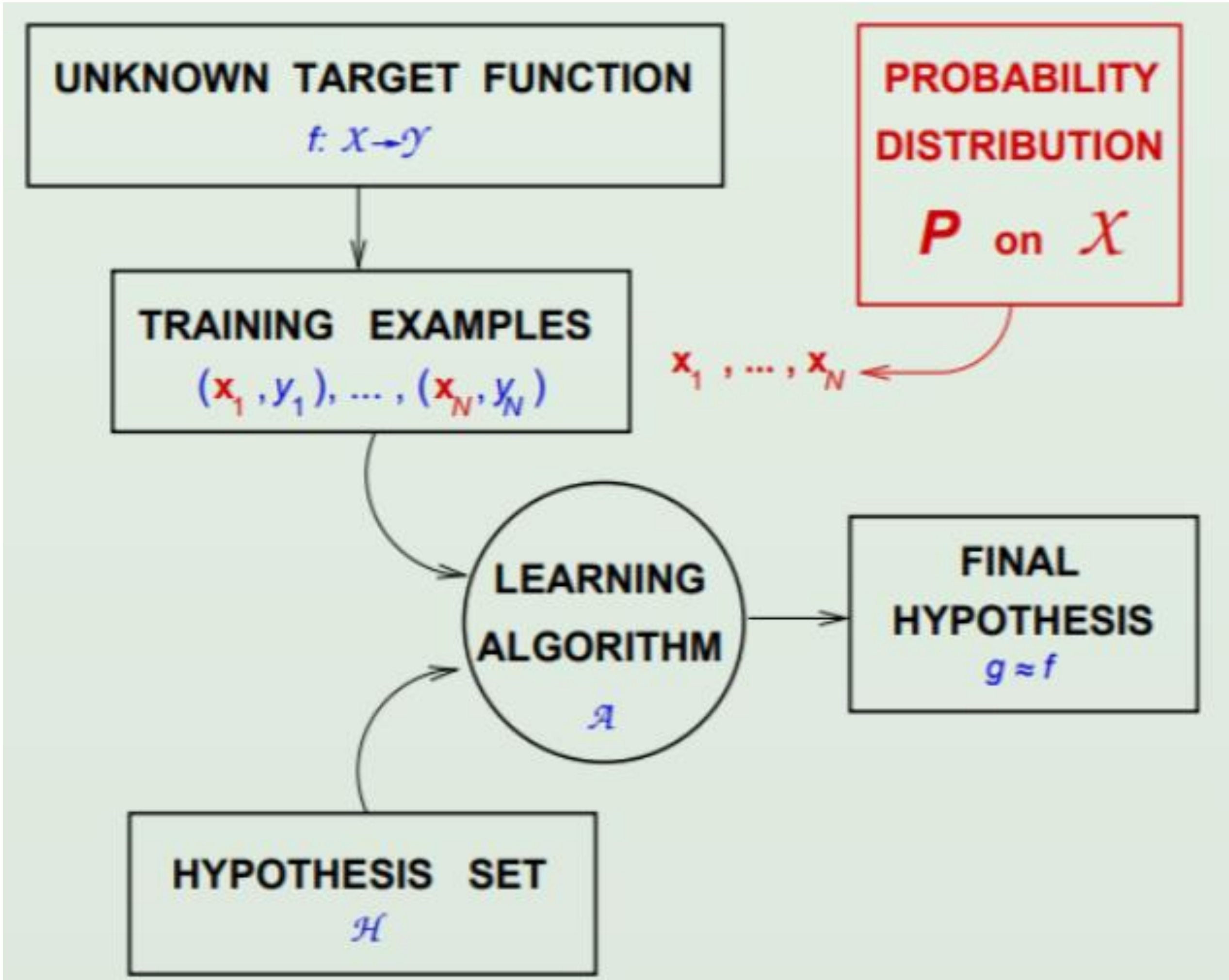
Fitting probability models


UNKNOWN TARGET FUNCTION
$f: \mathcal{X} \to \mathcal{Y}$
PROBABILITY DISTRIBUTION
P on $\mathcal{X}$
TRAINING EXAMPLES
$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
$\mathbf{x}_1, \ldots, \mathbf{x}_N$
LEARNING ALGORITHM
$\mathcal{A}$
FINAL HYPOTHESIS
$g \approx f$
HYPOTHESIS SET
$\mathcal{H}$

Probabilities model

Our assumption on how the data were generated?

Modeling-Learning-Inference

statistical model

Why probabilities modeling?

Inferences from data are intrinsically uncertain.

Probability theory: Model uncertainty instead of ignoring it

Inferences or prediction can be done by using probabilities

Uncertainty

We can't perfectly predict the exact output given the input, due to

- lack of knowledge of the input-output mapping (**model uncertainty**)
- and/or intrinsic (irreducible) stochasticity in the mapping (**data uncertainty**).

We capture our uncertainty using **conditional probability distribution**:

$$p(y = c|\boldsymbol{x}; \boldsymbol{\theta}) = f_c(\boldsymbol{x}; \boldsymbol{\theta})$$

where $f : X \to [0, 1]^C$ maps inputs to a probability distribution over the C possible output labels.

Basics of Probability Theory

- Space sample S
- Event E
- Space W of events
- Random variable
- Probability
- Joint probability
- Conditional probability

Bayes Rule

$$P(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{D})}$$

- P($\boldsymbol{\theta}$): *prior probability* (xác suất tiên nghiệm) of the variable $\boldsymbol{\theta}$.
  - Our uncertainty about $\boldsymbol{\theta}$ before observing data.
- P(**D**): prior probability that we can observe data **D**.
- P(**D** | $\boldsymbol{\theta}$): probability (*likelihood*) that we can observe data **D** provided that $\boldsymbol{\theta}$ is known.
- P($\boldsymbol{\theta}$ | **D**): *posterior probability* (xác suất hậu nghiệm) of $\boldsymbol{\theta}$ if we already have observed data **D**.

- Fitting probability distributions
  - Maximum Likelihood Estimation (ML Estimation or MLE)
  - Maximum A Posteriori Estimation (MAP Estimation or MAP)
  - Bayesian approach

$$P(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{D})}$$

Maximum Likelihood Estimation

As the name suggests we find the parameters under which the data $\mathbf{x}_{1\ldots I}$ is most likely.

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[Pr(\mathbf{x}_{1\ldots I}|\boldsymbol{\theta})\right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})\right]\end{aligned}$$

We have assumed that data was **independent** (hence product)

$$p(x,y) = p(x)p(y).$$

$$p(x,y|z) = p(x|z)p(y|z))$$

Maximum Likelihood Estimation

As the name suggests we find the parameters under which the data $\mathbf{x}_{1\ldots I}$ is most likely.

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[Pr(\mathbf{x}_{1\ldots I}|\boldsymbol{\theta})\right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})\right]\end{aligned}$$

We have assumed that data was **independent** (hence product)

Predictive Density:

Evaluate new data point $\mathbf{x}^*$ under probability distribution $Pr(\mathbf{x}^*|\hat{\boldsymbol{\theta}})$ with best parameters

Maximum a posteriori (MAP)

Fitting

As the name suggests we find the parameters which maximize the posterior probability $Pr(\boldsymbol{\theta}|\mathbf{x}_{1...I})$.

$$
\begin{aligned}
\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[Pr(\boldsymbol{\theta}|\mathbf{x}_{1...I})\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\frac{Pr(\mathbf{x}_{1...I}|\boldsymbol{\theta})Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_{1...I})}\right] \\
&= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\frac{\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_{1...I})}\right]
\end{aligned}
$$

Again we have assumed that data was independent

Maximum a posteriori (MAP)

Fitting

As the name suggests we find the parameters which maximize the posterior probability $Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})$.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\frac{\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta}) Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_{1\ldots I})} \right]$$

Since the denominator doesn't depend on the parameters we can instead maximize

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta}) Pr(\boldsymbol{\theta}) \right]$$

Maximum a posteriori

Predictive Density:

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

Bayesian Approach

Fitting

Compute the posterior distribution over possible parameter values using Bayes' rule:

$$Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I}) = \frac{\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_{1\ldots I})}$$

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

- Each possible parameter value makes a prediction
- Some parameters more probable than others

$$Pr(\mathbf{x}^*|\mathbf{x}_{1\ldots I}) = \int Pr(\mathbf{x}^*|\boldsymbol{\theta})Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})\, d\boldsymbol{\theta}$$

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 methods

Maximum likelihood:

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

Maximum a posteriori:

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

Bayesian:

Calculate weighted sum of predictions from all possible value of parameters

$$Pr(\mathbf{x}^*|\mathbf{x}_{1\ldots I}) = \int Pr(\mathbf{x}^*|\boldsymbol{\theta})Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})\, d\boldsymbol{\theta}$$

Predictive densities for 3 methods

How to rationalize different forms?

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

$$
\begin{aligned}
Pr(\mathbf{x}^*|\mathbf{x}_{1\ldots I}) &= \int Pr(\mathbf{x}^*|\boldsymbol{\theta})\delta[\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}]\, d\boldsymbol{\theta} \\
&= Pr(\mathbf{x}^*|\hat{\boldsymbol{\theta}}),
\end{aligned}
$$

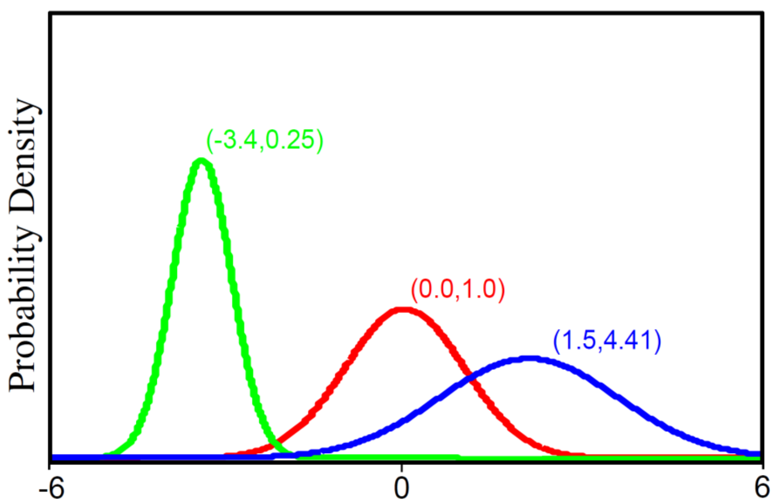
Example - Normal distribution

Univariate Normal Distribution

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

For short we write:

$$Pr(x) = \text{Norm}_x[\mu, \sigma^2]$$



Univariate normal distribution describes single continuous variable.

Takes 2 parameters μ and σ^2>0

Normal Inverse Gamma Distribution

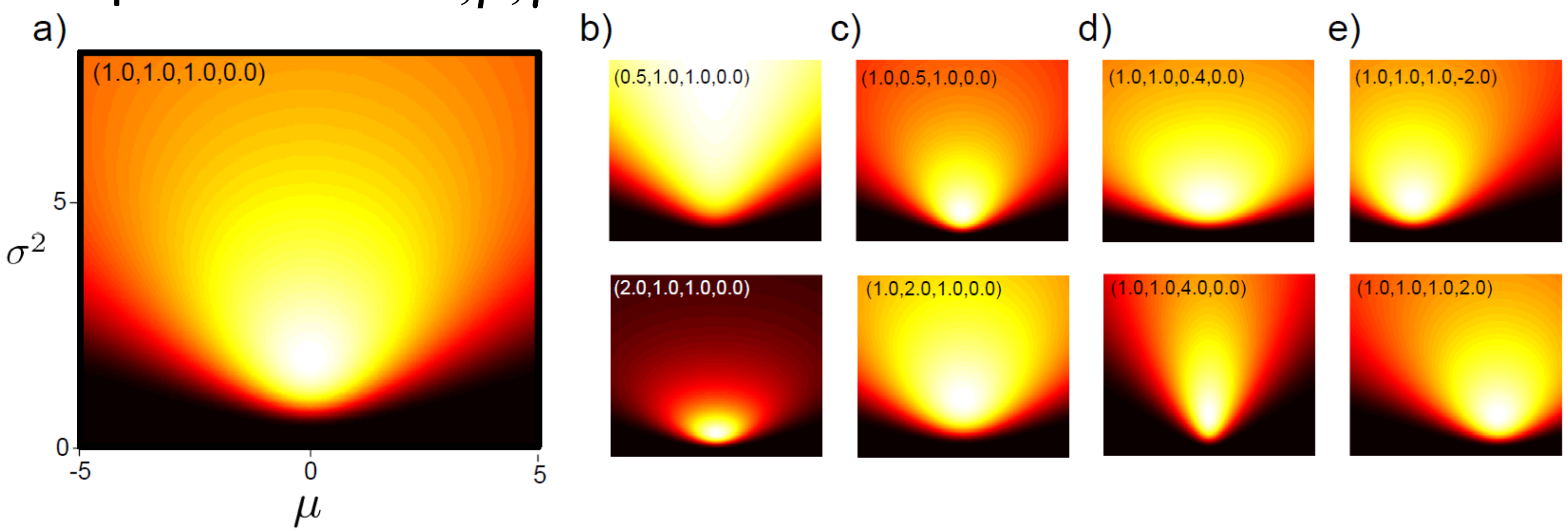
Defined on 2 variables μ and σ^2>0

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

or for short

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

Four parameters $\alpha,\beta,\gamma > 0$ and δ.



Fitting a normal distribution: MLE

As the name suggests we find the parameters under which the data $\mathbf{x}_{1\ldots I}$ is most likely.

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[Pr(\mathbf{x}_{1\ldots I}|\boldsymbol{\theta})\right] \\ &= \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left[\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})\right]\end{aligned}$$

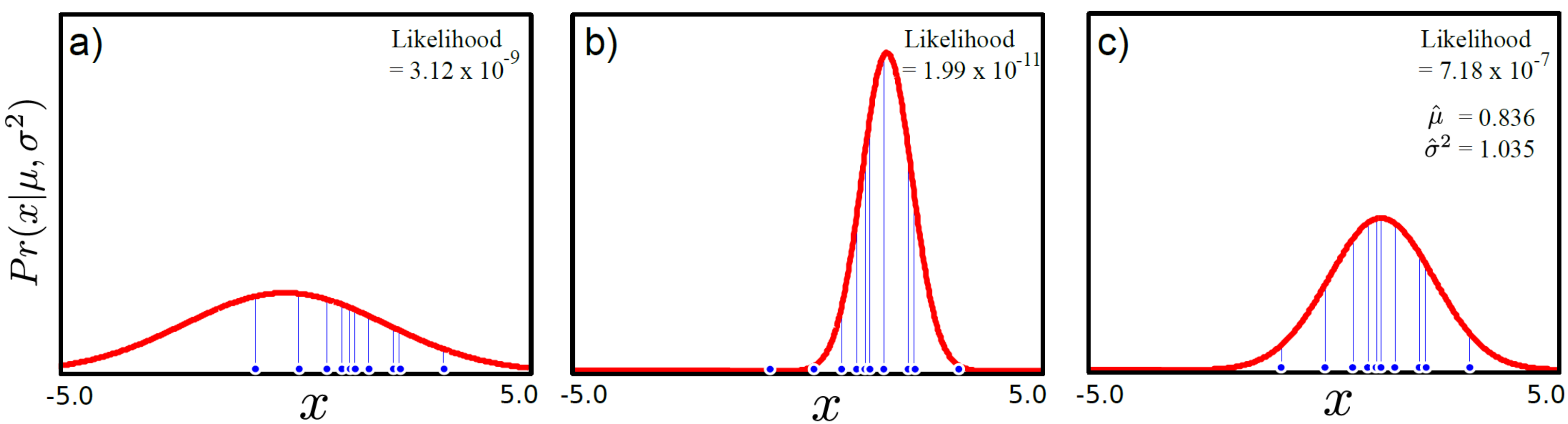
Likelihood given by pdf

$$Pr(x|\mu,\sigma^2) = \text{Norm}_x[\mu,\sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-0.5\frac{(x-\mu)^2}{\sigma^2}\right]$$

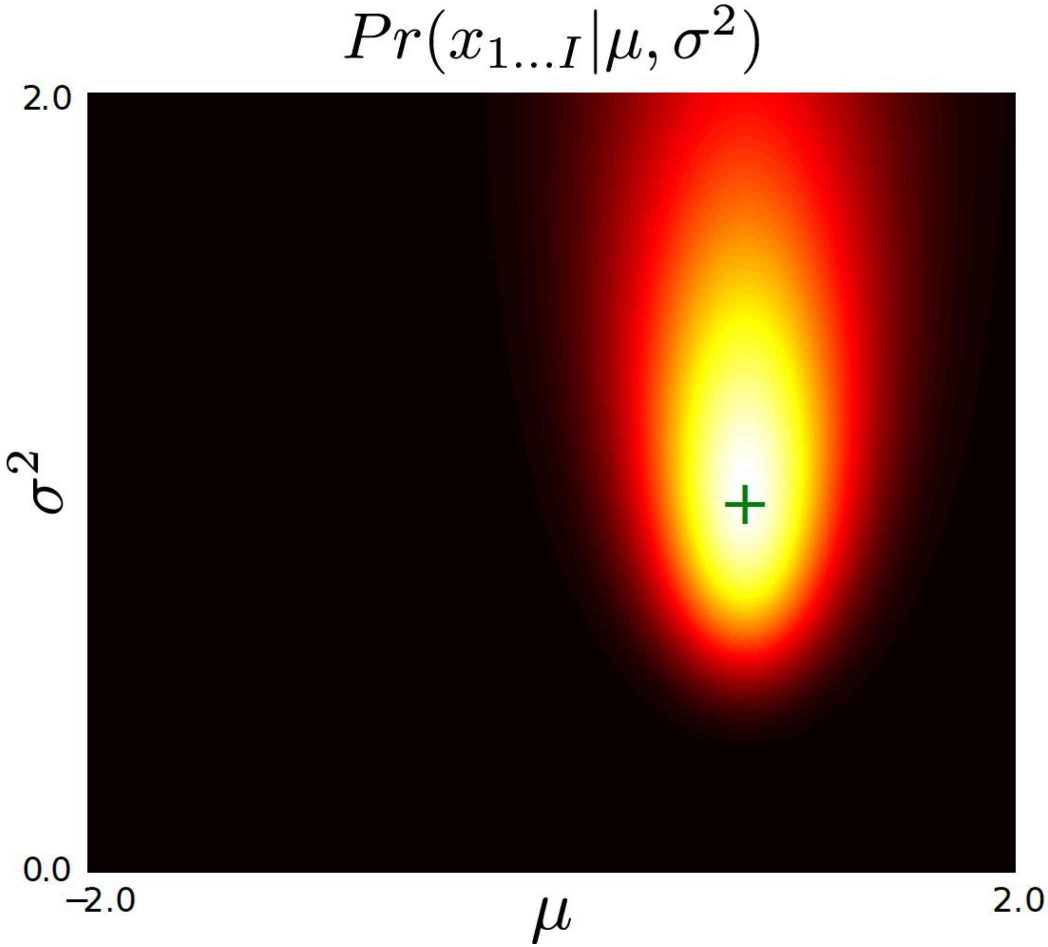
Fitting a normal distribution: MLE

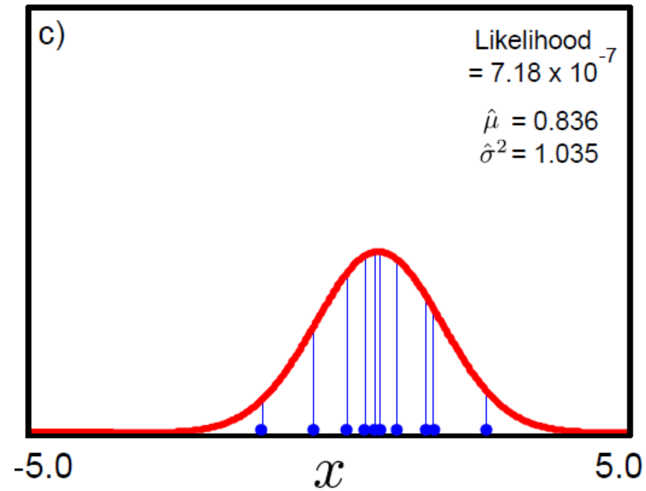
$$
\begin{aligned}
Pr(x_{1\ldots I}|\mu,\sigma^2) &= \prod_{i=1}^{I} Pr(x_i|\mu,\sigma^2) \\
&= \prod_{i=1}^{I} \text{Norm}_{x_i}[\mu,\sigma^2] \\
&= \frac{1}{(2\pi\sigma^2)^{I/2}} \exp\left[-0.5\sum_{i=1}^{I}\frac{(x_i-\mu)^2}{\sigma^2}\right]
\end{aligned}
$$

c) Better explains the data



Fitting a normal distribution: MLE





Plotted surface of likelihoods as a function of possible parameter values

ML Solution is at peak

Fitting normal distribution: MLE

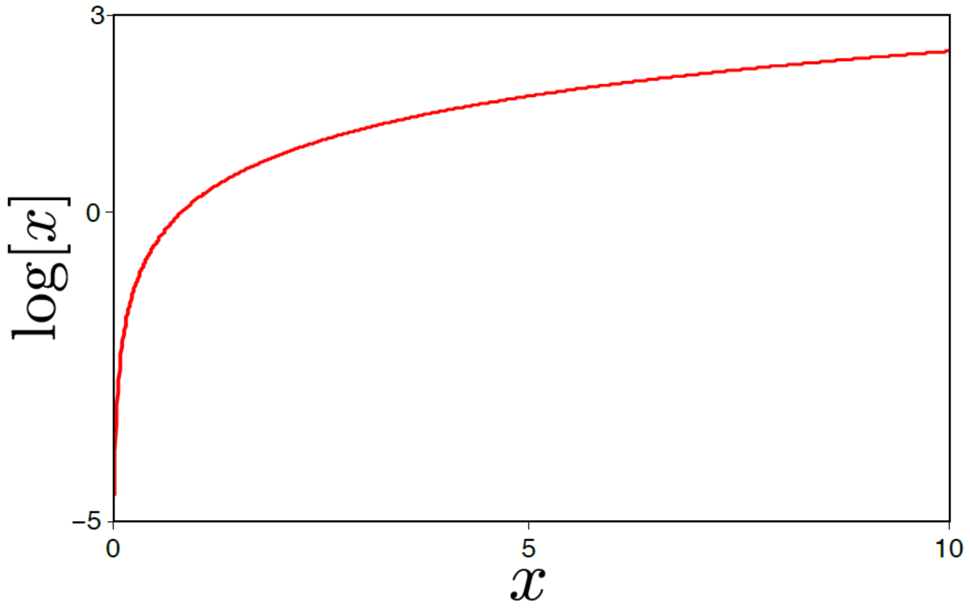
Algebraically:
$$\hat{\mu}, \hat{\sigma}^2 = \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[Pr(x_{1\ldots I}|\mu, \sigma^2) \right]$$

where:
$$Pr(x|\mu,\sigma^2) = \text{Norm}_x[\mu, \sigma^2]$$

or alternatively, we can maximize the logarithm

$$\begin{aligned}
\hat{\mu}, \hat{\sigma}^2 &= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \log\left[\text{Norm}_{x_i}[\mu, \sigma^2]\right] \right] \\
&= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[-0.5I \log[2\pi] - 0.5I \log \sigma^2 - 0.5 \sum_{i=1}^{I} \frac{(x_i - \mu)^2}{\sigma^2} \right]
\end{aligned}$$

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

Fitting normal distribution: MLE

$$
\begin{aligned}
\hat{\mu}, \hat{\sigma}^2 &= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \log\left[\operatorname{Norm}_{x_i}[\mu, \sigma^2]\right] \right] \\
&= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[-0.5 I \log[2\pi] - 0.5 I \log \sigma^2 - 0.5 \sum_{i=1}^{I} \frac{(x_i - \mu)^2}{\sigma^2} \right]
\end{aligned}
$$

How to maximize a function? Take derivative and set to zero.

$$
\begin{aligned}
\frac{\partial L}{\partial \mu} &= \sum_{i=1}^{I} \frac{(x_i - \mu)}{\sigma^2} \\
&= \frac{\sum_{i=1}^{I} x_i}{\sigma^2} - \frac{I\mu}{\sigma^2} = 0
\end{aligned}
$$

Solution:

$$
\hat{\mu} = \frac{\sum_{i=1}^{I} x_i}{I}
$$

Fitting normal distribution: MLE

Maximum likelihood solution:

$$\hat{\mu} = \frac{\sum_{i=1}^{I} x_i}{I}$$

$$\sigma^2 = \sum_{i=1}^{I} \frac{(x_i - \hat{\mu})^2}{I}$$

Should look familiar!

Least Squares

Maximum likelihood for the normal distribution...

$$\begin{aligned}\hat{\mu} &= \underset{\mu}{\operatorname{argmax}}\left[-0.5I\log[2\pi] - 0.5I\log\sigma^2 - 0.5\sum_{i=1}^{I}\frac{(x_i-\mu)^2}{\sigma^2}\right] \\ &= \underset{\mu}{\operatorname{argmax}}\left[-\sum_{i=1}^{I}(x_i-\mu)^2\right] \\ &= \underset{\mu}{\operatorname{argmin}}\left[\sum_{i=1}^{I}(x_i-\mu)^2\right],\end{aligned}$$

...gives `least squares' fitting criterion.

Fitting normal distribution: MAP

Fitting

As the name suggests we find the parameters which maximize the posterior probability $Pr(\boldsymbol{\theta}|\mathbf{x}_{1\ldots I})$.

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta}) Pr(\boldsymbol{\theta}) \right]$$

Likelihood is normal PDF

$$Pr(x|\mu,\sigma^2) = \text{Norm}_x[\mu,\sigma^2] = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-0.5\frac{(x-\mu)^2}{\sigma^2}\right]$$

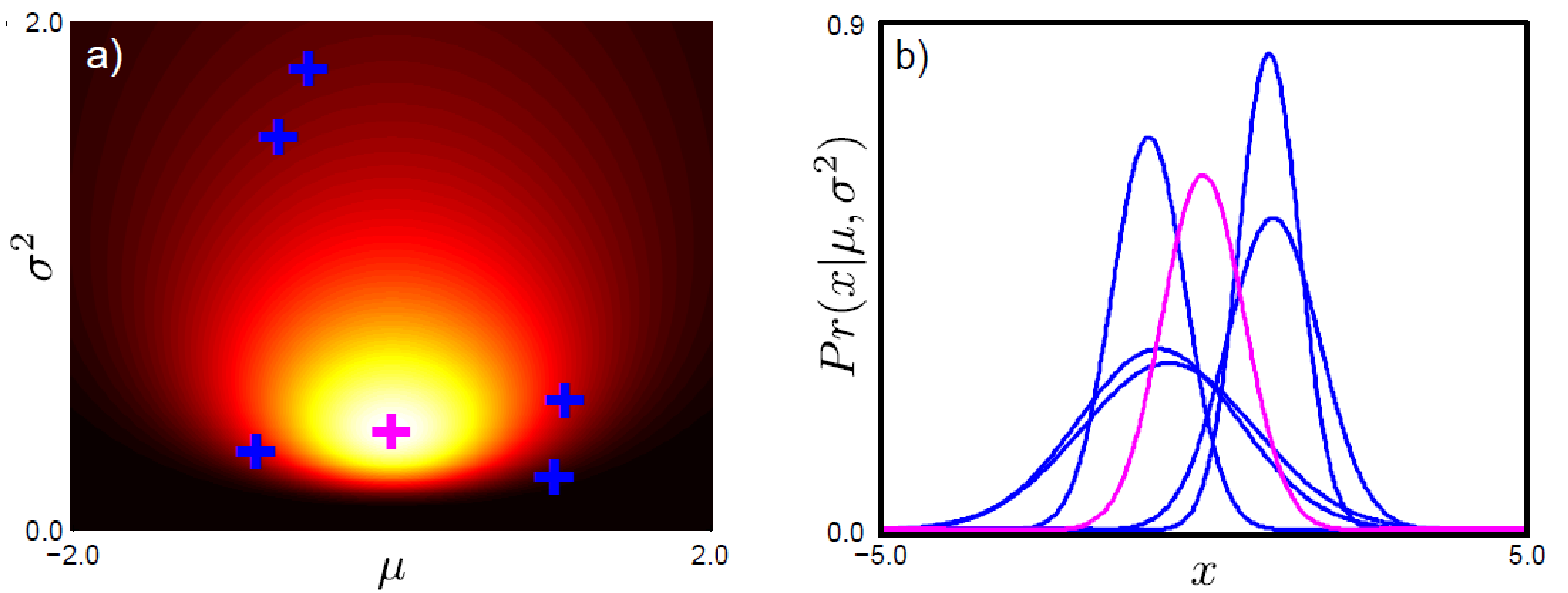
Fitting normal distribution: MAP

Prior

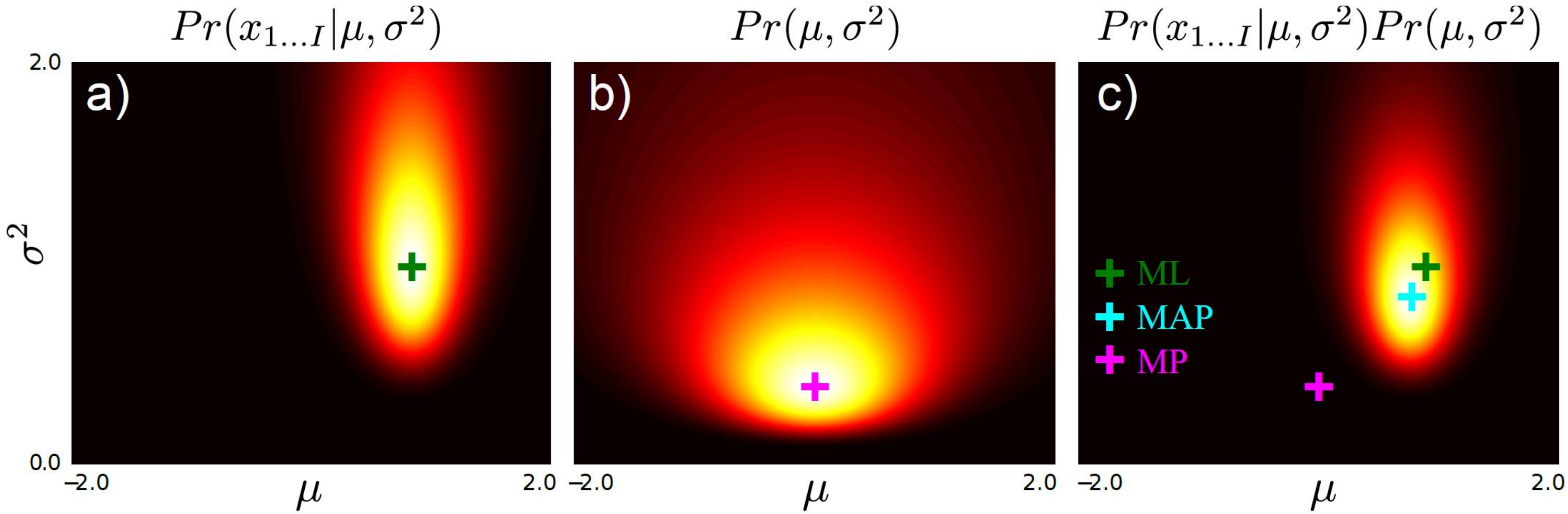
Use conjugate prior, normal scaled inverse gamma.

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

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



Fitting normal distribution: MAP



Likelihood

Prior

Posterior

$$
\begin{aligned}
\hat{\mu},\hat{\sigma}^2 &= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(x_i|\mu,\sigma^2)Pr(\mu,\sigma^2)\right] \\
&= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\prod_{i=1}^{I} \text{Norm}_{x_i}[\mu,\sigma^2]\text{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta]\right]
\end{aligned}
$$

Fitting normal distribution: MAP

$$\begin{aligned}\hat{\mu}, \hat{\sigma}^2 &= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(x_i|\mu,\sigma^2) Pr(\mu,\sigma^2) \right] \\ &= \underset{\mu,\sigma^2}{\operatorname{argmax}} \left[\prod_{i=1}^{I} \operatorname{Norm}_{x_i}[\mu,\sigma^2] \operatorname{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta] \right]\end{aligned}$$

Again maximize the log – does not change position of maximum

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

Fitting normal distribution: MAP

MAP solution:

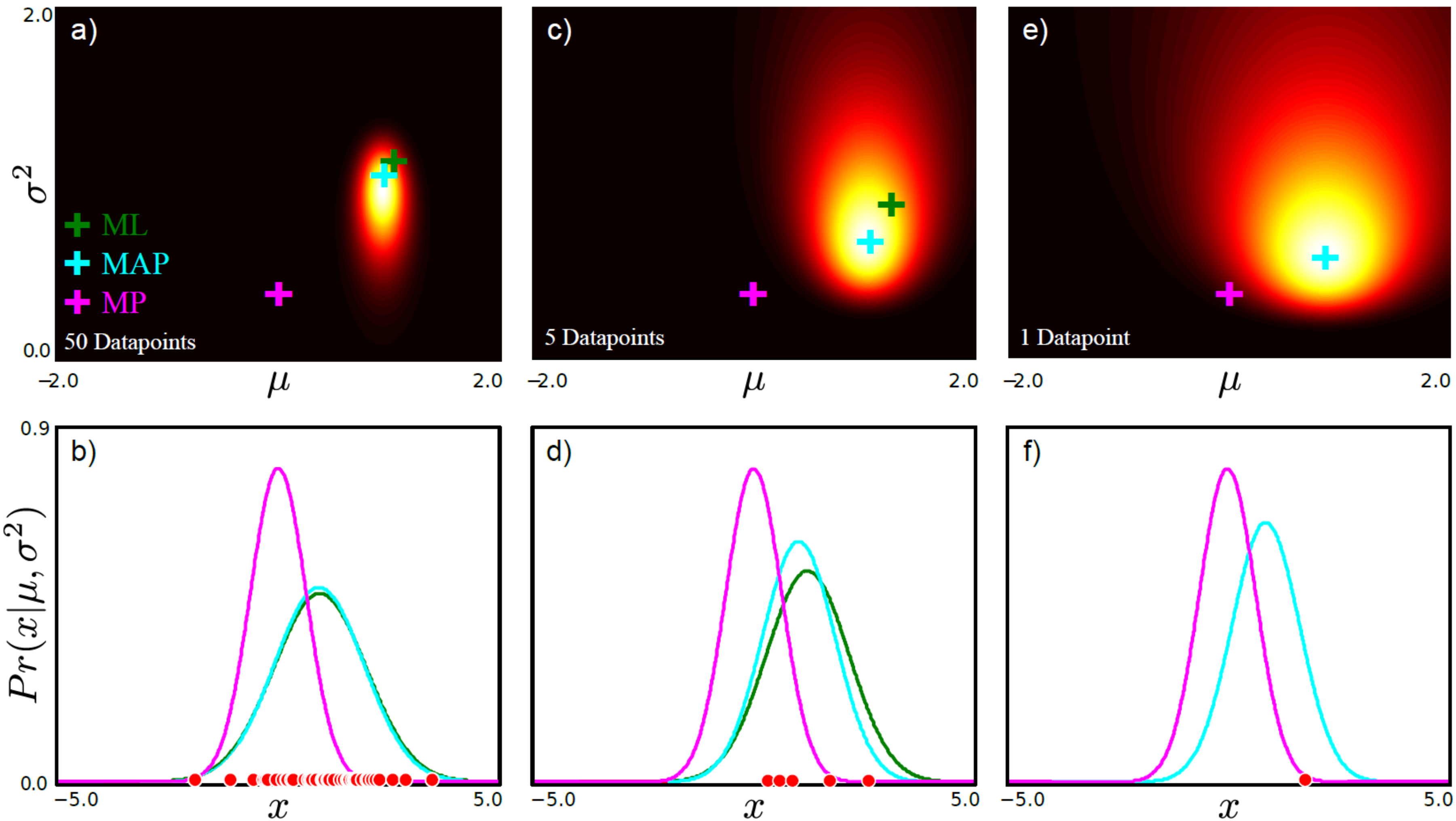
$$\hat{\mu} = \frac{\sum_{i=1} x_i + \gamma\delta}{I + \gamma}$$

$$\hat{\sigma^2} = \frac{\sum_{i=1}^{I}(x_i - \mu)^2 + 2\beta + \gamma(\delta - \mu)^2}{I + 3 + 2\alpha}$$

Mean can be rewritten as weighted sum of data mean and prior mean:

$$\hat{\mu} = \frac{I\overline{x} + \gamma\delta}{I + \gamma}$$

Fitting normal distribution: MAP



50 data points
The prior has little influence

5 data points
The prior has some influence

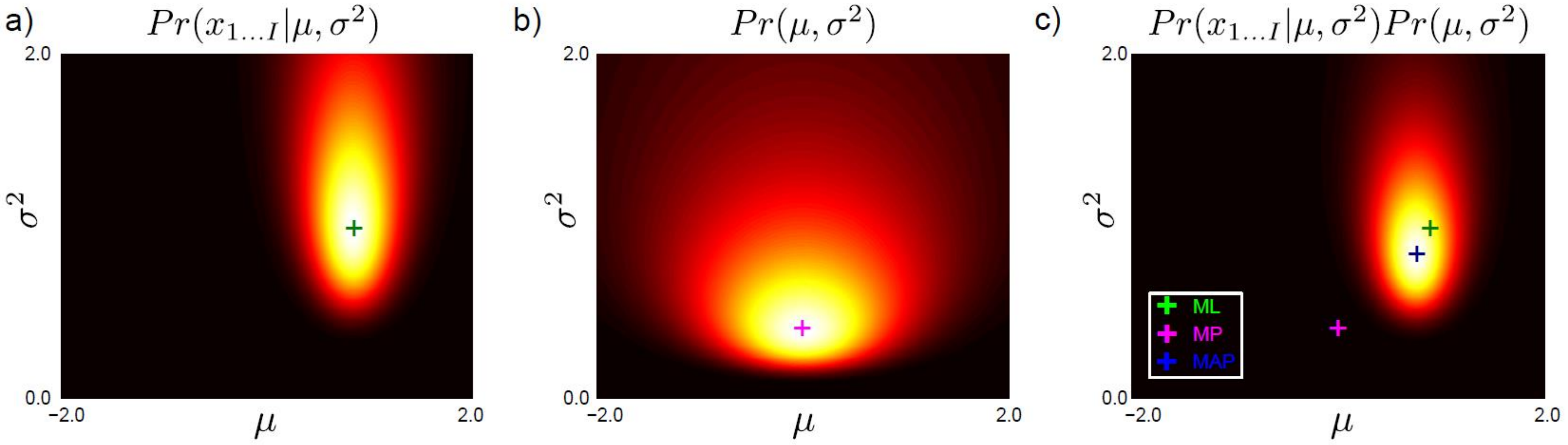
1 data point: No ML solution
Cannot compute variance

Fitting normal: Bayesian approach

Fitting

Compute the posterior distribution using Bayes' rule:

$$Pr(\boldsymbol{\theta}|\mathbf{x}_{1...I}) = \frac{\prod_{i=1}^{I} Pr(\mathbf{x}_i|\boldsymbol{\theta})Pr(\boldsymbol{\theta})}{Pr(\mathbf{x}_{1...I})}$$



Fitting normal: Bayesian approach

Fitting

Compute the posterior distribution using Bayes' rule:

$$
\begin{aligned}
Pr(\mu,\sigma^2|x_{1\ldots I}) &= \frac{\prod_{i=1}^{I} Pr(x_i|\mu,\sigma^2)Pr(\mu,\sigma^2)}{Pr(x_{1\ldots I})}\\
&= \frac{\prod_{i=1}^{I} \text{Norm}_{x_i}[\mu,\sigma^2]\text{NormInvGam}_{\mu,\sigma^2}[\alpha,\beta,\gamma,\delta]}{Pr(x_{1\ldots I})}\\
&= \frac{\kappa(\alpha,\beta,\gamma,\delta,x_{1\ldots I}).\text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}]}{Pr(x_{1\ldots I})},\\
&= \text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}]
\end{aligned}
$$

Two constants MUST cancel out or LHS not a valid pdf

Fitting normal: Bayesian approach

Fitting

Compute the posterior distribution using Bayes' rule:

$$Pr(\mu, \sigma^2 | x_{1\ldots I}) = \text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma}, \tilde{\delta}]$$

where

$$\tilde{\alpha} = \alpha + I/2, \qquad \tilde{\gamma} = \gamma + I \qquad \tilde{\delta} = \frac{(\gamma\delta + \sum_i x_i)}{\gamma + I}$$

$$\tilde{\beta} = \frac{\sum_i x_i^2}{2} + \beta + \frac{\gamma\delta^2}{2} - \frac{(\gamma\delta + \sum_i x_i)^2}{2(\gamma + I)}.$$

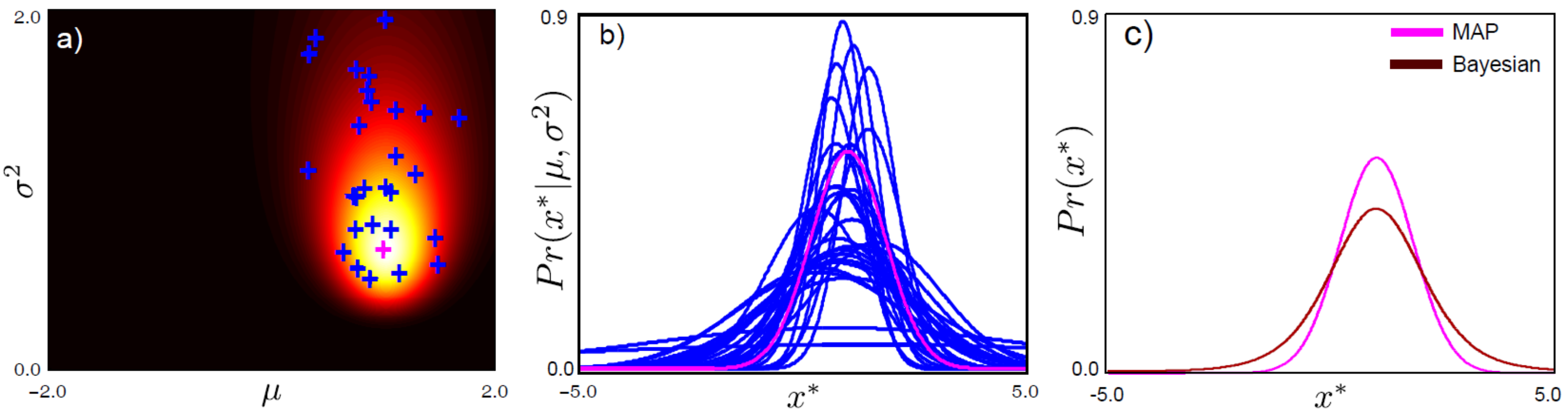
Fitting normal: Bayesian approach

Predictive density

Take weighted sum of predictions from different parameter values:

$$Pr(x^*|x_{1...I}) = \iint Pr(x^*|\mu,\sigma^2)Pr(\mu,\sigma^2|x_{1...I})d\mu d\sigma$$

The average of an infinite set of samples



Fitting normal: Bayesian approach

Predictive density

Take weighted sum of predictions from different parameter values:

$$
\begin{aligned}
Pr(x^*|x_{1\ldots I}) &= \iint Pr(x^*|\mu,\sigma^2)Pr(\mu,\sigma^2|x_{1\ldots I})d\mu d\sigma \\
&= \iint \text{Norm}_{x^*}[\mu,\sigma^2]\text{NormInvGam}_{\mu,\sigma^2}[\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta}]d\mu d\sigma \\
&= \iint \kappa(\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta},x_{1\ldots I}).\text{NormInvGam}_{\mu,\sigma^2}[\breve{\alpha},\breve{\beta},\breve{\gamma},\breve{\delta}]d\mu d\sigma \\
&= \kappa(\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta},x_{1\ldots I})\iint \text{NormInvGam}_{\mu,\sigma^2}[\breve{\alpha},\breve{\beta},\breve{\gamma},\breve{\delta}]d\mu d\sigma \\
&= \kappa(\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta},x_{1\ldots I})
\end{aligned}
$$

Fitting normal: Bayesian approach

Predictive density

Take weighted sum of predictions from different parameter values:

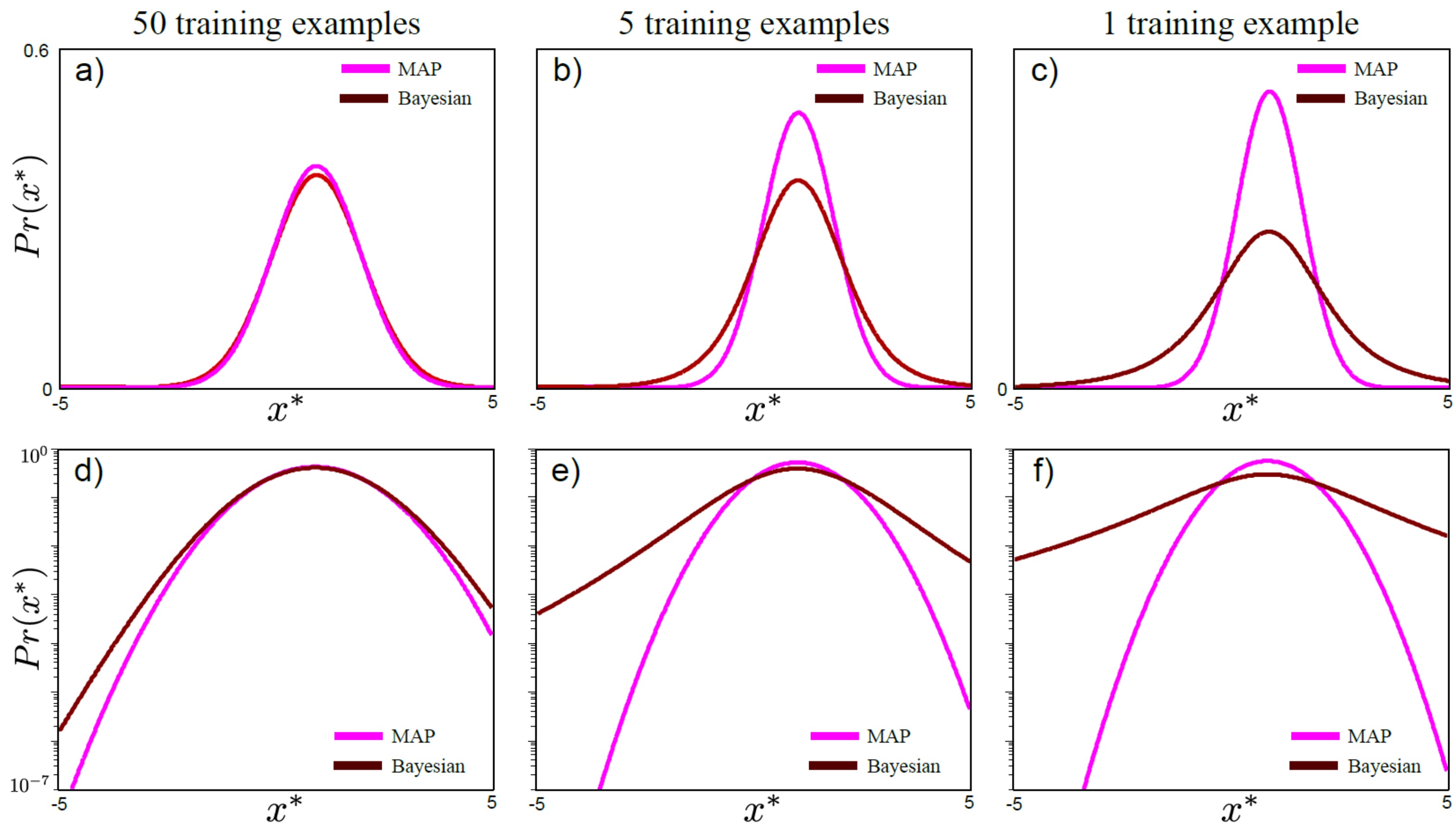
$$Pr(x^*|x_{1\ldots I}) = \kappa(\tilde{\alpha},\tilde{\beta},\tilde{\gamma},\tilde{\delta},x_{1\ldots I}) = \frac{1}{\sqrt{2\pi}}\frac{\sqrt{\tilde{\gamma}}\tilde{\beta}^{\tilde{\alpha}}}{\sqrt{\breve{\gamma}}\breve{\beta}^{\breve{\alpha}}}\frac{\Gamma[\breve{\alpha}]}{\Gamma[\tilde{\alpha}]}$$

where

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

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

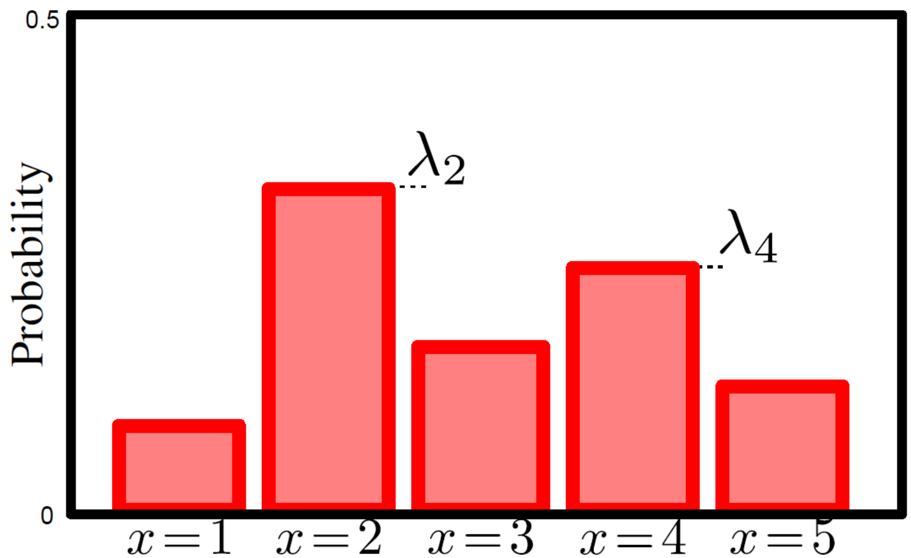
Fitting normal: Bayesian Approach



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

Categorical distribution

Categorical Distribution



$$Pr(x = k) = \lambda_k$$

or can think of data as vector with all elements zero except k[th] e.g. [0,0,0,1 0]

$$Pr(\mathbf{x} = \mathbf{e}_k) = \prod_{j=1}^{K} \lambda_j^{x_j} = \lambda_k$$

For short we write:

$$Pr(x) = \text{Cat}_x [\boldsymbol{\lambda}]$$

Categorical distribution describes situation where K possible outcomes y=1… y=k.

Takes a K parameters $\lambda_k \in [0, 1]$ where $\sum_k \lambda_k = 1$

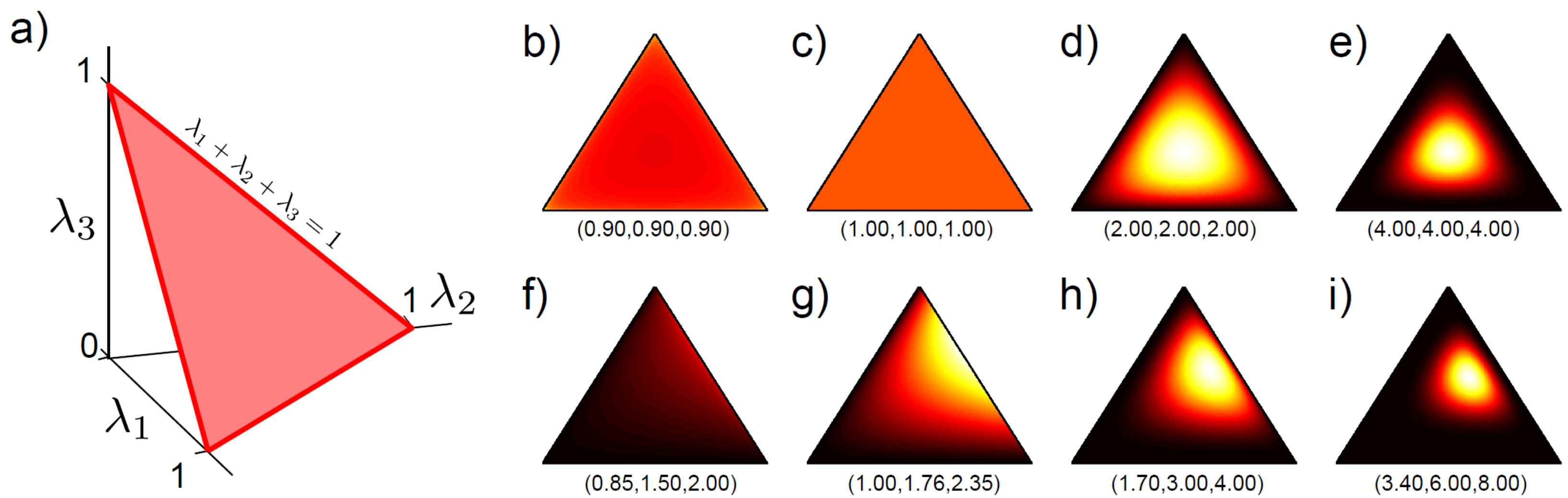
Dirichlet Distribution

Defined over K values $\lambda_k \in [0,1]$ where $\sum_k \lambda_k = 1$

$$Pr(\lambda_1 \ldots \lambda_K) = \frac{\Gamma[\sum_{k=1}^{K}\alpha_k]}{\prod_{k=1}^{K}\Gamma[\alpha_k]}\prod_{k=1}^{K}\lambda_k^{\alpha_k - 1}$$

Has k parameters α_k>0

Or for short: $Pr(\lambda_1 \ldots \lambda_K) = \text{Dir}_{\lambda_{1\ldots K}}[\alpha_1, \alpha_2, \ldots, \alpha_K]$



Categorical distribution: ML

I=6 simulates the roll of a dice

Maximize product of individual likelihoods

$$\begin{aligned}
\hat{\lambda}_{1\ldots 6} &= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(x_i|\lambda_{1\ldots 6}) \right] && \text{s.t.} \sum_k \lambda_k = 1 \\
&= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} \mathrm{Cat}_{x_i}[\lambda_{1\ldots 6}] \right] && \text{s.t.} \sum_k \lambda_k = 1 \\
&= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{k=1}^{6} \lambda_k^{N_k} \right] && \text{s.t.} \sum_k \lambda_k = 1
\end{aligned}$$

N_k is the total number of times we observed the k-th bin

Categorical distribution: ML

Instead maximize the log probability

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

Log likelihood

Lagrange multiplier to ensure that params sum to one

Take derivative, set to zero and re-arrange:

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

The proportion of times we observed bin k

Categorical distribution: MAP

MAP criterion:

$$
\begin{aligned}
\hat{\lambda}_{1\ldots 6} &= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} Pr(x_i|\lambda_{1\ldots 6}) Pr(\lambda_{1\ldots 6}) \right] \\
&= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{i=1}^{I} \mathrm{Cat}_{x_i}[\lambda_{1\ldots 6}] \mathrm{Dir}_{\lambda_{1\ldots 6}}[\alpha_{1\ldots 6}] \right] \\
&= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{k=1}^{6} \lambda_k^{N_k} \prod_{k=1}^{6} \lambda^{\alpha_k - 1} \right] \\
&= \underset{\lambda_{1\ldots 6}}{\operatorname{argmax}} \left[\prod_{k=1}^{6} \lambda_k^{N_k + \alpha_k - 1} \right].
\end{aligned}
$$

Categorical distribution: MAP

Take derivative, set to zero and re-arrange:

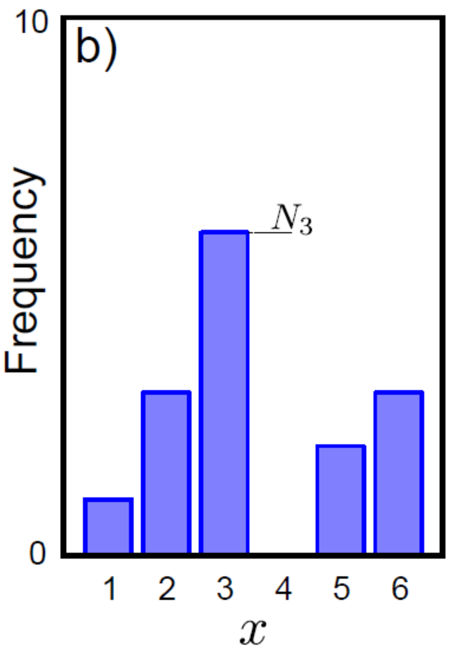
$$\hat{\lambda}_k = \frac{N_k + \alpha_k - 1}{\sum_{m=1}^{6}(N_m + \alpha_m - 1)}$$

With a uniform prior ($\alpha_{1..K}$=1), gives same result as maximum likelihood.

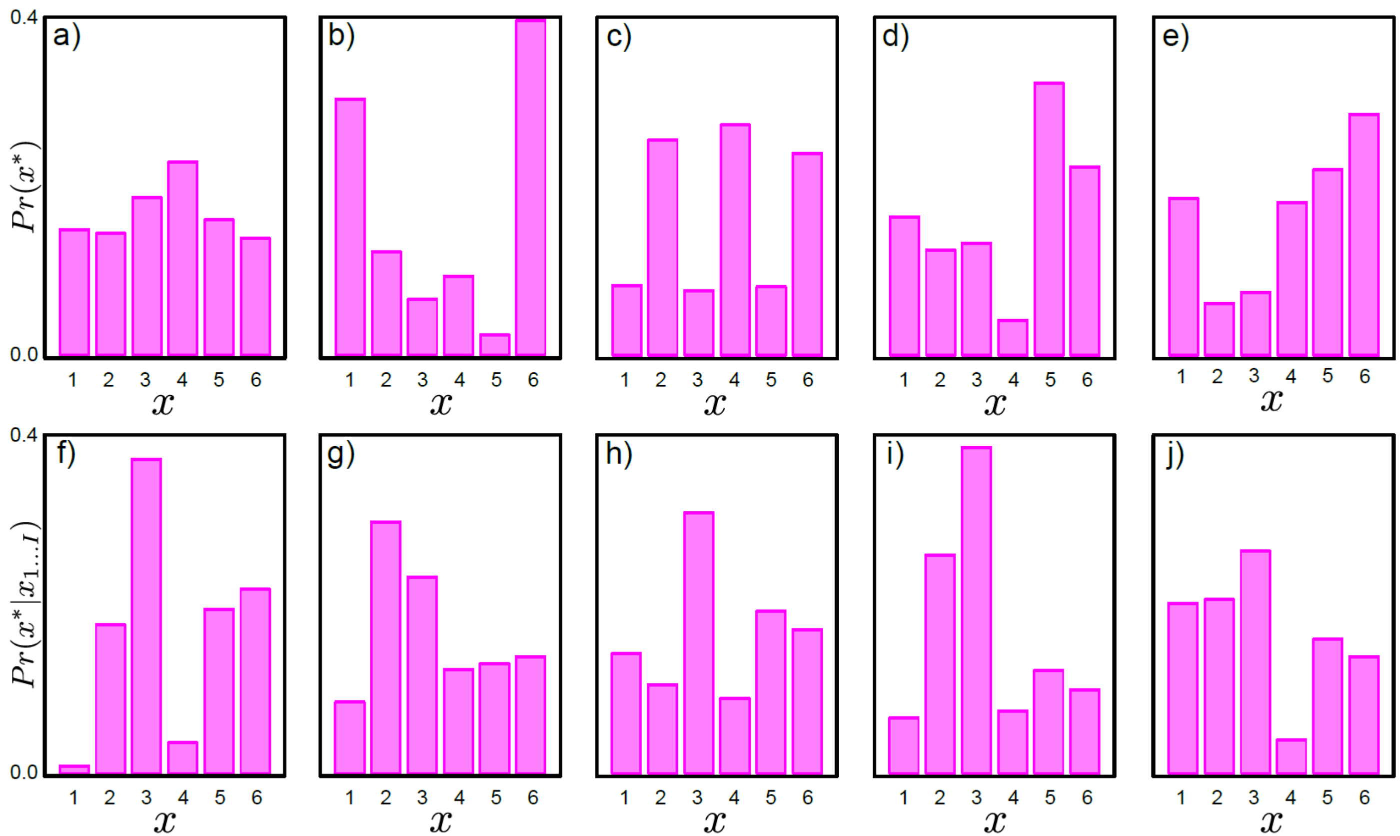
$$\hat{\lambda}_k = \frac{N_k}{\sum_{m=1}^{6} N_m}$$

Categorical Distribution

Five samples from Dirichlet prior with equal α_i (uniform distribution)



Observed data



Five samples from posterior (MAP)

The distribution favors histograms where bin three is larger and bin four is small as suggested by the data.

Categorical Distribution: Bayesian approach

Compute posterior distribution over parameters:

$$
\begin{aligned}
Pr(\lambda_1 \ldots \lambda_6|x_{1\ldots I}) &= \frac{\prod_{i=1}^{I} Pr(x_i|\lambda_{1\ldots 6})Pr(\lambda_{1\ldots 6})}{Pr(x_{1\ldots I})} \\
&= \frac{\prod_{i=1}^{I} \mathrm{Cat}_{x_i}[\lambda_{1\ldots 6}]\mathrm{Dir}_{\lambda_{1\ldots 6}}[\alpha_{1\ldots 6}]}{Pr(x_{1\ldots I})} \\
&= \frac{\kappa(\alpha_{1\ldots 6}, x_{1\ldots I})\mathrm{Dir}_{\lambda_{1\ldots 6}}[\tilde{\alpha}_{1\ldots 6}]}{Pr(x_{1\ldots I})} \\
&= \mathrm{Dir}_{\lambda_{1\ldots 6}}[\tilde{\alpha}_{1\ldots 6}],
\end{aligned}
$$

Two constants MUST cancel out or LHS not a valid pdf

Categorical Distribution: Bayesian approach

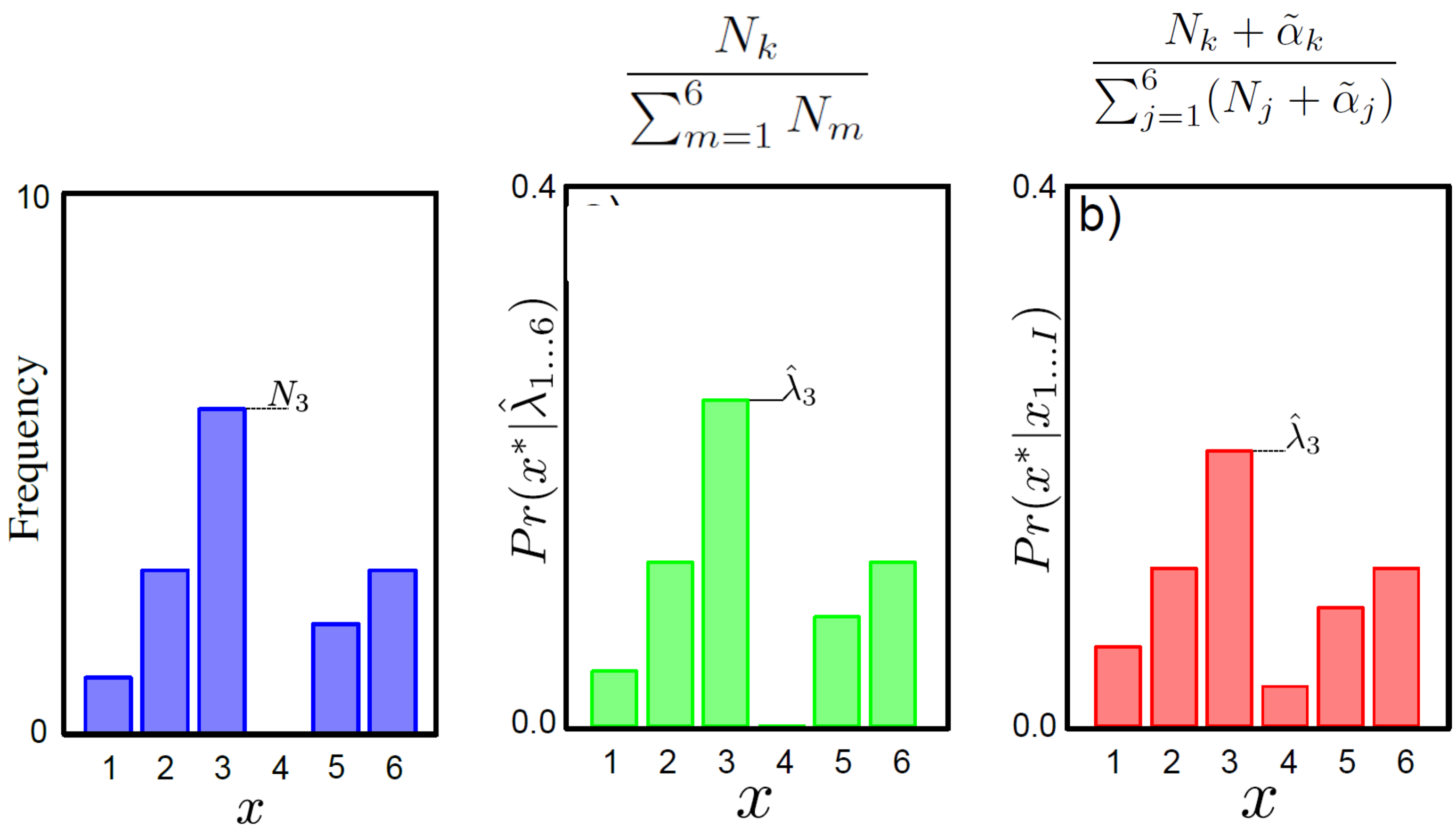
Compute predictive distribution:

$$\begin{aligned}
Pr(x^*|x_{1\ldots I}) &= \int Pr(x^*|\lambda_{1\ldots 6})Pr(\lambda_{1\ldots 6}|x_{1\ldots I})\, d\lambda_{1\ldots 6} \\
&= \int \text{Cat}_{x^*}[\lambda_{1\ldots 6}]\text{Dir}_{\lambda_{1\ldots 6}}[\tilde{\alpha}_{1\ldots 6}]\, d\lambda_{1\ldots 6} \\
&= \int \kappa(x^*, \tilde{\alpha}_{1\ldots 6})\text{Dir}_{\lambda_{1\ldots 6}}[\breve{\alpha}_{1\ldots 6}]\, d\lambda_{1\ldots 6} \\
&= \kappa(x^*, \tilde{\alpha}_{1\ldots 6}).
\end{aligned}$$

Two constants MUST cancel out or LHS not a valid pdf

$$Pr(x^* = k|x_{1\ldots I}) \quad = \quad \kappa(x^*, \tilde{\alpha}_{1\ldots 6}) = \frac{N_k + \tilde{\alpha}_k}{\sum_{j=1}^{6}(N_j + \tilde{\alpha}_j)}$$

ML / MAP vs. Bayesian



The ML/MAP approaches are confident w.r.t. the observed data.
The Bayesian approach predicts a more moderate distribution and allots some probability to the case x = 4 (we may have been unlucky in the data).

Conclusion

- Three ways to fit probability distributions
  - Maximum likelihood
  - Maximum a posteriori
  - Bayesian Approach

- Two worked example
  - Normal distribution (ML→ least squares)
  - Categorical distribution