

Selecting an estimate

Tim Bretl

Department of Aerospace Engineering
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign

AE498MPA
October 10, 2007

Outline

- What are the mean and covariance?
- Why are these terms useful?
- Why is the mean a good choice of estimate?
- How are the mean and covariance extracted from sampled data?
- What if the data come from a multi-modal distribution?

Expectation

- Let $x \in \mathbb{R}$ be a discrete random variable with probability mass function $p(x)$.
- The **expectation** or **mean** of x , often denoted μ , is

$$\mathbf{E}(x) = \sum_x xp(x).$$

- The mean is the **center of mass** of the distribution $p(x)$.

Expectation is linear

- Let $x \in \mathbb{R}$ and $y \in \mathbb{R}$ be discrete random variables.
- From the law of total probability,

$$p(x) = \sum_y p(x, y) \quad \text{and} \quad p(y) = \sum_x p(x, y).$$

- So expectation is **linear**:

$$\begin{aligned} \mathbf{E}(ax + by + c) &= \sum_x \sum_y (ax + by + c)p(x, y) \\ &= a \sum_x x \sum_y p(x, y) + b \sum_y y \sum_x p(x, y) + c \sum_{x, y} p(x, y) \\ &= a \sum_x xp(x) + b \sum_y yp(y) + c \\ &= a\mathbf{E}(x) + b\mathbf{E}(y) + c. \end{aligned}$$

Covariance

- Let $x \in \mathbb{R}$ be a discrete random variable with probability mass function $p(x)$.
- The **covariance** of x , often denoted Σ , is

$$\mathbf{E} \left((x - \mathbf{E}(x))^2 \right) = \sum_x (x - \mu)^2 p(x).$$

- Covariance measures the **mean square deviation** from the mean.
- It indicates how **wide** the distribution $p(x)$ is about the mean.

Covariance in terms of moments

- Let $\mu = \mathbf{E}(x)$ and $\Sigma = \mathbf{E} \left((x - \mathbf{E}(x))^2 \right)$.
- Since expectation is linear, we can write

$$\begin{aligned}\Sigma &= \mathbf{E} \left((x - \mu)^2 \right) \\ &= \mathbf{E} (x^2 - 2x\mu + \mu^2) \\ &= \mathbf{E}(x^2) - 2\mu\mathbf{E}(x) + \mu^2 \\ &= \mathbf{E}(x^2) - \mu^2.\end{aligned}$$

The mean and covariance give confidence bounds

- Let x be a discrete random variable. Then for *any* probability mass function $p(x)$, we can bound the probability that x is far from μ .

Chebyshev Inequality

$$P(|x - \mu| \geq a) \leq \frac{\Sigma}{a^2}$$

- From this inequality, we can construct a **confidence interval**.

$$P(x \in [\mu - a, \mu + a]) \geq 1 - \frac{\Sigma}{a^2}$$

- Define the **standard deviation** by $\sigma = \sqrt{\Sigma}$ and we can write

$$P(x \in [\mu - a, \mu + a]) \geq 1 - (\sigma/a)^2$$

So it is natural to define the confidence interval as some number of standard deviations, for example $a = 6\sigma$ gives probability 0.97.

The “best” estimate (mean square error)

- To decide on the “best” estimate \hat{x} , we need a **cost function**.
- A common choice of cost function is the **mean square error**, which penalizes the expected sum-square deviation from the estimate:

$$\mathbf{E}((x - \hat{x})^2)$$

- Notice that

$$\begin{aligned}\mathbf{E}((x - \hat{x})^2) &= \mathbf{E}(x^2 - 2x\hat{x} + \hat{x}^2) \\ &= \mathbf{E}(x^2) - 2\hat{x}\mathbf{E}(x) + \hat{x}^2.\end{aligned}$$

Differentiating with respect to \hat{x} gives

$$-2\mathbf{E}(x) + 2\hat{x} = 0$$

so the estimate \hat{x} that minimizes the mean square error is the **mean**

$$\hat{x} = \mu.$$

Another interpretation of minimum mean square error

- For a random variable z , we showed that

$$\Sigma_z = \mathbf{E}(z^2) - \mu_z^2.$$

- Apply this to the **error**

$$z = x - \hat{x}$$

and we have

$$\begin{aligned}\mathbf{E}((x - \hat{x})^2) &= \mathbf{E}(z^2) \\ &= \mu_z^2 + \Sigma_z \\ &= \mu_z^2 + \mathbf{E}((z - \mu_z)^2) \\ &= (\mu - \hat{x})^2 + \mathbf{E}(((x - \hat{x}) - (\mu - \hat{x}))^2) \\ &= (\mu - \hat{x})^2 + \mathbf{E}((x - \mu)^2) \\ &= (\mu - \hat{x})^2 + \Sigma\end{aligned}$$

- Our choice of \hat{x} does not change Σ . This is the error that we **cannot remove**. The best we can do is make the **bias** zero, choosing $\hat{x} = \mu$.

Extracting the mean from sampled data

- The **sample mean** of n samples x_1, \dots, x_n , each generated independently with mean μ and covariance Σ , is

$$s_n = \frac{1}{n} \sum_{i=1}^n x_i$$

- The sample mean is **unbiased** since $\mathbf{E}(s_n) = \mu$.
- Also, since each x_i is independent, you can verify that

$$\text{cov}(s_n) = \frac{\text{cov}(x_1) + \dots + \text{cov}(x_n)}{n^2} = \frac{\Sigma}{n}.$$

Applying the Chebyshev inequality, we have

$$p(|s_n - \mu| \geq \epsilon) \leq \frac{\Sigma}{n\epsilon^2}.$$

So as n gets large, the error probability goes to zero.

Extracting the covariance from sampled data

- The **sample covariance** of n samples x_1, \dots, x_n , each generated independently with mean μ and covariance Σ , is

$$Q_n = \frac{1}{n-1} \sum_{i=1}^n (x_i - s_n)^2$$

- You can verify that this estimate is **unbiased**, and that the error probability goes to zero as n gets large.

Density extraction for data sampled from a multi-modal distribution, using k -means clustering

Given n particles x_1, \dots, x_n and k initial sample means μ_1, \dots, μ_k , we iterate until the means stop moving:

- Initialize empty sets $L_1 = \{\}, \dots, L_k = \{\}$.
- Divide the particles into groups according to the nearest mean. So for each $i = 1, \dots, n$, find the mean μ_j that minimizes $\|x_i - \mu_j\|$, and add the index i to L_j .
- Recompute the means. So for each $j = 1, \dots, k$, let

$$\mu_j = \frac{1}{\text{length}(L_j)} \sum_{i \in L_j} x_i.$$

Afterward, we can compute the sample covariance $\Sigma_1, \dots, \Sigma_k$ corresponding to each mean. So we can represent the distribution of particles by k gaussians. (How do we choose k ?)