Weekly Overview Slides of Statistical Machine Learning CSE 575, Fall 2023

Moses A. Boudourides[1]

SPA and SCAI
Arizona State University

[1] Moses.Boudourides@asu.edu

Week 5

Statistical Models

February 9, 2023

Bayesian Statistics

Bayes Rule for Statistical Models

$$p(\theta|\mathcal{D}) = \frac{p(\theta)p(\mathcal{D}|\theta)}{p(\mathcal{D})} = \frac{p(\theta)p(\mathcal{D}|\theta)}{\int p(\theta')p(\mathcal{D}|\theta')d\theta'}.$$

The Bernoulli and the Binomial Distributions

Tossing a coin gives rise to the **Bernoulli distribution** $Y \sim \mathrm{Ber}(\theta)$, with parameter $\theta \in [0,1]$, defined as

$$\mathrm{Ber}(y|\theta) = \begin{cases} 1-\theta, & \text{if } y = 0, \\ \theta, & \text{if } y = 1, \end{cases}$$

which is written in a more concise form as

$$\mathrm{Ber}(y|\theta) = \theta^y(1-\theta)^{1-y}.$$

When tossing a coin integer $N \geq 1$ times (N is called **sample size**), the Bernoulli distribution is generalized to the **Binomial distribution** $Y \sim \mathrm{Bin}(N,\theta)$ defined as

$$\mathrm{Bin}(y|N,\theta) = \binom{N}{y}\theta^y(1-\theta)^{N-y}.$$

The Bernoulli and the Binomial Likelihoods

Suppose that data $\mathcal{D}$ are iid obtained by tossing a coin N times. Assuming the Bernoulli distribution at each toss, the likelihood is

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{N} \theta^{y_i}(1-\theta)^{1-y_i} = \theta^{N_1}(1-\theta)^{N_0},$$

where N_1 is the number of times for $y_i = 1$ and N_0 is the number of times for $y_i = 0$. The counts N_1 and N_0 are called **sufficient statistics** (obviously, $N = N_1 + N_0$).

Now, if data $\mathcal{D}$ were following a Binomial model, then the likelihood would be

$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_{i=1}^{N} \mathrm{Bin}(y_i|N,\theta) = \prod_{i=1}^{N} \binom{N}{y_i} \theta^{y_i}(1-\theta)^{N-y_i} \\ &= N^{N_1}\theta^{N_1}(1-\theta)^{N_0}. \end{aligned}$$

Since the scaling factor N^{N_1} (from the $\binom{N}{y}$) is independent of θ (and, thus, it can be ignored), the inferences about θ are the same for both the Bernoulli and the Binomial model.

The Beta–Binomial Model

Notice that the likelihoods for both the Bernoulli and the Binomial model are of the form of the **beta distribution**, defined for two parameters $a, b \in [0, 1]$, as

$$\mathrm{Beta}(x|a, b) = \frac{1}{B(a, b)} x^{a-1}(1-x)^{b-1}$$

where the **beta function** is defined through the **gamma function** $\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du$ as

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Thus, by taking the prior for the Bernoulli or Binomial model to follow the Beta distribution, i.e.,

$$p(\theta) \propto \theta^{\check{\alpha}-1}(1-\theta)^{\check{\beta}-1} = \mathrm{Beta}(\check{\alpha}, \check{\beta}),$$

we get the posterior to be

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto \theta^{N_1}(1-\theta)^{N_0}\theta^{\check{\alpha}-1}(1-\theta)^{\check{\beta}-1} \\ &\propto \mathrm{Beta}(\check{\alpha}+N_1, \check{\beta}+N_0). \end{aligned}$$

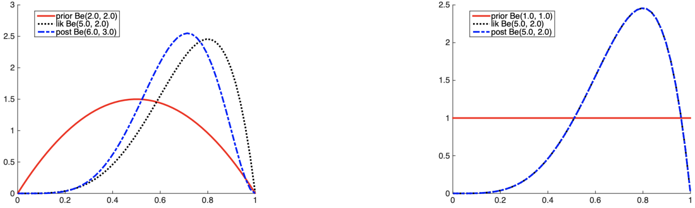
In other words, since the posterior has the same functional form as the prior, we say that the beta distribution is a conjugate prior for the Bernoulli–binomial likelihood.

Example

Updating a Beta prior with a Bernoulli likelihood with sufficient statistics $N_1 = 4, N_0 = 1$ in two cases:

(a) $\mathrm{Beta}(2, 2)$ prior.

(b) $\mathrm{Beta}(1, 1)$ prior, which is
$\mathrm{Beta}(\theta|1, 1) \propto \theta^0(1 - \theta)^0 = \mathrm{Unif}(\theta|0, 1)$.



(a)

(b)

The MAP Estimate

Since

$$\hat{\theta}_{\mathrm{map}} = \arg\max_{\theta} p(\theta|\mathcal{D}) = \arg\max_{\theta} \log p(\theta|\mathcal{D})$$
$$= \arg\max_{\theta} \log p(\theta) + \arg\max_{\theta} \log p(\mathcal{D}|\theta),$$

one can show that the MAP estimate is equal to

$$\hat{\theta}_{\mathrm{MAP}} = \frac{\check{\alpha} + N_1 - 1}{\check{\alpha} + \check{\beta} + N - 2}.$$

When we use the $\mathrm{Beta}(1, 1)$ prior, which amounts to the $\mathrm{Unif}(0, 1)$ distribution, the MAP estimates reduces to the MLE:

$$\hat{\theta}_{\mathrm{MLE}} = \frac{N_1}{N}.$$

Example (cont.)

(a) If we use the $\mathrm{Beta}(2, 2)$ prior amounts to **add–one smoothing**:

$$\hat{\theta}_{\mathrm{MAP}} = \frac{N_1 + 1}{N + 2}.$$

(b) If we use the $\mathrm{Beta}(1, 1)$ prior, $p(\theta) \propto 1$, the MAP estimate becomes the MLE, since $\log 1 = 0$:

$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} \log p(\theta).$$

Posterior Mean

If $p(\theta|\mathcal{D}) = \mathrm{Beta}(\check{\alpha}, \check{\beta})$, then the posterior mean is given by

$$\mathbb{E}[\theta|\mathcal{D}] = \frac{\check{\alpha}}{\check{\alpha} + \check{\beta}} = \frac{\check{\alpha}}{\check{N}},$$

where $\check{N} = \check{\alpha} + \check{\beta}$ is the strength (equivalent sample size) of the posterior. Moreover, since, as we have already seen, if the prior of a Bernoulli–binomial model follows the $\mathrm{Beta}(\check{\alpha}, \check{\beta})$ distribution, then the posterior follows the $\mathrm{Beta}(\check{\alpha} + N_1, \check{\beta} + N_0)$ distribution, we get, denoting the prior mean as $m = \check{\alpha}/\check{N}$,

$$\begin{aligned}\mathbb{E}[\theta|\mathcal{D}] = \frac{\check{\alpha} + N_1}{\check{\alpha} + N_1 + \check{\beta} + N_0} = \frac{\check{N}m + N_1}{n + \check{N}} = \frac{\check{N}}{N + \check{N}}m + \frac{N}{N + \check{N}}\frac{N_1}{N} \\ = \lambda m + (1 - \lambda)\hat{\theta}_{\mathrm{MLE}},\end{aligned}$$

where $\lambda = \frac{\check{N}}{N + \check{N}}$ is the ratio of the prior to posterior equivalent sample size. In other words, the posterior mean is a convex combination of the prior mean and the MLE. Therefore, the weaker the prior, the smaller is λ, and, hence, the closer the posterior mean is to the MLE.

Posterior Variance

To capture some notion of uncertainty in our estimate, the **standard error** of the estimate is defined as the posterior standard deviation:

$$\mathrm{se}(\theta) = \sqrt{\mathbb{V}[\theta|\mathcal{D}]}.$$

In the case of the Bernoulli–binomial model, we showed that the posterior follows a beta distribution $\mathrm{Beta}(\check{\alpha}, \check{\beta})$. The variance of the beta posterior is given by

$$\mathbb{V}[\theta|\mathcal{D}] = \frac{\hat{\alpha}\hat{\beta}}{(\hat{\alpha}+\hat{\beta})^2(1+\hat{\alpha}+\hat{\beta})} = \mathbb{E}[\theta|\mathcal{D}]^2 \frac{\hat{\beta}}{\hat{\alpha}(1+\hat{\alpha}+\hat{\beta})},$$

where $\hat{\alpha} = \check{\alpha} + N_1$ and $\hat{\beta} = \check{\beta} + N_0$. If $N \gg \check{\alpha} + \check{\beta}$, this simplifies to

$$\mathbb{V}[\theta|\mathcal{D}] \approx \frac{N_1 N_0}{N^3} = \frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N}.$$

Therefore, the standard error becomes

$$\mathrm{se}(\theta) \approx \sqrt{\frac{\hat{\theta}_{\mathrm{MLE}}(1-\hat{\theta}_{\mathrm{MLE}})}{N}}.$$

We see that the uncertainty goes down at a rate of $1/\sqrt{N}$. We also see that the uncertainty (variance) is maximized when $\hat{\theta} = 0.5$, and is minimized when $\hat{\theta}$ is close to 0 or 1. This makes sense, since it is easier to be sure that a coin is biased than to be sure that it is fair.

The Dirichlet Distribution

In the previous section, we discussed how to infer the probability that a coin comes up heads. In this section, we generalize these results to infer the probability that a dice with K sides comes up as face k.

Let S be a subset of a set X. Then the **characteristic function** of S is a function $\chi_s \colon X \to \{0,1\}$ defined as

$$\chi_s(x) = \begin{cases} 1, & \text{if } x \in S, \\ 0, & \text{if } x \notin S. \end{cases}$$

The **Dirichlet distribution** is a multivariate generalization of the beta distribution, which is defined on a **probability simplex**

$$S_K = \{\boldsymbol{x} = (x_1, \ldots, x_K) \colon 0 \le x_k \le 1, k = 1, \ldots, K, \sum_{k=1}^{K} x_k = 1\}$$

via the following PDF (w.r.t. the parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$, where $\alpha_k > 0$, for $k = 1, \ldots, K$):

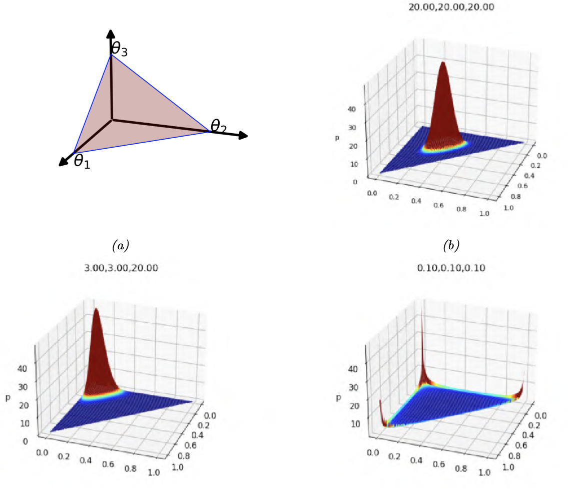
$$\mathrm{Dir}(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \chi_{S_K}(x_k),$$

where $\mathrm{B}(\boldsymbol{\alpha})$ is the multivariate beta function:

$$\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K} \alpha_k)}.$$

Example

(a) The Dirichlet distribution when $K = 3$ defines a distribution over the simplex, which can be represented by the triangular surface. Points on this surface satisfy $0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{3} \theta_k = 1$. (b) Plot of the Dirichlet density for $\boldsymbol{\alpha} = (20, 20, 20)$. (c) Plot of the Dirichlet density for $\boldsymbol{\alpha} = (3, 3, 20)$. (d) Plot of the Dirichlet density for $\boldsymbol{\alpha} = (0.1, 0.1, 0.1)$.



(a) *(b)* *(c)* *(d)*

The Dirichlet–Multinomial Model: Likelihood and Prior

Suppose we observe N dice roles, $\mathcal{D} = \{y_1, \ldots, y_N\}$, where $y_i \in \{1, \ldots, K\}$, for $i = 1, \ldots, N$. Assuming that $\mathcal{D}$ is iid, the likelihood has the form

$$p(\mathcal{D}|\theta) = \prod_{k=1}^{K} \theta_k^{N_k},$$

where N_k is the times for $y_i = k$, for $k = 1, \ldots, K$ (the N_k's are the sufficient statistics for this model and they sum up to N). The likelihood for the multinomial model has the same form, up to an irrelevant constant factor.

Since the parameter vector lives in the K–dimensional probability simplex $S_K = \{\boldsymbol{x} = (x_1, \ldots, x_K) : 0 \leq x_k \leq 1, k = 1, \ldots, K, \sum_{k=1}^{K} x_k = 1\}$, we need a prior that has support over this simplex (the support of a function is the subset of the domain where the function is not 0). Ideally, it would also be conjugate. Fortunately, the Dirichlet distribution satisfies both criteria. So, we will use the following prior:

$$\mathrm{Dir}(\boldsymbol{x}|\boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1} \chi_{S_K}(x_k).$$

The Dirichlet–Multinomial Model: Posterior

Combining the multinomial likelihood and the Dirichlet prior, we obtain the following posterior

$$\begin{aligned} p(\theta|\mathcal{D}) &\propto p(\mathcal{D}|\theta)\mathrm{Dir}(\boldsymbol{\theta}|\check{\boldsymbol{\alpha}}) \\ &= \left[\prod_{k=1}^{K}\theta_k^{N_k}\right]\left[\prod_{k=1}^{K}\theta_k^{\check{\alpha}_k-1}\right] \\ &\propto \mathrm{Dir}(\boldsymbol{\theta}|\check{\alpha}_1+N_1,\ldots,\check{\alpha}_K+N_K) \\ &= \mathrm{Dir}(\boldsymbol{\theta}|\hat{\boldsymbol{\alpha}}), \end{aligned}$$

where $\hat{\alpha}_k = \check{\alpha}_k + N_k$ are the parameters of the posterior.

Using the **Lagrange multiplier** technique in constrained optimization, the following MAP estimate can be computed (see details of proof in next slide):

$$\hat{\theta}_k = \frac{\hat{\alpha}_k - 1}{\sum_{k'=1}^{K}(\hat{\alpha}_{k'} - 1)}.$$

We can derive the mode of this posterior (i.e., the MAP estimate) by using calculus. However, we must enforce the constraint that $\sum_k \theta_k = 1$.[2]. We can do this by using a **Lagrange multiplier**. The constrained objective function, or **Lagrangian**, is given by the log likelihood plus log prior plus the constraint:

$$\ell(\boldsymbol{\theta}, \lambda) = \sum_k N_k \log \theta_k + \sum_k (\alpha_k - 1) \log \theta_k + \lambda \left(1 - \sum_k \theta_k\right) \tag{3.41}$$

To simplify notation, we define $N'_k \triangleq N_k + \alpha_k - 1$. Taking derivatives with respect to λ yields the original constraint:

$$\frac{\partial \ell}{\partial \lambda} = \left(1 - \sum_k \theta_k\right) = 0 \tag{3.42}$$

Taking derivatives with respect to θ_k yields

$$\begin{aligned} \frac{\partial \ell}{\partial \theta_k} &= \frac{N'_k}{\theta_k} - \lambda = 0 && (3.43) \\ N'_k &= \lambda \theta_k && (3.44) \end{aligned}$$

We can solve for λ using the sum-to-one constraint:

$$\begin{aligned} \sum_k N'_k &= \lambda \sum_k \theta_k && (3.45) \\ N + \alpha_0 - K &= \lambda && (3.46) \end{aligned}$$

where $\alpha_0 \triangleq \sum_{k=1}^K \alpha_k$ is the equivalent sample size of the prior. Thus the MAP estimate is given by

$$\hat{\theta}_k = \frac{N_k + \alpha_k - 1}{N + \alpha_0 - K} \tag{3.47}$$

The Gaussian Model

The Univariate Case: Prior and Posterior

Given N iid samples following the Gaussian distribution, $y_i \sim \mathcal{N}(\mu, \sigma^2)$, if the variance σ^2 is a known constant, the likelihood for the mean μ has the form:

$$p(\mathcal{D}|\mu) \propto \exp\left(-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \mu)^2 \right).$$

One can show that the conjugate prior is another Gaussian, $\mathcal{N}(\mu|\check{m}, \check{\tau}^2)$. Applying Bayes' rule for Gaussians, we find that the corresponding posterior is given by

$$p(\mu|\mathcal{D}, \sigma^2) = \mathcal{N}(\mu|\hat{m}, \hat{\tau}^2),$$

where

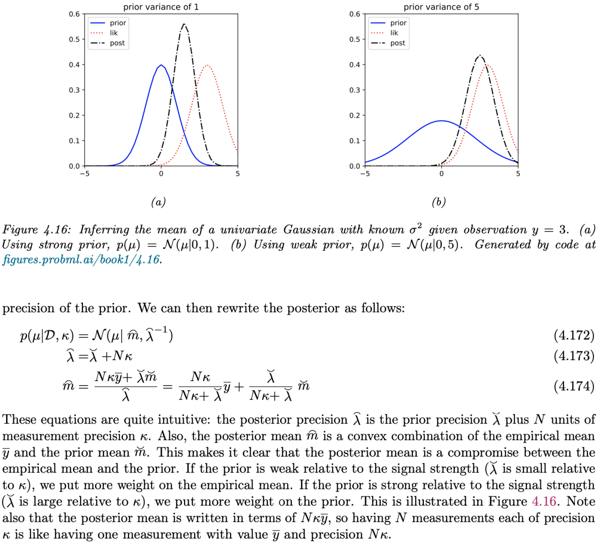
$$\hat{\tau}^2 = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\check{\tau}^2}} = \frac{\sigma^2 \check{\tau}^2}{N\check{\tau}^2 + \sigma^2},$$

$$\hat{m} = \hat{\tau}^2 \left(\frac{\check{m}}{\check{\tau}^2} + \frac{N\overline{y}}{\sigma^2} \right) = \frac{\sigma^2}{N\check{\tau}^2 + \sigma^2}\check{m} + \frac{N\check{\tau}^2}{N\check{\tau}^2 + \sigma^2}\overline{y},$$

where $\overline{y} = \frac{1}{N}\sum_{n=1}^{N} y_n$ is the empirical mean.

Interpretation

The previous result is easier to understand if we work in terms of the precision parameters, which are just inverse variances: $\kappa = 1/\sigma^2$, the observation precision, and $\hat{\lambda} = 1/\hat{\tau}^2$, the precision of the prior. Then the posterior is written as follows:



prior variance of 1
prior
lik
post
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−5
0
5
(a)
prior variance of 5
prior
lik
post
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−5
0
5
(b)

Figure 4.16: Inferring the mean of a univariate Gaussian with known σ^2 given observation $y = 3$. (a) Using strong prior, $p(\mu) = \mathcal{N}(\mu|0,1)$. (b) Using weak prior, $p(\mu) = \mathcal{N}(\mu|0,5)$. Generated by code at figures.probml.ai/book1/4.16.

precision of the prior. We can then rewrite the posterior as follows:

$$p(\mu|\mathcal{D},\kappa) = \mathcal{N}(\mu|\,\widehat{m}, \widehat{\lambda}^{-1}) \tag{4.172}$$

$$\widehat{\lambda} = \breve{\lambda} + N\kappa \tag{4.173}$$

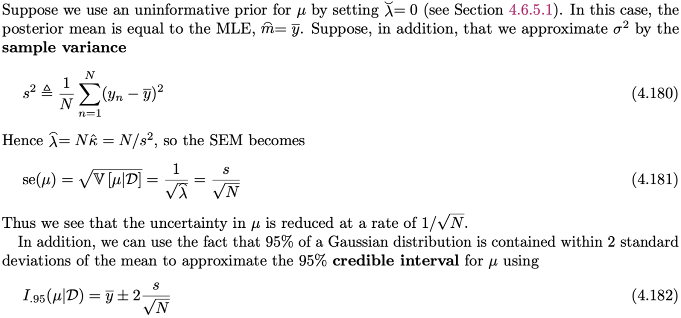
$$\widehat{m} = \frac{N\kappa\overline{y} + \breve{\lambda}\breve{m}}{\widehat{\lambda}} = \frac{N\kappa}{N\kappa + \breve{\lambda}}\overline{y} + \frac{\breve{\lambda}}{N\kappa + \breve{\lambda}}\breve{m} \tag{4.174}$$

These equations are quite intuitive: the posterior precision $\widehat{\lambda}$ is the prior precision $\breve{\lambda}$ plus N units of measurement precision κ. Also, the posterior mean $\widehat{m}$ is a convex combination of the empirical mean $\overline{y}$ and the prior mean $\breve{m}$. This makes it clear that the posterior mean is a compromise between the empirical mean and the prior. If the prior is weak relative to the signal strength ($\breve{\lambda}$ is small relative to κ), we put more weight on the empirical mean. If the prior is strong relative to the signal strength ($\breve{\lambda}$ is large relative to κ), we put more weight on the prior. This is illustrated in Figure 4.16. Note also that the posterior mean is written in terms of $N\kappa\overline{y}$, so having N measurements each of precision κ is like having one measurement with value $\overline{y}$ and precision $N\kappa$.

The univariate case: Posterior Variance

Remember that the posterior variance gives us a measure of confidence in our estimate and that the square root of this is called the **standard error of the mean**:

$$\mathrm{se}(\mu) = \sqrt{\mathbb{V}[\mu|\mathcal{D}]}.$$



Suppose we use an uninformative prior for μ by setting $\breve{\lambda}= 0$ (see Section 4.6.5.1). In this case, the posterior mean is equal to the MLE, $\widehat{m}= \overline{y}$. Suppose, in addition, that we approximate σ^2 by the **sample variance**

$$s^2 \triangleq \frac{1}{N}\sum_{n=1}^{N}(y_n - \overline{y})^2 \tag{4.180}$$

Hence $\widehat{\lambda}= N\hat{\kappa} = N/s^2$, so the SEM becomes

$$\mathrm{se}(\mu) = \sqrt{\mathbb{V}\left[\mu|\mathcal{D}\right]} = \frac{1}{\sqrt{\widehat{\lambda}}} = \frac{s}{\sqrt{N}} \tag{4.181}$$

Thus we see that the uncertainty in μ is reduced at a rate of $1/\sqrt{N}$.

In addition, we can use the fact that 95% of a Gaussian distribution is contained within 2 standard deviations of the mean to approximate the 95% **credible interval** for μ using

$$I_{.95}(\mu|\mathcal{D}) = \overline{y} \pm 2\frac{s}{\sqrt{N}} \tag{4.182}$$

The Multivariate Case: Notation

- We denote vectors by boldface lower case letters, such as $\boldsymbol{x}$.
- We denote matrices by boldface upper case letters, such as $\boldsymbol{X}$. We denote entries in a matrix by non-bold upper case letters, such as X_{ij}.
- All vectors are assumed to be column vectors unless noted otherwise. We use $[x_1, \ldots, x_D]$ to denote a column vector created by stacking D scalars. Similarly, if we write $\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D]$, where the left hand side is a tall column vector, we mean to stack the $\boldsymbol{x}_i$ along the rows. If we write $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_D]$, where the left hand side is a matrix, we mean to stack the $\boldsymbol{x}_i$ along the columns, creating a matrix.
- The **multivariate Gaussian**, otherwise called **multivariate normal**, will be abbreviated as **MVN**.

The Multivariate Case: Basics

The pdf of MVN in D dimensions is

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right].$$

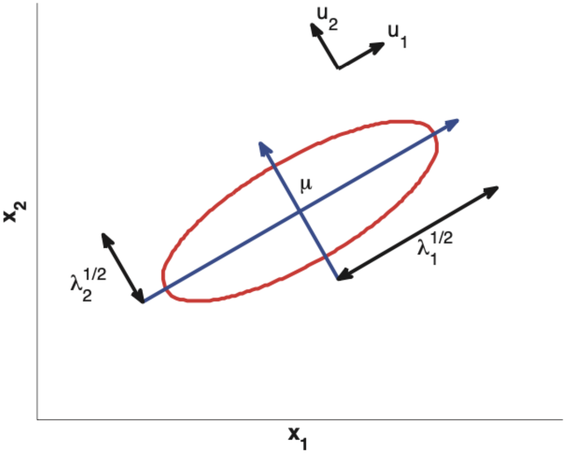
The expression inside the exponent is the **Mahalanobis distance** between a data vector $\boldsymbol{x}$ and the mean vector $\boldsymbol{\mu}$. We can gain a better understanding of this quantity by performing an **eigendecomposition** of $\boldsymbol{\Sigma}$. That is, we write $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$, where $\boldsymbol{U}$ is an orthonormal matrix of eigenvectors satsifying $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ (the identity matrix), and $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues. Using the eigendecomposition, we get $\boldsymbol{\Sigma}^{-1} = \boldsymbol{U}^{-T}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T = \sum_{i=1}^{D}\frac{1}{\lambda_i}\boldsymbol{u}_i\boldsymbol{u}_i^T$, where $\boldsymbol{u}_i$ is the i–th column of $\boldsymbol{U}$. Thus, the Mahalanobis distance is written as follows

$$\begin{aligned}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) &= (\boldsymbol{x}-\boldsymbol{\mu})^T\left(\sum_{i=1}^{D}\frac{1}{\lambda_i}\boldsymbol{u}_i\boldsymbol{u}_i^T\right)(\boldsymbol{x}-\boldsymbol{\mu}) \\ &= \sum_{i=1}^{D}\frac{1}{\lambda_i}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{u}_i\boldsymbol{u}_i^T(\boldsymbol{x}-\boldsymbol{\mu}) = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i},\end{aligned}$$

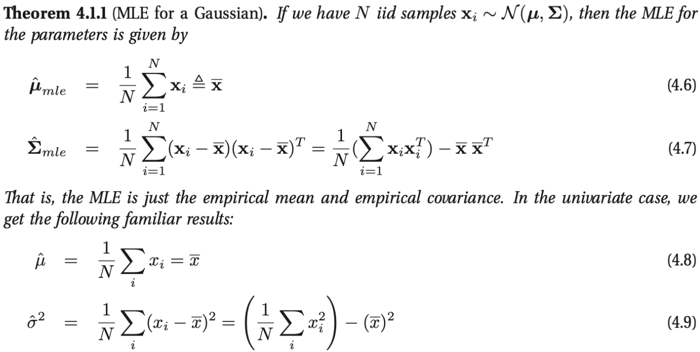
where $y_i = \boldsymbol{u}_i^T(\boldsymbol{x}-\boldsymbol{\mu})$. Recall that the equation for an ellipse in two dimensions is $\frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2} = 1$ and, thus, the interpretation that follows.

Geometric Interpretation

As we can see that the contours of equal probability density of a Gaussian lie along ellipses. The eigenvectors determine the orientation of the ellipse, and the eigenvalues determine how elogonated it is. In general, notice that the Mahalanobis distance corresponds to Euclidean distance in a transformed coordinate system, where we shift by $\boldsymbol{\mu}$ and rotate by $\boldsymbol{U}$.



MLE for an MVN



Theorem 4.1.1 (MLE for a Gaussian)**.** *If we have N iid samples $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the MLE for the parameters is given by*

$$\hat{\boldsymbol{\mu}}_{mle} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \triangleq \overline{\mathbf{x}} \tag{4.6}$$

$$\hat{\boldsymbol{\Sigma}}_{mle} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T = \frac{1}{N}(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T) - \overline{\mathbf{x}}\,\overline{\mathbf{x}}^T \tag{4.7}$$

That is, the MLE is just the empirical mean and empirical covariance. In the univariate case, we get the following familiar results:

$$\hat{\mu} = \frac{1}{N}\sum_i x_i = \overline{x} \tag{4.8}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - \overline{x})^2 = \left(\frac{1}{N}\sum_i x_i^2\right) - (\overline{x})^2 \tag{4.9}$$

The proof is in pages 99-100 of Murphy (2012).

MVN: Likelihood–Prior–Posterior

For D–dimensional (iid) data, the likelihood is:

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\mu}) &= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{y}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \frac{N}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y}_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_n - \boldsymbol{\mu}) \right] \\ &= \mathcal{N}(\overline{\boldsymbol{y}}|\boldsymbol{\mu}, \frac{1}{N}\boldsymbol{\Sigma}), \end{aligned}$$

where $\overline{\boldsymbol{y}} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{y}_n$. For simplicity, let us use a conjugate prior, which in this case is the following Gaussian

$$p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\check{\boldsymbol{m}}, \check{\boldsymbol{V}}).$$

Then we can derive the following Gaussian posterior for $\boldsymbol{\mu}$

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}|\hat{\boldsymbol{m}}, \hat{\boldsymbol{V}}),$$

where

$$\begin{aligned} \hat{\boldsymbol{V}}^{-1} &= \check{\boldsymbol{V}}^{-1} + N\boldsymbol{\Sigma}^{-1}, \\ \hat{\boldsymbol{\mu}} &= \hat{\boldsymbol{V}}\left(\boldsymbol{\Sigma}^{-1}(N\overline{\boldsymbol{y}}) + \check{\boldsymbol{V}}^{-1}\check{\boldsymbol{\mu}}\right). \end{aligned}$$