NUS National University of Singapore | School of Computing

CS5340
Uncertainty Modeling in AI

Lecture 2:
Fitting Probability Models

Asst. Prof. Lee Gim Hee

AY 2018/19

Semester 1

Course Schedule

Week	Date	Topic	Remarks
1	15 Aug	Introduction to probabilities and probability distributions	
2	22 Aug	Fitting probability models	Hari Raya Haji*
3	29 Aug	Bayesian networks (Directed graphical models)	
4	05 Sep	Markov random Fields (Undirected graphical models)	
5	12 Sep	I will be traveling	No Lecture
6	19 Sep	Variable elimination and belief propagation	
-	26 Sep	Recess week	No lecture
7	03 Oct	Factor graph and the junction tree algorithm	
8	10 Oct	Parameter learning with complete data	
9	17 Oct	Mixture models and the EM algorithm	
10	24 Oct	Hidden Markov Models (HMM)	
11	31 Oct	Monte Carlo inference (Sampling)	
12	07 Nov	Variational inference	
13	14 Nov	Graph-cut and alpha expansion	

* **Make-up lecture**: 25 Aug (Sat), 9.30am-12.30pm, LT 15

Acknowledgements

- A lot of slides and content of this lecture are adopted from:

1. "Computer Vision: Models, Learning, and Inference", Simon Prince.

2. "Pattern Recognition and Machine Learning", Christopher Bishop.

Learning Outcomes

- Students should be able to:

  1. Use the Maximum Likelihood, Maximum a Posterior and Bayesian approaches to learn the unknown parameters of probability distributions of a single random variable from data.

  2. Apply the concept of Naïve Bayes to simplify the parameter learning process.

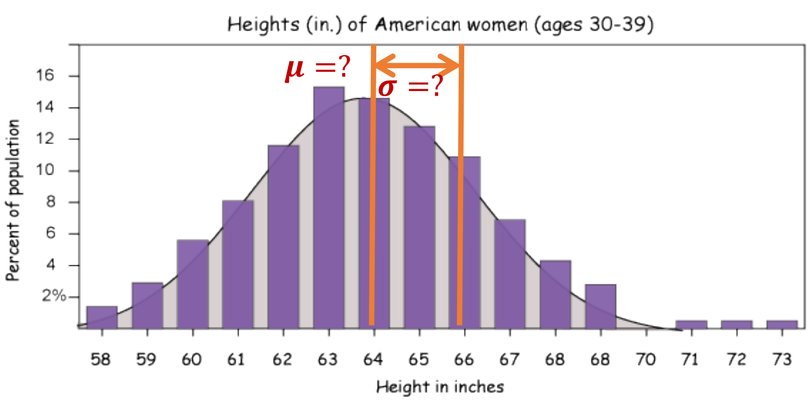
Fitting Probability Models

- In the last lecture, we have seen the definitions of some common parametric probability distributions $p(x|\theta)$.

- In this lecture, we will look at how to learn the unknown parameters θ from a set of given data, i.e. instances of the random variable, $X : \{x[1], \dots, x[N]\}$.

Example:

Fitting a Normal distribution to the height measurements of a population.



Given:
Height measurements $X : \{58.5, 60.1, 65, 64, \dots 72\}$, and probability distribution model

$$p(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\frac{(x-\mu)^2}{2\sigma^2}, \qquad \theta = \{\mu, \sigma^2\}$$

Find:
The unknown parameters θ!

Image source: http://www.drcruzan.com/ProbStat_Distributions.html

Fitting Probability Models

- Three approaches to learn the unknown parameters θ from a set of given data $X : \{x[1], \dots, x[N]\}$:

  1. Maximum likelihood estimate (MLE)
  2. Maximum a posteriori (MAP)
  3. Bayesian approach

Maximum Likelihood Estimate (MLE)

- **Fitting:** As the name suggests, we find the unknown parameters θ that maximize the likelihood $p(x|\theta)$.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}[p(x|\theta)]$$

$$= \underset{\theta}{\operatorname{argmax}}\left[\prod_{i=1}^{N} p(X = x[i]\ |\theta)\right] \qquad \text{(Naïve Bayes)}$$

- We have assumed that data was independent (hence product). Also known as the Naïve Bayes assumption.

- **Predictive Density:** Evaluate new data point x^* under the probability distribution with the best parameters $p(x^*|\hat{\theta})$.

Maximum a Posteriori (MAP)

- **Fitting:** As the name suggests, we find the unknown parameters θ that maximize the a posterior probability $p(\theta|x)$.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}[p(\theta|x)]$$

$$= \underset{\theta}{\operatorname{argmax}}\left[\frac{p(x|\theta)p(\theta)}{p(x)}\right]$$ (Bayes' rule)

$$= \underset{\theta}{\operatorname{argmax}}\left[\frac{\prod_{i=1}^{N} p(x[i] \mid \theta)\, p(\theta)}{p(x)}\right]$$ (Naïve Bayes)

$$= \underset{\theta}{\operatorname{argmax}}\left[\prod_{i=1}^{N} p(x[i] \mid \theta)\, p(\theta)\right]$$ ($p(x)$ is removed since it is independent of θ)

Maximum a Posteriori (MAP)

- **Fitting**: As the name suggests, we find the unknown parameters θ that maximize the a posterior probability $p(\theta|x)$.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left[\prod_{i=1}^{N} p(x[i] \mid \theta)\, p(\theta) \right]$$

- **Predictive Density**: Evaluate new data point x^* under the probability distribution with the best parameters $p(x^*|\hat{\theta})$.

Bayesian Approach

- **Fitting**: Instead of a point estimate $\hat{\theta}$, compute the posterior distribution over all possible parameter values using Bayes' rule:

$$p(\theta|x) = \frac{\prod_{i=1}^{N} p(x[i] \mid \theta) p(\theta)}{p(x)}$$

- **Principle**: why pick one set of parameters? There are many values that could have explained the data. Try to capture all of the possibilities.

Bayesian Approach

Predictive Density:

$$p(x^*|x) = \frac{p(x^*, x)}{p(x)}$$ (Conditional probability)

$$= \frac{\int p(x^*, x, \theta) d\theta}{p(x)}$$ (Marginal probability)

$$= \frac{\int p(x^*, \theta|x) \cancel{p(x)} d\theta}{\cancel{p(x)}}$$ (Conditional probability)

$$= \int p(x^*|x, \theta) p(\theta|x) d\theta$$ (Conditional probability)

$$= \int p(x^*|\theta) p(\theta|x) d\theta$$ (Conditional Independence)

Bayesian Approach

Predictive Density:

$$p(x^*|x) = \int p(x^*|\theta)p(\theta|x)d\theta$$

Prediction for each possible θ

Weights

Make a prediction that is an infinite weighted sum (integral) of the predictions for each parameter value, where weights are the probabilities.

Predictive Densities for 3 Approaches

Maximum Likelihood Estimate (MLE):

Evaluate new data point x^* under probability distribution with MLE parameters $p(x^*|\hat{\theta})$.

Maximum a Posteriori (MAP):

Evaluate new data point x^* under probability distribution with MAP parameters $p(x^*|\hat{\theta})$.

Bayesian:

Calculate weighted sum of predictions from all possible values of parameters

$$p(x^*|x) = \int p(x^*|\theta)p(\theta|x)d\theta$$

Predictive Densities for 3 Approaches

How to rationalize different forms?

Consider MLE and MAP estimates as probability distributions with zero probability everywhere except at estimate (i.e. delta functions):

$$p(x^*|x) = \int p(x^*|\theta)\delta[\theta - \hat{\theta}]d\theta$$

$$= p(x^*|\hat{\theta})$$

Examples

- Let's look at two examples on fitting probability model – univariate Normal distribution, and categorical distribution.

- Approach the same problem 3 different ways:
  1. Learn **MLE** parameters
  2. Learn **MAP** parameters
  3. Learn **Bayesian distribution** of parameters

- Will we get the same results?

Example 1: Univariate Normal Distribution

Problem:

Fit an univariate normal distribution model to a set of scalar data $X : \{x[1], \ldots x[N]\}$.

Recall that the univariate normal distribution is given by:

$$p(x) = \mathrm{Norm}_x[\mu, \sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$

Our goal is to find the two unknown parameters μ and σ^2.

Example 1: Univariate Normal Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}[p(x|\theta)]$$

$$= \underset{\theta}{\operatorname{argmax}}\left[\prod_{i=1}^{N} p(x[i] \mid \theta)\right]$$

(Naïve Bayes)

Likelihood given by pdf

$$p(x|\mu,\sigma^2) = \mathrm{Norm}_x[\mu,\sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2}$$

Example 1: Univariate Normal Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

$$p(x|\mu,\sigma^2) = \prod_{i=1}^{N} p(x[i] \mid \mu,\sigma^2)$$

(Naïve Bayes)

$$= \prod_{i=1}^{N} \mathrm{Norm}_{x[i]}\,[\mu,\sigma^2]$$

$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \prod_{i=1}^{N} \exp\left[-0.5\,\frac{(x[i]-\mu)^2}{\sigma^2}\right]$$

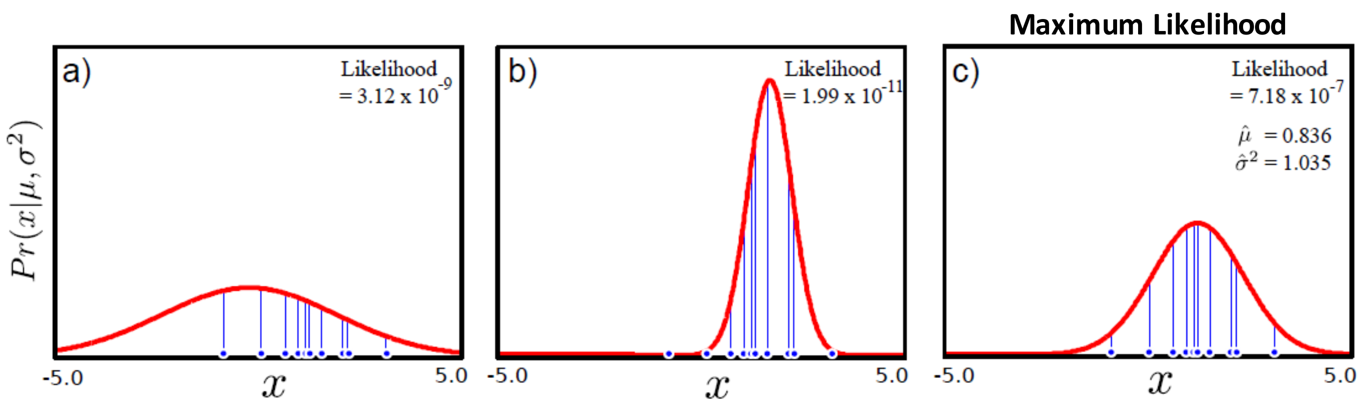
$$= \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-0.5 \sum_{i=1}^{N} \frac{(x[i]-\mu)^2}{\sigma^2}\right]$$

Example 1: Univariate Normal Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

$$p(x|\mu,\sigma^2) = \prod_{i=1}^{N} \text{Norm}_{x[i]}[\mu,\sigma^2], \qquad \hat{\mu},\ \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\text{argmax}}[p(x \mid \mu,\sigma^2)]$$

Intuition behind MLE:



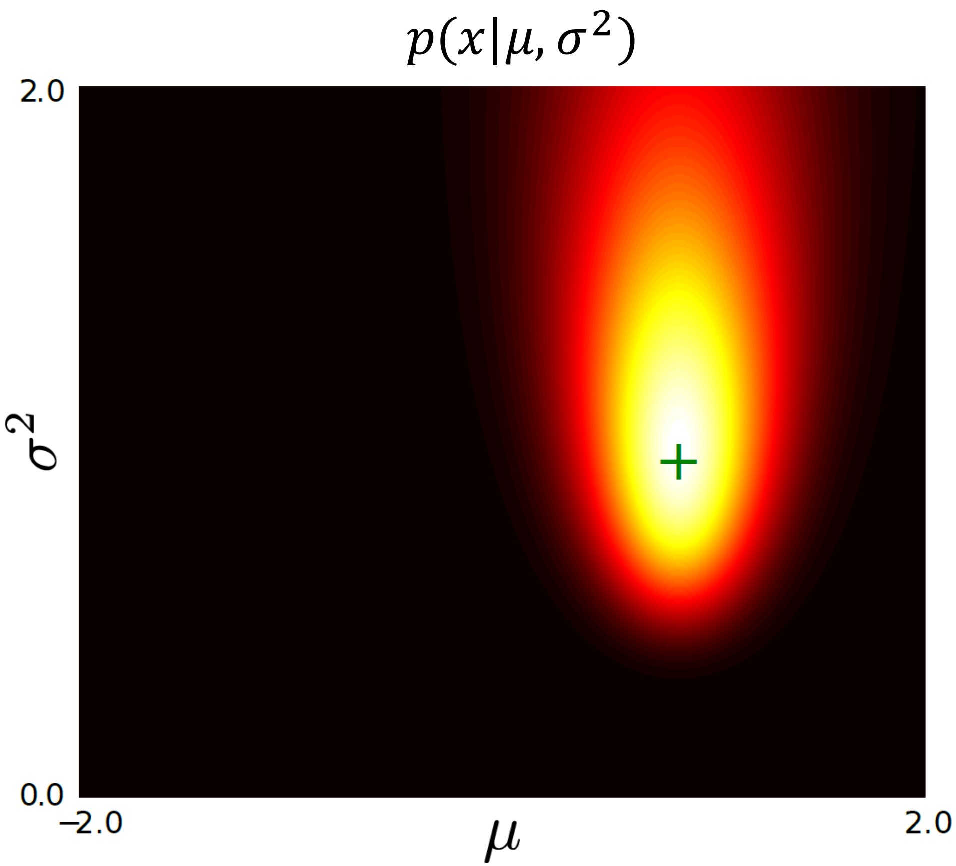
- Blue dots are the observed data $X : \{x[1], \ldots, x[N]\}$.
- Red curves are the Normal distribution for a possible μ and σ^2.
- The likelihood of a set of independently sampled data is the product of the individual likelihoods (blue vertical lines).
- The correct μ and σ^2 give the maximum likelihood.

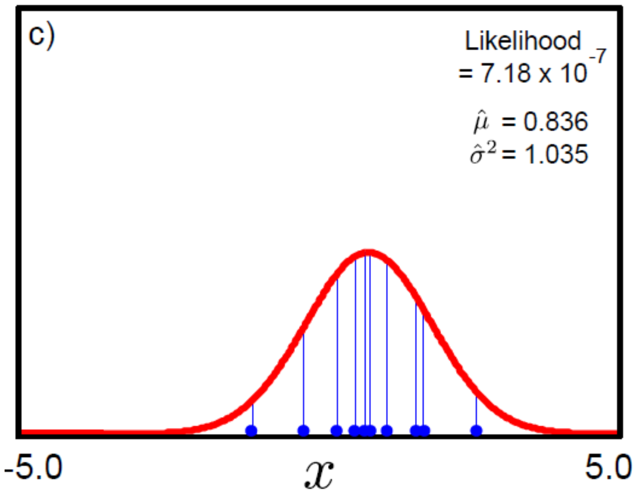
Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Example 1: Univariate Normal Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

Intuition behind MLE:





Plotted surface of likelihoods as a function of possible parameter values.

ML Solution is at the peak.

Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Example 1: Univariate Normal Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

Algebraically:

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}}[p(x|\mu, \sigma^2)]$$

where

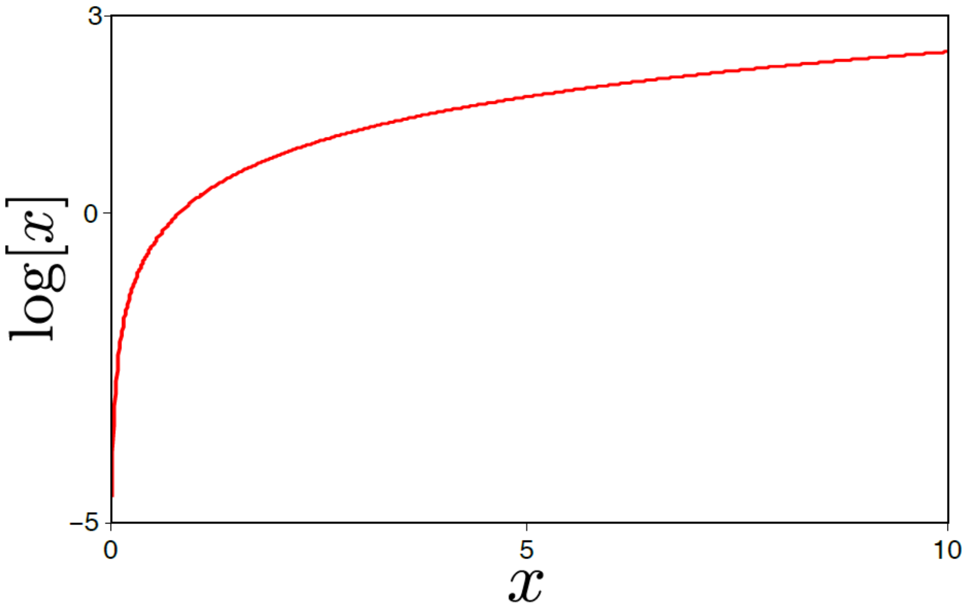
$$p(x|\mu, \sigma^2) = \prod_{i=1}^{N} \mathrm{Norm}_{x[i]}[\mu, \sigma^2],$$

or alternatively, we can maximize the logarithm:

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log\left[\mathrm{Norm}_{x[i]}[\mu, \sigma^2]\right]$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[-0.5N\log\,[2\pi] - 0.5N\log\sigma^2 - 0.5\sum_{i=1}^{N} \frac{(x[i]-\mu)^2}{\sigma^2}\right]$$

Why the Logarithm?



- The logarithm is a monotonic transformation.
- Hence, the position of the peak stays in the same place.
- But the log likelihood is easier to work with.

Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Example 1: Univariate Normal Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \sum_{i=1}^{N} \log\left[\text{Norm}_{x[i]}[\mu, \sigma^2]\right]$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \underbrace{\left[-0.5N\log\,[2\pi] - 0.5N\log\sigma^2 - 0.5 \sum_{i=1}^{N} \frac{(x[i]-\mu)^2}{\sigma^2}\right]}_{L}$$

Maximization can be done in closed-form by taking derivative w.r.t. the variable and equate to zero:

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{N} \frac{(x[i]-\mu)}{\sigma^2} = \frac{\sum_{i=1}^{N} x[i]}{\sigma^2} - \frac{N\mu}{\sigma^2} = 0, \qquad \frac{\partial L}{\partial \sigma^2} = -\frac{N}{\sigma^2} + \sum_{i=1}^{N} \frac{(x[i]-\mu)^2}{\sigma^4} = 0$$

$$\boxed{\Rightarrow \quad \hat{\mu} = \frac{\sum_{i=1}^{N} x[i]}{N} = \bar{x}, \qquad \Rightarrow \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{N} (x[i]-\mu)^2}{N}}$$

Least Squares

Maximum likelihood for the normal distribution...

$$\hat{\mu} = \underset{\mu}{\operatorname{argmax}} \left[-0.5N\log[2\pi] - 0.5N\log\sigma^2 - 0.5 \sum_{i=1}^{N} \frac{(x[i]-\mu)^2}{\sigma^2} \right]$$

$$= \underset{\mu}{\operatorname{argmax}} \left[-\sum_{i=1}^{N} (x[i]-\mu)^2 \right]$$

$$= \underset{\mu}{\operatorname{argmin}} \left[\sum_{i=1}^{N} (x[i]-\mu)^2 \right]$$

...gives `least squares' fitting criterion.

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left[\prod_{i=1}^{N} p(x[i] \mid \theta)\, p(\theta) \right]$$

Likelihood Prior

Likelihood: univariate Normal distribution

$$p(x|\mu, \sigma^2) = \prod_{i=1}^{N} \mathrm{Norm}_{x[i]}[\mu, \sigma^2],$$

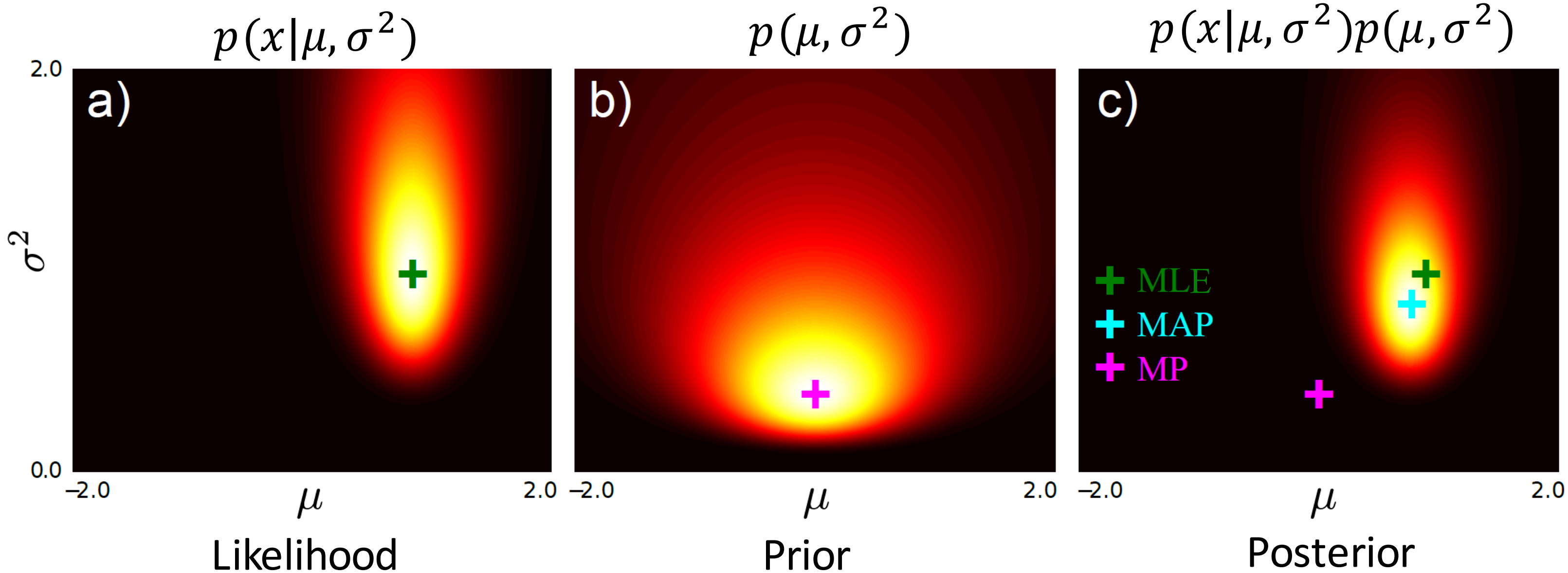
Prior: conjugate prior – normal inverse gamma distribution

$$p(\mu, \sigma^2) = \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$$

$$= \frac{\sqrt{\gamma}}{\sigma\sqrt{2\pi}} \frac{\beta^{\alpha}}{\Gamma[\alpha]} \left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left[-\frac{2\beta + \gamma(\delta - \mu)^2}{2\sigma^2}\right]$$

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)



$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\prod_{i=1}^{N} p(x[i]|\mu,\sigma^2)\, p(\mu,\sigma^2) \right]$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\prod_{i=1}^{N} \mathrm{Norm}_{x[i]}[\mu,\sigma^2]\, \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta] \right]$$

Image Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\prod_{i=1}^{N} p(x[i] \mid \mu, \sigma^2)\, p(\mu, \sigma^2)\right]$$

$$= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\prod_{i=1}^{N} \mathrm{Norm}_{x[i]}[\mu, \sigma^2]\, \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]\right]$$

Maximize the logarithm:

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\sum_{i=1}^{N} \log\left[\mathrm{Norm}_{x[i]}[\mu, \sigma^2]\right] + \log\left[\mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]\right]\right]$$

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)

$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu, \sigma^2}{\operatorname{argmax}} \left[\underbrace{\sum_{i=1}^{N} \log \left[\operatorname{Norm}_{x[i]}[\mu, \sigma^2] \right] + \log \left[\operatorname{NormInvGam}_{\mu, \sigma^2}[\alpha, \beta, \gamma, \delta] \right]}_{L} \right]$$

Taking derivatives and setting to zero:

$$\frac{\partial L}{\partial \mu} = 0, \qquad \frac{\partial L}{\partial \sigma^2} = 0$$

We get:

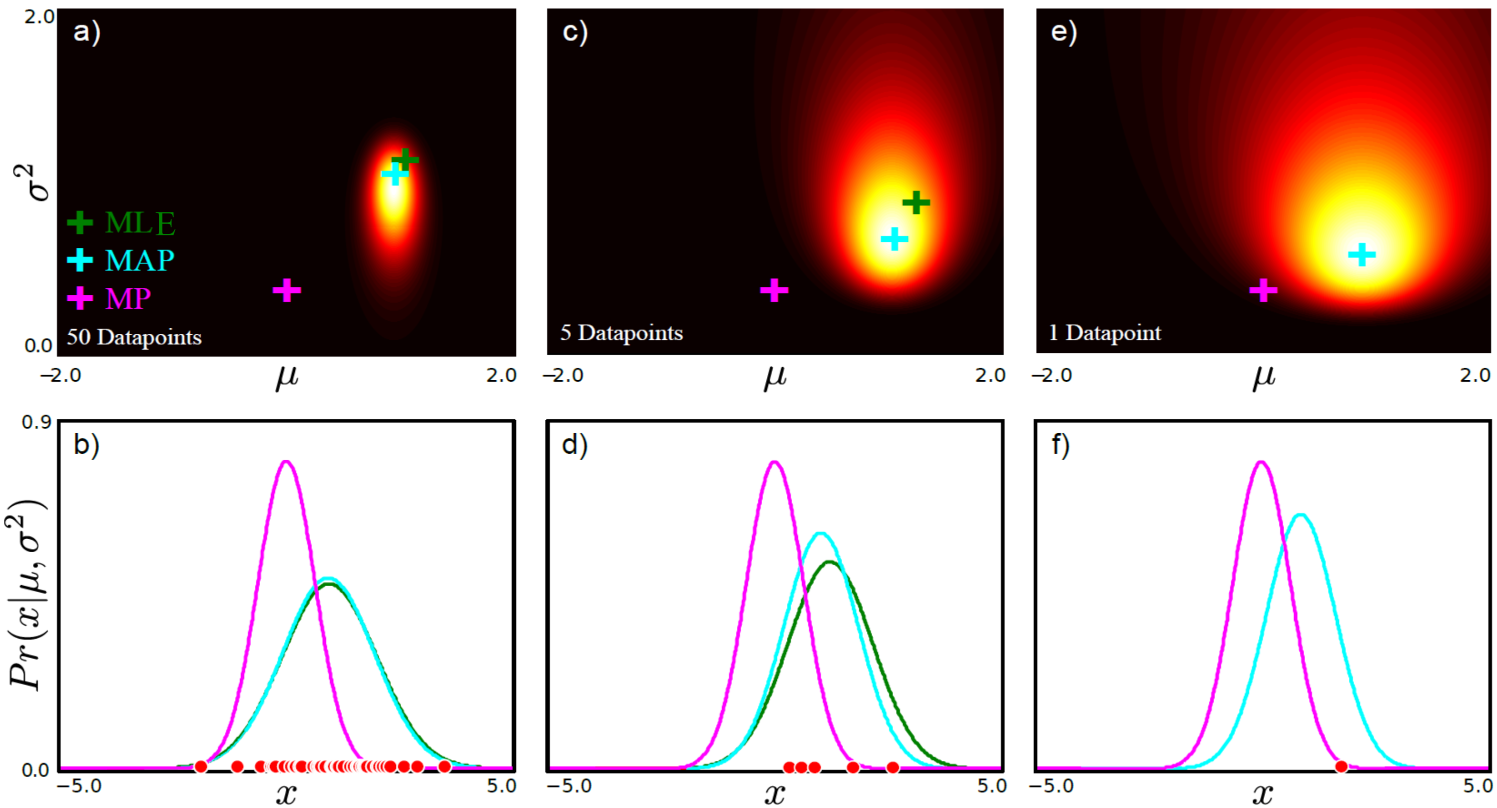
$$\hat{\mu} = \frac{\sum_i x[i] + \gamma\delta}{N + \gamma}, \qquad \hat{\sigma}^2 = \frac{\sum_i (x[i] - \mu)^2 + 2\beta + \gamma(\delta - \mu)^2}{N + 3 + 2\alpha}$$

$$= \frac{N\bar{x} + \gamma\delta}{N + \gamma}$$

Example 1: Univariate Normal Distribution

Approach 2: Maximum a Posteriori (MAP)

More data points → MAP is closer to MLE
Fewer data points → MAP is closer to MP



50 data points | 5 data points | 1 data points

Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Compute the posterior distribution using Bayes' rule:

$$p(\theta|x) = \frac{\prod_{i=1}^{N} p(x[i] \mid \theta) p(\theta)}{p(x)}$$

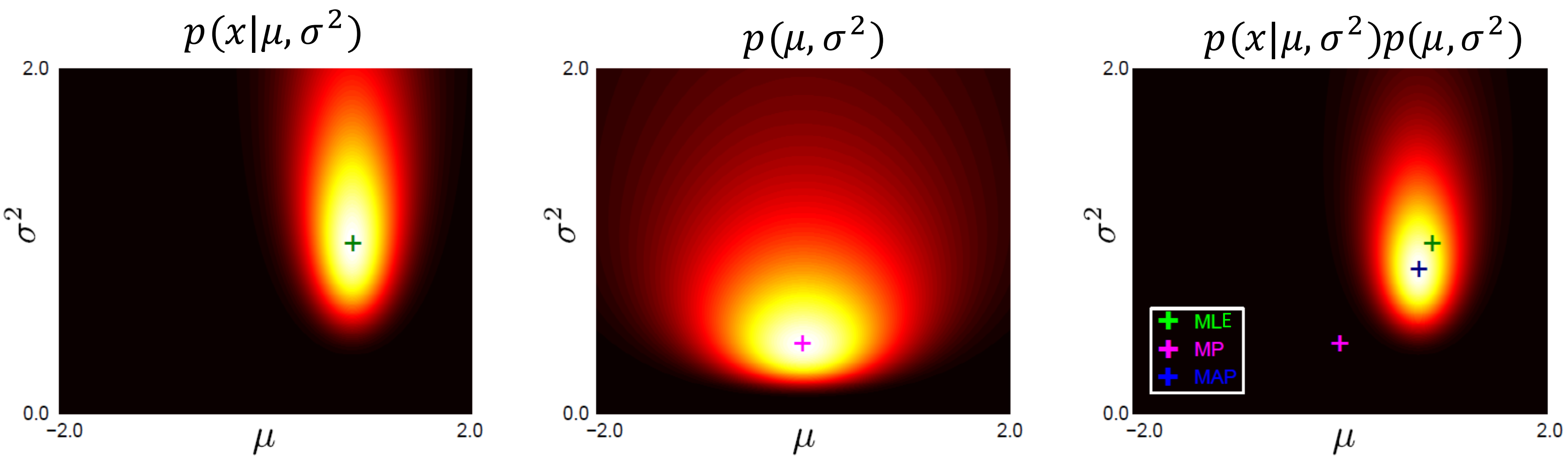


Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Compute the posterior distribution using Bayes' rule:

$$p(\theta|x) = \frac{\prod_{i=1}^{N} p(x[i] \mid \theta)p(\theta)}{p(x)} = \frac{\prod_{i=1}^{N} p(x[i] \mid \theta)p(\theta)}{\int \prod_{i=1}^{N} p(x[i] \mid \theta)p(\theta)\, d\theta}$$

where:

$$\prod_{i=1}^{N} p(x[i] \mid \theta)p(\theta) = \prod_{i=1}^{N} \mathrm{Norm}_{x[i]}[\mu, \sigma^2]\, \mathrm{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$$

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

$$\prod_{i=1}^{N} p(x[i]|\theta)p(\theta) = \prod_{i=1}^{N} \text{Norm}_{x[i]}[\mu, \sigma^2]\,\text{NormInvGam}_{\mu,\sigma^2}[\alpha, \beta, \gamma, \delta]$$

Rearranging:

$$\prod_{i=1}^{N} p(x[i]|\theta)p(\theta) = \underbrace{\kappa[\alpha, \beta, \gamma, \delta, x]}_{\text{Constant}}\text{NormInvGam}_{\mu,\sigma^2}\left[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}\right]$$

where

$$\tilde{\alpha} = \alpha + \frac{N}{2}, \qquad \tilde{\delta} = \frac{(\gamma\delta + \sum_i x[i])}{\gamma + N},$$

$$\tilde{\gamma} = \gamma + N, \qquad \tilde{\beta} = \frac{\sum_i x[i]^2}{2} + \beta + \frac{\gamma\delta^2}{2} - \frac{(\gamma\delta + \sum_i x[i])^2}{2(\gamma + N)}.$$

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Compute the posterior distribution using Bayes' rule:

$$p(\theta|x) = \frac{\prod_{i=1}^{N} p(x[i]|\theta)p(\theta)}{p(x)} = \frac{\prod_{i=1}^{N} p(x[i]|\theta)p(\theta)}{\int \prod_{i=1}^{N} p(x[i]|\theta)p(\theta)\,d\theta}$$

$$p(\theta|x) = \frac{\cancel{\kappa[\alpha,\beta,\gamma,\delta,x]}\mathrm{NormInvGam}_{\mu,\sigma^2}\left[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}\right]}{\cancel{\kappa[\alpha,\beta,\gamma,\delta,x]}\underbrace{\int\int \mathrm{NormInvGam}_{\mu,\sigma^2}\left[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}\right]d\mu d\sigma^2}_{=1}}$$

$$\boxed{p(\theta|x) = \mathrm{NormInvGam}_{\mu,\sigma^2}\left[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}\right]}$$

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Predictive density

Take weighted sum of predictions from different parameter values:

$$p(x^*|x) = \int\int p(x^*|\mu,\sigma^2)p(\mu,\sigma^2|x)d\mu d\sigma^2$$

Posterior: $p(\mu,\sigma^2|x)$

Samples from posterior

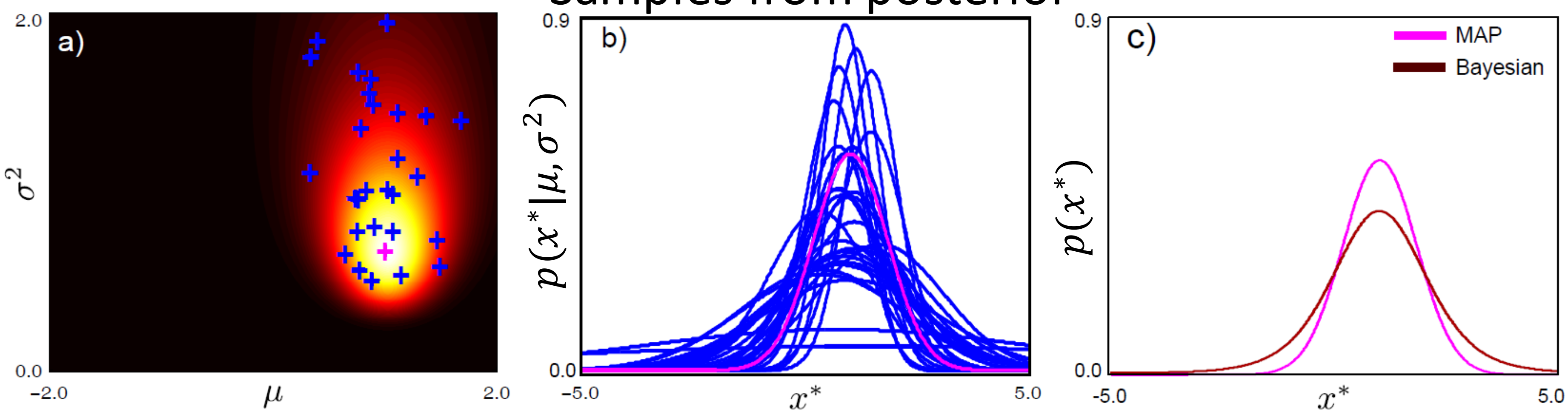


Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Predictive density

Take weighted sum of predictions from different parameter values:

$$p(x^*|x) = \int\int p(x^*|\mu,\sigma^2)p(\mu,\sigma^2|x)d\mu d\sigma^2$$

$$= \int\int \text{Norm}_{x^*}[\mu,\sigma^2]\text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}]d\mu d\sigma^2$$

$$= \kappa[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta},x^*]\underbrace{\int\int \text{NormInvGam}_{\mu,\sigma^2}[\breve{\alpha},\breve{\beta},\breve{\gamma},\breve{\delta}]d\mu d\sigma^2}_{=1}$$

$$= \kappa[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta},x^*]$$

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

Predictive density

Take weighted sum of predictions from different parameter values:

$$p(x^*|x) = \kappa[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}, x^*] = \frac{1}{\sqrt{2\pi}} \frac{\tilde{\beta}^{\tilde{\alpha}} \sqrt{\tilde{\gamma}}}{\breve{\beta}^{\breve{\alpha}} \sqrt{\breve{\gamma}}} \frac{\Gamma[\breve{\alpha}]}{\Gamma[\tilde{\alpha}]}$$

where

$$\breve{\alpha} = \tilde{\alpha} + 1/2, \qquad \breve{\gamma} = \tilde{\gamma} + 1$$

$$\breve{\beta} = \frac{x^{*2}}{2} + \tilde{\beta} + \frac{\tilde{\gamma}\tilde{\delta}^2}{2} - \frac{(\tilde{\gamma}\tilde{\delta} + x^*)^2}{2(\tilde{\gamma} + 1)}.$$

Example 1: Univariate Normal Distribution

Approach 3: Bayesian

As the training data decreases, the Bayesian prediction becomes less certain but the MAP prediction is erroneously overconfident.

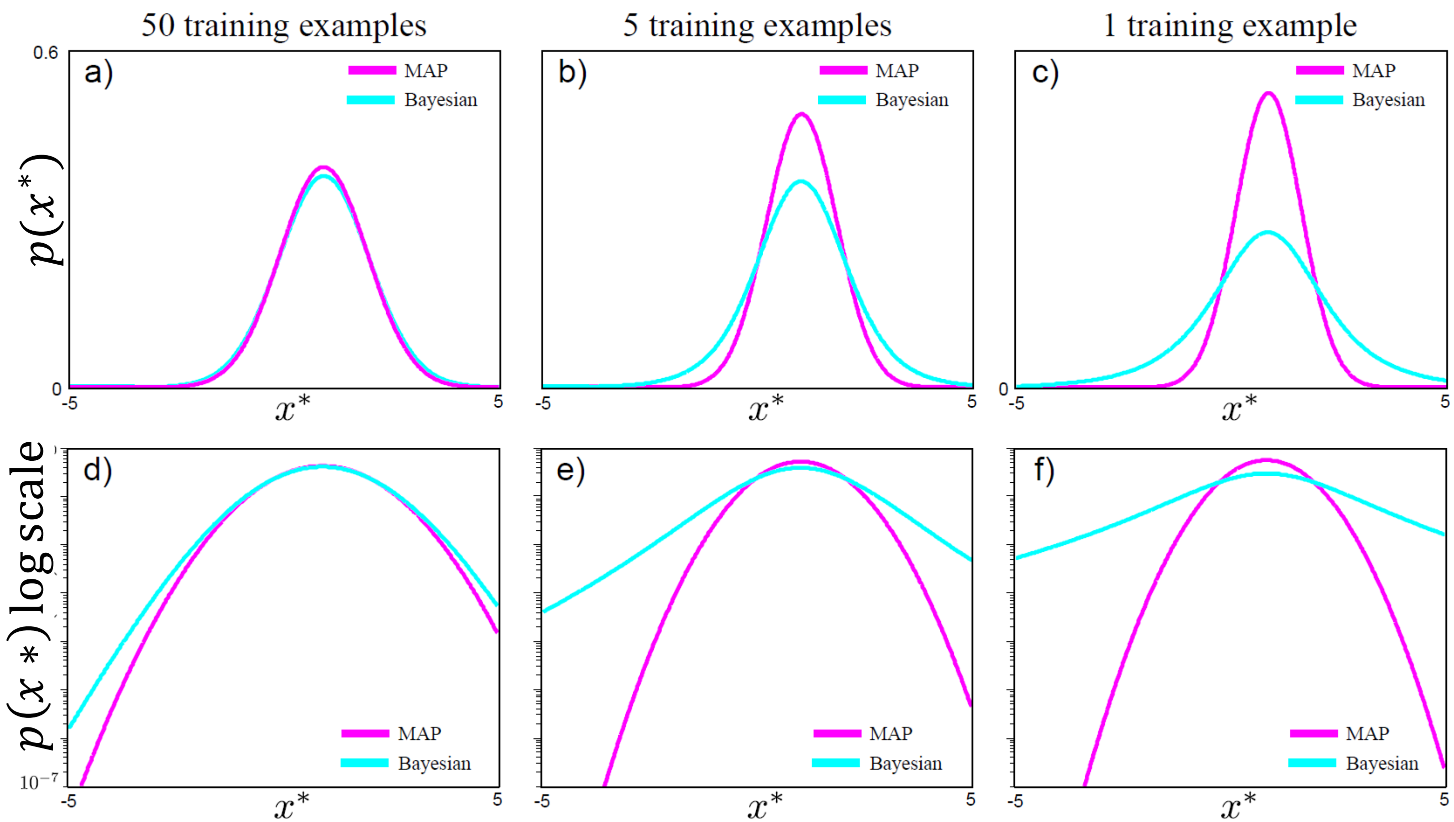
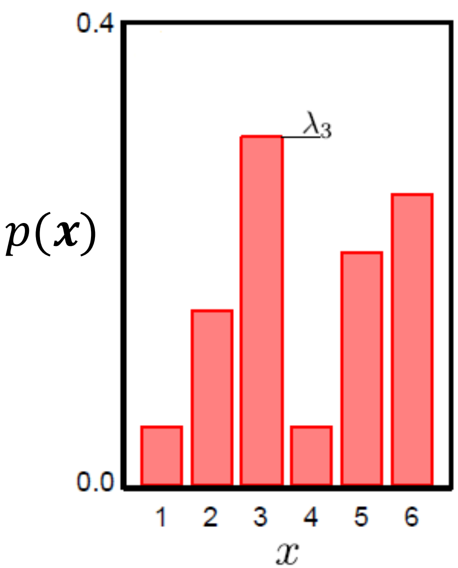


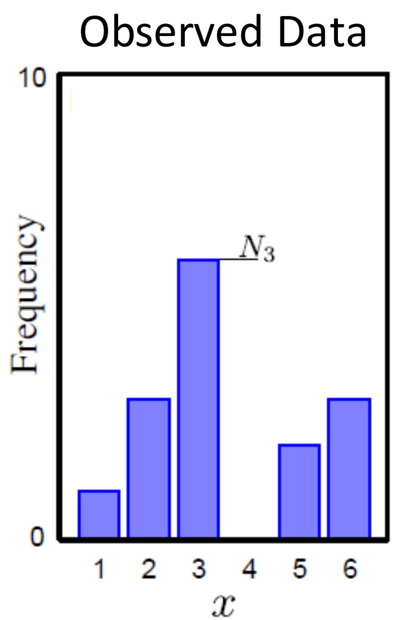
Image Source: “Computer Vision: Models, Learning, and Inference”, Simon Prince

Example 2: Categorical Distribution

Problem:

Fit a categorical distribution model (**6 categories**) to a set of data $\boldsymbol{X} : \{\boldsymbol{x}[1], \dots, \boldsymbol{x}[N]\}$, where $\boldsymbol{x}[i] = \mathbf{e}_k$ is a vector with all zero elements expect *k*th e.g. $[0,0,0,1,0,0]$.





- Throwing a 6-faced die *N* times.
- $\boldsymbol{X} : \{\boldsymbol{x}[1], \dots, \boldsymbol{x}[N]\}$ is obtained from all the outcomes.
- $\boldsymbol{x}[i] = [0,0,0,1,0,0]$ when the outcome is 4.

Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince

Example 2: Categorical Distribution

Problem:

Fit a categorical distribution model (**6 categories**) to a set of data $\boldsymbol{X} : \{\boldsymbol{x}[1], \dots, \boldsymbol{x}[N]\}$, where $\boldsymbol{x}[i] = \mathbf{e}_k$ is a vector with all zero elements expect k^{th} e.g. $[0,0,0,1,0,0]$.

Recall that the categorical distribution is given by:

k^{th} element of $\mathbf{e}_k$

$$p(\boldsymbol{X} = \mathbf{e}_k) = \text{Cat}_{\boldsymbol{x}}[\lambda] = \prod_{k=1}^{K} \lambda_k^{x_k} = \lambda_k$$

Our goal is to find the K=6 unknown parameters $\lambda_k \in [0,1]$, where $\sum_k \lambda_k = 1$.

Example 2: Categorical Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}[p(\boldsymbol{x}|\theta)]$$

$$= \underset{\theta}{\operatorname{argmax}}\left[\prod_{i=1}^{N} p(\boldsymbol{x}[i]|\theta)\right]$$ (Naïve Bayes)

Likelihood given by pdf

$$p(\boldsymbol{x}|\lambda) = \mathrm{Cat}_{\boldsymbol{x}}[\lambda] = \prod_{k=1}^{K} \lambda_k^{x_k} = \lambda_k$$

Example 2: Categorical Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

$$\hat{\lambda}_{1\ldots6} = \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\boldsymbol{x}[i] \mid \lambda_{1\ldots6}), \qquad s.t. \quad \sum_k \lambda_k = 1$$

$$= \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \prod_{i=1}^{N} \mathrm{Cat}_{\boldsymbol{x}[i]}[\lambda_{1\ldots6}], \qquad s.t. \quad \sum_k \lambda_k = 1$$

k^{th} element of $\boldsymbol{x}_i$, $x_{ik} \in \{0,1\}$

$$= \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \prod_{i=1}^{N} \prod_{k=1}^{6} \lambda_k^{x_{ik}}, \qquad s.t. \quad \sum_k \lambda_k = 1$$

N_k = # times we observed bin k

$$= \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \prod_{k=1}^{6} \lambda_k^{N_k}, \qquad s.t. \quad \sum_k \lambda_k = 1$$

Example 2: Categorical Distribution

Approach 1: Maximum Likelihood Estimation (MLE)

Applying log probability and Lagrange multiplier ν on the constraint, we get the auxiliary function :

$$\mathcal{L} = \sum_{k=1}^{6} N_k \log[\lambda_k] + \upsilon \left(\sum_{k=1}^{6} \lambda_k - 1 \right)$$

Take derivative of $\mathcal{L}$ w.r.t λ_k and υ, set to zero and solve for λ_k:

$$\boxed{\hat{\lambda}_k = \frac{N_k}{\sum_{m=1}^{6} N_m}}$$

Normalized counts of # times we observed bin *k*

Example 2: Categorical Distribution

Approach 2: Maximum a Posteriori (MAP)

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left[\prod_{i=1}^{N} \underbrace{p(\boldsymbol{x}[i]|\theta)}_{\text{Likelihood}} \, \underbrace{p(\theta)}_{\text{Prior}} \right]$$

Likelihood: categorical distribution

$$p(\boldsymbol{x}|\lambda) = \prod_{i=1}^{N} \mathrm{Cat}_{\boldsymbol{x}[i]}[\lambda_{1\dots K}] = \prod_{i=1}^{N} \prod_{k=1}^{K} \lambda_k^{x_{ik}} = \prod_{k=1}^{K} \lambda_k^{N_k}$$

Prior: conjugate prior – Dirichlet distribution

$$p(\lambda_1, \dots, \lambda_K) = \mathrm{Dir}_{\lambda_{1\dots K}}[\alpha_1, \dots \alpha_K]$$

$$= \frac{\Gamma[\sum_{k=1}^{K} \alpha_k]}{\prod_{k=1}^{K} \Gamma[\alpha_k]} \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1}, \qquad \text{s.t. } \lambda_k \in [0,1], \ \sum_k \lambda_k = 1$$

Example 2: Categorical Distribution

Approach 2: Maximum a Posteriori (MAP)

$$\hat{\lambda}_{1\ldots6} = \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\boldsymbol{x}[i]|\lambda_{1\ldots6})p(\lambda_{1\ldots6}), \qquad s.t. \quad \sum_k \lambda_k = 1$$

$$= \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \prod_{i=1}^{N} \mathrm{Cat}_{\boldsymbol{x}[i]}[\lambda_{1\ldots6}]\mathrm{Dir}_{\lambda_{1\ldots6}}[\alpha_1, \ldots \alpha_6], \qquad s.t. \quad \sum_k \lambda_k = 1$$

Independent of $\lambda \Rightarrow$ can be ignored

$$= \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \cancel{\frac{\Gamma[\sum_{k=1}^{6}\alpha_k]}{\prod_{k=1}^{6}\Gamma[\alpha_k]}} \prod_{k=1}^{6} \lambda_k^{N_k} \prod_{k=1}^{6} \lambda_k^{\alpha_k - 1}, \qquad s.t. \quad \sum_k \lambda_k = 1$$

$$= \underset{\lambda_{1\ldots6}}{\operatorname{argmax}} \prod_{k=1}^{6} \lambda_k^{N_k + \alpha_k - 1}, \qquad s.t. \quad \sum_k \lambda_k = 1$$

Example 2: Categorical Distribution

Approach 2: Maximum a Posteriori (MAP)

Applying log probability and Lagrange multiplier ν on the constraint, we get the auxiliary function:

$$\mathcal{L} = \sum_{k=1}^{6} (N_k + \alpha_k - 1) \log \lambda_k + \upsilon \left(\sum_{k=1}^{6} \lambda_k - 1 \right)$$

Take derivative of $\mathcal{L}$ w.r.t λ_k and υ, set to zero and solve for λ_k:

$$\boxed{\hat{\lambda}_k = \frac{N_k + \alpha_k - 1}{\sum_{m=1}^{6} (N_m + \alpha_m - 1)}}$$

Same result as MLE with a uniform prior $\alpha_{1\ldots k} = 1$

$$\hat{\lambda}_k = \frac{N_k}{\sum_{m=1}^{6} N_m}$$

Example 2: Categorical Distribution

Approach 3: Bayesian

Compute the posterior distribution using Bayes' rule:

$$p(\theta|\boldsymbol{x}) = \frac{\prod_{i=1}^{N} p(\boldsymbol{x}[i]|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{\prod_{i=1}^{N} p(\boldsymbol{x}[i]|\theta)p(\theta)}{\int \prod_{i=1}^{N} p(\boldsymbol{x}[i]|\theta)p(\theta)\, d\theta}$$

where:

$$\prod_{i=1}^{N} p(\boldsymbol{x}[i]|\theta)p(\theta) = \prod_{i=1}^{N} \mathrm{Cat}_{\boldsymbol{x}[i]}[\lambda_{1\ldots6}]\,\mathrm{Dir}_{\lambda_{1\ldots6}}[\alpha_1, \ldots \alpha_6]$$

Example 2: Categorical Distribution

Approach 3: Bayesian

$$\prod_{i=1}^{N} p(\boldsymbol{x}[i]|\theta)p(\theta) = \prod_{i=1}^{N} \text{Cat}_{\boldsymbol{x}[i]}[\lambda_{1\ldots6}]\,\text{Dir}_{\lambda_{1\ldots6}}[\alpha_1, \ldots \alpha_6]$$

Rearranging:

$$\prod_{i=1}^{N} p(\boldsymbol{x}[i]|\theta)p(\theta) = \underbrace{\kappa[\alpha_{1..6}, \boldsymbol{x}[1], \ldots, \boldsymbol{x}[N]]}_{\text{Constant}}\text{Dir}_{\lambda_{1\ldots6}}[\tilde{\alpha}_1, \ldots \tilde{\alpha}_6]$$

where

$$\tilde{\alpha}_k = \alpha_k + N_k,$$

Example 2: Categorical Distribution

Approach 3: Bayesian

Compute the posterior distribution using Bayes' rule:

$$p(\theta|\boldsymbol{x}) = \frac{\prod_{i=1}^{N} p(\boldsymbol{x}[i]|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{\prod_{i=1}^{N} p(\boldsymbol{x}[i]|\theta)p(\theta)}{\int \prod_{i=1}^{N} p(\boldsymbol{x}[i]|\theta)p(\theta)\, d\theta}$$

$$p(\theta|\boldsymbol{x}) = \frac{\cancel{\kappa[\alpha_{1\ldots6}, \boldsymbol{x}[1], \ldots, \boldsymbol{x}[N]]}\,\mathrm{Dir}_{\lambda_{1\ldots6}}[\tilde{\alpha}_1, \ldots \tilde{\alpha}_6]}{\cancel{\kappa[\alpha_{1\ldots6}, \boldsymbol{x}[1], \ldots, \boldsymbol{x}[N]]}\,\underbrace{\int \mathrm{Dir}_{\lambda_{1\ldots6}}[\tilde{\alpha}_1, \ldots \tilde{\alpha}_6]\, d\lambda_{1\ldots6}}_{=1}}$$

$$\boxed{p(\theta|\boldsymbol{x}) = \mathrm{Dir}_{\lambda_{1\ldots6}}[\tilde{\alpha}_1, \ldots \tilde{\alpha}_6]}$$

Example 2: Categorical Distribution

Approach 3: Bayesian

Predictive density

Take weighted sum of predictions from different parameter values:

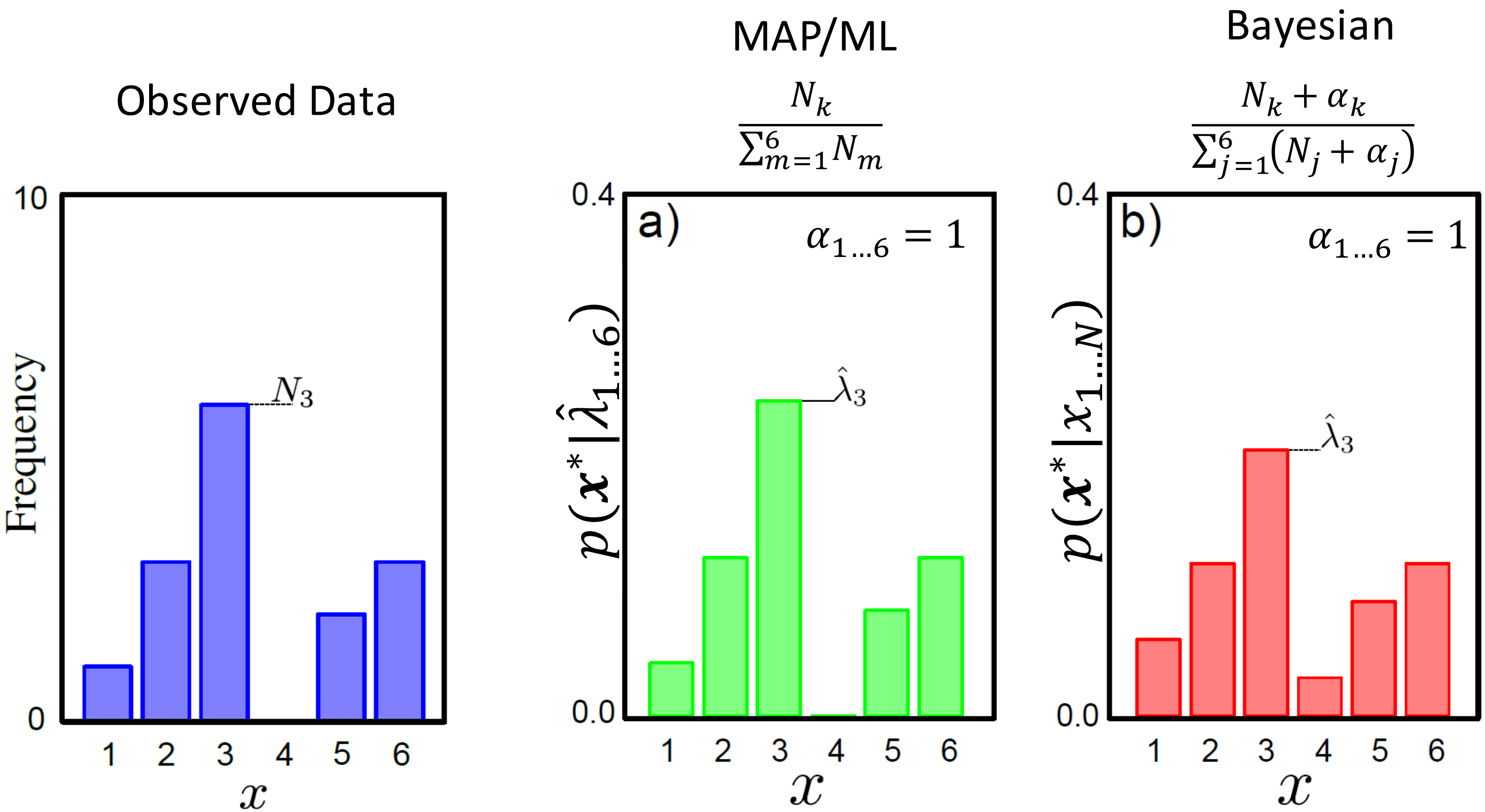
$$p(\boldsymbol{x}^* = \mathbf{e}_{\boldsymbol{k}}|\boldsymbol{x}) = \int p(\boldsymbol{x}^*|\lambda_{1\dots6})p(\lambda_{1\dots6}|\boldsymbol{x})d\lambda_{1\dots6}$$

$$= \int \mathrm{Cat}_{\boldsymbol{x}^*}[\lambda_{1\dots6}]\mathrm{Dir}_{\lambda_{1\dots6}}[\tilde{\alpha}_1, \dots \tilde{\alpha}_6]d\lambda_{1\dots6}$$

$$= \kappa[\tilde{\alpha}_{1..6}, \boldsymbol{x}^*]\underbrace{\int \mathrm{Dir}_{\lambda_{1\dots6}}[\breve{\alpha}_1, \dots \breve{\alpha}_6]d\lambda_{1\dots6}}_{=1}$$

$$= \kappa[\tilde{\alpha}_{1\dots6}, \boldsymbol{x}^*] = \frac{N_k + \alpha_k}{\sum_{j=1}^{6}(N_j + \alpha_j)}$$

Example 2: Categorical Distribution



The Bayesian approach predicts a more moderate distribution and allots some probability to the case $x = 4$ despite having seen no training examples in this category.

Image Source: "Computer Vision: Models, Learning, and Inference", Simon Prince