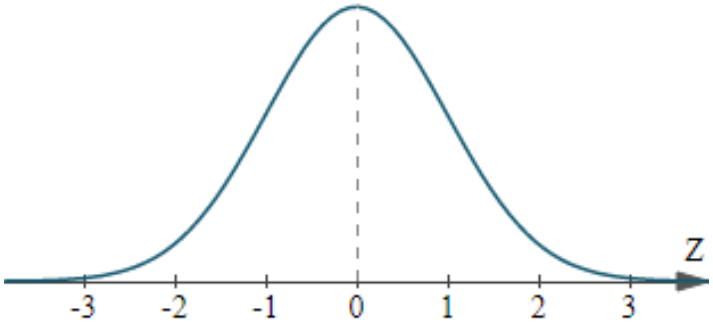
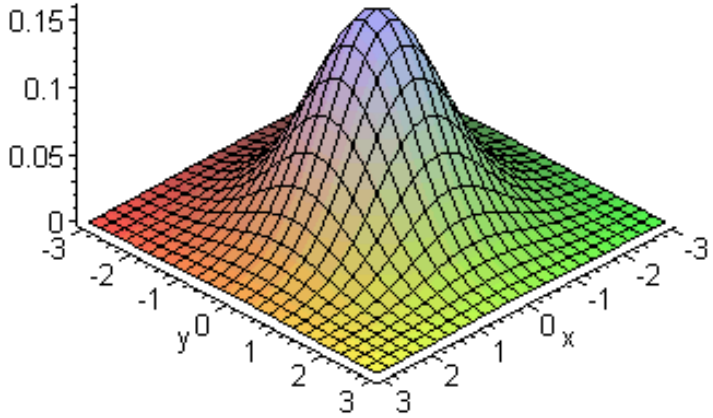
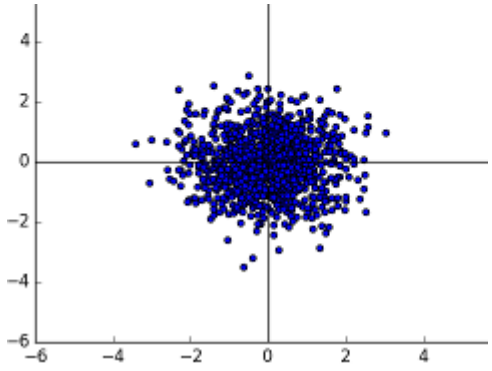
Generative Models

Multivariate Gaussian







one-dimensional

two-dimensional

Multivariate Gaussian

States. $x \in \mathbb{R}^d$

Parameters. $\mu \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$,
Σ positive definite.

Probability Density Function.

$X \sim N_d(\mu, \Sigma)$

μ is the mean vector of the observed data.

Σ is the covariance matrix

$$p(x \mid \mu, \Sigma) = (2\pi)^{-d/2} \, (\det \Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right\}$$

Computation of covariant matrix

- $X_1, X_2, \ldots, X_n$ are random variables.

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

Computation of covariant matrix

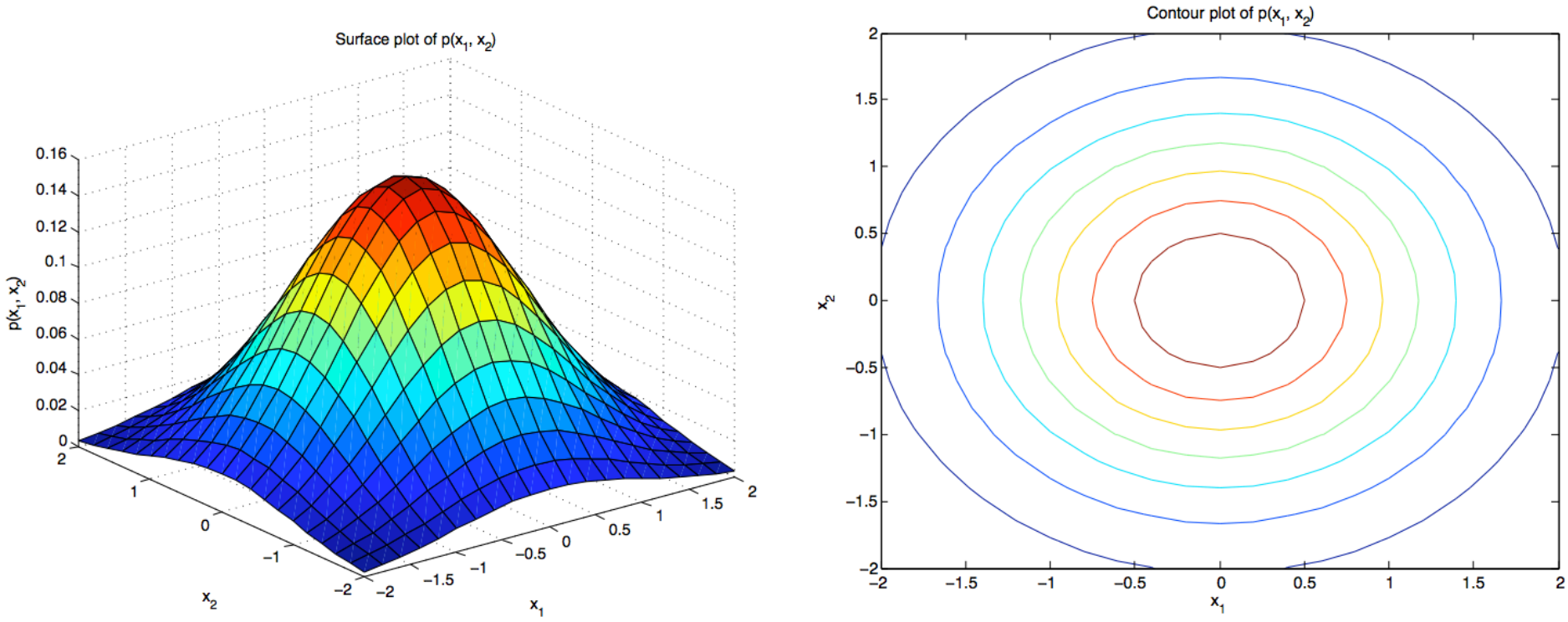
- $X = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 1 \end{bmatrix}$
- $\mu = [\mu_1, \mu_2, \mu_3] = [2, 1.5, 2];$
- $X - \mu = \begin{bmatrix} -1 & 0.5 & 1 \\ 1 & -0.5 & -1 \end{bmatrix}$
- $\sum = \frac{1}{2}(X - \mu)^T (X - \mu) = \frac{1}{2}\begin{bmatrix} 2 & -1 & -2 \\ -1 & 0.5 & 1 \\ -2 & 1 & 2 \end{bmatrix}$

Spherical Gaussian

- $p(x \mid \mu, \sigma^2) = (2\pi\sigma^2)^{-d/2} \exp\left\{-\frac{1}{2\sigma^2}\|x-\mu\|^2\right\}, \quad \mu \in \mathbb{R}^d, \sigma \in \mathbb{R}.$

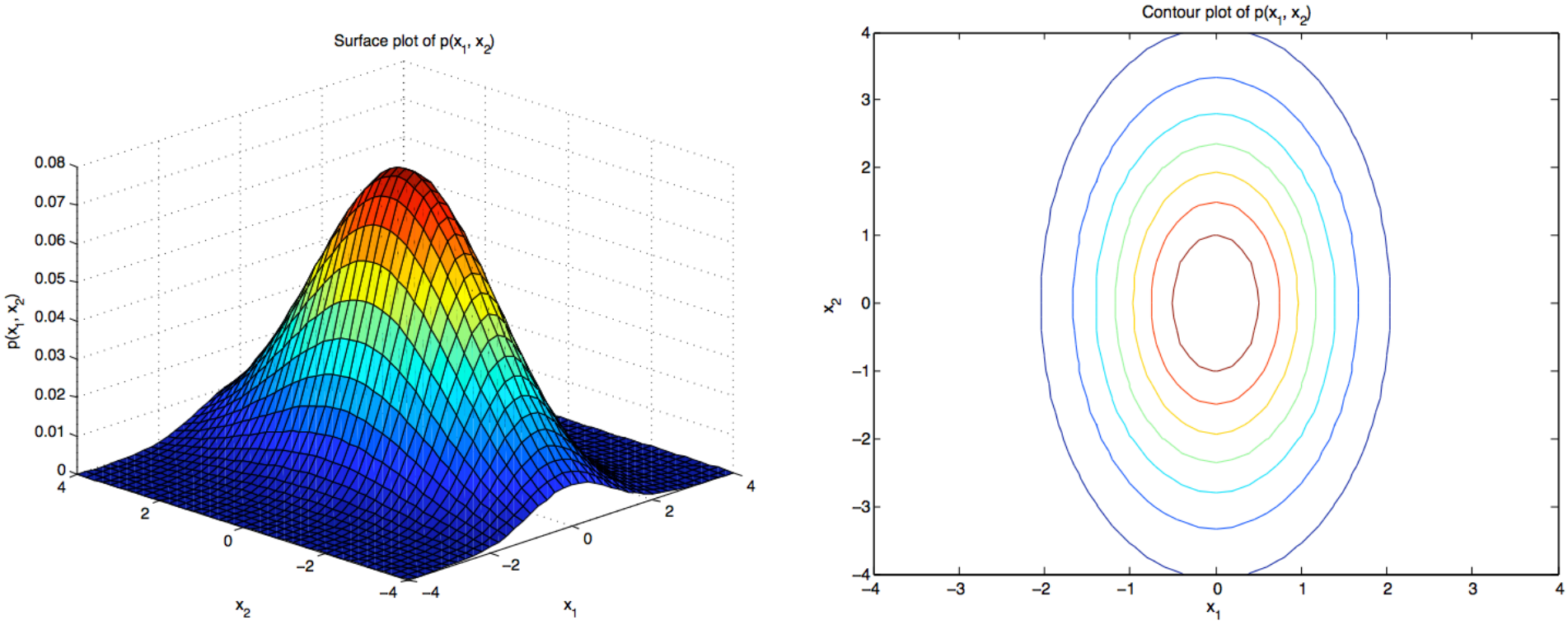
$$\sum = \begin{bmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{bmatrix}$$ diagonal covariance, equal variances

Spherical Gaussian



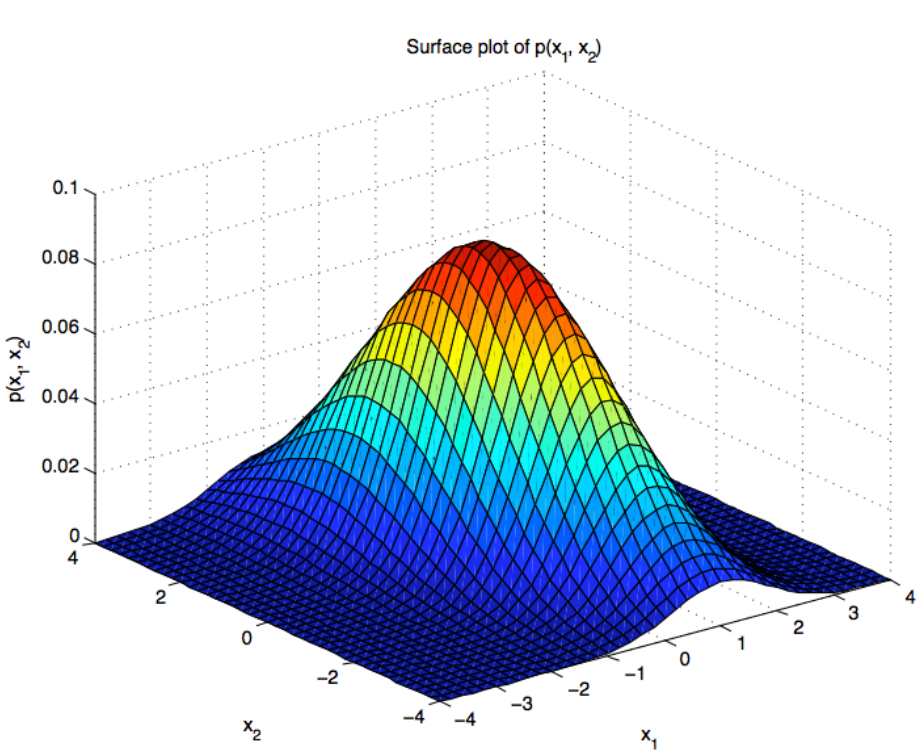
(a) Spherical Gaussian (diagonal covariance, equal variances)

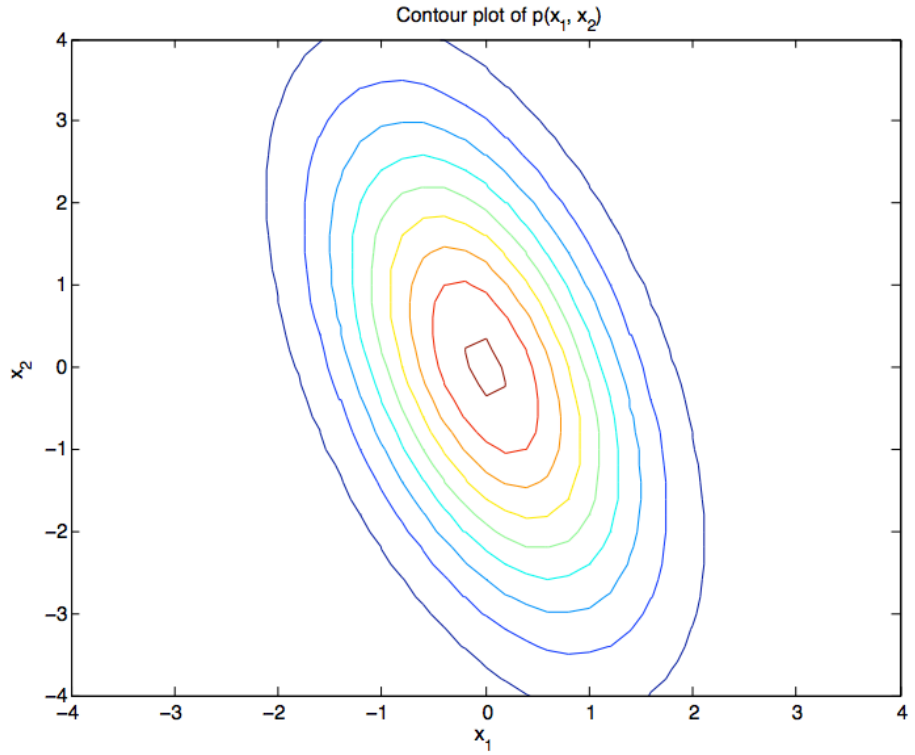
Gaussian with diagonal covariance matrix



(b) Gaussian with diagonal covariance matrix

Gaussian with full covariance matrix





(c) Gaussian with full covariance matrix

Maximum Likelihood

Maximum likelihood Estimate (MLE)

Model.
For each parameter $\theta \in \mathbb{R}^d$, we have a distribution $p(x|\theta)$.

Data.
Set of observations $\mathcal{S} = \{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$

Parameter estimation.
Find the parameter $\hat{\theta} \in \mathbb{R}^d$ that 'best' describes the data $\mathcal{S}$.

Likelihood. $p(\mathcal{S}|\theta) = \prod_{x \in \mathcal{S}} p(x|\theta)$

Maximum Likelihood Estimate. $\hat{\theta} = \operatorname{argmax}_\theta p(\mathcal{S}|\theta)$

Log Likelihood

Maximizing the likelihood

$$p(\mathcal{S}|\theta) = \prod_{x \in \mathcal{S}} p(x|\theta)$$

is the same as minimizing the negative log likelihood

$$-\frac{1}{\mathrm{n}} \log p(\mathcal{S}|\theta) \quad = -\frac{1}{n} \log \prod_{x \in \mathcal{S}} p(x|\theta)$$

$$= \frac{1}{n} \sum_{x \in \mathcal{S}} -\log p(x|\theta).$$

In fact, we define this to be the *training loss.*

It is the average of the point loss $-\log p(x|\theta)$.

MLE

Example.

Treat a document $\mathcal{S}$ as a *bag of words*, counting only how many times $n(w)$ each dictionary word $w \in W$ appears in $\mathcal{S}$.

Assume that the document was *generated* one word at a time, each word independently from the others.

Assume that words are drawn from same distribution $\theta = (\theta_w)$ where $w \in W$ appears with probability θ_w.

independent and identically distributed

Likelihood. $p(\mathcal{S}|\theta) = \prod_{w \in W} \theta_w^{n(w)}$

Training Loss. $\mathcal{L}_n(\theta) = \frac{1}{n} \sum_{w \in W} -n(w) \log \theta_w$

MLE

minimize $\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{w \in W} -n(w) \log \theta_w$

subject to $\theta_w \geq 0$ for all $w \in W$

$\sum_{w \in W} \theta_w = 1$

Using Lagrange multipliers, we showed that the minimum is attained at the following point.

MLE. $\hat{\theta}_w = \frac{n(w)}{\sum_{w' \in W} n(w')}$

MLE – Gaussian

Example.

Assume that data $\mathcal{S} = \{x^{(1)}, x^{(2)}, \dots, x^{(n)}\}$ is independent and identically distributed to a spherical Gaussian with mean $\mu \in \mathbb{R}^d$ and variance σ^2.

Training Loss.

$$\mathcal{L}_n(\mu, \sigma^2) = -\frac{1}{n}\log p(\mathcal{S}|\mu, \sigma^2) = -\frac{1}{n}\sum_{x\in\mathcal{S}} \log p(x|\mu, \sigma^2)$$
$$= \frac{d}{2}\log(2\pi\sigma^2) + \frac{1}{2n\sigma^2}\sum_{x\in\mathcal{S}}\|x-\mu\|^2$$

MLE. $$\hat{\mu} = \frac{1}{n}\sum_{x\in\mathcal{S}} x\,, \quad \hat{\sigma}^2 = \frac{1}{nd}\sum_{x\in\mathcal{S}}\|x-\hat{\mu}\|^2$$

solve
$\frac{\partial \ell}{\partial \mu} = 0$
$\frac{\partial \ell}{\partial \sigma} = 0$

Summary

- Distributions
  - Multinomial
  - Multivariate Gaussian
  - Spherical Gaussian
- Maximum Likelihood
  - Training Loss
  - Multinomial MLE
  - Gaussian MLE

Intended Learning Outcomes

Probabilistic Models

- Explain how probabilistic modeling is useful in machine learning.
- Define the multinomial and the multivariate Gaussian distributions.
- Define *independent and identically distributed* (i.i.d.)

Maximum Likelihood

- Given a probabilistic model and data, describe a training loss for parameter estimation. Define maximum likelihood estimate (MLE).
- Given a probabilistic model and data, derive the MLE. Write down the MLE for multinomial and multivariate Gaussian distributions.