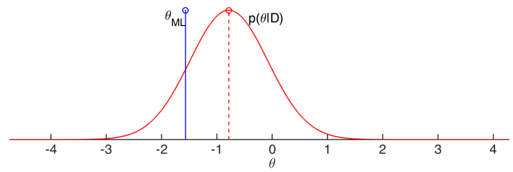
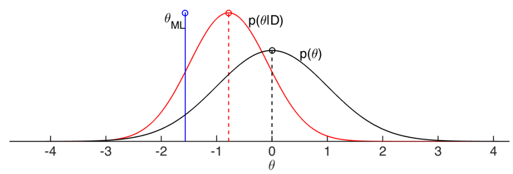
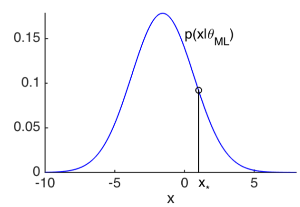
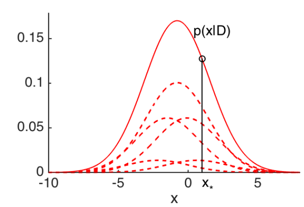
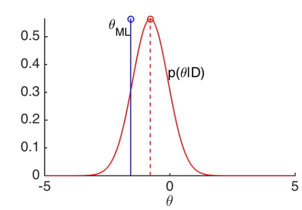
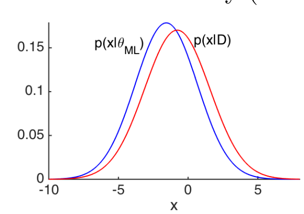
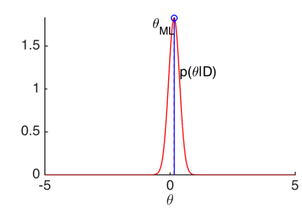
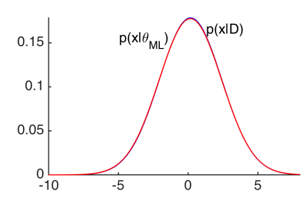
CS5487 - Comparisons of Maximum Likelihood and Bayesian Estimation

Antoni B. Chan
Department of Computer Science
City University of Hong Kong

Sep 21, 2016

1 Comparison between MLE and Bayesian Estimation

	MLE	Bayesian
1) Parameter estimate	Single number $\hat{\theta}_{ML}$. 	Distribution $p(\Theta\|D)$.
2) Characterization	One "best" estimate. 	Complete characterization including uncertainty.
3) Flexibility	WYSIWYG, "agnostic" data-driven.	Estimate influenced by prior $p(\theta)$; "biased by beliefs".
4) Predictions	$p(x_*\|\hat{\theta}_{ML})$ from one model. 	$p(x_*\|D)$ averaged over all models (regularization effect). 
5) Computation	OK, no integration, just optimization	Requires integrals: $\int p(D\|\theta)p(\theta)d\theta$, $\int p(x\|\theta)p(\theta\|D)d\theta$. When no closed-form solution, requires approximations (sampling, Laplace, EP, variational, etc.).
6) "small" data	Can overfit to the data. 	Regularizes the ML estimate when there is uncertainty (little data). 
7) "big" data	Asymptotically unbiased and efficient. 	Ignores the prior, and converges to the MLE when there is certainty (more data). 

2 Gaussian Distribution

Gaussian observation model with Gaussian prior distribution:

- prior distribution: $p(\mu) = \mathcal{N}(\mu|\mu_0, \sigma_0^2)$
- observation likelihood: $p(x|\mu) = \mathcal{N}(x|\mu, \sigma^2)$.
- The mean μ is unknown, while $\{\mu_0, \sigma_0^2, \sigma^2\}$ are known.

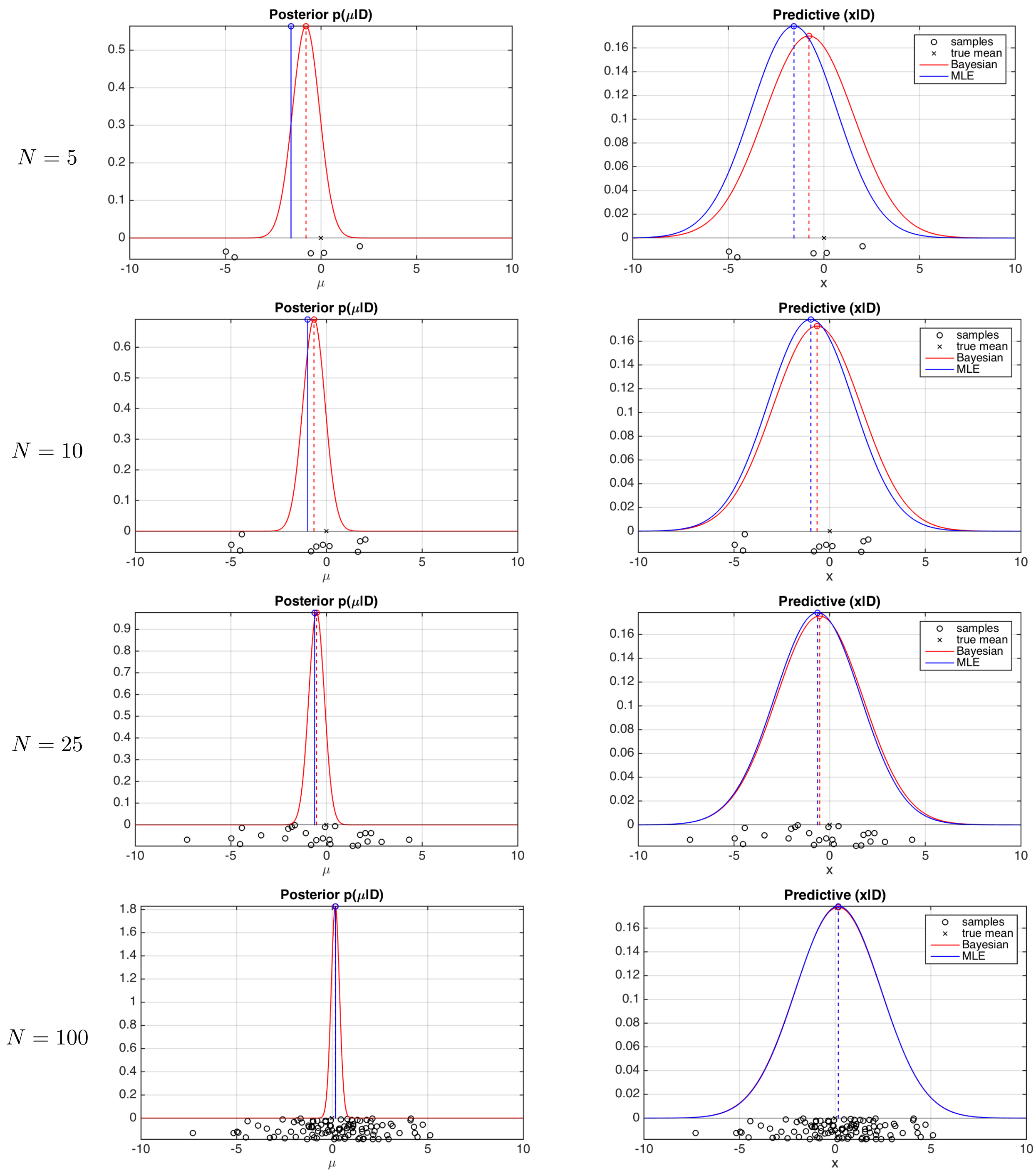
The maximum likelihood estimate and various Bayesian estimates are summarized below:

	parameter estimate	posterior distribution $p(\mu\|\mathcal{D})$	predictive distribution $p(x\|\mathcal{D})$
MLE	$\hat{\mu}_{\mathrm{ML}} = \frac{1}{n}\sum_i x_i$	$\delta(\mu - \hat{\mu}_{\mathrm{ML}})$	$\mathcal{N}(x\|\hat{\mu}_{\mathrm{ML}}, \sigma^2)$
Bayesian	$\begin{cases} \hat{\mu}_n = \alpha\hat{\mu}_{\mathrm{ML}} + (1-\alpha)\mu_0 \\ \alpha = \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2} \\ \frac{1}{\hat{\sigma}_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \end{cases}$	$\mathcal{N}(\mu\|\hat{\mu}_n, \hat{\sigma}_n^2)$	$\mathcal{N}(x\|\hat{\mu}_n, \hat{\sigma}_n^2 + \sigma^2)$
MAP	$\hat{\mu}_{\mathrm{MAP}} = \hat{\mu}_n$	$\delta(\mu - \hat{\mu}_{\mathrm{MAP}})$	$\mathcal{N}(x\|\hat{\mu}_{\mathrm{MAP}}, \sigma^2)$
Bayesian (non-informative; $\sigma_0^2 \to \infty$)	$\hat{\mu}_n = \hat{\mu}_{\mathrm{ML}}$	$\mathcal{N}(\mu\|\hat{\mu}_n, \frac{1}{n}\sigma^2)$	$\mathcal{N}(x\|\hat{\mu}_{\mathrm{ML}}, (1+\frac{1}{n})\sigma^2)$

different for small n,
same for large n

2.1 Gaussian Example

The true distribution is $p(x) = \mathcal{N}(x|0, 5)$, and $N = \{5, 10, 25, 100\}$ samples are drawn. The prior is $p(\mu) = \mathcal{N}(\mu|0, 1)$. The below plots show the posterior for μ and the predictive distribution for MLE and Bayesian Estimation. The points are plotted below the densities, and are randomly scattered in the y-direction for visualization. The two methods differ when there are few examples, with the Bayesian method biased towards the prior. When there are many examples ($N = 100$), the two methods have the similar estimates.



3 Bernoulli Distribution (Problem 3.7)

Bernoulli observation model with different priors:

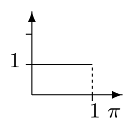
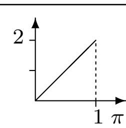
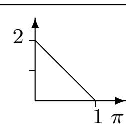
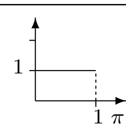
- prior distribution: $p(\pi)$
- observation likelihood: $p(x|\pi) = \pi^x (1-\pi)^{1-x}$
- data likelihood: $p(\mathcal{D}|\pi) = \pi^s (1-\pi)^{n-s}$, where $s = \sum_i x_i$.

It can be shown that the *predictive distribution* has the form:

$$p(x|\mathcal{D}) = \hat{\pi}^x (1-\hat{\pi})^{1-x}. \tag{1}$$

Hence, the predictive distribution is also a Bernoulli distribution, but with a modified parameter $\hat{\pi}$.

The maximum likelihood estimate and various Bayesian estimates for different priors are summarized below:

	prior $p(\pi)$	predictive distribution $p(x\|\mathcal{D})$	# of tosses	# of 1's	interpretation
MLE	–	$\hat{\pi} = \frac{s}{n}$	n	s	–
MAP (uniform)		$\hat{\pi} = \frac{s}{n}$	n	s	"same as MLE"
MAP (favor 1's)		$\hat{\pi} = \frac{s+1}{n+1}$	$n+1$	$s+1$	"add a 1"
MAP (favor 0's)		$\hat{\pi} = \frac{s}{n+1}$	$n+1$	s	"add a 0"
Bayesian (uniform)		$\hat{\pi} = \frac{s+1}{n+2}$	$n+2$	$s+1$	"add one of each"

can fix empty bins by filling with extra samples (*regularization*)

Note: the Bayesian estimate is consistent with the non-informative prior (1 is equally as likely as 0).