Bayesian Non-parametrics and Variational Inference

A Brief Introduction

Melanie F. Pradier

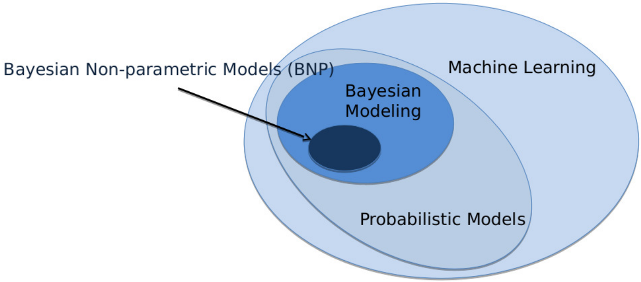
Universidad Carlos III in Madrid

09 July 2015

In this Talk...

Machine Learning

"Machine learning explores the construction and study of algorithms that can learn from data".



Outline

1. BNP framework
2. Basic Models
3. Some Applications
4. Overview of Variational Inference

Outline

1. BNP framework
2. Basic Models
3. Some Applications
4. Overview of Variational Inference

Bayesian Modeling

- 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: $p(X)$

- Combine Prior Knowledge with Data Evidence

Advantage 1: Whole Information in the Posterior

Estimators

- ML estimator

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

- MAP estimator

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

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

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

Advantage 1: Whole Information in the Posterior

Estimators

- ML estimator

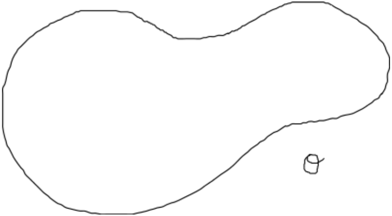
$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p\left(X|\theta\right)$$

- MAP estimator

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

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

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



Advantage 1: Whole Information in the Posterior

Estimators

- ML estimator

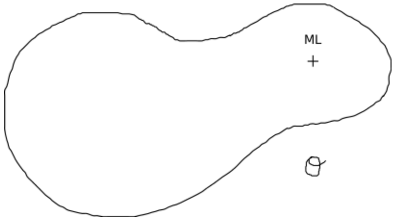
$$\hat{\theta}_{ML} = \operatorname*{argmax}_{\theta} p\left(X|\theta\right)$$

- MAP estimator

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

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

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



Advantage 1: Whole Information in the Posterior

Estimators

- ML estimator

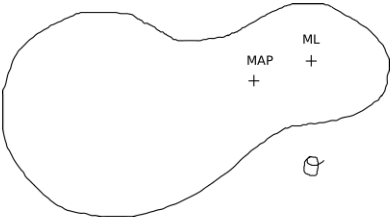
$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p(X|\theta)$$

- MAP estimator

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

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

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



Advantage 1: Whole Information in the Posterior

Estimators

- ML estimator

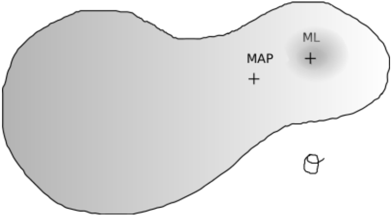
$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p\left(X|\theta\right)$$

- MAP estimator

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

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

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



Advantage 1: Whole Information in the Posterior

Estimators

- ML estimator

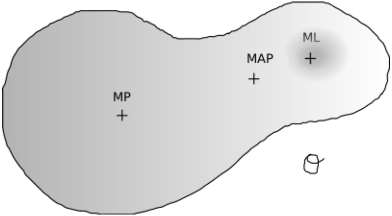
$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, p\left(X|\theta\right)$$

- MAP estimator

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

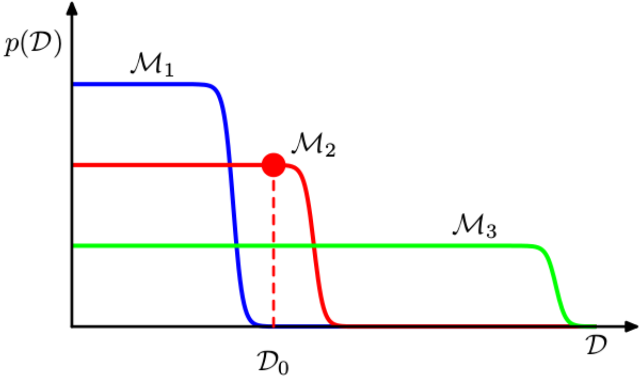
- **Posterior distribution** $\rightarrow$ Mean Posterior estimator (MP)

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



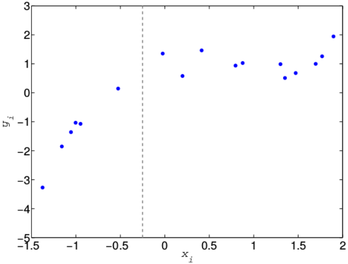
Advantage 2: Model Selection

- Occam´s Razor [Bishop]



Advantage 2: Model Selection

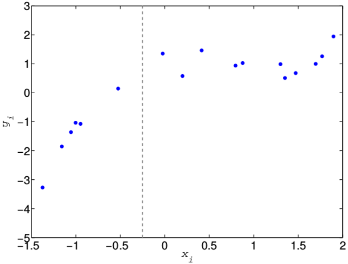
Regression Problem [Bishop]



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

Advantage 2: Model Selection

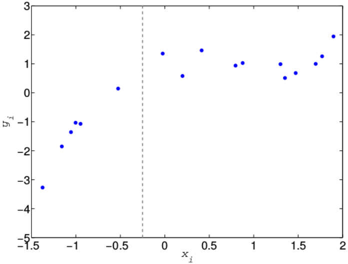
Regression Problem [Bishop]

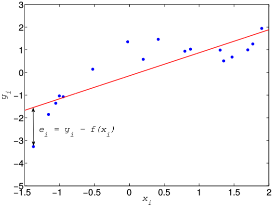


$$y_i = \beta^T \phi(x_i) + \epsilon$$

Advantage 2: Model Selection

Regression Problem [Bishop]

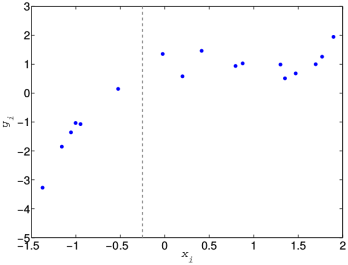




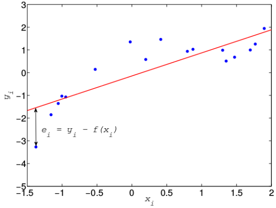
$$y_i = \beta^T \phi(x_i) + \epsilon$$

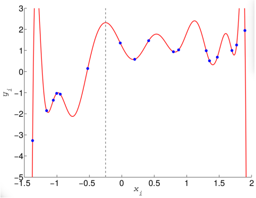
Advantage 2: Model Selection

Regression Problem [Bishop]



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





Bayesian Perspective of Regularization

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

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

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

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

Regularization actually equivalent to putting a prior!

Bayesian Perspective of Regularization

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

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

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

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

Regularization actually equivalent to putting a prior!

Bayesian Perspective of Regularization

1. Do cross-validation
2. Put some regularization: $\min_{\beta} \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \lambda \left\| \beta \right\|^2$
3. 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!

Bayesian Perspective of Regularization

1. Do cross-validation
2. Put some regularization: $\min_{\beta} \sum_i \left\| y_i - \beta^T \phi(x_i) \right\|^2 - \lambda \|\beta\|^2$
3. 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!

What does Non-Parametric mean?

- 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, k-nearest neighbors...

What does Non-Parametric mean?

- 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, k-nearest neighbors...

What does Non-Parametric mean?

- 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, k-nearest neighbors...

What does Non-Parametric mean?

- 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, k-nearest neighbors...

What does Non-Parametric mean?

- 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, k-nearest neighbors...

Advantages for Bayesian Non-Parametrics

Automatic Model Selection

- Train multiple models and select/average (Bayesian approach)
- Train a single model that can adapt complexity

More Flexibility

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

Advantages for Bayesian Non-Parametrics

Automatic Model Selection

- Train multiple models and select/average (Bayesian approach)
- Train a single model that can adapt complexity

More Flexibility

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

Advantages for Bayesian Non-Parametrics

Automatic Model Selection

- Train multiple models and select/average (Bayesian approach)
- Train a single model that can adapt complexity

More Flexibility

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

Outline

1. BNP framework
2. Basic Models
3. Some Applications
4. Overview of Variational Inference

Mixture Model

$$\pi_k : \text{ mixture weights}$$
$$\phi_k : \text{ mixture parameters}$$

$$p(x) = \sum_{k=1}^{K} \pi_k F(x; \phi_k)$$

e.g.

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

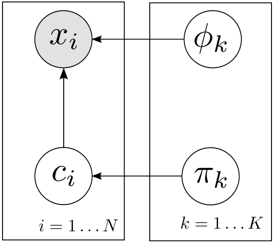
$$x_i | c_i, \phi \sim F(\phi_{c_i})$$

$$\phi_k \sim G_0$$

$$c_i \sim \text{Discrete}(\pi_1, \ldots, \pi_K)$$

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

Mixture Model



π_k : mixture weights
ϕ_k : mixture parameters

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

$$\phi_k \sim G_0$$

$$c_i \sim \text{Discrete}(\pi_1, \dots, \pi_K)$$

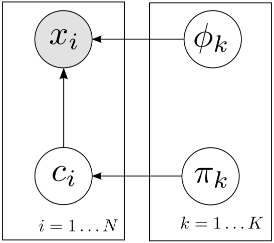
$$\pi_{1:K} \sim \text{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

$$p(x) = \sum_{k=1}^{K} \pi_k F(x; \phi_k)$$

e.g.

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

Mixture Model



π_k : mixture weights
ϕ_k : mixture parameters

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

$$\phi_k \sim G_0$$

$$c_i \sim \mathrm{Discrete}\left(\pi_1, \dots, \pi_K\right)$$

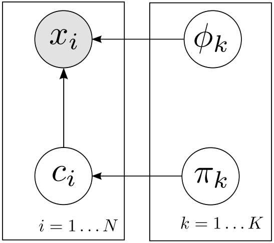
$$\pi_{1:K} \sim \mathrm{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

$$p(x) = \sum_{k=1}^{K} \pi_k F\left(x; \phi_k\right)$$

e.g.

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

Mixture Model



π_k : mixture weights
ϕ_k : mixture parameters

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

$$\phi_k \sim G_0$$

$$c_i \sim \mathrm{Discrete}(\pi_1, \dots, \pi_K)$$

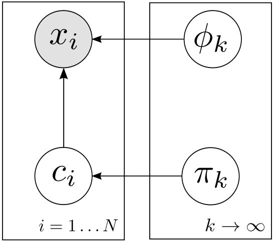
$$\pi_{1:K} \sim \mathrm{Dirichlet}\left(\frac{\alpha}{K}, \dots, \frac{\alpha}{K}\right)$$

$$p(x) = \sum_{k=1}^{K} \pi_k F(x; \phi_k)$$

e.g.

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

Infinite Mixture Model



π_k : mixture weights

ϕ_k : mixture parameters

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

$$\theta_i | G \sim G$$

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

$$p(x) = \sum_{k=1}^{K^+} \pi_k F(x; \phi_k)$$

e.g.

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

Latent Feature Model

$$x_n = Zb_n + \epsilon_n \qquad \text{where typically} \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

Assumptions lead to different models

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

Latent Feature Model

$$x_n = Zb_n + \epsilon_n \qquad \text{where typically} \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

Assumptions lead to different models

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

Infinite Latent Feature Model

$$x_n = Zb_n + \epsilon_n \qquad \text{where typically} \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

- Underlying process: Indian Buffet Process
- IBP places a prior distribution over binary matrices where the number of columns (latent features) $K \to \infty$.
- Matrix $Z_{N \times K} \sim \mathrm{IBP}\left(\alpha\right)$ 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.

Infinite Latent Feature Model

$$x_n = Zb_n + \epsilon_n \qquad \text{where typically} \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$$

- Underlying process: Indian Buffet Process
- IBP places a prior distribution over binary matrices where the number of columns (latent features) $K \to \infty$.
- Matrix $\boldsymbol{Z}_{N \times K} \sim \mathrm{IBP}\left(\alpha\right)$ 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.

Non-parametric Regression Model

$$y_i = f(x_i) + \epsilon$$

We use a Gaussian Process as prior, define as

$$f \sim GP\left(\mu(x), K(x, x')\right)$$

where

- $\mu(x)$: mean function
- $K(x, x')$: covariance matrix.

Summary: the BNP framework

Tools: Probabilistic Models -> BNPs

- A priori knowledge
- Model Selection through data
- Interpretability
- Generalization

Summary: the BNP framework

Tools: Probabilistic Models -> BNPs

- A priori knowledge
- Model Selection through data
- Interpretability
- Generalization

Summary: the BNP framework

Tools: Probabilistic Models -> BNPs

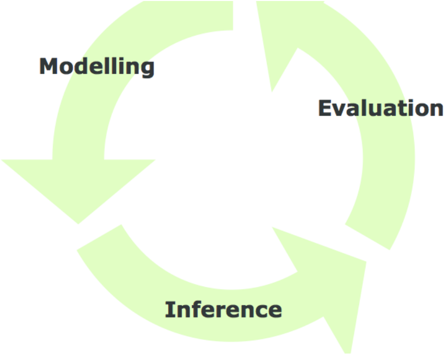
- A priori knowledge
- Model Selection through data
- Interpretability
- Generalization

Objectives

- Data Exploration
- Regression
- Classification

Summary: the BNP framework

- How? Statistical Learning Cycle [Gelman 2004]

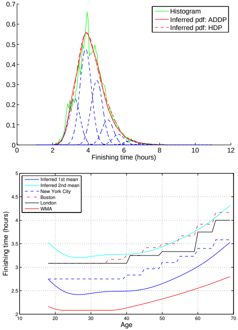


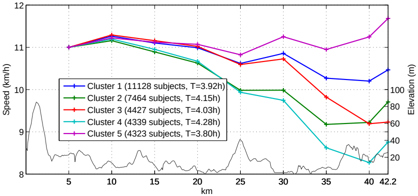
Outline

1. BNP framework
2. Basic Models
3. Some Applications
4. Overview of Variational Inference

Marathon Modeling

M. F. Pradier, F. J. R. Ruiz and F. Perez-Cruz. **Prior Design for Dependent Dirichlet Processes: An Application to Marathon Modeling.** Submitted to PlosONE. 2015.

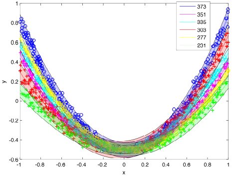


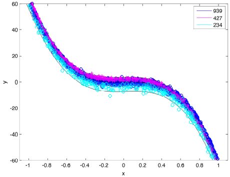


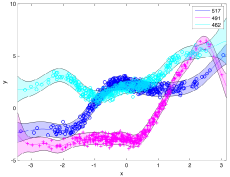
M. F. Pradier, P. G. Moreno, F. J.R. Ruiz, I. Valera, H. Mollina-Bulla and F. Perez-Cruz, **Map/Reduce Uncollapsed Gibbs Sampling for Bayesian Non Parametric Models.** Workshop in Software Engineering for Machine Learning (NIPS). 2014.

Non-stationary Non-linear Regression

M. F. Pradier and F. Perez-Cruz. **Infinite Mixture of Global Gaussian Processes.** Submitted to Neural Information Processing Systems (NIPS). 2015.



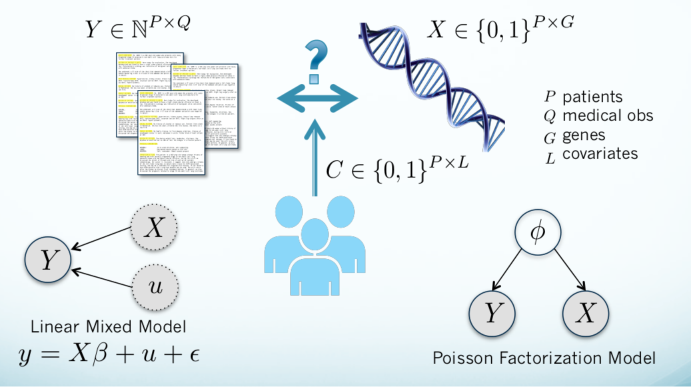




		Toy dBs			Real dBs		
		Heteroscedasticity	*Non-Gaussianity*	*Multimodality*	*Concrete*	*Marathon*	*RSSI*
PLLH	sGP	-0.0217	-3.4920	-3.3030	-0.5855	-1.6373	0.2033
	IMoE	0.7017	-2.1248	-2.1604	1.9452	-1.6308	**0.9943**
	IMoGGP	**0.9008**	**-1.2575**	**-2.1237**	**2.3587**	**-1.5723**	0.9846

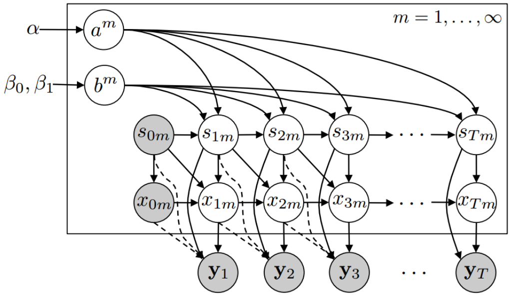
Genetic Associations in Cancer

M. F. Pradier, F. Perez-Cruz and G. Rätsch. **Sparse Poisson Factorization Model for Genetic Associations with Clinical Features in Cancer.** Working paper. 2015.



Source Separation Problem

I. Valera, F. J. R. Ruiz, L. Svensson and F. Perez-Cruz. **Infinite Factorial Dynamical Model.** Submitted to Neural Information Processing Systems (NIPS). 2015.



Scenarios

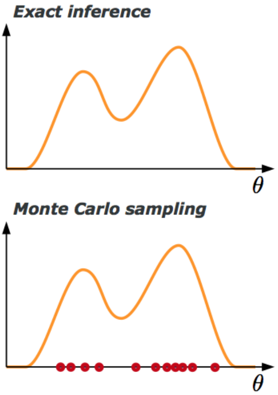
- Multitarget Tracking
- Power Dissagregation
- Multiuser Detection

Outline

1. BNP framework
2. Basic Models
3. Some Applications
4. Overview of Variational Inference

Inference in BNPs

[Schmidt, MLSS, DTU]



Scenario

- $X = x_{1:n}$: observations
- $Z = z_{1:n}$: hidden variables
- We want to compute posterior

$$p(Z|X) = \frac{p(Z,X)}{\int_Z p(Z,X)\, dZ}$$

- Computation Intractable!

Example: Gaussian Mixture Model

- $\mu_k \sim \mathcal{N}\left(0, \tau^2\right)$ for $k = 1 \ldots K$
- for $i = 1 \ldots n$:
    - $c_i \sim \mathrm{Discrete}\left(\boldsymbol{\pi}\right)$
    - $x_i \sim \mathcal{N}\left(\mu_{z_i}, \sigma^2\right)$
- In this case, $Z = \{\mu_{1:K}, c_{1:N}\}$ and

$$p\left(Z|X\right) = \frac{\prod_{k=1}^{K} p\left(\mu_k\right) \prod_{i=1}^{n} p\left(c_i\right) p\left(x_i|c_i, \mu_{1:K}\right)}{\int_{\mu_{1:K}} \sum_{c_{1:n}} \prod_{k=1}^{K} p\left(\mu_k\right) \prod_{i=1}^{n} p\left(c_i\right) p\left(x_i|c_i, \mu_{1:K}\right) d\mu_{1:K} dc_{1:n}}$$

Variational Inference

1. Main Idea: Pick a simple/tractable family of distributions

$$q\left(Z|\boldsymbol{\nu}\right) \qquad \text{where } \boldsymbol{\nu}: \text{ variational parameters}$$

2. Find $\boldsymbol{\nu}$ such that q is close to posterior p

$$\min_{\boldsymbol{\nu}} KL\left(q||p\right) = \mathrm{E}_q\left[\log \frac{q\left(Z\right)}{p\left(Z|X\right)}\right]$$

Using Bayes Rule,

$$KL\left(q||p\right) = \mathrm{E}_q\left[\log q\left(Z\right)\right] - \mathrm{E}_q\left[\log p\left(Z,X\right)\right] + \log p\left(X\right)$$

Variational Inference

- Inference tackled as an optimization problem
- Generalization of EM algorithm
- If we reverse measure, i.e. $KL\left(p||q\right)$, we get Expectation Propagation

Mean Field Approximation

- Which family $q\left(Z|\boldsymbol{\nu}\right)$ should we pick?
- Mean Field Approx assumes independence:

$$q\left(Z|\boldsymbol{\nu}\right) = \prod_{j=1}^{M} q\left(z_j|\nu_j\right)$$

- Allows for an efficient coordinate ascent optimization

When to use Variational Inference?

Advantages

- Fast, easy to compute
- Scalable

Drawbacks

- Hard to derive (in BNP, infinite expectations)
- Approximation to the posterior
- Convergence to local minimum

When to use Variational Inference?

Advantages

- Fast, easy to compute
- Scalable

Drawbacks

- Hard to derive (in BNP, infinite expectations)
- Approximation to the posterior
- Convergence to local minimum

When to use Variational Inference?

Advantages

- Fast, easy to compute
- Scalable

Drawbacks

- Hard to derive (in BNP, infinite expectations)
- Approximation to the posterior
- Convergence to local minimum

Conclusion: In this Talk...

1. Bayesian Thinking
2. Bayesian Non-Parametrics
    - Adapts complexity
    - Flexible model
3. Overview of Variational Inference

Quote from Z. Ghahramani

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

Conclusion: In this Talk...

1. Bayesian Thinking
2. Bayesian Non-Parametrics
    - Adapts complexity
    - Flexible model
3. Overview of Variational Inference

Quote from Z. Ghahramani

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

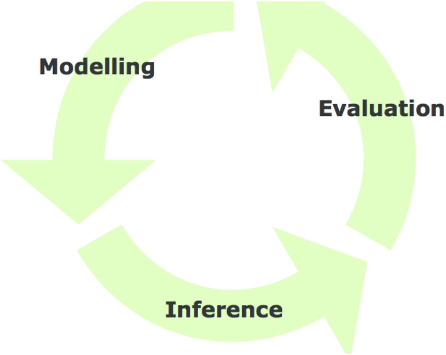
Sources and References

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.

The End

Statistical Learning Cycle [Gelman 2004]



Thank you!

Looking forward to your questions. . .