Bayesian Non-parametrics for Biomedical Applications

Melanie F. Pradier

Memorial Sloan-Kettering Cancer Center, Universidad Carlos III in Madrid

05 February 2015

Outline

1. Bayesian Modeling
2. Bayesian Non-parametrics
3. Biomedical Applications
4. Conclusions

Sources and References

Parts of these slides are adapted from the following sources

- C. Bishop: *Pattern Recognition and Machine Learning*, 2006.
- K. P. Murphy: *Machine Learning: a Probabilistic Perspective*, 2012.
- D. J.C. MacKay: *Information Theory, Inference, and Learning Algorithms*, 2003.
- Z. Ghahramani & C. E. Rasmussen, Slides for *Machine Learning Course* at Cambridge University.
- S. J. Gershman, D.M. Blei: A tutorial on Bayesian nonparametric models, 2012.
- Y.W. Teh: Slides for *Probabilistic and Bayesian Machine Learning*, UC3M, 2010.
- M. N. Schmidt & M. Morup: *Advanced Topics in Machine Learning*, MLSS, DTU, 2013.
- D. B. Dunson: *Nonparametric Bayes Applications to Biostatistics*, 2010.

Outline

Example 1: Medical Diagnosis

Problem Formulation

- 1% of scanned women have breast cancer
- 80% of women with breast cancer get positive mammography
- 9.6% of women without breast cancer also get positive mammography

Question: A random women gets a positive scan, what is the probability that she has breast cancer?

1. less than 1%
2. around 10%
3. around 90%
4. more than 99%

Example 1: Medical Diagnosis

Problem Formulation

- 1% of scanned women have breast cancer
- 80% of women with breast cancer get positive mammography
- 9.6% of women without breast cancer also get positive mammography

Question: A random women gets a positive scan, what is the probability that she has breast cancer?

1. less than 1%
2. around 10%
3. around 90%
4. more than 99%

Example 1: Medical Diagnosis

Problem Formulation

- 1% of scanned women have breast cancer
- 80% of women with breast cancer get positive mammography
- 9.6% of women without breast cancer also get positive mammography

$C/\bar{C}$ = has cancer or not
$M/\bar{M}$ = positive scan or not

- $p(C) = 0.01$
- $p(M|C) = 0.8$
- $p\left(M|\bar{C}\right) = 0.096$

$p(C|M)$?

Considering 10.000 subjects

	M	$\bar{M}$
C	80	20
$\bar{C}$	950	8950

$p(C|M) = \frac{p(C,M)}{p(C,M)+p\left(\bar{C},M\right)} = \frac{p(C,M)}{p(M)} \simeq 7.8\%$

Example 1: Medical Diagnosis

Problem Formulation

- 1% of scanned women have breast cancer
- 80% of women with breast cancer get positive mammography
- 9.6% of women without breast cancer also get positive mammography

$C/\bar{C}$ = has cancer or not
$M/\bar{M}$ = positive scan or not

- $p(C) = 0.01$
- $p(M|C) = 0.8$
- $p\left(M|\bar{C}\right) = 0.096$

$p(C|M)$?

Considering 10.000 subjects

	M	$\bar{M}$
C	80	20
$\bar{C}$	950	8950

$$p(C|M) = \frac{p(C,M)}{p(C,M)+p\left(\bar{C},M\right)} = \frac{p(C,M)}{p(M)} \simeq 7.8\%$$

Example 1: Medical Diagnosis

Problem Formulation

- 1% of scanned women have breast cancer
- 80% of women with breast cancer get positive mammography
- 9.6% of women without breast cancer also get positive mammography

$C/\bar{C}$ = has cancer or not
$M/\bar{M}$ = positive scan or not

- $p(C) = 0.01$
- $p(M|C) = 0.8$
- $p(M|\bar{C}) = 0.096$

$p(C|M)$?

Considering 10.000 subjects

	M	$\bar{M}$
C	80	20
$\bar{C}$	950	8950

$$p(C|M) = \frac{p(C,M)}{p(C,M)+p(\bar{C},M)} = \frac{p(C,M)}{p(M)} \simeq 7.8\%$$

Example 1: Medical Diagnosis

Problem Formulation

- 1% of scanned women have breast cancer
- 80% of women with breast cancer get positive mammography
- 9.6% of women without breast cancer also get positive mammography

$C/\bar{C}$ = has cancer or not
$M/\bar{M}$ = positive scan or not

- $p(C) = 0.05$?
- $p(M|C) = 0.8$
- $p(M|\bar{C}) = 0.096$

$p(C|M)$?

Considering 10.000 subjects

	M	$\bar{M}$
C	80	20
$\bar{C}$	950	8950

$$p(C|M) = \frac{p(C,M)}{p(C,M)+p(\bar{C},M)} = \frac{p(C,M)}{p(M)} \simeq 7.8\%$$

Example 1: Medical Diagnosis

Problem Formulation

- 1% of scanned women have breast cancer
- 80% of women with breast cancer get positive mammography
- 9.6% of women without breast cancer also get positive mammography

$C/\bar{C}$ = has cancer or not
$M/\bar{M}$ = positive scan or not

- $p(C) = 0.05$
- $p(M|C) = 0.8$
- $p(M|\bar{C}) = 0.096$

$p(C|M)$?

Considering 10.000 subjects

	M	$\bar{M}$
C	400	100
$\bar{C}$	912	8588

$$p(C|M) = \frac{p(C,M)}{p(C,M)+p(\bar{C},M)} = \frac{p(C,M)}{p(M)} \simeq 52.5\%$$

Bayesian Statistics

- Probability = degree of belief (in contrast with frequentist definition)

Bayes Rule

$$p(\theta|X) = \frac{p(X|\theta)\, p(\theta)}{p(X)}$$

- posterior: $p(\theta|X)$
- likelihood: $p(X|\theta)$
- prior: $p(\theta)$
- evidence: $p(X)$

Bayesian Statistics

1. Sum rule: $p(A) = \sum_B p(A, B)$ or $p(A) = \int p(A, B)\, dB$
2. Product rule: $p(A, B) = p(A|B)\, p(B)$

Evidence = marginal likelihood $p(X) = \int p(X, \theta)\, d\theta = \int p(X|\theta)\, p(\theta)\, d\theta$

Question: What is $p(X|\theta)$?

- Likelihood?
- Conditional distribution?

Example 2: Coin Flipping

The Frequentist Approach

Problem Formulation

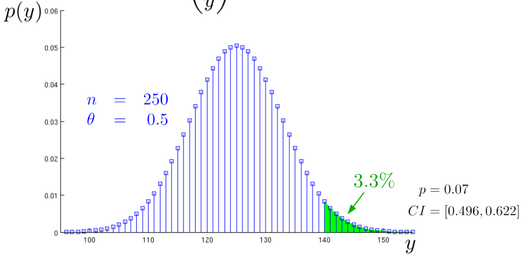
- Imagine you want to know if a coin is biased.
- Imagine you see 140 times Head and 110 times Tail.
- Is the coin well balanced or not?

Example 2: Coin Flipping
The Frequentist Approach

Problem Formulation

- Imagine you want to know if a coin is biased.
- Imagine you see 140 times Head and 110 times Tail.
- Is the coin well balanced or not?

$$p(y) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$$



Example 2: Coin Flipping
The Bayesian Approach

- Both data y and parameter θ as random variables

$$y|\theta \sim Binomial\left(y|N, \theta\right)$$

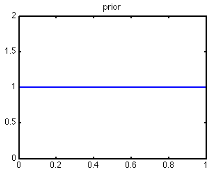
$$\theta \sim Beta\left(\theta|\alpha, \beta\right)$$

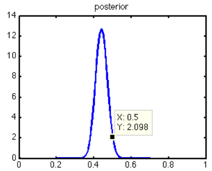
- Joint distribution

$$p\left(y, \theta\right) = p\left(y|\theta\right) p\left(\theta\right)$$

- Posterior distribution

$$p\left(\theta|y\right) = \frac{p\left(y|\theta\right) p\left(\theta\right)}{p\left(y\right)}$$





Example 2: Coin Flipping
The Bayesian Approach

- Both data y and parameter θ as random variables

$$y|\theta \sim Binomial\left(y|N, \theta\right)$$

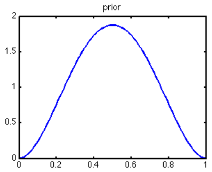
$$\theta \sim Beta\left(\theta|\alpha, \beta\right)$$

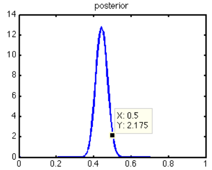
- Joint distribution

$$p\left(y, \theta\right) = p\left(y|\theta\right) p\left(\theta\right)$$

- Posterior distribution

$$p\left(\theta|y\right) = \frac{p\left(y|\theta\right) p\left(\theta\right)}{p\left(y\right)}$$





Parameter Estimation

Estimators

- ML estimator

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p\,(y|\theta)$$

- MAP estimator

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p\,(\theta|y)$$

- **Posterior distribution** $\rightarrow$ Posterior Mean estimator (MP)

$$\hat{\theta}_{PM} = \int \theta p\,(\theta|X)\, d\theta$$

Parameter Estimation

Estimators

- ML estimator

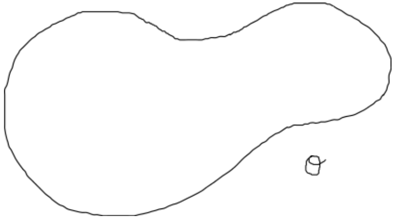
$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p\left(y|\theta\right)$$

- MAP estimator

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p\left(\theta|y\right)$$

- **Posterior distribution** $\rightarrow$ Posterior Mean estimator (MP)

$$\hat{\theta}_{PM} = \int \theta p\left(\theta|X\right) d\theta$$



Parameter Estimation

Estimators

- ML estimator

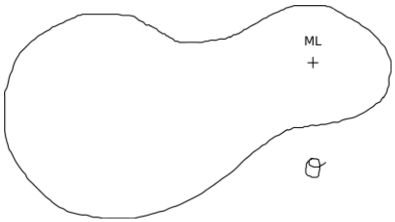
$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p\left(y|\theta\right)$$

- MAP estimator

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p\left(\theta|y\right)$$

- **Posterior distribution** $\rightarrow$ Posterior Mean estimator (MP)

$$\hat{\theta}_{PM} = \int \theta p\left(\theta|X\right) d\theta$$



Parameter Estimation

Estimators

- ML estimator

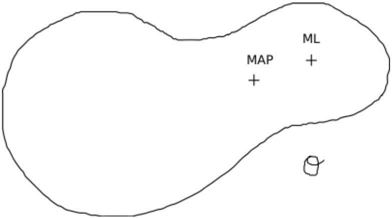
$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p\,(y|\theta)$$

- MAP estimator

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p\,(\theta|y)$$

- **Posterior distribution** $\rightarrow$ Posterior Mean estimator (MP)

$$\hat{\theta}_{PM} = \int \theta p\,(\theta|X)\, d\theta$$



Parameter Estimation

Estimators

- ML estimator

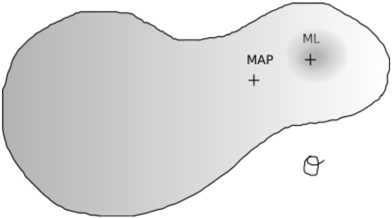
$$\hat{\theta}_{ML} = \operatorname*{argmax}_{\theta} p(y|\theta)$$

- MAP estimator

$$\hat{\theta}_{MAP} = \operatorname*{argmax}_{\theta} p(\theta|y)$$

- **Posterior distribution** → Posterior Mean estimator (MP)

$$\hat{\theta}_{PM} = \int \theta p(\theta|X)\, d\theta$$



Parameter Estimation

Estimators

- ML estimator

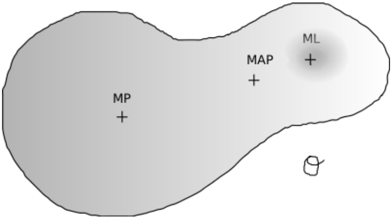
$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p\left(y|\theta\right)$$

- MAP estimator

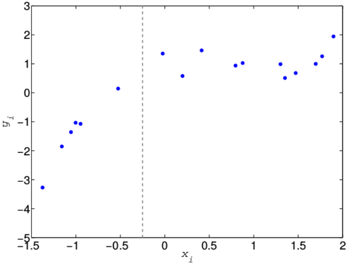
$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p\left(\theta|y\right)$$

- **Posterior distribution** $\rightarrow$ Posterior Mean estimator (MP)

$$\hat{\theta}_{PM} = \int \theta p\left(\theta|X\right) d\theta$$

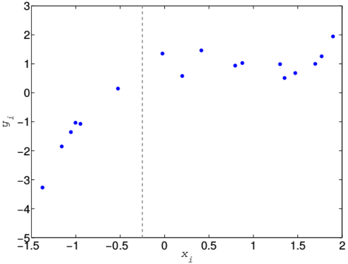


Example 3: Temporal Regression



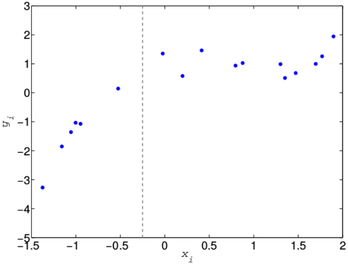
$$y_i = \beta^T \phi(x_i) + \epsilon$$

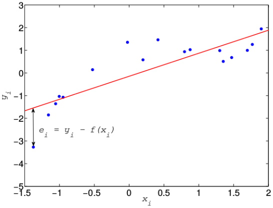
Example 3: Temporal Regression



$$y_i = \beta^T \phi(x_i) + \epsilon$$

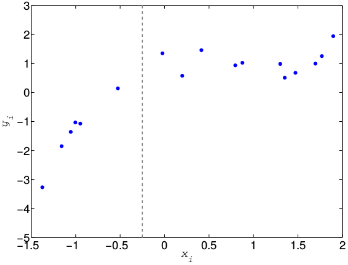
Example 3: Temporal Regression



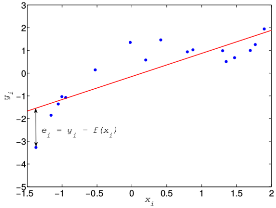


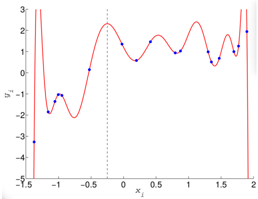
$$y_i = \beta^T \phi(x_i) + \epsilon$$

Example 3: Temporal Regression



$$y_i = \beta^T \phi(x_i) + \epsilon$$





How to avoid overfitting?

- Do cross-validation
- Put some regularization: $\min_{\beta} \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \lambda \left\| \beta \right\|^2$
- Put a prior on the coefficients β

$\beta \sim N\left(0, \tau^2 I\right)$

$y_i | \beta \sim N\left(\beta^T \phi(x_i), \sigma^2\right)$

$$\begin{aligned} \log p(Y, \beta) &= \sum_i p(y_i|\beta) + p(\beta) \\ &\propto \sum_i -\frac{1}{2\sigma^2} \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \frac{1}{2\tau^2} \|\beta\|^2 \\ &\propto \left(-\frac{1}{2\sigma^2}\right) \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 + \frac{\sigma^2}{\tau^2} \|\beta\|^2 \end{aligned}$$

$$\begin{aligned} p(Y, \beta) &= p(Y|\beta)\, p(\beta) \\ &= \prod_i p(y_i|\beta)\, p(\beta) \end{aligned}$$

Regularization actually equivalent to putting a prior!

How to avoid overfitting?

- Do cross-validation
- Put some regularization: $\min_{\beta} \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \lambda \left\| \beta \right\|^2$
- Put a prior on the coefficients β

$$\beta \sim N\left(0, \tau^2 I\right)$$

$$\log p\left(Y, \beta\right) = \sum_i p\left(y_i|\beta\right) + p\left(\beta\right)$$

$$y_i|\beta \sim N\left(\beta^T \phi\left(x_i\right), \sigma^2\right)$$

$$\propto \sum_i -\frac{1}{2\sigma^2} \left\| y_i - \beta^T \phi\left(x_i\right) \right\|^2 - \frac{1}{2\tau^2} \left\| \beta \right\|^2$$

$$\propto \left(-\frac{1}{2\sigma^2}\right) \sum_i \left\| y_i - \beta^T \phi\left(x_i\right) \right\|^2 + \frac{\sigma^2}{\tau^2} \left\| \beta \right\|^2$$

$$p\left(Y, \beta\right) = p\left(Y|\beta\right) p\left(\beta\right) = \prod_i p\left(y_i|\beta\right) p\left(\beta\right)$$

Regularization actually equivalent to putting a prior!

How to avoid overfitting?

- Do cross-validation
- Put some regularization: $\min\limits_{\beta} \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \lambda \left\| \beta \right\|^2$
- Put a prior on the coefficients β

$$\beta \sim N\left(0, \tau^2 I\right)$$

$$y_i|\beta \sim N\left(\beta^T \phi(x_i), \sigma^2\right)$$

$$p(Y, \beta) = p(Y|\beta)\, p(\beta) = \prod_i p(y_i|\beta)\, p(\beta)$$

$$\log p(Y, \beta) = \sum_i p(y_i|\beta) + p(\beta)$$

$$\propto \sum_i -\frac{1}{2\sigma^2} \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \frac{1}{2\tau^2} \left\| \beta \right\|^2$$

$$\propto \left(-\frac{1}{2\sigma^2}\right) \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 + \frac{\sigma^2}{\tau^2} \left\| \beta \right\|^2$$

Regularization actually equivalent to putting a prior!

How to avoid overfitting?

- Do cross-validation
- Put some regularization: $\min_{\beta} \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \lambda \left\| \beta \right\|^2$
- Put a prior on the coefficients β

$$\beta \sim N\left(0, \tau^2 I\right)$$

$$y_i|\beta \sim N\left(\beta^T \phi(x_i), \sigma^2\right)$$

$$\begin{aligned} p(Y,\beta) &= p(Y|\beta)\, p(\beta) \\ &= \prod_i p(y_i|\beta)\, p(\beta) \end{aligned}$$

$$\begin{aligned} \log\ p(Y,\beta) &= \sum_i p(y_i|\beta) + p(\beta) \\ &\propto \sum_i -\frac{1}{2\sigma^2} \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \frac{1}{2\tau^2} \left\| \beta \right\|^2 \\ &\propto (-\frac{1}{2\sigma^2}) \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 + \frac{\sigma^2}{\tau^2} \left\| \beta \right\|^2 \end{aligned}$$

Regularization actually equivalent to putting a prior!

How to avoid overfitting?

- Do cross-validation
- Put some regularization: $\min_{\beta} \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \lambda \left\| \beta \right\|^2$
- Put a prior on the coefficients β

$$\beta \sim N\left(0, \tau^2 I\right)$$

$$y_i | \beta \sim N\left(\beta^T \phi(x_i), \sigma^2\right)$$

$$\begin{aligned} p(Y, \beta) &= p(Y|\beta)\, p(\beta) \\ &= \prod_i p(y_i|\beta)\, p(\beta) \end{aligned}$$

$$\begin{aligned} \log\ p(Y, \beta) &= \sum_i p(y_i|\beta) + p(\beta) \\ &\propto \sum_i -\frac{1}{2\sigma^2} \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \frac{1}{2\tau^2} \left\| \beta \right\|^2 \\ &\propto (-\frac{1}{2\sigma^2}) \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 + \frac{\sigma^2}{\tau^2} \left\| \beta \right\|^2 \end{aligned}$$

Regularization actually equivalent to putting a prior!

How to avoid overfitting?

- Do cross-validation
- Put some regularization: $\min_{\beta} \sum_i \left\|y_i - \beta^T \phi(x_i)\right\|^2 - \lambda \|\beta\|^2$
- Put a prior on the coefficients β

$$\beta \sim N\left(0, \tau^2 I\right)$$

$$y_i|\beta \sim N\left(\beta^T \phi(x_i), \sigma^2\right)$$

$$\begin{aligned} p(Y, \beta) &= p(Y|\beta)\, p(\beta) \\ &= \prod_i p(y_i|\beta)\, p(\beta) \end{aligned}$$

$$\begin{aligned} \log\ p(Y, \beta) &= \sum_i p(y_i|\beta) + p(\beta) \\ &\propto \sum_i -\frac{1}{2\sigma^2}\left\|y_i - \beta^T \phi(x_i)\right\|^2 - \frac{1}{2\tau^2}\|\beta\|^2 \\ &\propto (-\frac{1}{2\sigma^2}) \sum_i \left\|y_i - \beta^T \phi(x_i)\right\|^2 + \frac{\sigma^2}{\tau^2}\|\beta\|^2 \end{aligned}$$

Regularization actually equivalent to putting a prior!

How to avoid overfitting?

- Do cross-validation
- Put some regularization: $\min_{\beta} \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \lambda \|\beta\|^2$
- Put a prior on the coefficients β

$$\beta \sim N\left(0, \tau^2 I\right)$$

$$y_i | \beta \sim N\left(\beta^T \phi(x_i), \sigma^2\right)$$

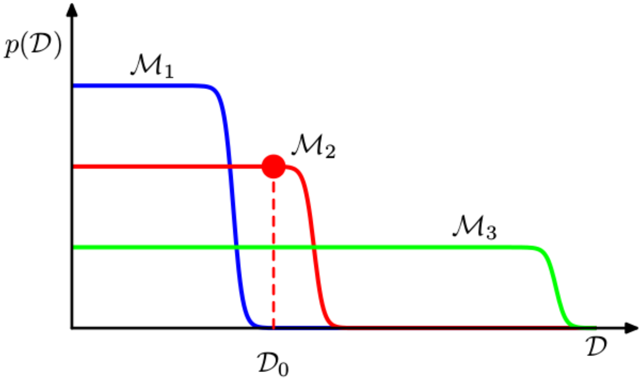
$$\begin{aligned} p(Y, \beta) &= p(Y|\beta)\, p(\beta) \\ &= \prod_i p(y_i|\beta)\, p(\beta) \end{aligned}$$

$$\begin{aligned} \log\ p(Y, \beta) &= \sum_i p(y_i|\beta) + p(\beta) \\ &\propto \sum_i -\frac{1}{2\sigma^2} \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \frac{1}{2\tau^2} \|\beta\|^2 \\ &\propto (-\frac{1}{2\sigma^2}) \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 + \frac{\sigma^2}{\tau^2} \|\beta\|^2 \end{aligned}$$

Regularization actually equivalent to putting a prior!

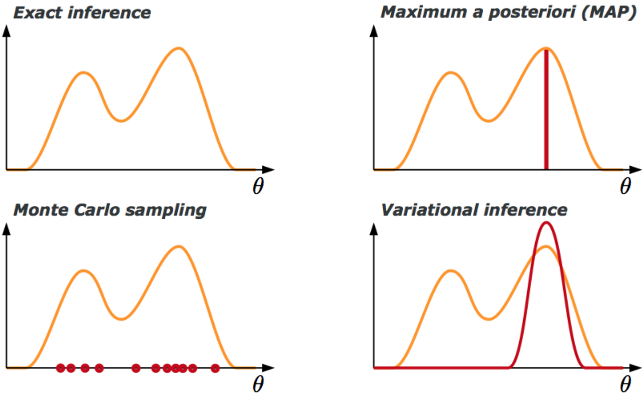
Occam´s Razor

[Bishop]



A few words about Inference

[Schmidt, MLSS, DTU]



Outline

1. Bayesian Modeling
2. Bayesian Non-parametrics
3. Biomedical Applications
4. Conclusions

Motivation

- Models are almost never correct for real world data
- How to deal with model misfit?
  - Quantify closeness to true model (ground truth)
  - Model Selection or averaging
  - Increase flexibility of your model

Motivation

- Models are almost never correct for real world data
- How to deal with model misfit?
  - Quantify closeness to true model (ground truth)
  - Model Selection or averaging
  - Increase flexibility of your model

Motivation

- Models are almost never correct for real world data
- How to deal with model misfit?
  - Quantify closeness to true model (ground truth)
  - Model Selection or averaging
  - Increase flexibility of your model

Motivation

- Models are almost never correct for real world data
- How to deal with model misfit?
  - Quantify closeness to true model (ground truth)
  - Model Selection or averaging
  - Increase flexibility of your model

Motivation

- Models are almost never correct for real world data

- How to deal with model misfit?
  - Quantify closeness to true model (ground truth)
  - Model Selection or averaging
  - Increase flexibility of your model

Motivation

- Models are almost never correct for real world data
- How to deal with model misfit?
  - Quantify closeness to true model (ground truth)
  - Model Selection or averaging
  - Increase flexibility of your model

Motivation

Avoid model selection

- Train multiple models and select/average
- Train a single model that can adapt complexity

Flexibility

- Hidden structure assumed to grow with the data
- Complexity included in the posterior

Motivation

Avoid model selection

- Train multiple models and select/average
- Train a single model that can adapt complexity

Flexibility

- Hidden structure assumed to grow with the data
- Complexity included in the posterior

Motivation

Avoid model selection

- Train multiple models and select/average
- Train a single model that can adapt complexity

Flexibility

- Hidden structure assumed to grow with the data
- Complexity included in the posterior

What is Bayesian Non-Parametrics?

- Bayesian: Combine Prior Knowledge with Data Evidence
- Non-parametric
  - really large parametric model
  - hidden structure assumed to grow with the data
  - model over infinite dimensional function or measure space
  - Notice: successful methods often nonparametric: kernel methods, SVM, deep networks, k-nearest neighbors...

What is Bayesian Non-Parametrics?

- Bayesian: Combine Prior Knowledge with Data Evidence
- Non-parametric
  - really large parametric model
  - hidden structure assumed to grow with the data
  - model over infinite dimensional function or measure space
  - Notice: successful methods often nonparametric: kernel methods, SVM, deep networks, k-nearest neighbors...

What is Bayesian Non-Parametrics?

- Bayesian: Combine Prior Knowledge with Data Evidence

- Non-parametric
  - really large parametric model
  - hidden structure assumed to grow with the data
  - model over infinite dimensional function or measure space
  - Notice: successful methods often nonparametric: kernel methods, SVM, deep networks, k-nearest neighbors...

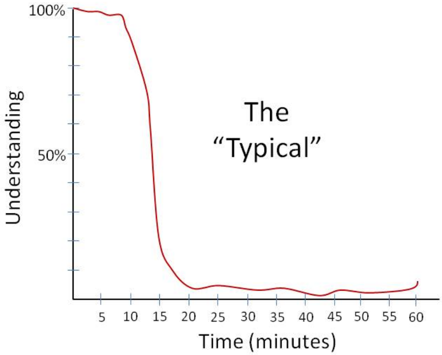
What is Bayesian Non-Parametrics?

- Bayesian: Combine Prior Knowledge with Data Evidence
- Non-parametric
  - really large parametric model
  - hidden structure assumed to grow with the data
  - model over infinite dimensional function or measure space
  - Notice: successful methods often nonparametric: kernel methods, SVM, deep networks, k-nearest neighbors...

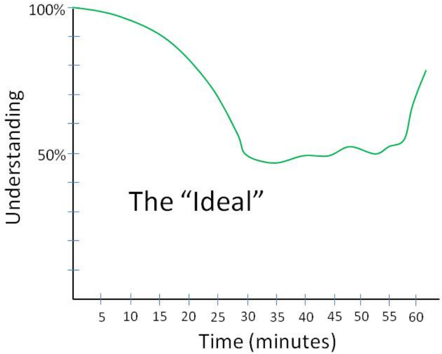
What is Bayesian Non-Parametrics?

- Bayesian: Combine Prior Knowledge with Data Evidence
- Non-parametric
  - really large parametric model
  - hidden structure assumed to grow with the data
  - model over infinite dimensional function or measure space
  - Notice: successful methods often nonparametric: kernel methods, SVM, deep networks, k-nearest neighbors...

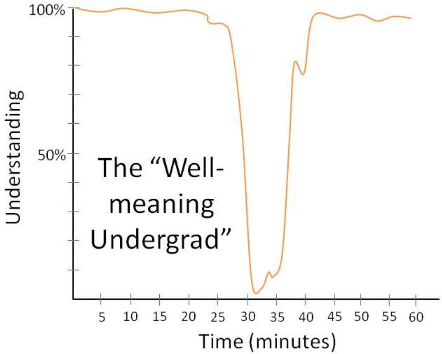
Short break


100%
50%
Understanding
The "Typical"
5
10
15
20
25
30
35
40
45
50
55
60
Time (minutes)

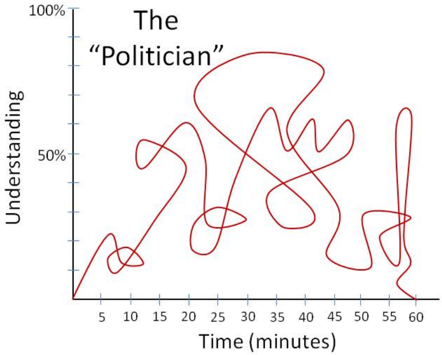
Short break


100%
50%
Understanding
The "Ideal"
5 10 15 20 25 30 35 40 45 50 55 60
Time (minutes)

Short break


100%
50%
Understanding
The "Well-meaning Undergrad"
5
10
15
20
25
30
35
40
45
50
55
60
Time (minutes)

Short break


The
"Politician"
100%
50%
Understanding
5
10
15
20
25
30
35
40
45
50
55
60
Time (minutes)

Finite Gaussian Mixture Model

$$p(x) = \sum_{k=1}^{K} \pi_k N(x; \mu_k, \Sigma_k)$$

π_k : mixture weights

ϕ_k : mixture parameters

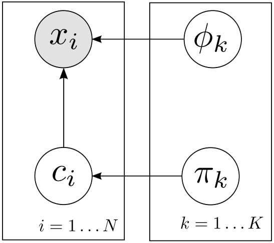
$$x_i | c_i, \phi_{c_i} \sim F(\phi_{c_i})$$

$$\phi_k \sim G_0$$

$$c_i \sim Cat(\pi_1, \dots, \pi_K)$$

$$\pi_{1:K} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

Finite Gaussian Mixture Model



$$p(x) = \sum_{k=1}^{K} \pi_k N(x; \mu_k, \Sigma_k)$$

π_k : mixture weights

ϕ_k : mixture parameters

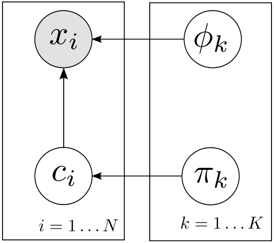
$$x_i | c_i, \phi_{c_i} \sim F(\phi_{c_i})$$

$$\phi_k \sim G_0$$

$$c_i \sim Cat(\pi_1, \ldots, \pi_K)$$

$$\pi_{1:K} \sim \mathrm{Dirichlet}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right)$$

Finite Gaussian Mixture Model



π_k : mixture weights

ϕ_k : mixture parameters

$$x_i | c_i, \phi_{c_i} \sim F\left(\phi_{c_i}\right)$$

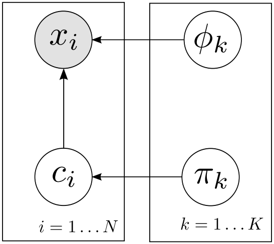
$$\phi_k \sim G_0$$

$$c_i \sim Cat\left(\pi_1, \dots, \pi_K\right)$$

$$\pi_{1:K} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

$$p\left(x\right) = \sum_{k=1}^{K} \pi_k N\left(x; \mu_k, \Sigma_k\right)$$

Finite Gaussian Mixture Model



π_k : mixture weights
ϕ_k : mixture parameters

$$x_i | c_i, \phi_{c_i} \sim F\left(\phi_{c_i}\right)$$

$$\phi_k \sim G_0$$

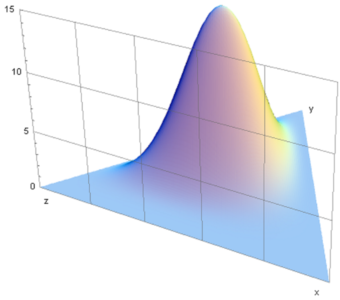
$$c_i \sim Cat\left(\pi_1, \dots, \pi_K\right)$$

$$\pi_{1:K} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

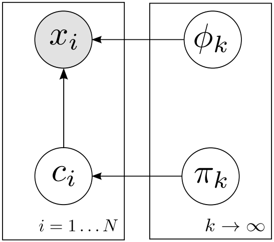
$$p\left(x\right) = \sum_{k=1}^{K} \pi_k N\left(x; \mu_k, \Sigma_k\right)$$

Dirichlet Distribution

$$f\left(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K\right) = \frac{1}{B\left(\alpha\right)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$



Infinite Gaussian Mixture Model



π_k : mixture weights
ϕ_k : mixture parameters

$$x_i|\theta_i \sim F(\theta_i)$$

$$\theta_i|G \sim G$$

$$G \sim \mathrm{DP}(\alpha, \mathrm{G}_0)$$

$$p(x) = \sum_{k=1}^{K^+} \pi_k N(x; \mu_k, \Sigma_k)$$

Dirichlet Process

Dirichlet Process

- stochastic process whose realization is a probability distribution

$$G \sim \mathrm{DP}(\alpha, \mathrm{H})$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

H : base measure
α : concentration parameter

Dirichlet Process

Dirichlet Process

- stochastic process whose realization is a probability distribution

$$G \sim \mathrm{DP}\left(\alpha, \mathrm{H}\right)$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

H : base measure

α : concentration parameter

Dirichlet Process

Dirichlet Process

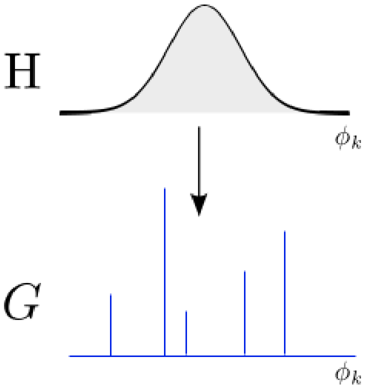
- stochastic process whose realization is a probability distribution

$$G \sim \mathrm{DP}(\alpha, \mathrm{H})$$

$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

H : base measure

α : concentration parameter



Dirichlet Process

- What about ϕ_k? $\phi_k \sim G_0$
- What about π_k? $\Longrightarrow$
  Stick Breaking Process

$$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$$

$$v_k \sim \mathrm{Beta}(1, \alpha)$$

- What about c_i? $\Longrightarrow$ Chinese Restaurant Process

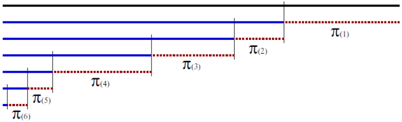
Dirichlet Process

- What about ϕ_k? $\phi_k \sim G_0$

- What about π_k? $\Longrightarrow$ Stick Breaking Process

$$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$$

$$v_k \sim \mathrm{Beta}(1, \alpha)$$



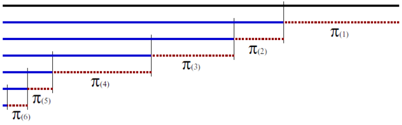
- What about c_i? $\Longrightarrow$ Chinese Restaurant Process

Dirichlet Process

- What about ϕ_k? $\phi_k \sim G_0$
- What about π_k? $\Longrightarrow$ Stick Breaking Process

$$\pi_k = v_k \prod_{i=1}^{k-1} (1 - v_i)$$

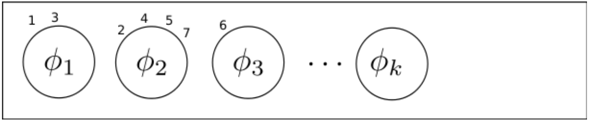
$$v_k \sim \mathrm{Beta}(1, \alpha)$$



- What about c_i? $\Longrightarrow$ Chinese Restaurant Process

Chinese Restaurant Process

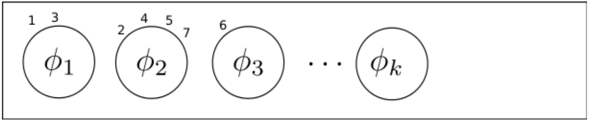
- Imagine a restaurant with countable infinitely many tables



- First customer always chooses the first table
- The i^{th} customer chooses:
  - unoccupied table with probability: $\alpha/(i-1+\alpha)$
  - occupied table with probability: $m_k/(i-1+\alpha)$ where m_k is number of people sitting at that table.

Chinese Restaurant Process

- Imagine a restaurant with countable infinitely many tables



- First customer always chooses the first table
- The i^{th} customer chooses:
  - unoccupied table with probability: $\alpha / (i - 1 + \alpha)$
  - occupied table with probability: $m_k / (i - 1 + \alpha)$ where m_k is number of people sitting at that table.

Latent Factor Model

$$y_n = Gx_n + \epsilon_n$$

Assumptions lead to different models

- Factor Analysis
- Principal Component Analysis
- Independent Component Analysis
- ...

Latent Factor Model

$$y_n = Gx_n + \epsilon_n$$

Assumptions lead to different models

- Factor Analysis
- Principal Component Analysis
- Independent Component Analysis
- ...

Indian Buffet Process

$$y_n = Zx_n + \epsilon_n$$

- IBP places a prior distribution over binary matrices where the number of columns (latent features) $K \to \infty$.
- Matrix $\boldsymbol{Z}_{N \times K} \sim \mathrm{IBP}(\alpha)$ with α : concentration parameter.
- Each element $z_{nk} \in \{0, 1\}$ indicates whether the k^{th} feature contributes to the n^{th} data point.
- For finite number of data points N, number of non-zero columns K^+ is finite.

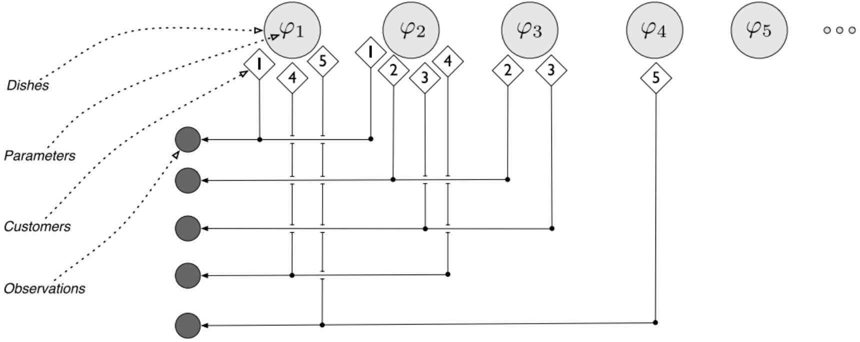
Indian Buffet Process

$$y_n = Zx_n + \epsilon_n$$

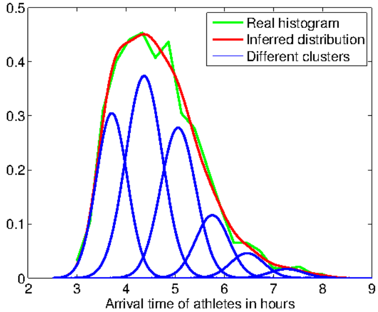
- IBP places a prior distribution over binary matrices where the number of columns (latent features) $K \to \infty$.
- Matrix $\boldsymbol{Z}_{N \times K} \sim \mathrm{IBP}(\alpha)$ with α : concentration parameter.
- Each element $z_{nk} \in \{0, 1\}$ indicates whether the k^{th} feature contributes to the n^{th} data point.
- For finite number of data points N, number of non-zero columns K^+ is finite.

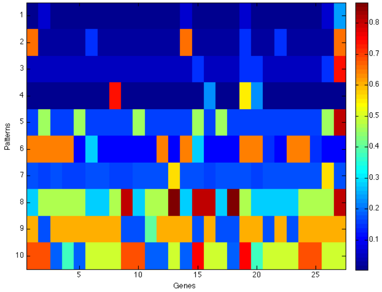
Indian Buffet Process

Culinary Metaphor [Gershman & Blei, 2012]



Bayesian Nonparametrics in Action





Outline

1. Bayesian Modeling
2. Bayesian Non-parametrics
3. Biomedical Applications
4. Conclusions

Linear Mixed Model

$$y_{ij} = x_{ij}\beta + z_{ij} b_i + \epsilon_{ij}, \;\; \epsilon \sim N\left(0, \sigma^2\right)$$

$$b_i \sim P$$

$$P \sim DP\left(\alpha,\; P_0\right)$$

- P = random effects distribution

[Bush & MacEachern (1996), Müller & Rosner (1997), Kleinman & Ibrahim (1998), Ishwaran & Takahara (2002),...]

Functional Data Analysis

$$y_{ij} \sim N\left(f_i\left(t_{ij}\right), \sigma^2\right)$$

$$f_i\left(t\right) = \sum_{h=1}^{K^+} \beta_{ih}\phi_h\left(t\right)$$

$$\beta_i \sim P$$

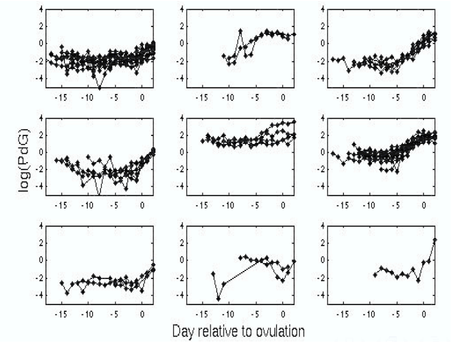
$\{\phi_h\}_{h=1}^{K^+}$: basis functions

- β_{ih}: subject-specific coefficients, all coefficients assigned a joint distribution

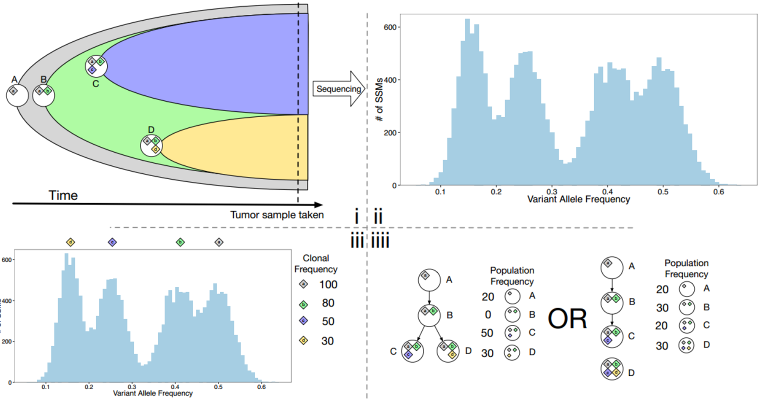
Clustering Hormone Curves

[Ray & Mallick (2006)]

- Progesterone measured across menstrual cycle (172 women)
- One approach: multivariate spline model with DP on distribution of basis coefficients



Evolution of cancer subpopulation



[Deshwar, 2014]

Outline

1. Bayesian Modeling
2. Bayesian Non-parametrics
3. Biomedical Applications
4. Conclusions

Summary of the talk

1. Bayesian Thinking
2. Basics on Bayesian Non-Parametrics
    - Adapts complexity
    - Flexible model
3. Some Biomedical Applications

Quote from Z. Ghahramani

- Why Bayesian?
    - Simplicity (of the Framework)
- Why Non-Parametrics?
    - Complexity (of the Real World)

Summary of the talk

1. Bayesian Thinking
2. Basics on Bayesian Non-Parametrics
    - Adapts complexity
    - Flexible model
3. Some Biomedical Applications

Quote from Z. Ghahramani

- Why Bayesian?
    - Simplicity (of the Framework)
- Why Non-Parametrics?
    - Complexity (of the Real World)

Summary: Bayesian Non-Parametrics

Advantages

- good predictive performance
- flexible
- robust to overfitting
- model-based
- interpretability
    - borrowing information
    - dimensionality reduction

Limitations

- Scalability
- Expert knowledge into priors difficult
- Some inconsistencies

Summary: Bayesian Non-Parametrics

Advantages

- good predictive performance
- flexible
- robust to overfitting
- model-based
- interpretability
  - borrowing information
  - dimensionality reduction

Limitations

- Scalability
- Expert knowledge into priors difficult
- Some inconsistencies

Summary: Bayesian Non-Parametrics

Advantages

- good predictive performance
- flexible
- robust to overfitting
- model-based
- interpretability
  - borrowing information
  - dimensionality reduction

Limitations

- Scalability
- Expert knowledge into priors difficult
- Some inconsistencies

Software Available

[Gershman & Blei, 2012]

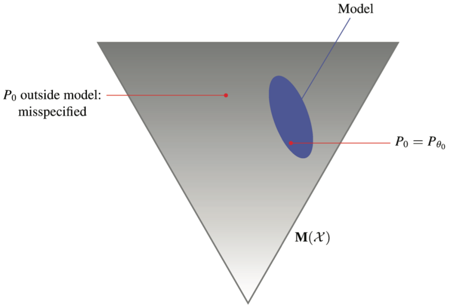
Software packages implementing various Bayesian nonparametric models.

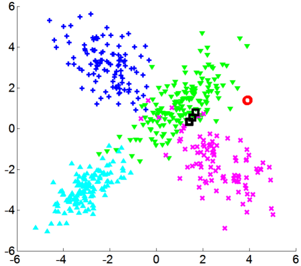
Model	Algorithm	Language	Author	Link
CRP mixture model	MCMC	Matlab	Jacob Eisenstein	http://people.csail.mit.edu/jacobe/software.html
CRP mixture model	MCMC	R	Matthew Shotwell	http://cran.r-project.org/web/packages/profdpm/index.html
CRP mixture model	Variational	Matlab	Kenichi Kurihara	http://sites.google.com/site/kenichikurihara/academic-software
IBP latent factor model	MCMC	Matlab	David Knowles	http://mlg.eng.cam.ac.uk/dave
IBP latent factor model	Variational	Matlab	Finale Doshi-Velez	http://people.csail.mit.edu/finale/new-wiki/doku.php?id=publications_posters_presentations_code

Discussion

Large Support despite of Inconsistency Issues

Miller, Harrison, *Inconsistency of Pitman-Yor Process Mixtures for the Number of Components*.





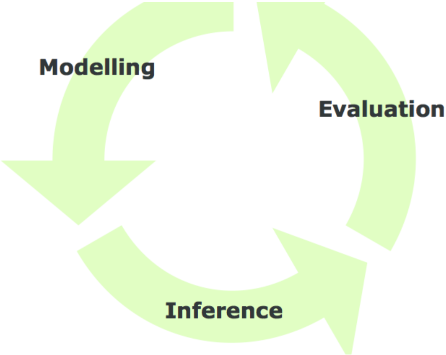
[Peter Orbanz & Yee Whye Teh, MLSS 2011]

Discussion

Automatic Vs Tailored Models: What do we want?

The End

Statistical Learning Cycle [Gelman 2004]



Thank you!

Looking forward to your questions...