1) = 1-q$.

$$\begin{aligned}
p(y) &= q^{\frac{1+y}{2}} (1-q)^{\frac{1-y}{2}} \\
&= (q(1-q))^{1/2} \left(\frac{q}{1-q}\right)^{y/2} \\
&= \exp\left(\left(\frac{1}{2}\log\frac{q}{1-q}\right) y + \frac{1}{2}\log(q(1-q))\right).
\end{aligned}$$

We let $\theta := \frac{1}{2}\log\frac{q}{1-q}$, which implies

$$\begin{aligned}
\frac{q}{1-q} &= e^{2\theta} \\
q &= e^{2\theta} - qe^{2\theta} \\
q &= \frac{e^{2\theta}}{1+e^{2\theta}}.
\end{aligned}$$

Thus,

$$\begin{aligned}
A(\theta) &= -\frac{1}{2}\log(q(1-q)) \\
&= -\frac{1}{2}\log\left(\frac{e^{2\theta}}{1+e^{2\theta}} \cdot \frac{1}{1+e^{2\theta}}\right) \\
&= \frac{1}{2}(\log(1+e^{2\theta}) + \log(1+e^{-2\theta})).
\end{aligned}$$

Letting $\tau := 1$ and $h_\tau(y) := 1$ gives the EDF form.

Example 3.4 (Poisson distribution belongs to the EDF)**.**

$$p(y) = \frac{1}{y!}e^{-\lambda}\lambda^y = \frac{1}{y!}\exp(-\lambda + y\log\lambda).$$

This is clear with $\theta := \log\lambda$, $A(\theta) := e^\theta$, $\tau := 1$, and $h_\tau(y) := 1/y!$.

Most familiar distributions are in the EDF with the exception of the uniform distribution.

We call θ the canonical parameter but call η the **natural parameter**. The exponential dispersion family requires a special link function called the **canonical link function** which is necessary to lead to convex optimization in inference.

Canonical link principle: we always use the link function

$$\eta = g(\mu) := (A')^{-1}(\mu).$$

For instance, in the Gaussian distribution, $A(\theta) = \frac{1}{2}\theta^2$, so $A'(\theta) = \theta$, and finally $g(\mu) = (A')^{-1}(\mu) = \mu$, as seen in the beginning of the section.

In the Bernoulli distribution,

$$\begin{aligned}
A(\theta) &= \frac{1}{2}(\log(1+e^{2\theta}) + \log(1+e^{-2\theta})) \\
A'(\theta) &= \frac{1}{2}\left(\frac{2e^{2\theta}}{1+e^{2\theta}} + \frac{-2e^{-2\theta}}{1+e^{-2\theta}}\right) \\
A'(\theta) &= \frac{e^{2\theta}-1}{e^{2\theta}+1} \\
\mu &= \frac{e^{2\theta}-1}{e^{2\theta}+1} && \text{inverting} \\
\mu e^{2\theta} + \mu &= e^{2\theta} - 1 \\
\frac{1+\mu}{1-\mu} &= e^{2\theta} \\
\theta &= \frac{1}{2}\log\frac{1+\mu}{1-\mu}.
\end{aligned}$$

This is almost the same as the link function considered at the beginning; the factor of $1/2$ is inconsequential.

In the Poisson distribution, $A(\theta) = A'(\theta) = e^\theta$, so $g(\mu) = \log\mu$.

Again, the reason for choosing the canonical link instead of an arbitrary link function is to ensure tractability: we want to avoid non-convex optimization problems.

Note that we only need to involve μ (mean of distribution) in link function.

The goal of canonical link theory is to have $g(\mu)$ be the canonical parameter θ.

$$\theta = \eta = g(\mu)$$

The following theorem shows that we achieve this goal if we take $g(\mu) := (A')^{-1}(\mu)$.

Theorem 3.5. *If Y follows a distribution $p(y) = h_\tau(y)\exp\left(\frac{\theta y - A(\theta)}{\tau}\right)$ from the EDF, we have*

$$\mu := \mathbb{E}[Y] = A'(\theta).$$

Consequently, $(A')^{-1}(\mu) = \theta$.

Proof.

$$
\begin{aligned}
1 &= \int p(y)\,dy \\
1 &= \int h_\tau(y) \exp\left(\frac{\theta y - A(\theta)}{\tau}\right) dy \\
1 &= e^{-A(\theta)/\tau} \int h_\tau(y) e^{\theta y/\tau}\,dy \\
A(\theta) &= \tau \log \int h_\tau(y) e^{\theta y/\tau}\,dy \\
A'(\theta) &= \tau \frac{\int h_\tau(y) \frac{y}{\tau} e^{\theta y/\tau}\,dy}{\int h_\tau(y) e^{\theta y/\tau}\,dy} && \text{Leibniz rule or Fubini's theorem} \\
&= \tau \frac{\int h_\tau(y) \frac{y}{\tau} e^{\theta y/\tau}\,dy}{e^{A(\theta)/\tau}} \\
&= \int y h_\tau(y) \exp\left(\frac{\theta y - A(\theta)}{\tau}\right) dy \\
&= \mathbb{E}[Y]
\end{aligned}
$$

□

To fit the GLM, maximize the log-likelihood $\log \prod_{i=1}^n \mathbb{P}(Y_i = y_i \mid X_i = x_i)$; to predict, report $\mu_i = g^{-1}(\eta_i)$.

4 Exploratory analysis (unsupervised learning)

4.1 Graphical models

In this section we introduce Gaussian graphical models. Note that these are different from "Bayesian network" graphical models that appear in other literature.

Let $X_1, \dots, X_n \sim \mathcal{N}(\mu, \Sigma)$ be i.i.d. random variables, with $\mu \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$. We aim to find a sparse estimate for $\Theta := \Sigma^{-1}$, called the **precision matrix**.

Given $X_1, \dots, X_d$ and $A \subset \{1, \dots, d\}$, we let $X_A := \{X_j : j \in A\}$ and $X_{\setminus A} := \{X_j : j \notin A\}$.

Definition 4.1. Let $A, B \subset [1, \dots, d]$. We say X_A and X_B are **independent** if for any realizations x_A and x_B, we have $p(x_A, x_B) = p(x_A) \cdot p(x_B)$. We denote this $X_A \perp\!\!\!\perp X_B$.

We say X_A and X_B are **conditionally independent** given X_C if for any realizations x_A, x_B, and x_C, then $p(x_A, x_B \mid x_C) = p(x_A \mid x_C) \cdot p(x_B \mid x_C)$. In some sense, X_C "explains" the dependency between X_A and X_B.

Independent random variables need not always be conditionally independent: if X and Y are independent, they are not independent given $Z := X + Y$.

Conditionally independent random variables need not be independent either: let X, Y_1, Y_2 be independent. Then $X + Y_1$ and $X + Y_2$ are conditionally independent given X, but are not independent.

If we have the influences $A \longrightarrow B \longrightarrow C$, then $A \not\perp\!\!\!\perp B$, but $A \perp\!\!\!\perp C \mid B$.

Conditional independence relationships help represent joint distributions more efficiently. For instance, if $X_1, \dots, X_d$ are all binary random variables, we need $2^d - 1$ numbers to store the joint distribution. If we know that there are conditional independence relationships, we can factorize the joint distribution and use less space to store it.

Theorem 4.2.

$$\Theta_{j,k} = 0 \iff X_j \perp\!\!\!\perp X_k \mid X_{\setminus\{j,k\}}.$$

Proof. We want to show that $\Theta_{1,2} = 0 \implies X_1 \perp\!\!\!\perp X_2 \mid X_{\setminus\{1,2\}}$. It suffices to show the factorization $p(x_1, x_2 \mid x_3, \dots, x_d) = p(x_1 \mid x_3, \dots, x_d) p(x_2 \mid x_3, \dots, x_d)$. Without loss of generality we assume $\mu = \mathbf{0}$.

$$p(x_1, \dots, x_d) = \frac{1}{(2\pi)^{d/2}} |\Theta|^{1/2} \exp\left(-\frac{1}{2} x^\top \Theta x\right)$$

$$p(x) \propto \exp\left(-\frac{1}{2}\Theta_{1,1}x_1^2 - \frac{1}{2}\Theta_{2,2}x_2^2 - \sum_{j=3}^{d} \Theta_{1,j}x_1x_j - \sum_{k=3}^{d} \Theta_{2,k}x_2x_k - f(\Theta_{j,k}, x_j, x_k; 3 \le j \le k \le d)\right),$$

where the function f contains the remaining terms in the expansion of $x^\top \Theta x$, which will cancel out in the calculation below because they are free of x_1 and x_2. Note that we used the fact that $\Theta_{1,2} = 0$.

$$\begin{aligned}
&p(x_1, x_2 \mid x_3, \dots, x_d) \\
&= \frac{p(x_1, \dots, x_d)}{\int\int p(x_1, \dots, x_d)\, dx_1\, dx_2} \\
&= \frac{\exp\left(-\frac{1}{2}\Theta_{1,1}x_1^2 - \frac{1}{2}\Theta_{2,2}x_2^2 - \sum_{j=3}^{d} \Theta_{1,j}x_1x_j - \sum_{k=3}^{d} \Theta_{2,k}x_2x_k\right)}{\int\int \exp\left(-\frac{1}{2}\Theta_{1,1}x_1^2 - \frac{1}{2}\Theta_{2,2}x_2^2 - \sum_{j=3}^{d} \Theta_{1,j}x_1x_j - \sum_{k=3}^{d} \Theta_{2,k}x_2x_k\right) dx_1\, dx_2} \\
&= \frac{\exp\left(-\frac{1}{2}\Theta_{1,1}x_1^2 - \frac{1}{2}\Theta_{2,2}x_2^2 - \sum_{j=3}^{d} \Theta_{1,j}x_1x_j - \sum_{k=3}^{d} \Theta_{2,k}x_2x_k\right)}{\int \exp\left(-\frac{1}{2}\Theta_{1,1}x_1^2 - \sum_{j=3}^{d} \Theta_{1,j}x_1x_j\right) dx_1 \cdot \int \exp\left(-\frac{1}{2}\Theta_{2,2}x_2^2 - \sum_{k=3}^{d} \Theta_{2,j}x_2x_k\right) dx_2}
\end{aligned}$$

This gives the desired factorization.

In Homework 3 a proof for both directions is outlined; we sketch the argument here. Again we assume without loss of generality that $\mu = \mathbf{0}$. Letting $A := \{1, 2\}$, we can partition Σ into

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{AA} & \mathbf{\Sigma}_{AA^c} \\ \mathbf{\Sigma}_{A^cA} & \mathbf{\Sigma}_{A^cA^c} \end{bmatrix}.$$

It turns out that

$$X_A \mid X_{A^c} \sim \mathcal{N}(\mathbf{0}, \Sigma_{AA} - \Sigma_{AA^c}\Sigma_{A^cA^c}^{-1}\Sigma_{A^cA}).$$

From the fact that $\Sigma\Theta = I$, we have $\Theta_{AA}^{-1} = \Sigma_{AA} - \Sigma_{AA^c}\Sigma_{A^cA^c}^{-1}\Sigma_{A^cA}$, so $X_A \mid X_{A^c} \sim \mathcal{N}(\mathbf{0}, \Theta_{AA}^{-1})$. Using the fact that the components of a multivariate Gaussian are independent if and only if their covariance is zero, we easily have $\Theta_{1,2} = 0$ if and only if $X_1 \perp\!\!\!\perp X_2 \mid X_{\setminus\{1,2\}}$. □

We construct a graphical representation of Θ. Let $G := (V, E)$ with $V := \{1, \ldots, d\}$ and $E \subset V \times V$. We place an edge $(j, k) \in E$ (where $j \neq k$) whenever $\Theta_{j,k} \neq 0$. This is called graphical model learning or graph estimation.

Estimating the precision matrix: maximum penalized likelihood estimation (MPLE).

$$p(x) = \frac{1}{(2\pi)^{d/2}}|\Theta|^{1/2}\exp\left(-\frac{1}{2}(x-\mu)^\top\Theta(x-\mu)\right)$$

We already know $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$.

$$\begin{aligned}
\mathcal{L}_n(\Theta) &:= \frac{1}{(2\pi)^{d/2}}|\Theta|^{n/2}\exp\left(-\frac{1}{2}\sum_{i=1}^n(X_i-\overline{X})^\top\Theta(X_i-\overline{X})\right) \\
\ell(\Theta) &= -\frac{1}{2}\sum_{i=1}^n(X_i-\overline{X})^\top\Theta(X_i-\overline{X}) + \frac{n}{2}\log|\Theta| - \frac{nd}{2}\log(2\pi) \\
&= C + \frac{n}{2}\log|\Theta| - \frac{1}{2}\sum_{i=1}^n(X_i-\overline{X})^\top\Theta(X_i-\overline{X}) \\
&= C + \frac{n}{2}\log|\Theta| - \frac{1}{2}\sum_{i=1}^n\operatorname{Tr}\big((X_i-\overline{X})^\top\Theta(X_i-\overline{X})\big) \\
&= C + \frac{n}{2}\log|\Theta| - \frac{1}{2}\sum_{i=1}^n\operatorname{Tr}\big(\Theta(X_i-\overline{X})(X_i-\overline{X})^\top\big) \\
&= C + \frac{n}{2}\log|\Theta| - \frac{n}{2}\operatorname{Tr}\left(\Theta\frac{1}{n}\sum_{i=1}^n(X_i-\overline{X})(X_i-\overline{X})^\top\right) \\
&= C + \frac{n}{2}\log|\Theta| - \frac{n}{2}\operatorname{Tr}(\Theta S_n).
\end{aligned}$$

Note that the sample covariance matrix S_n is the sufficient statistic for Θ. We impose sparsity by imposing a constraint on Θ.

$$\begin{aligned}
\widehat{\Theta} &:= \operatorname*{argmax}_\Theta \ell(\Theta) \quad \text{s.t. } \|\Theta\|_1 := \sum_{j,k}|\Theta_{j,k}| \leq L \\
\widehat{\Theta} &= \operatorname*{argmax}_\Theta \ell(\Theta) - \lambda\|\Theta\|_1 \\
&= \operatorname*{argmin}_\Theta \operatorname{Tr}(\Theta S_n) - \log|\Theta| + \lambda\|\Theta\|_1
\end{aligned}$$

This optimization problem is called the **graphical lasso**.

We now define the graph induced by Θ and discuss how it represents conditional independence relationships. Let G be a graph with vertices $\{1, \ldots, d\}$, and let there be an [undirected] edge between distinct vertices i and j if and only if $\Theta_{i,j} \neq 0$.

Let $A, B, C \subset V$. We say C **separates** A and B (denoted $A \perp B \mid C$) if removing C separates A and B (no path between A and B after removing C). This turns out to be equivalent to $X_A \perp\!\!\!\perp X_B \mid X_C$.

Graph theory	Probability theory
$G = (V, E)$; $V = [d]$	$p(x_1, \ldots, x_d)$
separation $A \perp B \mid C$	conditional independence $X_A \perp\!\!\!\perp X_B \mid X_C$

Separation and conditional independence are connected by the Markov properties

Definition 4.3 (Pairwise Markov Property)**.**

$$X_j \perp\!\!\!\perp X_k \mid X_{\setminus\{j,k\}} \iff (j,k) \notin E.$$

For the Gaussian model we have defined, Theorem 4.2 and the definition of the graph immediately imply that the Gaussian model has the pairwise Markov property.

Definition 4.4 (Global Markov Property)**.** For $A, B, C \subset V = [d]$,

$$X_A \perp\!\!\!\perp X_B \mid X_C \iff A \perp B \mid C.$$

Clearly, the global Markov property implies the pairwise Markov property, but surprisingly, the converse often holds.

Theorem 4.5 (Lauritzen, 1996)**.** *If $p(x) > 0$, then the pairwise Markov property is equivalent to the global Markov property.*

This theorem is proved by induction on the number of nodes d.

4.2 Clustering, mixture models, and latent variable models

Clustering is a classification problem but with hidden/unobserved/latent class labels. Our goal is to recover the latent class labels.

Example 4.6 (Mixture of two Gaussians)**.** Let Z equal 1 with probability η and 2 otherwise. Let $X \mid (Z = 1) \sim \mathcal{N}(\mu_1, 1)$ and $X \mid (Z = 2) \sim \mathcal{N}(\mu_2, 1)$. We observe only the observed random samples $X_1, \ldots, X_n$, and not the latent variables $Z_1, \ldots, Z_n$.

Given $X_1, \ldots, X_n$, we aim to infer η, μ_1, and μ_2, and $Z_1, \ldots, Z_n$. We would also like to compute the probability $\mathbb{P}(Z_i = 1 \mid X_1, \ldots, X_n)$, the probability that X_i is sampled from distribution 1. [Note that if the Z_i are observed, we can do Gaussian discriminant analysis as before, but here the Z_i are unobserved.]

For the mixture of two Gaussians, the parameter is $\theta := (\eta, \mu_1, \mu_2)^\top$ and the model is the family of distributions of the form

$$p_\theta(x) = \sum_{j=1}^{2} p_\theta(x, Z = j) = \sum_{j=1}^{2} p_\theta(x \mid Z = j)\mathbb{P}(Z = j) = \eta p_{\mu_1}(x) + (1 - \eta)p_{\mu_2}(x),$$

for all $\eta \in [0, 1]$, $\mu_1, \mu_2 \in \mathbb{R}^d$. Here, η is called the **mixing coefficient**.

The above example is a type of mixture model.

Definition 4.7. A **mixture model** has a density that is a convex combination of a set of component densities. It is **finite** if there are finitely many components and **infinite** if there are infinitely many components.

We continue with the mixture of two Gaussians before moving to general mixture models.

The log-likelihood does not involve the Z_i because they are unobserved. The following is the **marginal log-likelihood**; this is contrast to the **complete log-likelihood** which replaces $p_\theta(x_i)$ with $p_\theta(x_i, z_i)$.

$$\begin{aligned}
\ell^{\mathrm{M}}(\theta) &:= \sum_{i=1}^{n} \log p_\theta(x_i) \\
&= \sum_{i=1}^{n} \log(\eta p_{\mu_1}(x_i) + (1 - \eta)p_{\mu_2}(x_i)) \\
&= \sum_{i=1}^{n} \log\left(\eta \frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu_1)^2/2} + (1 - \eta)\frac{1}{\sqrt{2\pi}} e^{-(x_i - \mu_2)^2/2}\right)
\end{aligned}$$

However, this function is non-convex and difficult to maximize using previous approaches. What alternatives do we have? Guess and check is a simple method. Gradient ascent is better (but requires several trials to attempt to avoid local optima). In applications, we typically use the EM algorithm.

Remark: the EM algorithm is not theoretically better than guess and check or gradient ascent, but in practice it is better. Moreover, it relies on the structure of the log-likelihood and the statistical model; although unlike the other two methods, the EM algorithm cannot be used to maximize general functions, it does utilize the structure of the log-likelihood much better.

We provide the intuition of the EM algorithm in two equivalent forms.

- **Block coordinate ascent.** Suppose we want to solve $\max_{x,y} f(x,y)$. Initialize $y^{(1)}$ randomly. For $t = 1, 2, \ldots$, perform the following updates repeatedly until convergence.

$$x^{(t+1)} \leftarrow \operatorname*{argmax}_x f(x, y^{(t)})$$
$$y^{(t+1)} \leftarrow \operatorname*{argmax}_y f(x^{(t+1)}, y)$$

  By the monotone convergence theorem, convergence is guaranteed if f is bounded from above. Note however this only guarantees convergence to a local maximum and not necessarily to the global maximum.

- **Minorization-maximization strategy.** Let f be a function we want to maximize. Initialize $x^{(1)}$, find a function ψ_1 that is easy to maximize and everywhere less than or equal to f. Let $x^{(2)}$ be the maximizer of ψ_1, and continue.

We now provide the rough intuition behind the EM algorithm for the mixture of two Gaussians.

Part 1. If $Z_1, \ldots, Z_n$ are known, how do we update θ? Then this is simply classification; use MLE with the complete likelihood. Let $n_1 := \sum_{i=1}^n \mathbf{1}\{Z_i = 1\}$ and $n_2 := n - n_1$.

$$\begin{aligned}
\ell^{\mathrm{C}}(\theta) &:= \sum_{i=1}^n \log p_\theta(X_i, Z_i) \\
&= \sum_{i:Z_i=1} [\log p_\theta(X_i \mid Z_i = 1)\mathbb{P}_\eta(Z_i = 1)] + \sum_{i:Z_i=2} [\log p_\theta(X_i \mid Z_i = 2)\mathbb{P}_\eta(Z_i = 2)] \\
&= \sum_{i:Z_i=1} \log(\eta p_{\mu_1}(X_i)) + \sum_{i:Z_i=2} \log((1-\eta) p_{\mu_2}(X_i)) \\
&= n_1 \log \eta + n_2 \log(1-\eta) + \sum_{i:Z_i=1} \log p_{\mu_1}(X_i) + \sum_{i:Z_i=2} \log p_{\mu_2}(X_i)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \ell^{\mathrm{C}}(\theta)}{\partial \eta} &= \frac{n_1}{\eta} - \frac{n_2}{1-\eta} \\
\frac{\partial \ell^{\mathrm{C}}(\theta)}{\partial \mu_1} &= \sum_{i:Z_i=1} \frac{\partial}{\partial \mu_1}\left(-\frac{1}{2}(X_i - \mu_1)^2 + \log\frac{1}{\sqrt{2\pi}}\right) \\
&= \sum_{i=1}^n \mathbf{1}\{Z_i = 1\}(X_i - \mu_1) \\
&= -n_1\mu_1 + \sum_{i=1}^n \mathbf{1}\{Z_i = 1\} X_i
\end{aligned}$$

Setting the partial derivatives to zero gives

$$\begin{aligned}
\widehat{\eta} &:= \frac{n_1}{n} = \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{Z_i = 1\} \\
\widehat{\mu}_1 &:= \frac{\sum_{i=1}^n \mathbf{1}\{Z_i = 1\} X_i}{n_1} = \frac{\sum_{i=1}^n \mathbf{1}\{Z_i = 1\} X_i}{\sum_{i=1}^n \mathbf{1}\{Z_i = 1\}} \\
\widehat{\mu}_2 &:= \frac{\sum_{i=1}^n \mathbf{1}\{Z_i = 2\} X_i}{n_2} = \frac{\sum_{i=1}^n \mathbf{1}\{Z_i = 2\} X_i}{\sum_{i=1}^n \mathbf{1}\{Z_i = 2\}}
\end{aligned}$$

Part 2. If we are only given θ, how can we update $Z_1, \ldots, Z_n$? An obstacle is that the $Z_1, \ldots, Z_n$ are random variables, not fixed quantities like parameters. Instead, we could seek something like $\mathbb{P}(Z_i = 1 \mid X_1, \ldots, X_n)$ or $\mathbb{E}[Z_i \mid X_1, \ldots, X_n]$. Note that if we know the probability distribution we can compute the expectation, but the reverse is only possible in the case when Z_i takes on two values. In the EM algorithm we seek the probability distribution, not the expectation.[5]

$$\begin{aligned}
\mathbb{P}_\theta(Z_i = 1 \mid X_1, \ldots, X_n) &= \mathbb{P}_\theta(Z_i = 1 \mid X_i) \\
&= \frac{p_{\mu_1}(X_i \mid Z_i = 1)\mathbb{P}_\eta(Z_i = 1)}{p_{\mu_1}(X_i \mid Z_i = 1)\mathbb{P}_\eta(Z_i = 1) + p_{\mu_2}(X_i \mid Z_i = 2)\mathbb{P}_\eta(Z_i = 2)} \\
&= \frac{\eta p_{\mu_1}(X_i)}{\eta p_{\mu_1}(X_i) + (1-\eta)p_{\mu_2}(X_i)} \\
&= \frac{\eta e^{-(x_i-\mu_1)^2/2}}{\eta e^{-(x_i-\mu_1)^2/2} + (1-\eta)e^{-(x_i-\mu_2)^2/2}} \\
\mathbb{P}_\theta(Z_i = 2 \mid X_1, \ldots, X_n) &= \frac{(1-\eta) e^{-(x_i-\mu_2)^2/2}}{\eta e^{-(x_i-\mu_1)^2/2} + (1-\eta)e^{-(x_i-\mu_2)^2/2}}
\end{aligned}$$

So, we have found $\mathbb{P}(Z_i = 1 \mid X_1, \ldots, X_n) = \mathbb{E}[\mathbf{1}\{Z_i = 1\} \mid X_1, \ldots, X_n]$.

Part 3. If we only know the value of $\mathbb{P}(Z_i = 1 \mid X_1, \ldots, X_n) = \mathbb{E}[\mathbf{1}\{Z_i = 1\} \mid X_1, \ldots, X_n]$, how do we update θ? If we make a choice of Z_i (either by choosing the more likely outcome or by sampling) and repeat Part 1, convergence of the whole algorithm is unfortunately not guaranteed. Instead, we take the results of Part 1, and replace all the indicator random variables with expectations. [We will justify this later.]

$$\begin{aligned}
\widehat{\eta} &:= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\mathbf{1}\{Z_i = 1\} \mid X_1, \ldots, X_n] \\
\widehat{\mu}_1 &:= \frac{\sum_{i=1}^n \mathbb{E}[\mathbf{1}\{Z_i = 1\} \mid X_1, \ldots, X_n] X_i}{\sum_{i=1}^n \mathbb{E}[\mathbf{1}\{Z_i = 1\} \mid X_1, \ldots, X_n]} \\
\widehat{\mu}_2 &:= \frac{\sum_{i=1}^n \mathbb{E}[\mathbf{1}\{Z_i = 1\} \mid X_1, \ldots, X_n] X_i}{\sum_{i=1}^n \mathbb{E}[\mathbf{1}\{Z_i = 1\} \mid X_1, \ldots, X_n]}
\end{aligned}$$

Example 4.8 (Complete EM algorithm for mixture of two Gaussians)**.** Initialize $\theta^{(1)} := (\eta^{(1)}, \mu_1^{(1)}, \mu_2^{(1)})^\top$. For $t = 1, 2, \ldots,$

- **E step.**

$$\begin{aligned}
\gamma_{i,1}^{(t+1)} &:= \mathbb{P}_{\theta^{(t)}}(Z_i = 1 \mid X_1, \ldots, X_n) = \frac{\eta^{(t)} p_{\mu_1^{(t)}}(X_i)}{\eta^{(t)} p_{\mu_1^{(t)}}(X_i) + (1-\eta^{(t)}) p_{\mu_2^{(t)}}(X_i)} \\
\gamma_{i,2}^{(t+1)} &:= 1 - \gamma_{i,1}^{(t+1)}.
\end{aligned}$$

- **M step.**

$$\begin{aligned}
\eta^{(t+1)} &:= \frac{1}{n}\sum_{i=1}^n \gamma_{i,1}^{(t+1)} \\
\mu_1^{(t+1)} &:= \frac{\sum_{i=1}^n \gamma_{i,1}^{(t+1)} X_i}{\sum_{i=1}^n \gamma_i^{(t+1)}} \\
\mu_2^{(t+1)} &:= \frac{\sum_{i=1}^n \gamma_{i,2}^{(t+1)} X_i}{\sum_{i=1}^n \gamma_{i,2}^{(t+1)}}
\end{aligned}$$

[5] Why is the EM algorithm called "expectation maximization" then? It may be because $\mathbb{P}(Z_i = 1 \mid X_1, \ldots, X_n) = \mathbb{E}[\mathbf{1}\{Z_i = 1\} \mid X_1, \ldots, X_n]$. It may also be because we maximize the expected complete log likelihood, as we will see later.

Repeat until convergence of the marginal log-likelihood.

Example 4.9 (Mixture of K Gaussians)**.** Let $Z \sim \text{Multi}(\eta_0, \dots, \eta_{K-1})$, i.e., $\mathbb{P}(Z = j) = \eta_j$ for $j = 0, \dots, K-1$ where $\eta_j \geq 0$ and $\sum_{j=0}^{K-1} \eta_j = 1$.

We will also extend to multivariate Gaussian distributions with unknown covariance. $X \mid (Z = j) \sim \mathcal{N}_d(\mu_j, \Sigma_j)$.

Note that taking $K = 2$ will give exactly the same algorithm as the previous example.

Let

$$\theta := (\eta_0, \dots, \eta_{K-1}, \mu_0, \dots, \mu_{K-1}, \Sigma_0, \dots, \Sigma_{K-1}).$$

Initialize $\theta^{(1)}$. For $t = 1, 2, 3 \dots$,

- **E step.** For all $1 \leq i \leq n$ and $0 \leq j \leq K-1$, define

$$\gamma_{i,j}^{(t+1)} := \mathbb{P}_{\theta^{(t)}}(Z_i = j \mid X_1, \dots, X_n) = \frac{\eta_j^{(t)} p_{\mu_j^{(t)}, \Sigma_j^{(t)}}(X_i)}{\sum_{\ell=0}^{K-1} \eta_\ell^{(t)} p_{\mu_\ell^{(t)}, \Sigma_\ell^{(t)}}(X_i)}.$$

- **M step.** For each $j = 0, \dots, K-1$,

$$\begin{aligned}
\eta_j^{(t+1)} &:= \frac{1}{n} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} \\
\mu_j^{(t+1)} &:= \frac{\sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} X_i}{\sum_{i=1}^{n} \gamma_{i,j}^{(t+1)}} \\
\Sigma_{i,j}^{(t+1)} &:= \frac{\sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} (X_i - \mu_j^{(t+1)})(X_i - \mu_j^{(t+1)})^\top}{\sum_{i=1}^{n} \gamma_{i,j}^{(t+1)}}
\end{aligned}$$

Repeat until convergence of the marginal log-likelihood.

Formal derivation of the EM algorithm (minorization-maximization perspective). We want to maximize $\ell^{\text{M}}(\psi)$. We initialize $\psi^{(0)}$. We find a "lower bound" function $F_0(\psi)$ that satisfies

- $F_0 \leq \ell^{\text{M}}$,
- $F_0(\psi^{(0)}) = \ell^{\text{M}}(\psi^{(0)})$.

Let $\psi^{(1)} := \text{argmax}_\psi F_0(\psi)$. Define F_1 in a similar way, repeat.

Our goal is, given random samples $X_1, \dots, X_n$, to fit a finite mixture model where $Z \sim \text{Multi}(\eta_0, \dots, \eta_{K-1})$ and $X \mid (Z = j) \sim p_{\theta_j}$.

$$\begin{aligned}
p_\psi(X) &= \sum_{j=0}^{K-1} p_{\theta_j}(X) \cdot \eta_j \\
\eta_j &= \mathbb{P}(Z = j) \\
p_{\theta_j}(x) &= p(x \mid Z = j).
\end{aligned}$$

We want to infer $\psi := (\theta_0, \eta_0, \theta_1, \eta_1, \dots, \theta_{K-1}, \eta_{K-1})^\top$ and $Z_1, \dots, Z_n$ by maximizing the marginal log-likelihood

$$\ell^{\text{M}}(\psi) := \sum_{i=1}^{n} \log p_\psi(X_i).$$

For $j = 0, \dots, K-1$ we will define

$$\gamma_{i,j} := \mathbb{P}_\psi(Z_i = j \mid X_i). \tag{3}$$

Note that $\sum_{j=0}^{K-1} \gamma_{i,j} = 1$, and moreover, for each fixed i, the vector $(\gamma_{i,0}, \dots, \gamma_{i,K-1})$ is a distribution for Z_j.

Theorem 4.10 (Jensen's inequality)**.**

$$\log \mathbb{E}[X] \geq \mathbb{E}[\log X].$$

For intuition, it is easy to see from a plot of the log function that $\log \frac{x_1+x_2}{2} \geq \frac{\log(x_1)+\log(x_2)}{2}$, due to the concavity of the log function.

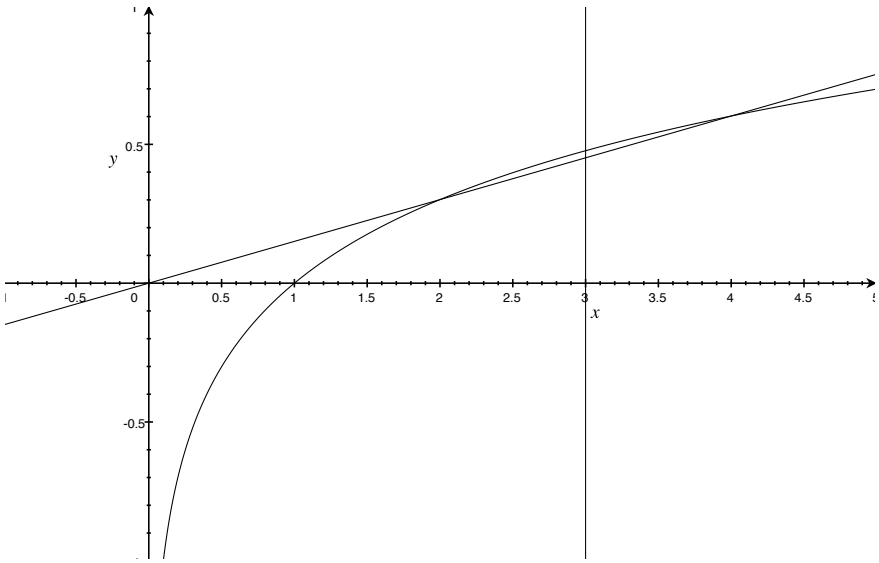

Figure 7: $\log \frac{2+4}{2} \geq \frac{\log(2)+\log(4)}{2}$

Given the current parameter $\psi^{(t)}$, we want to find a lower bound function $F_t(\psi)$ such that

- $F_t \leq \ell^{\mathrm{M}}$,
- $F_t(\psi^{(t)}) = \ell^{\mathrm{M}}(\psi^{(t)})$.

The following holds for any $\gamma := \{\{\gamma_{i,j}\}_{i=1}^n\}_{j=0}^{K-1}$ satisfying $\gamma_{i,j} \geq 0$ and $\sum_{j=0}^{K-1} \gamma_{i,j} = 1$, so in particular it holds for $\gamma_{i,j}$ as defined in (3).

$$\begin{aligned}
\ell^{\mathrm{M}}(\psi) &= \sum_{i=1}^n \log p_\psi(X_i) \\
&= \sum_{i=1}^n \log \left(\sum_{j=0}^{K-1} p_\psi(X_i, Z_i = j) \right) \\
&= \sum_{i=1}^n \log \left(\sum_{j=0}^{K-1} \gamma_{i,j} \frac{p_\psi(X_i, Z_i = j)}{\gamma_{i,j}} \right) \\
&= \sum_{i=1}^n \log \mathbb{E}_{Z_i} \left[\frac{p_\psi(X_i, Z_i)}{\gamma_{i,j}} \right] && \text{“let” } Z_i \sim \mathrm{Mult}(\gamma_{i,0}, \ldots, \gamma_{i,K-1}) \\
&\geq \sum_{i=1}^n \mathbb{E}_{Z_i} \left[\log \frac{p_\psi(X_i, Z_i)}{\gamma_{i,j}} \right] && \text{Jensen's inequality} \\
&= \sum_{i=1}^n \sum_{j=0}^{K-1} \gamma_{i,j} \log \frac{p_\psi(X_i, Z_i = j)}{\gamma_{i,j}} \\
&=: Q(\gamma, \psi)
\end{aligned}$$

Thus, $Q(\gamma, \psi)$ is a lower bound for ℓ^M, for any ψ and any γ satisfying $\gamma_{i,j} \geq 0$ and $\sum_{j=0}^{K-1} \gamma_{i,j} = 1$.

In general, $Q(\gamma, \psi)$ is easy to maximize, especially for distributions from the exponential family.

We define $F_t(\psi) := Q(\gamma^{(t+1)}, \psi)$, where $\gamma_{i,j}^{(t+1)} := \mathbb{P}_{\psi^{(t)}}(Z_i = j \mid X_i)$. Note that we have shown that $F_t \leq \ell^{\mathrm{M}}$, but we have not yet shown that the choice of $\gamma^{(t+1)}$ gives $F_t(\psi^{(t)}) = \ell^{\mathrm{M}}(\psi^{(t)})$.

$$\begin{aligned}
F_t(\psi^{(t)}) &= \sum_{i=1}^{n} \sum_{j=0}^{K-1} \mathbb{P}_{\psi^{(t)}}(Z_i = j \mid X_i) \log \frac{p_{\psi^{(t)}}(X_i, Z_i = j)}{\mathbb{P}_{\psi^{(t)}}(Z_i = j \mid X_i)} \\
&= \sum_{i=1}^{n} \sum_{j=0}^{K-1} \mathbb{P}_{\psi^{(t)}}(Z_i = j \mid X_i) \log p_{\psi^{(t)}}(X_i) \\
&= \sum_{i=1}^{n} \log p_{\psi^{(t)}}(X_i) \sum_{j=0}^{K-1} \mathbb{P}_{\psi^{(t)}}(Z_i = j \mid X_i) \\
&= \sum_{i=1}^{n} \log p_{\psi^{(t)}}(X_i) \\
&= \ell^{\mathrm{M}}(\psi^{(t)}).
\end{aligned}$$

In particular, our definition of $\gamma_{i,j}^{(t+1)} := \mathbb{P}_{\psi^{(t)}}(Z_i = j \mid X_i)$ is precisely $\operatorname{argmax}_\gamma Q(\gamma, \psi^{(t)})$.

This yields the block coordinate ascent interpretation of the EM algorithm. Initialize $\psi^{(0)}$. For $t = 1, 2, \ldots,$

- **E step.**
$$\gamma^{(t+1)} := \operatorname*{argmax}_{\gamma} Q(\gamma, \psi^{(t)}) = \mathbb{P}_{\psi^{(t)}}(Z_i = j \mid X_i).$$
- **M step.**
$$\psi^{(t+1)} := \operatorname*{argmax}_{\psi} Q(\gamma^{(t+1)}, \psi).$$

The minorization-maximization perspective is similar.

- **E step.**
$$\gamma^{(t+1)} := \operatorname*{argmax}_{\gamma} Q(\gamma, \psi^{(t)}) = \mathbb{P}_{\psi^{(t)}}(Z_i = j \mid X_i).$$
We define $F_t(\psi) := Q(\gamma^{(t+1)}, \psi)$, and note that we have shown that $F_t \leq \ell^{\mathrm{M}}$ and that $F_t(\psi^{(t)}) = \ell^{\mathrm{M}}(\psi^{(t)})$.
- **M step.**
$$\psi^{(t+1)} := \operatorname*{argmax}_{\psi} F_t(\psi).$$

Theorem 4.11. *Under the finite mixture model $p_\psi(x) = \sum_{j=0}^{K-1} \eta_j \cdot p_{\theta_j}(x)$ (see above), at the t^{th} iteration of the EM algorithm (where we know $\psi^{(t)}$) we do the following.*

- ***E step.*** *Compute $\gamma_{i,j}^{(t+1)} = \mathbb{P}_{\psi^{(t)}}(Z_i = j \mid X_i)$.*
- ***M step.*** *Set*
$$\eta_j^{(t+1)} := \frac{1}{n} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)}$$
$$\theta_j^{(t+1)} := \operatorname*{argmax}_{\theta_j} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} \log p_{\theta_j}(X_i)$$

[Note that $p_{\theta_j}(X_i) := p_\psi(X_i \mid Z = j)$.]

Note that the M step is simply maximizing the expected complete log-likelihood.

Proof. The E step is clear from the above derivation. For the M step,

$$F_t(\psi) = \sum_{i=1}^{n} \sum_{j=0}^{K-1} \gamma_{i,j}^{(t+1)} \log \frac{p_\psi(X_i, Z_i = j)}{\gamma_{i,j}^{(t+1)}}.$$

We can optimize the θ_j separately.

$$\begin{aligned}
\theta_j^{(t+1)} &:= \operatorname*{argmax}_{\theta_j} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} \log p_{\theta_j,\eta_j}(X_i, Z_i = j) \\
&= \operatorname*{argmax}_{\theta_j} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} \log p_{\theta_j}(X_i \mid Z_i = j) \cdot \eta_j \\
&= \operatorname*{argmax}_{\theta_j} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} \log p_{\theta_j}(X_i \mid Z_i = j) \\
&= \operatorname*{argmax}_{\theta_j} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} \log p_{\theta_j}(X_i). && \text{definition}
\end{aligned}$$

For η,

$$F_t(\psi) = \sum_{i=1}^{n} \sum_{j=0}^{K-1} \gamma_{i,j}^{(t+1)} \log \frac{p_\psi(X_i \mid Z_i = j)\eta_j}{\gamma_{i,j}^{(t+1)}},$$

so

$$\eta^{(t+1)} = \operatorname*{argmax}_{\eta_0,\ldots,\eta_{K-1}} \sum_{i=1}^{n} \sum_{j=0}^{K+1} \gamma_{i,j}^{(t+1)} \log \eta_j \quad \text{subject to } \sum_{j=0}^{K-1} \eta_j = 1.$$

The Lagrangian form is

$$\sum_{i=1}^{n} \sum_{j=0}^{K-1} \gamma_{i,j}^{(t+1)} \log \eta_j - \alpha \left(-1 + \sum_{j=0}^{K-1} \eta_j \right).$$

Taking the derivative w.r.t. η_j and setting it equal to zero gives

$$\alpha = \frac{1}{\widehat{\eta}_j} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)},$$

and this holds for all j. Noting that

$$1 = \sum_{j=0}^{K-1} \widehat{\eta}_j = \frac{1}{\alpha} \sum_{j=0}^{K-1} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} = \frac{n}{\alpha},$$

we have

$$\widehat{\eta}_j = \frac{1}{n} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)}.$$

□

Finally, we would like to show that the EM algorithm converges.

Theorem 4.12. *Let $(\psi^{(t)})_{t=0}^{\infty}$ be the sequence generated by the EM algorithm. Then*

$$\ell^M(\psi^{(t)}) \leq \ell^M(\psi^{(t+1)}).$$

Proof.

$$\begin{aligned} \ell^{\mathrm{M}}(\psi^{(t)}) &= F_t(\psi^{(t)}) && \text{def. of } F_t \\ &\leq F_t(\psi^{(t+1)}) && \psi^{(t+1)} := \operatorname*{argmax}_{\psi} F_t(\psi) \\ &\leq \ell^{\mathrm{M}}(\psi^{(t+1)}). && \text{def. of } F_t \end{aligned}$$

□

Corollary 4.13 (Convergence of the EM algorithm)**.** *When the marginal likelihood ℓ^M is bounded from above, then the EM algorithm converges.*

Proof. This is immediate from the previous theorem and the monotone convergence theorem. □

Example 4.14 (EM algorithm for the mixture of K Gaussians)**.**

$$\begin{aligned} Z &\sim \mathrm{Multi}(\eta_0, \ldots, \eta_{K-1}) \\ X \mid (Z = j) &\sim \mathcal{N}(\mu_J, \Sigma_j). \end{aligned}$$

Want to infer $\theta := \{\eta_0, \ldots, \eta_{K-1}, \mu_0, \ldots, \mu_{K-1}, \Sigma_0, \ldots, \Sigma_{K-1}\}$.

Initialize $\theta^{(0)}$.

- E step.

$$\gamma_{i,j}^{(t+1)} := \mathbb{P}_{\theta^{(t)}}(Z_i = j \mid X_i) = \frac{\eta^{(t)} p_{\mu_j^{(t)}, \Sigma_j^{(t)}}(X_i)}{\sum_{\ell=0}^{K-1} \eta^{(t)} p_{\mu_\ell^{(t)}, \Sigma_\ell^{(t)}}(X_i)}$$

- M step.

$$\begin{aligned} \eta_j^{(t+1)} &:= \frac{1}{n} \sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} \\ \mu_j^{(t+1)} &:= \frac{\sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} X_i}{\sum_{i=1}^{n} \gamma_{i,j}^{(t+1)}} \\ \Sigma_{i,j}^{(t+1)} &:= \frac{\sum_{i=1}^{n} \gamma_{i,j}^{(t+1)} (X_i - \mu_j^{(t+1)})(X_i - \mu_j^{(t+1)})^\top}{\sum_{i=1}^{n} \gamma_{i,j}^{(t+1)}} \end{aligned}$$

4.3 K-means algorithm

Definition 4.15. The **K-means algorithm** is the limiting procedure by applying the EM algorithm on a sequence of mixtures of K isotropic (spherical) Gaussians which become degenerate (variance shrinks to zero).

An **isotropic (spherical)** Gaussian distribution is a Gaussian distribution whose covariance matrix is of the form $\sigma^2 I_d$.

Let

$$\begin{aligned} Z &\sim \mathrm{Multi}(\eta_0, \ldots, \eta_{K-1}) \\ X \mid (Z = j) &\sim \mathcal{N}(\mu_j, \sigma^2 I). \end{aligned}$$

Note that not only is each conditional distribution isotropic, but also all the conditional distributions have the same covariance matrix $\sigma^2 I$.

Taking $\sigma^2 \to 0$ will make the mixture of Gaussians tend to K point masses (degenerate).

For the "E step" of the K-means algorithm,

$$\begin{aligned}\gamma_{i,j}^{(t+1)} &:= \frac{\eta_j \exp\left(\frac{-\|X_i-\mu_j\|_2^2}{2\sigma^2}\right)}{\sum_{\ell=0}^{K-1} \eta_\ell \exp\left(\frac{-\|X_i-\mu_\ell\|_2^2}{2\sigma^2}\right)} \\ &= \eta_j \left(\sum_{\ell=0}^{K-1} \eta_\ell \exp\left(\frac{\|X_i-\mu_j\|_2^2 - \|X_i-\mu_\ell\|_2^2}{2\sigma^2}\right)\right)^{-1}\end{aligned}$$

Taking $\sigma^2 \to 0$ gives

$$\gamma_{i,j} = \mathbf{1}\{\|X_i-\mu_j\|_2^2 < \|X_i-\mu_\ell\|_2^2 \text{ for every } \ell \neq j\},$$

i.e., let $\gamma_{i,j}$ be 1 if and only if μ_j is the closest cluster center to X_i. [In the unlikely case where there is more than one index that give the smallest 2-norm, then just randomly pick one.] Then, $\mathbb{P}(Z_i = j \mid X_i)$ is either 0 or 1, so we can just "set" the value of Z_i at this step. This is a "hard" assignment of Z_j in contrast to the "soft" assignment in the EM algorithm, where the condition distribution is more general.

So, we define $\gamma_{i,j} := \mathbf{1}\{Z_i = j\}$. Note that to do the E-step, the only information we need is $\mu_0^{(t)}, \ldots, \mu_{K-1}^{(t)}$. We do not need $\eta_0^{(t)}, \ldots, \eta_{K-1}^{(t)}$ or σ^2.

For the "M step" of the K-means algorithm, the updating rule is the same as the EM algorithm, but only needed for $\mu_0, \ldots, \mu_{K-1}$.

$$\mu_j^{(t+1)} := \frac{\sum_{i=1}^n \gamma_{i,j} X_i}{\sum_{i=1}^n \gamma_{i,j}} = \frac{\sum_{i=1}^n \mathbf{1}\{Z_i = j\} X_i}{\sum_{i=1}^n \mathbf{1}\{Z_i = j\}}$$

The formal description of the K-means algorithm is as follows. Initialize $\mu_0^{(0)}, \ldots, \mu_{K-1}^{(0)}$. For $t = 0, 1, 2, 3 \ldots$,

- **E step.** For $i = 1, \ldots, n$ and $j = 0, \ldots, K-1$, set

$$\gamma_{i,j}^{(t+1)} := \mathbf{1}\{\|X_i-\mu_j\|_2^2 < \|X_i-\mu_\ell\|_2^2 \text{ for every } \ell \neq j\}.$$

- **M step.** For $j = 0, \ldots, K-1$, set

$$\mu_j^{(t+1)} := \frac{\sum_{i=1}^n \gamma_{i,j} X_i}{\sum_{i=1}^n \gamma_{i,j}}$$

Repeat until convergence. [We have not yet proved that this converges.]

	K-means	EM
Model	mixture of K isotropic Gaussians limiting operation (mysterious)	mixture of K Gaussians
Initialization	randomly initialize $\mu_0, \ldots, \mu_{K-1}$	Initialize $\mu_0, \ldots, \mu_{K-1}$ using K-means. For the η_j and Σ_j, assign points to closest μ_j take sample proportion and sample covariance

We have just described the model-based derivation of the K-means algorithm. It is not clear from the limiting formulation that this is guaranteed to converge. We now present the risk-based perspective, which is equivalent to the model-based perspective, and from which we can prove convergence.

The goal is to choose K cluster centers to minimize

$$\widehat{R}(\mu_0, \ldots, \mu_{K-1}) := \frac{1}{n} \sum_{i=1}^n \min_{0 \leq j \leq K-1} \|X_i - \mu_j\|_2^2$$

This objective function is very natural: given cluster centers, we assign each point to the closest cluster center and its "badness" is the sum of the squared distances to the corresponding cluster centers. The minimum makes this optimization problem NP-hard.

Note that this corresponds to the population risk

$$R(\mu_0, \dots, \mu_{K-1}) := \mathbb{E}\left[\min_{0 \le j \le K-1} \|X_i - \mu_j\|_2^2\right]$$

The following is an equivalent form of the K-means optimization.

$$\min_{\substack{Z_i \in \{0,\dots,K-1\},\ i=1,\dots,n \\ \mu_0,\dots,\mu_K}} \frac{1}{n} \sum_{i=1}^{n} \|X_i - \mu_{Z_i}\|_2^2.$$

Introducing the extra parameters Z_i allows us to use block coordinate ascent. Let this objective function be $F(Z, \mu)$ where $Z = (Z_1, \dots, Z_n)$ and $\mu = (\mu_0, \dots, \mu_{K-1})$. We show that the risk-based perspective is equivalent to the model-based perspective.

Initialize $\mu_0^{(0)}, \dots, \mu_{K_1}^{(0)}$. For $t = 0, 1, 2, \dots$

- **E step.** If we have μ and want to update Z, then let Z_i be such that μ_{Z_i} is the closest cluster center to X_i. For all $i = 1, \dots, n$,
$$Z_i^{(t+1)} := \operatorname*{argmin}_{0 \le j \le K-1} \|X_i - \mu_j^{(t)}\|_2^2$$
- **M step.** If we have Z and want to update μ, then just let μ_j be the center of the points X_i that belong to cluster Z_j. For $j = 0, \dots, K-1$,
$$\mu_j^{(t+1)} := \frac{\sum_{i=1}^n \mathbf{1}\{Z_i^{(t+1)} = j\} X_i}{\sum_{i=1}^n \mathbf{1}\{Z_i^{(t+1)} = j\}}.$$

Theorem 4.16. *The K-mean algorithm converges.*

Proof. It is easy to show that the objective function $\frac{1}{n}\sum_{i=1}^n \|X_i - \mu_{Z_i}\|_2^2$ is nonincreasing with each E step and M step. Since it is bounded from below by zero, the monotone convergence theorem implies that it converges. □

4.4 Extensions

Example 4.17 (Hidden Markov model)**.**

$$\begin{aligned} Z_i &\sim \mathrm{Ber}(\eta) \\ X_i \mid (Z_i = j) &\sim \mathcal{N}(\mu_j, \Sigma_j) && j \in \{0, 1\} \\ \mathbb{P}(Z_k = j \mid Z_{k-1} = \ell) &= \theta_{j,\ell} && \text{Markov chain} \end{aligned}$$

Can compute MLE using EM.

Example 4.18 (Factor model)**.**

$$\begin{aligned} Z &\sim \mathcal{N}(0, I_k) \\ X \mid Z &\sim \mathcal{N}(\mu + \Lambda Z, \psi) && \psi \text{ is diagonal} \end{aligned}$$

This is an infinite mixture model because Z is continuous. Still can find MLE using EM.

Example 4.19 (Principal Component Analysis (PCA))**.**

$$\begin{aligned} Z &\sim \mathcal{N}(0, I_k) \\ X \mid Z &\sim \mathcal{N}(\mu + \Lambda Z, \sigma^2 I_d) \end{aligned}$$

Taking $\sigma^2 \to 0$, the MLE from EM is PCA.