Are you exploiting your assumptions?

Towards expressive priors for biomarker discovery and functional prediction

Melanie F. Pradier

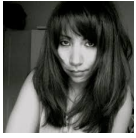
Harvard University

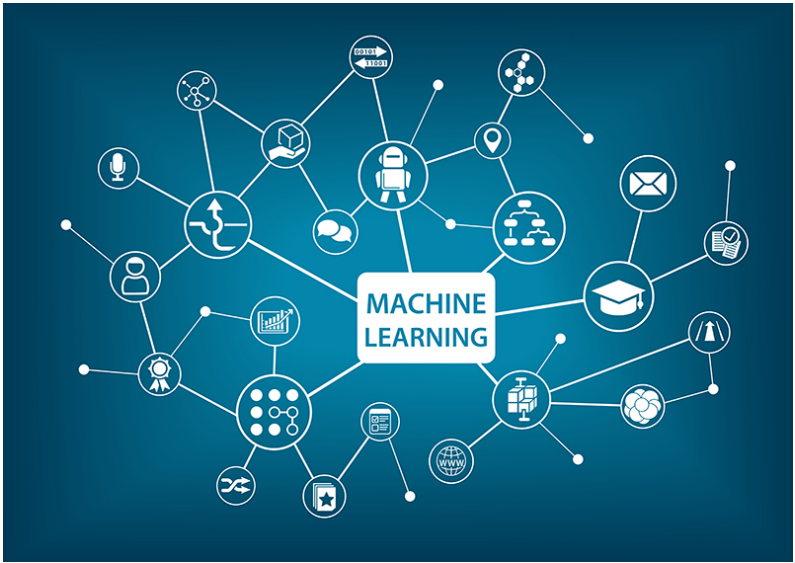
February 18th, 2020

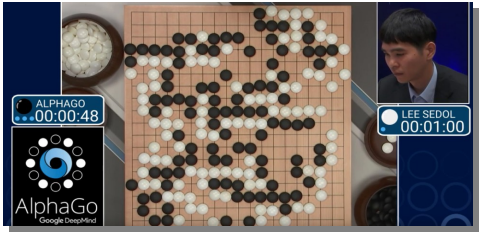
Are you exploiting your assumptions?

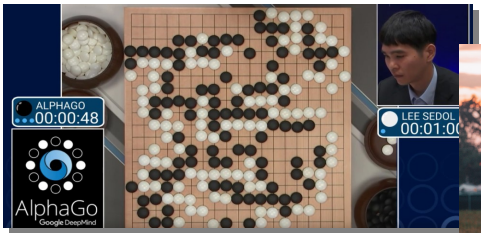
Towards expressive priors for biomarker discovery and functional prediction

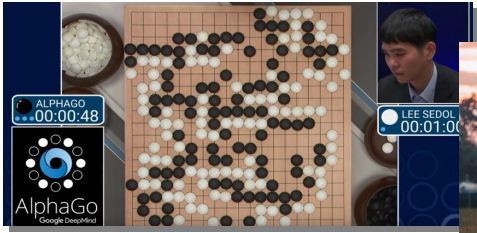
Research in collaboration with...

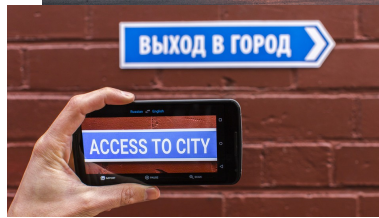



MACHINE
LEARNING

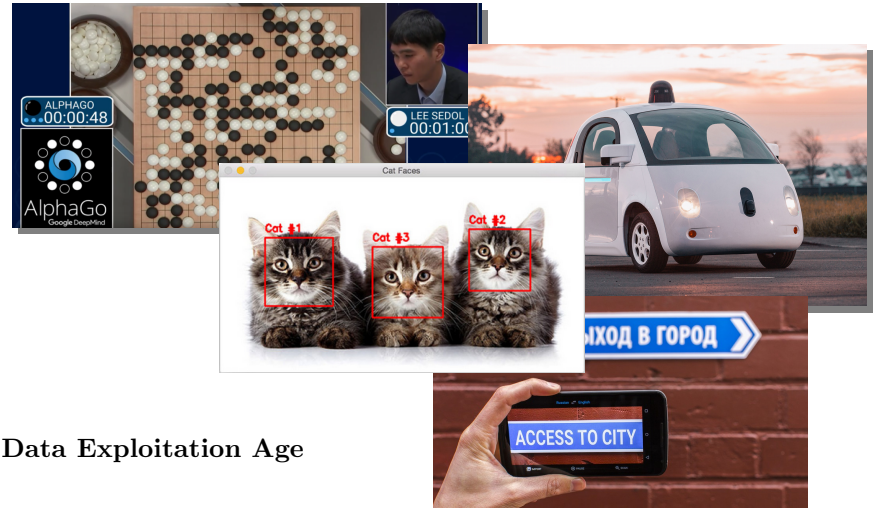

ALPHAGO
00:00:48
AlphaGo
Google DeepMind
LEE SEDOL
00:01:00


ALPHAGO
00:00:48
AlphaGo
Google DeepMind
LEE SEDOL
00:01:0

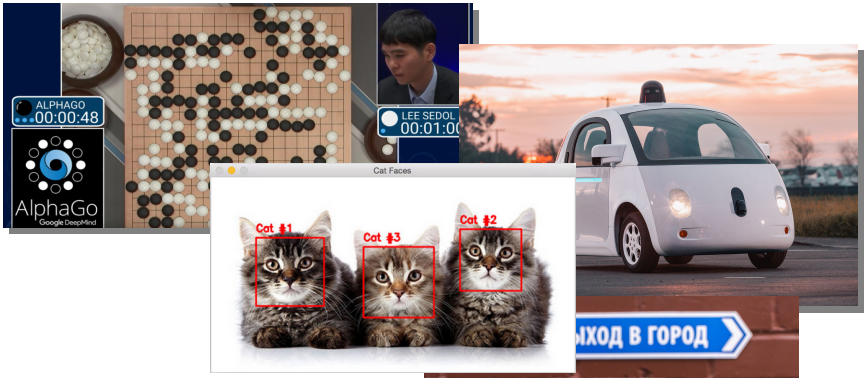

ALPHAGO
00:00:48
LEE SEDOL
AlphaGo
Google DeepMind


ВЫХОД В ГОРОД
ACCESS TO CITY


ALPHAGO
00:00:48
LEE SEDOL
AlphaGo
Google DeepMind
Cat Faces
Cat #1
Cat #3
Cat #2
ВЫХОД В ГОРОД
ACCESS TO CITY



Data Exploitation Age



Data Exploitation Age

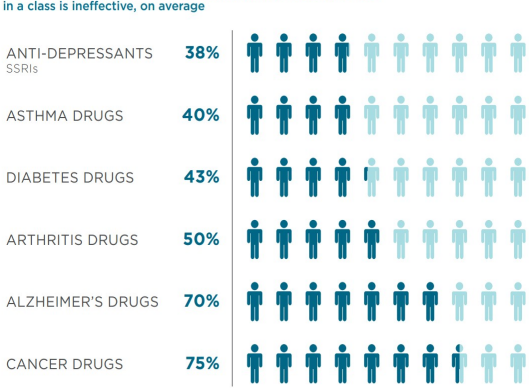
$\cdots$ but are we making the utmost out of data?

Are we making the utmost out of data?

An example: personalized medicine

Are we making the utmost out of data?

An example: personalized medicine

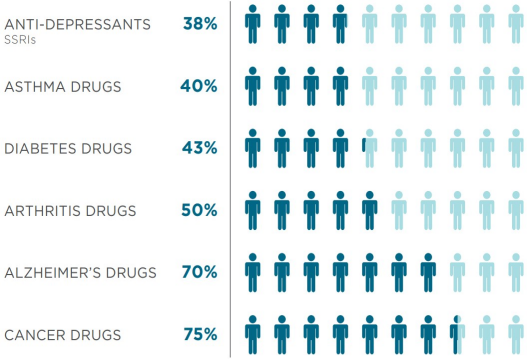


Source: Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, "Clinical Trends in Molecular Medicine," Volume 7, Issue 5, 1 May 2001, pages 201-204.

Are we making the utmost out of data?

An example: personalized medicine

Percentage of the patient population for which a particular drug in a class is ineffective, on average

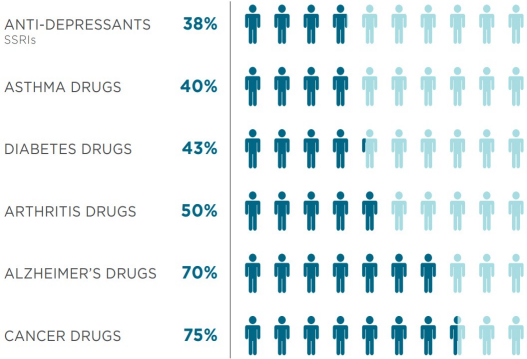


Source: Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, "Clinical Trends in Molecular Medicine," Volume 7, Issue 5, 1 May 2001, pages 201-204.

Are we making the utmost out of data?

An example: personalized medicine

Percentage of the patient population for which a particular drug in a class is ineffective, on average



Source: Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, "Clinical Trends in Molecular Medicine," Volume 7, Issue 5, 1 May 2001, pages 201-204.

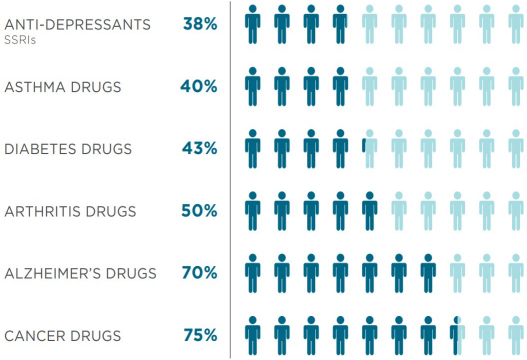
Challenges

- Complexity

Are we making the utmost out of data?

An example: personalized medicine

Percentage of the patient population for which a particular drug in a class is ineffective, on average



Source: Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, "Clinical Trends in Molecular Medicine," Volume 7, Issue 5, 1 May 2001, pages 201-204.

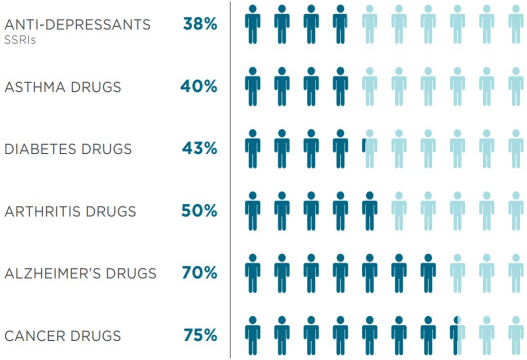
Challenges

- Complexity
- Noise, missing data

Are we making the utmost out of data?

An example: personalized medicine

Percentage of the patient population for which a particular drug in a class is ineffective, on average



Source: Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, "Clinical Trends in Molecular Medicine," Volume 7, Issue 5, 1 May 2001, pages 201-204.

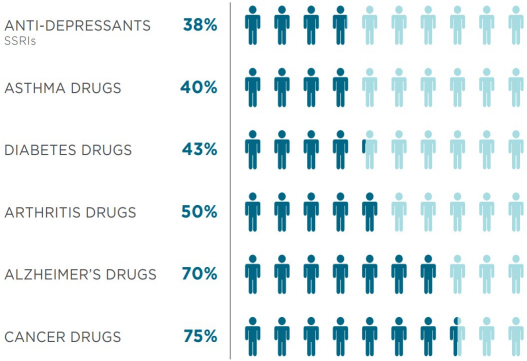
Challenges

- Complexity
- Noise, missing data
- *Small data within big data*
- ...

Are we making the utmost out of data?

An example: personalized medicine

Percentage of the patient population for which a particular drug in a class is ineffective, on average



Source: Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, "Clinical Trends in Molecular Medicine," Volume 7, Issue 5, 1 May 2001, pages 201-204.

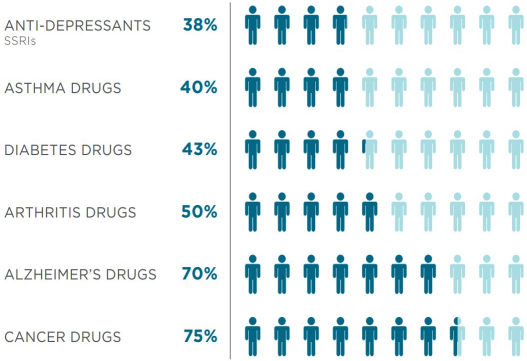
Challenges

- Complexity
- Noise, missing data
- *Small data within big data*
- ...
- Need to understand
  → data exploration

Are we making the utmost out of data?

An example: personalized medicine

Percentage of the patient population for which a particular drug in a class is ineffective, on average



Source: Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, "Clinical Trends in Molecular Medicine," Volume 7, Issue 5, 1 May 2001, pages 201-204.

Challenges

- Complexity
- Noise, missing data
- *Small data within big data*
- ...
- Need to understand
  → data exploration

How can ML systems help experts "understand" data?

Expressivity-Interpretability Loop

Interpretability

- "ability to explain or to present in understandable terms to a human" (Doshi-Velez and Kim, 2017)
- 2018 EU General Data Protection Regulation (Goodman et.al. 2016)

Expressivity-Interpretability Loop

Interpretability

- "ability to explain or to present in understandable terms to a human" (Doshi-Velez and Kim, 2017)
- 2018 EU General Data Protection Regulation (Goodman et.al. 2016)

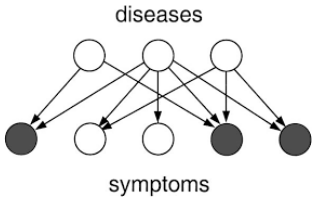
Expressivity

- ability to encode assumptions/desiderata into the model

In this talk, expressivity/interpretability via probabilistic graphical models

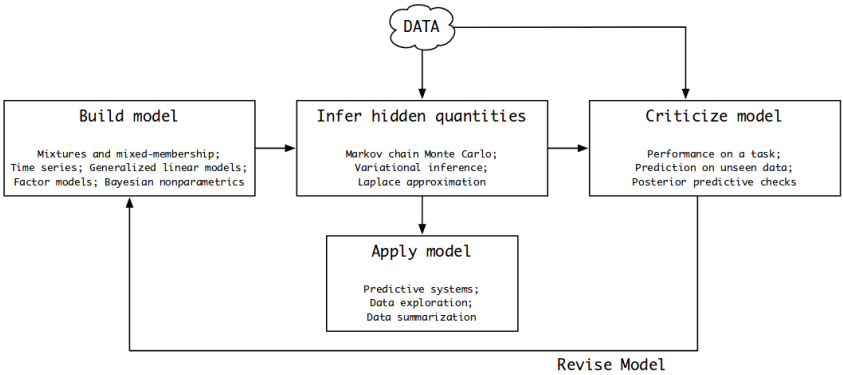
Why probabilistic graphical models?

- Generative model ≡ unsupervised approach, goal is to model p(Data)
- Graphical model for multidisciplinary research
- Assumptions and desiderata through *latent variables*



Why probabilistic graphical models?

The "Box's loop" (Blei, 2014)



Outline

- Overview
- Goal I: Biomarker discovery
- Goal II: Functional prediction
- Wrap-up

Our Focus: Biomarker discovery

Def: "any variable that can be used as an indicator of a particular disease state".

Our Focus: Biomarker discovery

Def: "any variable that can be used as an indicator of a particular disease state".

Biomarkers are used everywhere!!

Our Focus: Biomarker discovery

Def: "any variable that can be used as an indicator of a particular disease state".

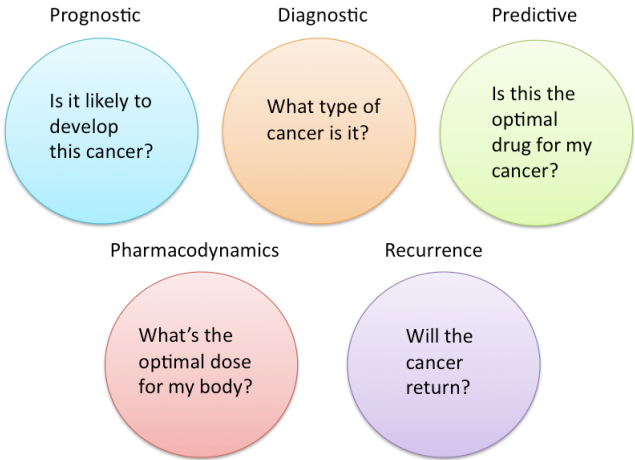
Biomarkers are used everywhere!!

Some examples

- Prostate-specific antigen (PSA) to diagnose prostate cancer
- Estrogen / progesterone to predict sensitivity to endocrine therapy in breast cancer
- KRAS mutation to predict resistance to EGFr antibody treatment

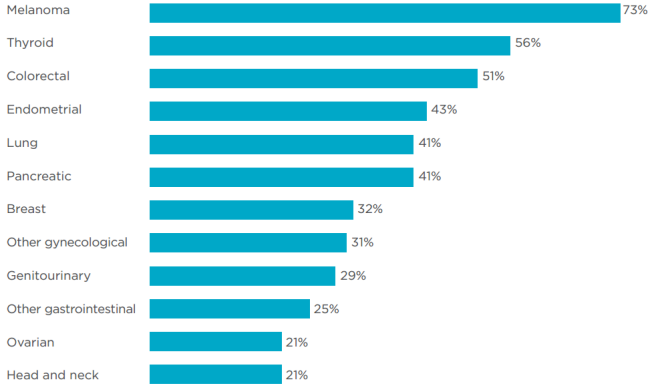
Our Focus: Biomarker discovery

Def: "any variable that can be used as an indicator of a particular disease state".



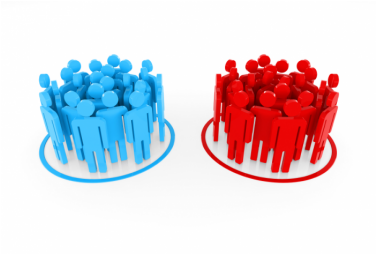
Biomarkers as potential targets for new drugs

TACKLING TUMORS: Percentage of patients whose tumors were driven by certain genetic mutations that could be targets for specific drugs, by types of cancer.



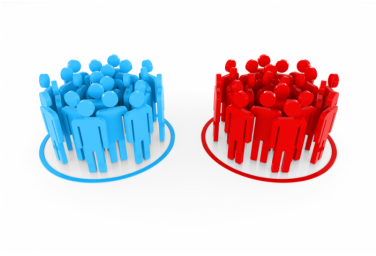
Problem Formulation

Biomarker discovery in clinical trials



Problem Formulation

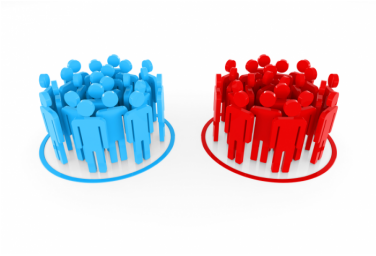
Biomarker discovery in clinical trials



We want to discover:

Problem Formulation

Biomarker discovery in clinical trials

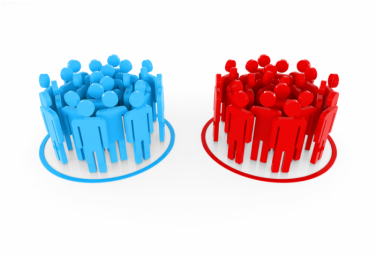


We want to discover:

1. Indicators of disease progression: prognostic biomarkers

Problem Formulation

Biomarker discovery in clinical trials

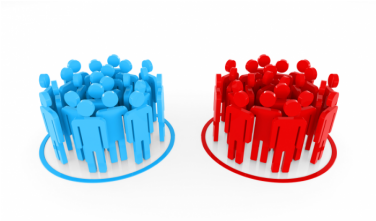


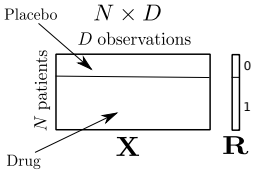
We want to discover:

1. Indicators of disease progression: prognostic biomarkers
2. Indicators of (positive) drug response: predictive biomarkers

Problem Formulation

Biomarker discovery in clinical trials





We want to discover:

1. Indicators of disease progression: prognostic biomarkers
2. Indicators of (positive) drug response: predictive biomarkers

Application: Immunotherapy for liver cancer

[Abou-Alfa et.al, 2016]

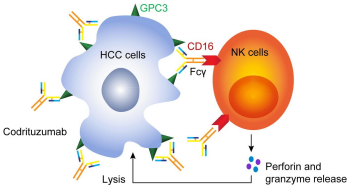


Diagram: Mechanism of codrituzumab-induced antibody-dependent cytotoxicity through the interaction of Fc CD16 in NK cells

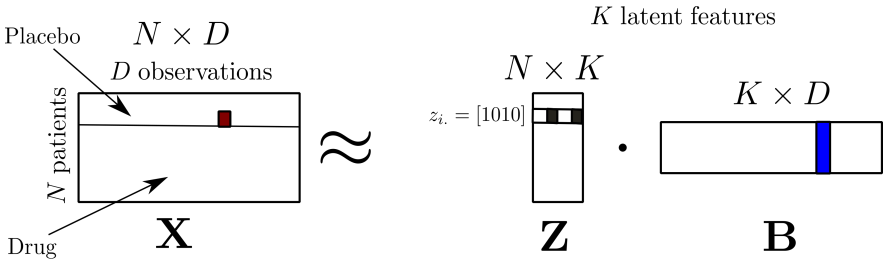
So far...

- No evidence for treatment effectiveness
- Hypothesis: drug exposure as confounder

How to deal with...

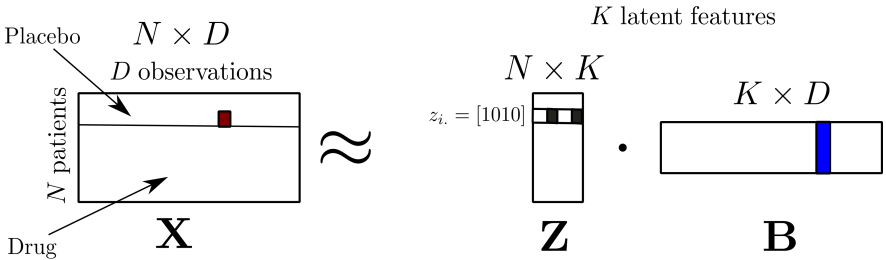
- data complexity?
- data heterogeneity?
- drug effect vs natural response?

Our Approach: Latent Feature Model



- $x_{id} = 173\,\text{ml/dL} = 73 + 0 + 100\,\text{ml/dL}$

Our Approach: Latent Feature Model



- $x_{id} = 173\,\text{ml/dL} = 73 + 0 + 100\,\text{ml/dL}$

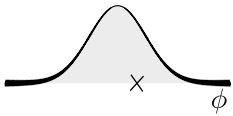
Note: Correlation does not imply causality!

How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]

How to deal with complexity and patient heterogeneity?

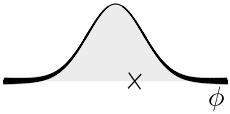
Indian Buffet Process [Ghahramani et.al, 2006]



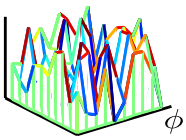
$$x \sim \mathcal{N}(\mu_0, \sigma_0^2) \longrightarrow x = 0.356$$

How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]



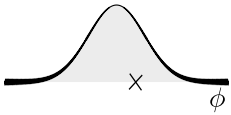
$$\longrightarrow \quad x \sim \mathcal{N}(\mu_0, \sigma_0^2) \qquad x = 0.356$$



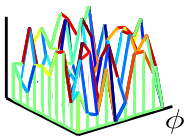
$$\longrightarrow \quad Z \sim \text{IBP}(\alpha) \qquad Z = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \vdots$$

How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]



$$\longrightarrow \quad x \sim \mathcal{N}(\mu_0, \sigma_0^2) \qquad x = 0.356$$



$$\longrightarrow \quad Z \sim \text{IBP}(\alpha) \qquad Z = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \vdots$$

- Prior over binary matrices with infinite number of columns
- Rows $\equiv$ observations; columns $\equiv$ features
- $\mathbf{Z} \sim \text{IBP}(\alpha)$
- α: concentration parameter

How to deal with complexity and patient heterogeneity?

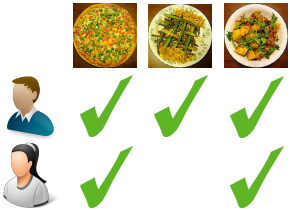
Indian Buffet Process [Ghahramani et.al, 2006]

How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]

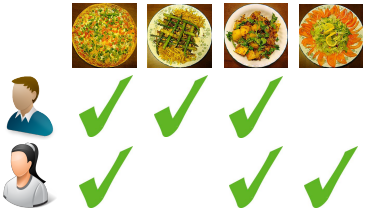
How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]



How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]



How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]

How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]

How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]

How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]

1	1	1	0	0	0
1	0	1	1	0	0
0	1	1	0	1	1

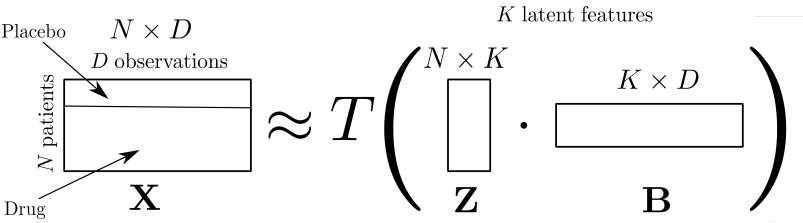
How to deal with complexity and patient heterogeneity?

Indian Buffet Process [Ghahramani et.al, 2006]

1	1	0	1	0	1
1	0	1	0	0	1
0	0	1	0	1	1

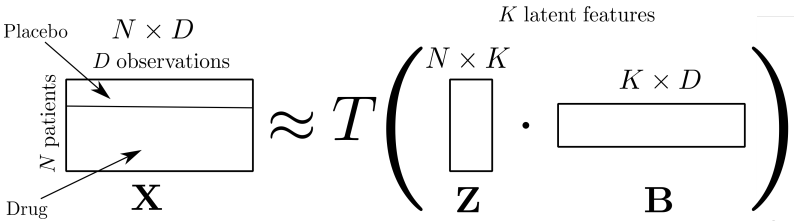
How about feature heterogeneity?

General latent feature model (GLFM) [Valera et.al, 2020]



How about feature heterogeneity?

General latent feature model (GLFM) [Valera et.al, 2020]

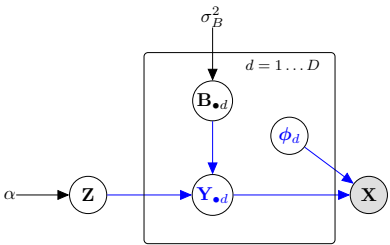


- K potentially unbounded
- Link functions T_d for each feature d

How about feature heterogeneity?

General latent feature model (GLFM) [Valera et.al, 2020]

Latent feature model for heterogeneous datasets



- Link functions T_d depend on type of data for each feature d

$$
\begin{aligned}
x_{nd} &= T_d(y_{nd}; \boldsymbol{\phi_d}) \\
y_{nd}|\mathbf{Z}, \mathbf{B} &\sim \mathcal{N}(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}, \sigma_y^2) \\
B_{kd} &\sim \mathcal{N}(0, \sigma_B^2) \\
\mathbf{Z} &\sim \text{IBP}(\alpha)
\end{aligned}
$$

GLFM package

- Open-source python/matlab/R code
- Deals with Heterogeneous datasets

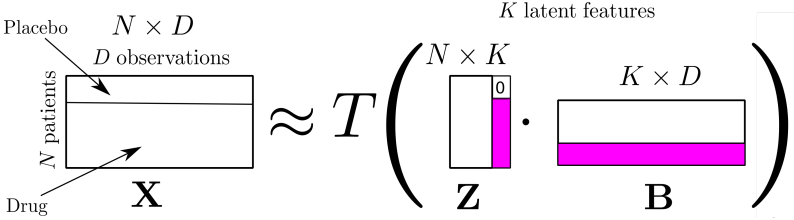
https://github.com/ivaleraM/GLFM

How to distinguish drug effect vs natural response?

Case-control Indian Buffet Process (C-IBP) [Pradier et.al, 2019]

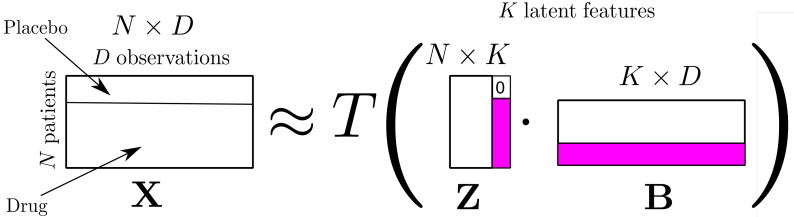
How to distinguish drug effect vs natural response?

Case-control Indian Buffet Process (C-IBP) [Pradier et.al, 2019]



How to distinguish drug effect vs natural response?

Case-control Indian Buffet Process (C-IBP) [Pradier et.al, 2019]

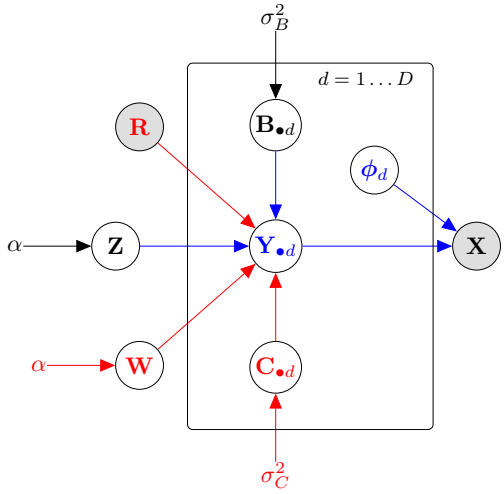


Treatment-specific latent features

- can only activate for patients in treatment arm
- number learned automatically

How to distinguish drug effect vs natural response?

Case-control Indian Buffet Process (C-IBP) [Pradier et.al, 2019]



R_n: drug indicator por patient n

$$x_{nd} = T_d(y_{nd}; \boldsymbol{\phi_d})$$

$$y_{nd}|\mathbf{Z}, \mathbf{W}, \mathbf{B}, \mathbf{C}, \mathbf{R} \sim \mathcal{N}(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d} + \mathbb{1}[\mathbf{R}_n = 1]\mathbf{W}_{n\bullet}\mathbf{C}_{\bullet d}, \sigma_y^2)$$

$$B_{kd} \sim \mathcal{N}(0, \sigma_B^2)$$

$$\mathbf{Z} \sim \text{IBP}(\alpha)$$

$$C_{kd} \sim \mathcal{N}(0, \sigma_C^2)$$

$$\mathbf{W} \sim \text{IBP}(\alpha)$$

- **Inference**: MCMC approach with accelerated Gibbs sampling
- **Biomarker discovery**: statistical multiple hypothesis testing

Results: subpopulations

GPC3 Antibody Treatment against Liver Cancer (J. Hepatology. 2016 Apr, Abou-Alfa et.al.)

- 180 patients: 60 took a placebo, 120 took the drug
- PFS: Progression Free Survival

Sub-population	Drug Identifier	F1	F2	F3	Size (number of patients)	Mean PFS (months)	Median PFS (months)
1.	0	0	0	0	33.37	3.06	1.65
2.	0	0	1	0	4.07	2.29	2.24
3.	0	1	0	0	17.84	2.72	1.81
4.	0	1	1	0	4.72	7.05	7.18
5.	1	0	0	0	51.52	3.22	2.55
6.	1	0	0	1	16.77	4.17	3.65
7.	1	0	1	0	8.38	1.74	1.33
8.	1	0	1	1	2.07	2.69	2.65
9.	1	1	0	0	29.88	3.36	2.03
10.	1	1	0	1	4.90	4.44	4.34
11.	1	1	1	0	4.53	6.31	5.31
12.	1	1	1	1	1.94	10.04	10.01

Results: subpopulations

GPC3 Antibody Treatment against Liver Cancer (J. Hepatology. 2016 Apr, Abou-Alfa et.al.)

- 180 patients: 60 took a placebo, 120 took the drug
- PFS: Progression Free Survival

Sub-population	Drug Identifier	F1	F2	F3	Size (number of patients)	Mean PFS (months)	Median PFS (months)
1.	0	0	0	0	33.37	3.06	1.65
2.	0	0	1	0	4.07	2.29	2.24
3.	0	1	0	0	17.84	2.72	1.81
4.	0	1	1	0	4.72	7.05	7.18
5.	1	0	0	0	51.52	3.22	2.55
6.	1	0	0	1	16.77	4.17	3.65
7.	1	0	1	0	8.38	1.74	1.33
8.	1	0	1	1	2.07	2.69	2.65
9.	1	1	0	0	29.88	3.36	2.03
10.	1	1	0	1	4.90	4.44	4.34
11.	1	1	1	0	4.53	6.31	5.31
12.	1	1	1	1	1.94	10.04	10.01

Results: subpopulations

GPC3 Antibody Treatment against Liver Cancer (J. Hepatology. 2016 Apr, Abou-Alfa et.al.)

- 180 patients: 60 took a placebo, 120 took the drug
- PFS: Progression Free Survival

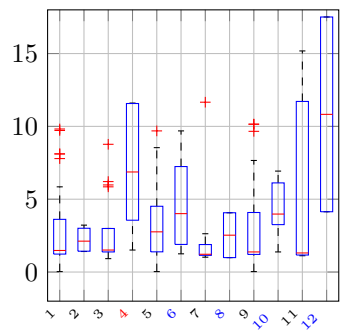
Sub-population	Drug Identifier	F1	F2	F3	Size (number of patients)	Mean PFS (months)	Median PFS (months)
1.	0	0	0	0	33.37	3.06	1.65
2.	0	0	1	0	4.07	2.29	2.24
3.	0	1	0	0	17.84	2.72	1.81
4.	0	1	1	0	4.72	7.05	7.18
5.	1	0	0	0	51.52	3.22	2.55
6.	1	0	0	1	16.77	4.17	3.65
7.	1	0	1	0	8.38	1.74	1.33
8.	1	0	1	1	2.07	2.69	2.65
9.	1	1	0	0	29.88	3.36	2.03
10.	1	1	0	1	4.90	4.44	4.34
11.	1	1	1	0	4.53	6.31	5.31
12.	1	1	1	1	1.94	10.04	10.01

Results: subpopulations

GPC3 Antibody Treatment against Liver Cancer (J. Hepatology. 2016 Apr, Abou-Alfa et.al.)

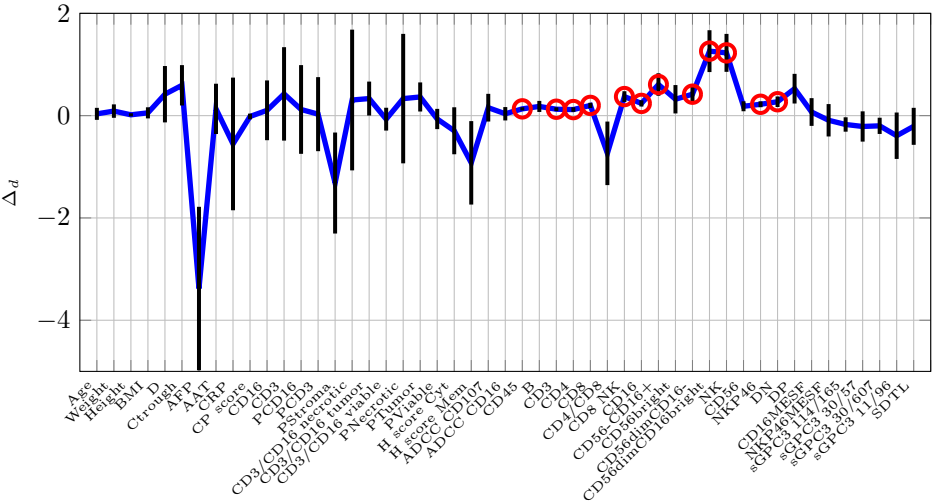
- 180 patients: 60 took a placebo, 120 took the drug
- PFS: Progression Free Survival

Sub-population	Drug Identifier	F1	F2	F3	Size (number of patients)	Mean PFS (months)	Median PFS (months)
1.	0	0	0	0	33.37	3.06	1.65
2.	0	0	1	0	4.07	2.29	2.24
3.	0	1	0	0	17.84	2.72	1.81
4.	0	1	1	0	4.72	7.05	7.18
5.	1	0	0	0	51.52	3.22	2.55
6.	1	0	0	1	16.77	4.17	3.65
7.	1	0	1	0	8.38	1.74	1.33
8.	1	0	1	1	2.07	2.69	2.65
9.	1	1	0	0	29.88	3.36	2.03
10.	1	1	0	1	4.90	4.44	4.34
11.	1	1	1	0	4.53	6.31	5.31
12.	1	1	1	1	1.94	10.04	10.01



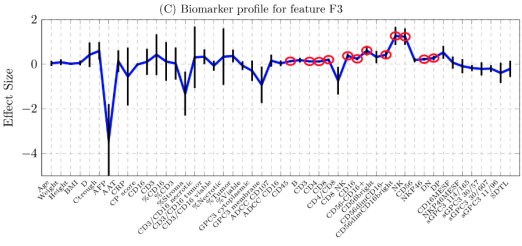
Results: biomarker discovery

Treatment-specific feature F3

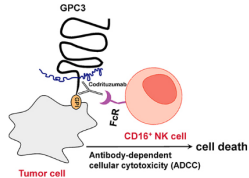


Results: biomarker discovery

Treatment-specific feature F3



[Pradier et.al, 2019]



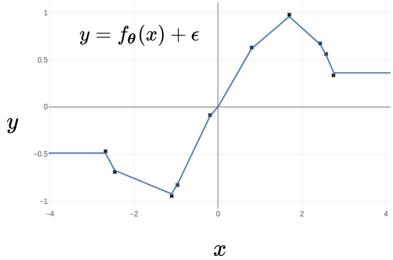
What did we found?

- Subgroup for which treatment is especially effective
- Relevant biomarkers (drug acting as expected)

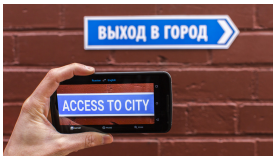
Outline

- Overview
- Goal I: Biomarker discovery
- Goal II: Functional prediction
- Wrap-up

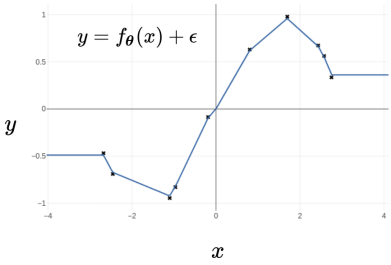
Neural networks (NNs) as universal approximators



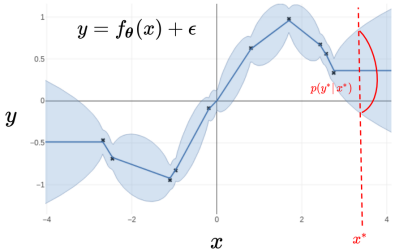
Several success stories...



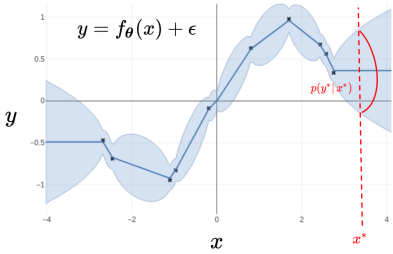
But what if stakes are high?



But what if stakes are high?

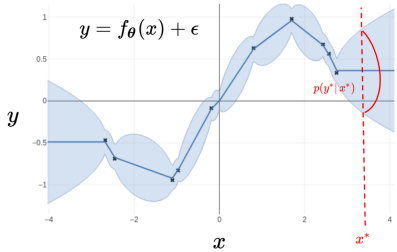


But what if stakes are high?



Uncertainty estimation becomes crucial!

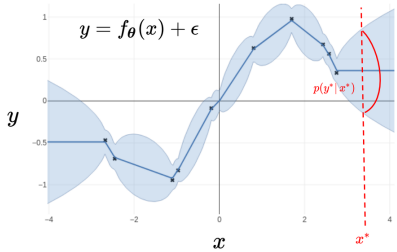
Sometimes we have a priori functional knowledge...



Some examples of assumptions:

- Range of heart rate at rest between 60-100 bpm.
- Slow/fast variation of air pollutant
- Volatility of stock market

Sometimes we have a priori functional knowledge...



Some examples of assumptions:

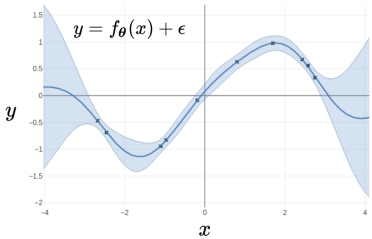
- Range of heart rate at rest between 60-100 bpm.
- Slow/fast variation of air pollutant
- Volatility of stock market

How can we incorporate such functional desiderata into the model?

An easy way to specify functional desiderata: Gaussian Processes (GPs)

Definition: *a Gaussian process is a collection of random variables, any finite number of which have (consistent) Gaussian distributions.*

$$f \sim \mathcal{N}\big(\mu(\cdot), k\,(\cdot,\cdot)\,\big)$$



Example: RBF kernel as covariance function:

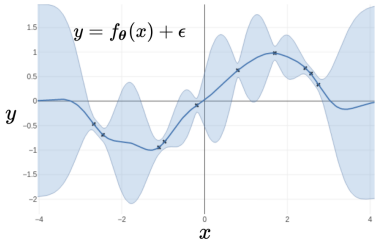
$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\gamma^2}\right)$$

- Stationarity
- Lengthscale
- Amplitude variance

An easy way to specify functional desiderata: Gaussian Processes (GPs)

Definition: *a Gaussian process is a collection of random variables, any finite number of which have (consistent) Gaussian distributions.*

$$f \sim \mathcal{N}\big(\mu(\cdot), k\,(\cdot,\cdot)\,\big)$$



Example: RBF kernel as covariance function:

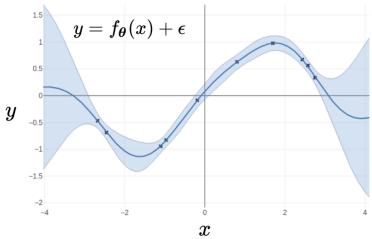
$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\gamma^2}\right)$$

- Stationarity
- Lengthscale
- Amplitude variance

An easy way to specify functional desiderata: Gaussian Processes (GPs)

Definition: *a Gaussian process is a collection of random variables, any finite number of which have (consistent) Gaussian distributions.*

$$f \sim \mathcal{N}\big(\mu(\cdot), k\,(\cdot,\cdot)\,\big)$$



Example: RBF kernel as covariance function:

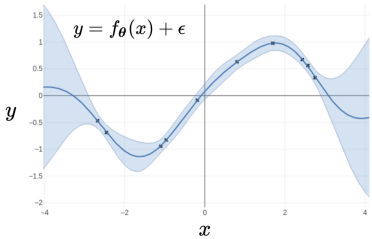
$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\gamma^2}\right)$$

- Stationarity
- Lengthscale
- Amplitude variance

An easy way to specify functional desiderata: Gaussian Processes (GPs)

Definition: *a Gaussian process is a collection of random variables, any finite number of which have (consistent) Gaussian distributions.*

$$f \sim \mathcal{N}\big(\mu(\cdot), k\,(\cdot,\cdot)\,\big)$$



Example: RBF kernel as covariance function:

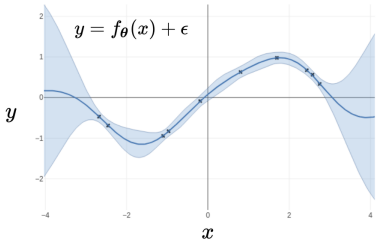
$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\gamma^2}\right)$$

- Stationarity
- Lengthscale
- Amplitude variance

An easy way to specify functional desiderata: Gaussian Processes (GPs)

Definition: *a Gaussian process is a collection of random variables, any finite number of which have (consistent) Gaussian distributions.*

$$f \sim \mathcal{N}\big(\mu(\cdot), k\,(\cdot,\cdot)\,\big)$$



Example: RBF kernel as covariance function:

$$k(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\gamma^2}\right)$$

- Stationarity
- Lengthscale
- Amplitude variance

GPs are great, but what if I still want a NN?

Benefits of NN approaches:

- widely used (many tools available)
- parametric expression
- fast at evaluation time

GPs are great, but what if I still want a NN?

Benefits of NN approaches:

- widely used (many tools available)
- parametric expression
- fast at evaluation time

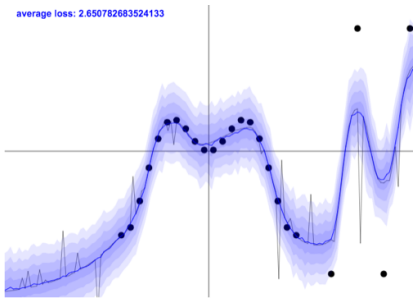
Key Research Questions:

1. Can we design Bayesian NN priors that encode **stationarity properties** like a GP while retaining the benefits of neural networks?
2. Can we easily specify lengthscale and amplitude variance in a **decoupled** fashion?

Bayesian Neural Networks

- Assume prior on network parameters
- Most common, i.i.d Gaussians

$$
\begin{aligned}
\mathbf{y} &= f_{\boldsymbol{\theta}}(\mathbf{x}) + \boldsymbol{\epsilon} \\
\epsilon &\sim \mathcal{N}(0, \sigma_y^2 I) \\
\theta_i &\sim \mathcal{N}(0, \sigma_\theta^2 I) \quad \forall i
\end{aligned}
$$



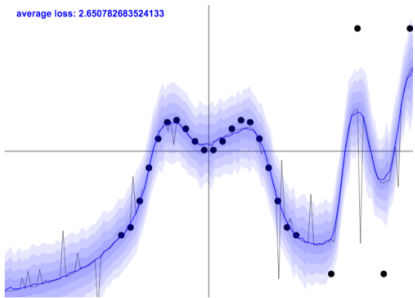
(Yarin Gal blog)

- $p(\boldsymbol{\theta}) \implies p(f)$

Bayesian Neural Networks

- Assume prior on network parameters
- Most common, i.i.d Gaussians

$$\begin{aligned} \mathbf{y} &= f_{\boldsymbol{\theta}}(\mathbf{x}) + \boldsymbol{\epsilon} \\ \epsilon &\sim \mathcal{N}(0, \sigma_y^2 I) \\ \theta_i &\sim \mathcal{N}(0, \sigma_\theta^2 I) \quad \forall i \end{aligned}$$



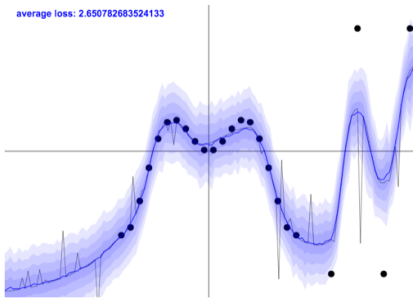
(Yarin Gal blog)

- $p(\boldsymbol{\theta}) \implies p(f)$
- But what does a prior over weights mean in function space?

Bayesian Neural Networks

- Assume prior on network parameters
- Most common, i.i.d Gaussians

$$\begin{aligned} \mathbf{y} &= f_{\boldsymbol{\theta}}(\mathbf{x}) + \boldsymbol{\epsilon} \\ \epsilon &\sim \mathcal{N}(0, \sigma_y^2 I) \\ \theta_i &\sim \mathcal{N}(0, \sigma_\theta^2 I) \quad \forall i \end{aligned}$$

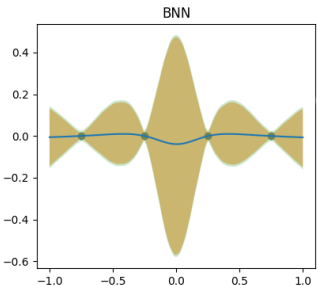


(Yarin Gal blog)

- $p(\boldsymbol{\theta}) \implies p(f)$
- But what does a prior over weights mean in function space? Hard to know!

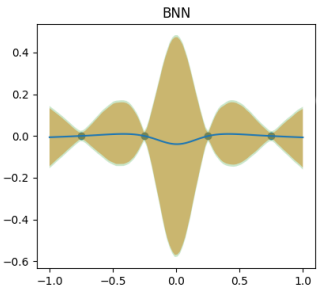
Not only hard to encode functional properties with BNNs; some properties are impossible to get

- For example, a BNN (with RBF activations) is nonstationary in amplitude variance (Williams, 1997)



Not only hard to encode functional properties with BNNs; some properties are impossible to get

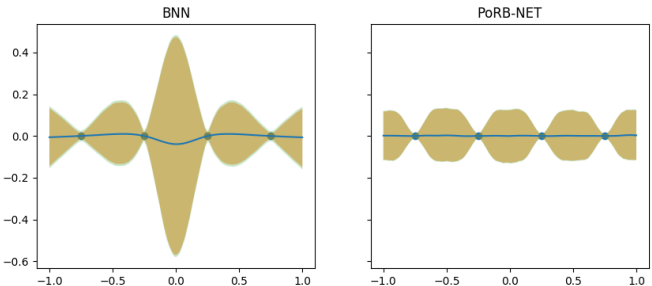
- For example, a BNN (with RBF activations) is nonstationary in amplitude variance (Williams, 1997)



Question: can we design a Bayesian NN that exhibits stationarity?

Not only hard to encode functional properties with BNNs; some properties are impossible to get

- For example, a BNN (with RBF activations) is nonstationary in amplitude variance (Williams, 1997)



Question: can we design a Bayesian NN that exhibits stationarity? Yes!

Related works

Expressive priors for Bayesian NNs

- Functional BNNs (Flam-Shepherd, et.al 2017; Sun et.al, 2019): sample-based optimization w.r.t. reference functional distribution
- Neural processes (Garnelo et al., 2018): meta-learning to identify functional properties based on many prior examples
- (Pearce et al., 2019) BNN architectures that recover equivalent GP kernel combinations in the infinite width limit

Related works

Expressive priors for Bayesian NNs

- Functional BNNs (Flam-Shepherd, et.al 2017; Sun et.al, 2019): sample-based optimization w.r.t. reference functional distribution
- Neural processes (Garnelo et al., 2018): meta-learning to identify functional properties based on many prior examples
- (Pearce et al., 2019) BNN architectures that recover equivalent GP kernel combinations in the infinite width limit

	user specs	optim. free	finite width	deep
Sun et.al, 2019	yes	no	yes	yes
Garnelo et.al, 2018	no	no	yes	yes
Pearce et.al, 2019	yes	yes	no	yes
PoRB-NET (this work)	yes	yes	yes	not yet

Radial Basis Function Networks (RBFNs)

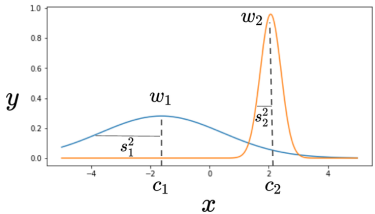
- Around since the 90s (Gyorfi et.al, 2002), recently renewed attention (Taghi et.al, 2004; Zadeh et.al, 2018)

Radial Basis Function Networks (RBFNs)

- Around since the 90s (Gyorfi et.al, 2002), recently renewed attention (Taghi et.al, 2004; Zadeh et.al, 2018)
- NN based on radial basis $\phi(\cdot)$, e.g., $\phi(x) = \exp(-x^2)$

$$f_{\boldsymbol{\theta}}(x) = b + \sum_{k=1}^{K} w_k \phi\big(s_k(x - c_k)\big),$$

- $s_k^2 \in \mathbb{R}$: scale
- $c_k \in \mathbb{R}$: center
- $w_k \in \mathbb{R}$: output weight
- $b \in \mathbb{R}$: output bias



Comparison RBFN versus BNN formulation (D=1)

$$f_{\boldsymbol{\theta}}(x) = b + \sum_{k=1}^{K} w_k \phi\big(s_k(x - c_k)\big)$$

- $s_k^2 \in \mathbb{R}$: scale
- $c_k \in \mathbb{R}$: center
- $w_k \in \mathbb{R}$: output weight
- $b \in \mathbb{R}$: output bias

$$f_{\boldsymbol{\theta}}(x) = b + \sum_{k=1}^{K} w_k \phi\big(v_k x + d_k\big)$$

- $v_k \in \mathbb{R}$: input weight
- $d_k \in \mathbb{R}$: input bias
- $w_k \in \mathbb{R}$: output weight
- $b \in \mathbb{R}$: output bias

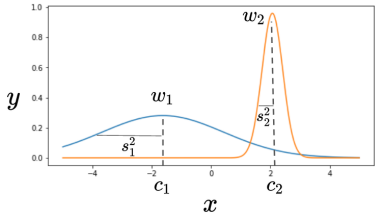
Take-away: priors on different random quantities, RBFN more intuitive

Bayesian RBFNs (Barber et.al, 1998)

$$
\begin{aligned}
c_k &\sim \mathcal{N}(0, \sigma_c^2) \\
s_k^2 &\sim \text{Gamma}(\alpha_s, \beta_s) \\
w_k &\sim \mathcal{N}\left(0, \sigma_w^2 I\right) \\
B &\sim \mathcal{N}(0, \sigma_b^2) \\
y_n \,|\, x_n, \boldsymbol{\theta} &\sim \mathcal{N}(f_{\boldsymbol{\theta}}(x_n), \sigma_y^2)
\end{aligned}
$$

where

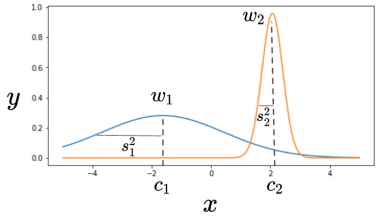
$$
f_{\boldsymbol{\theta}}(x) = b + \sum_{k=1}^{K} w_k \exp\left(-s_k^2 (x - c_k)^2\right)
$$



- $s_k^2 \in \mathbb{R}$: scale
- $c_k \in \mathbb{R}$: center
- $w_k \in \mathbb{R}$: output weight
- $b \in \mathbb{R}$: output bias

Bayesian RBFNs (Barber et.al, 1998)

$$
\begin{aligned}
c_k &\sim \mathcal{N}(0, \sigma_c^2) \\
s_k^2 &\sim \text{Gamma}(\alpha_s, \beta_s) \\
w_k &\sim \mathcal{N}\left(0, \sigma_w^2 I\right) \\
B &\sim \mathcal{N}(0, \sigma_b^2) \\
y_n \,|\, x_n, \boldsymbol{\theta} &\sim \mathcal{N}(f_{\boldsymbol{\theta}}(x_n), \sigma_y^2)
\end{aligned}
$$



where

$$
f_{\boldsymbol{\theta}}(x) = b + \sum_{k=1}^{K} w_k \exp\left(-s_k^2 (x - c_k)^2\right)
$$

- $s_k^2 \in \mathbb{R}$: scale
- $c_k \in \mathbb{R}$: center
- $w_k \in \mathbb{R}$: output weight
- $b \in \mathbb{R}$: output bias

Functional properties still hard or impossible to encode!

Functional properties still hard or impossible

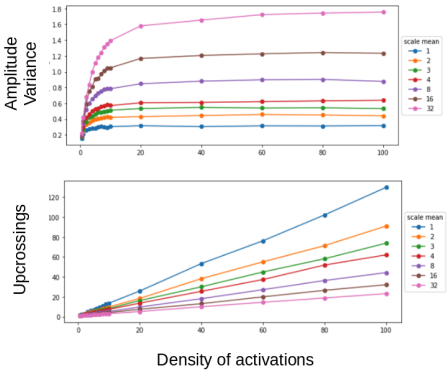
Issues:

- non-stationary covariance function (Williams, 1997)
- lengthscale and variance are **coupled**

Functional properties still hard or impossible

Issues:

- non-stationary covariance function (Williams, 1997)
- lengthscale and variance are **coupled**



- As RBFs concentrate in same region:
  - summation $\implies$ higher variance
  - increase in expressivity $\implies$ more upcrossings

Poisson Process Radial Basis Function Networks (PoRB-NET)

$$
\begin{aligned}
c_k &\sim \mathcal{N}(0, \sigma_c^2) \\
s_k^2 &\sim \text{Gamma}(\alpha_s, \beta_s) \\
w_k &\sim \mathcal{N}\left(0, \sigma_w^2 I\right) \\
B &\sim \mathcal{N}(0, \sigma_b^2) \\
y_n \,|\, x_n, \boldsymbol{\theta} &\sim \mathcal{N}(f_{\boldsymbol{\theta}}(x_n), \sigma_y^2)
\end{aligned}
$$

where

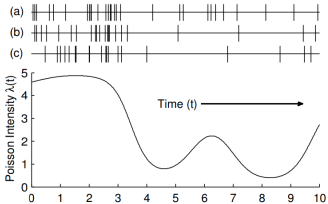
$$
f_{\boldsymbol{\theta}}(x) = B + \sum_{k=1}^{K} w_k \exp\left(- s_k^2 (x - c_k)^2 \right)
$$

Poisson Process Radial Basis Function Networks (PoRB-NET)

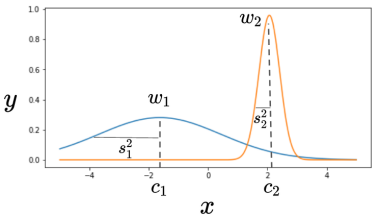
$$
\begin{aligned}
\mathbf{c} \,|\, \lambda &\sim \text{Poisson Process}\,(\lambda) \\
s_k^2 &\sim \text{Gamma}(\alpha_s, \beta_s) \\
w_k &\sim \mathcal{N}\left(0, \tilde{\sigma}_w^2 I\right) \\
B &\sim \mathcal{N}(0, \tilde{\sigma}_b^2) \\
y_n \,|\, x_n, \boldsymbol{\theta} &\sim \mathcal{N}(f_{\boldsymbol{\theta}}(x_n), \sigma_y^2)
\end{aligned}
$$

where

$$
f_{\boldsymbol{\theta}}(x) = B + \sum_{k=1}^{K} w_k \exp\left(-s_k^2 (x - c_k)^2\right)
$$



(Adams et.al, 2009)

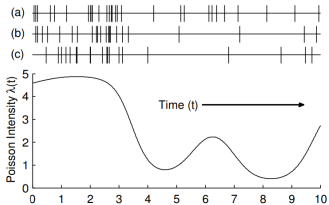


Poisson Process Radial Basis Function Networks (PoRB-NET)

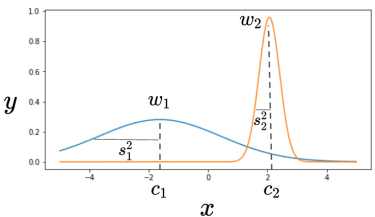
$$
\begin{aligned}
\mathbf{c} \,|\, \lambda &\sim \text{Poisson Process}\,(\lambda) \\
s_k^2 &= \lambda^2(c_k) \\
w_k &\sim \mathcal{N}\left(0, \tilde{\sigma}_w^2 I\right) \\
B &\sim \mathcal{N}(0, \tilde{\sigma}_b^2) \\
y_n \,|\, x_n, \boldsymbol{\theta} &\sim \mathcal{N}(f_{\boldsymbol{\theta}}(x_n), \sigma_y^2)
\end{aligned}
$$

where

$$
f_{\boldsymbol{\theta}}(x) = B + \sum_{k=1}^{K} w_k \exp\left(-s_k^2 (x - c_k)^2\right)
$$



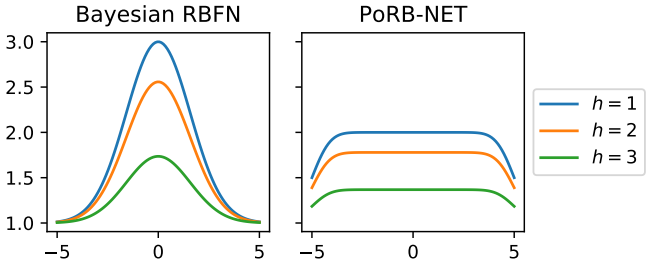
(Adams et.al, 2009)



- We have proposed an expressive prior for NNs
- We show desirable properties:
  1. Stationarity
  2. Decoupling of lengthscale and ampliture variance
  3. Consistency
- We demonstrate successful behavior empirically

Stationarity

$$\mathrm{Cov}(f(x), f(x+h)) = \sigma_b^2 + \sigma_w^2 \mathbb{E}\left[K\right] \underbrace{\mathbb{E}_\theta\left[\rho(x;\theta)\rho(x+h;\theta)\right]}_{:=U(x,x+h)}$$



$$U(x_1, x_2) \propto \underbrace{\exp\left(-\frac{(x_1 - x_2)^2}{2(2\sigma_s^2 + \sigma_s^4/\sigma_c^2)}\right)}_{Stationary} \underbrace{\exp\left(-\frac{x_1^2 + x_2^2}{2(2\sigma_c^2 + \sigma_s^2)}\right)}_{Nonstationary}$$

$$U(x_1, x_2) = \frac{\lambda}{\Lambda}\sqrt{\frac{\pi}{s^2}} \exp\left\{-s^2\left(\frac{x_1 - x_2}{2}\right)^2\right\} \left[\Phi((C_1 - x_m)\sqrt{2s^2}) - \Phi((C_0 - x_m)\sqrt{2s^2}\lambda)\right]$$

Decoupled lengthscale and amplitude variance

- **Homogeneous Poisson Process**
  - We derive closed-form expression for covariance function
  - Poisson process defined over finite region $\mathcal{C}$
  - As size of $\mathcal{C}$ tends to infinity,

$$\mathrm{Cov}\left(f(x_1), f(x_2)\right) \approx \sigma_b^2 + \tilde{\sigma}_w^2 \exp\left\{-\lambda^2 \left(\frac{x_1 - x_2}{2}\right)^2\right\}$$

- **Non-homogeneous Poisson Process**
  - Empirical stationarity

Consistency

- Estimator $\hat{g}_n(x)$ is said to be consistent with respect to the true regression function $g_0(x)$ if, as n tends to infinity:

$$\int (\hat{g}_n(x) - g_0(x))^2 \, dx \xrightarrow{p} 0.$$

- Posterior consistent over Hellinger neighborhoods if $\forall \epsilon > 0$,

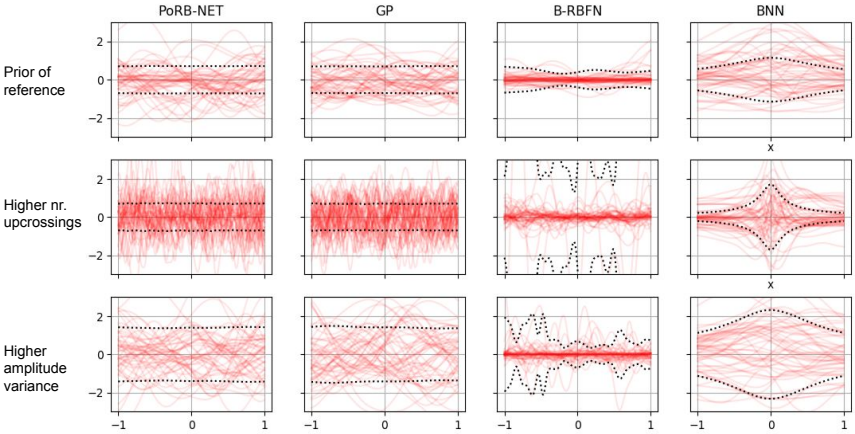
$$p(\{f : D_H(f, f_0) \leq \epsilon\}) \xrightarrow{p} 1.$$

- (Lee, 2000) shows that Hellinger consistency implies frequentist consistency.

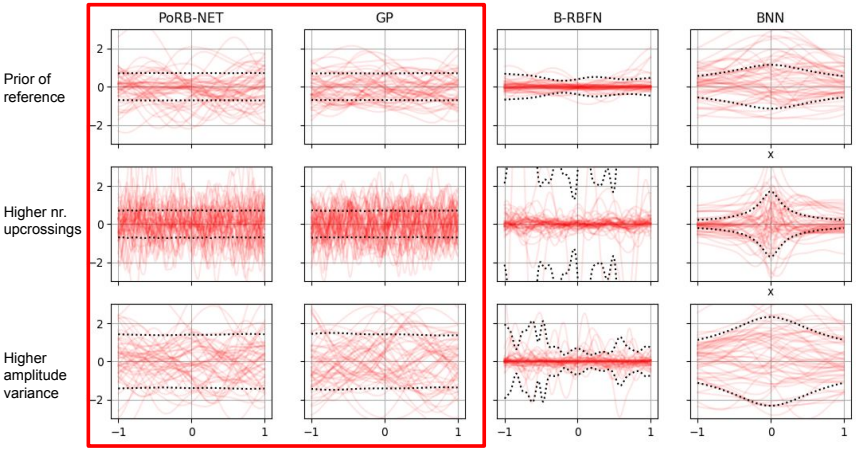
Theorem (Consistency of PoRB-NETs)

A PoRB-NET with uniform intensity function is Hellinger consistent as the number of observations goes to infinity.

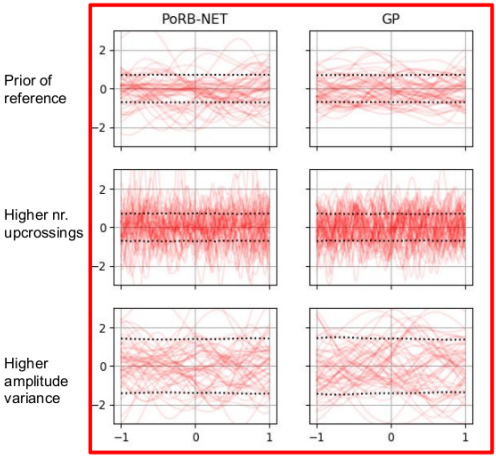
PoRB-NET allows for easy specification of lengthscale and signal variance like a GP



PoRB-NET allows for easy specification of lengthscale and signal variance like a GP

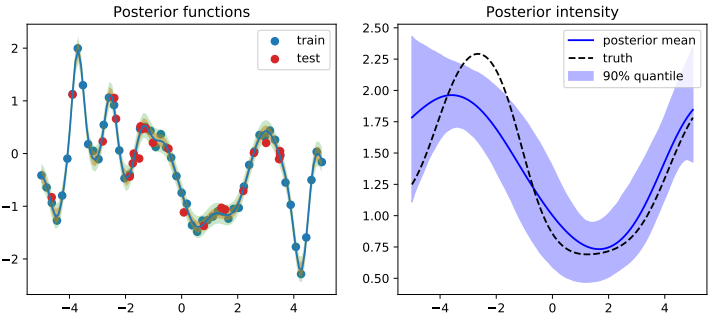


PoRB-NET allows for easy specification of lengthscale and signal variance like a GP



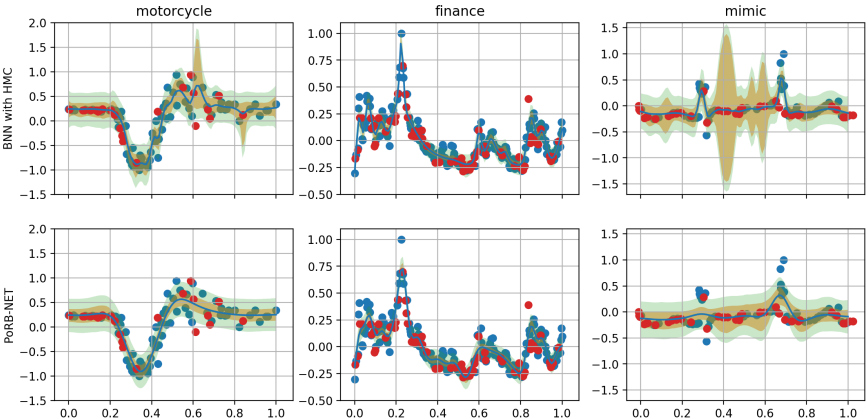
1. stationarity
2. easy specification in a decoupled manner

PoRB-NET is able to learn input-dependent lengthscale information



PoRB-NET adds more hidden units wherever needed, and adapts architecture width based on the data.

PoRB-NET is able to capture non-stationary patterns in real scenarios, adapting the lengthscale locally



Goal II: functional prediction

Take-Aways...

In this talk, we have...

- highlighted incapacity of BNNs to express functional properties
- introduced PoRB-NET, a Bayesian NN prior to encode functional desiderata like a GP
- proposed an inference scheme to learn input-dependent lengthscale
- showed theoretical properties: (i) consistency, (ii) decoupling of amplitude and lengthscale
- validated empirically in synthetic and real datasets

All information online: https://arxiv.org/abs/1912.05779

Outline

- Overview
- Goal I: Biomarker discovery
- Goal II: Functional prediction
- Wrap-up

Conclusion

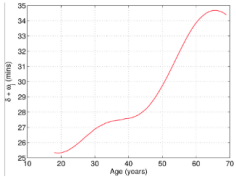
In this talk...

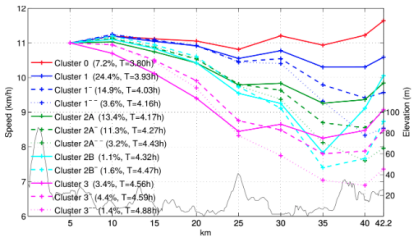
Expressive priors to encode assumptions/desiderata

1. Structured latent feature model
   - subpopulation learning
   - biomarker discovery
2. Novel Bayesian prior for Neural Networks
   - encoding of stationarity
   - decoupling of amplitude variance and lengtscale

Other projects...

Sport Science

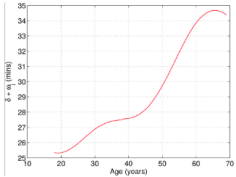


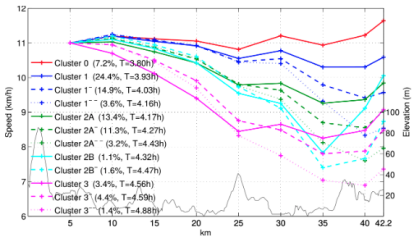


M. F. Pradier, F. J. R. Ruiz, and F. Perez-Cruz. **Prior Design for Dependent Dirichlet Processes: An Application to Marathon Modeling**. *PlosONE*. 2016.

Other projects...

Sport Science

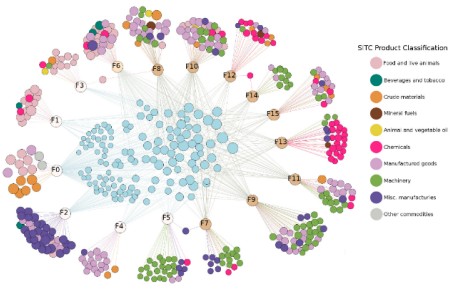


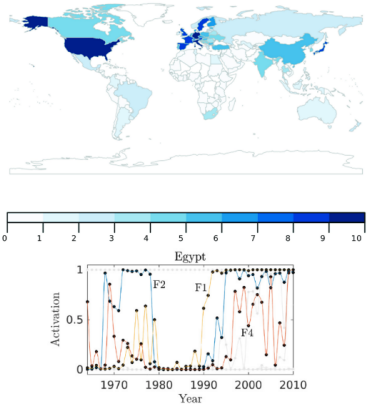


M. F. Pradier, F. J. R. Ruiz, and F. Perez-Cruz. **Prior Design for Dependent Dirichlet Processes: An Application to Marathon Modeling**. *PlosONE*. 2016.

Other projects...

Economics

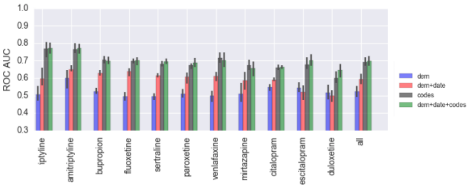


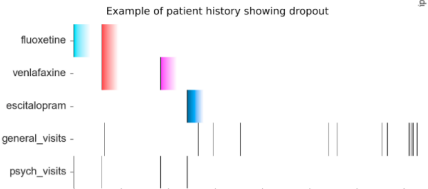


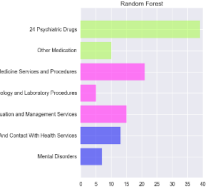
M. F. Pradier*, Z. Utkovski*, V. Stojkoski, L. Kocarev and F. Perez-Cruz. **Economic Complexity Unfolded: An Interpretable Model for the Productive Structure of Economies**. *PlosONE*. 2018.

Other projects...

Medicine: healthcare in psychiatry



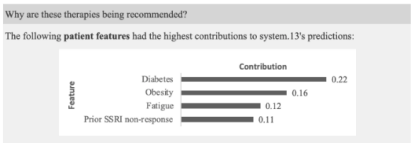




M. F. Pradier, T. H. McCoy, M. Hughes, R. H. Perlis and F. Doshi-Velez. **Predicting Treatment Discontinuation after Antidepressant Initiation**. Nature Translational *Psychiatry*. 2019.
M. F. Pradier, M. Hughes, T. H. McCoy, S. Barroilhet, F. Doshi-Velez and R. H. Perlis. **Predicting Transition from Mayor Depression to Bipolar Disorder after Antidepressant Initiation**. *Submitted to American Journal of Psychiatry*. 2019.

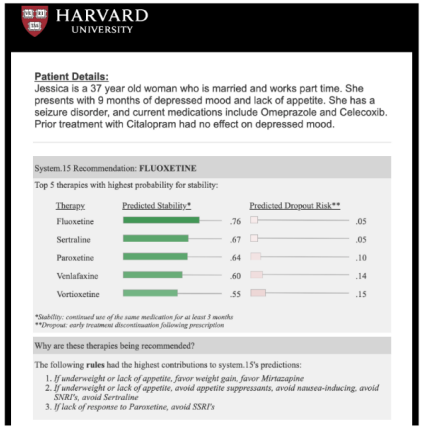
From the lab to the clinic

- Ongoing user study at MGH, Boston
  - Impact of explanations
  - Usefulness, trust...



Which antidepressant medication would you be most likely to prescribe in this situation?

Ongoing [M. Jacobs et.al 2019]



Acknowledgements

Special thanks to:

- Beau Coker
- Finale Doshi-Velez
- All members of DTAK!

- Oscar Puig
- Francesca Milletti
- Fernando Perez-Cruz

- Isabel Valera
- Maria Lomeli
- Zoubin Ghahramani

Thank you for listening!

Looking forward to your questions!
https://melaniefp.github.io/

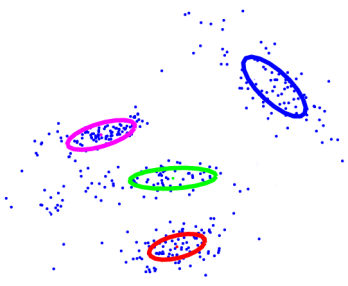
Appendix

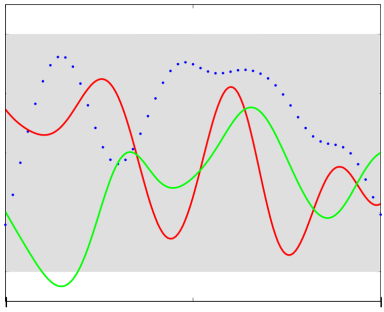
Bayesian nonparametrics

- Bayesian: to handle uncertainty

$$p(\text{parameters}|\text{data}) \propto p(\text{data}|\text{parameters})p(\text{parameters})$$

- Nonparametric: to adapt model complexity (hypothesis generation)



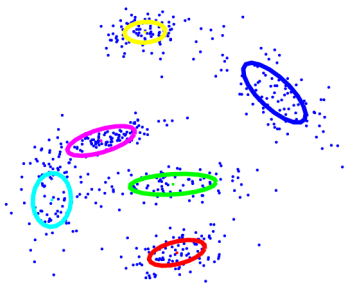


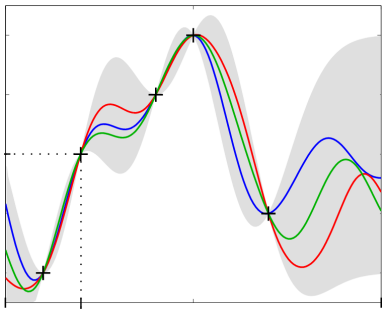
Bayesian nonparametrics

- Bayesian: to handle uncertainty

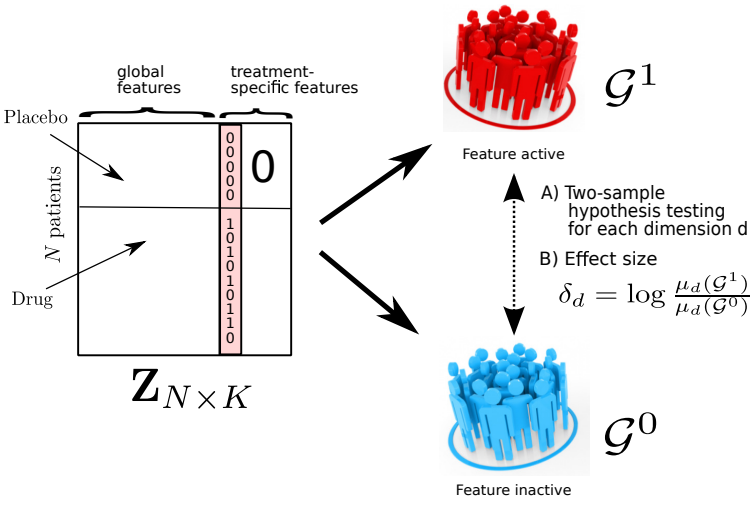
$$p(\text{parameters}|\text{data}) \propto p(\text{data}|\text{parameters})p(\text{parameters})$$

- Nonparametric: to adapt model complexity (hypothesis generation)

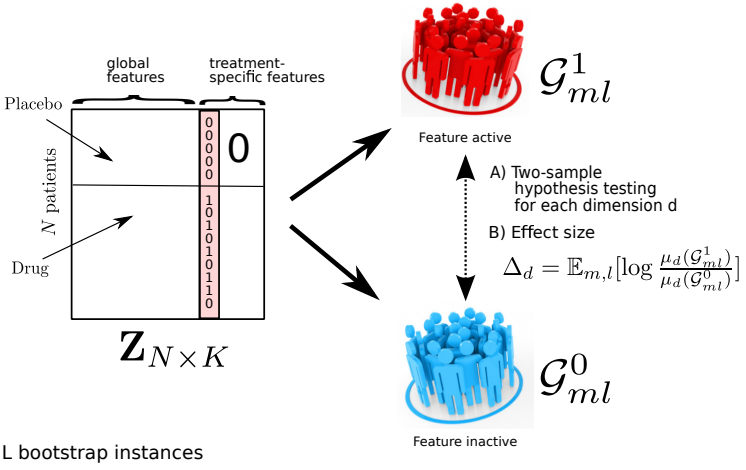




Statistical procedure for biomarker discovery

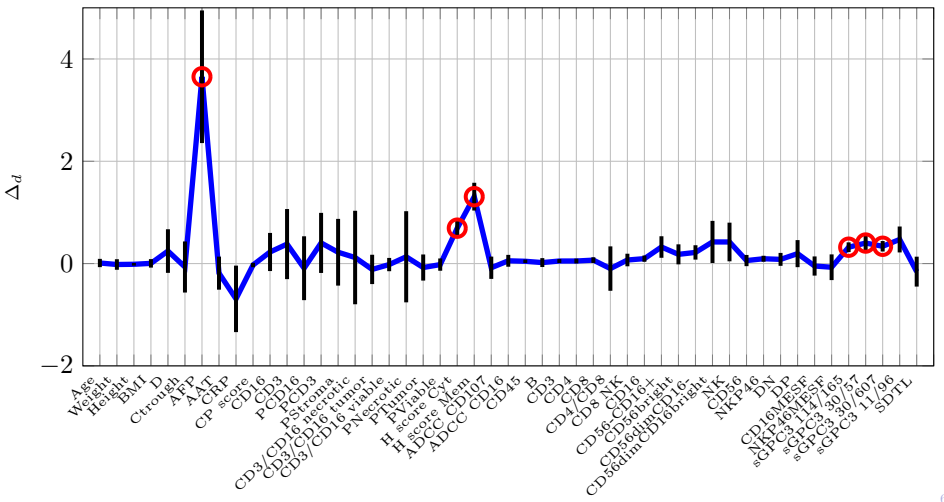


Statistical procedure for biomarker discovery



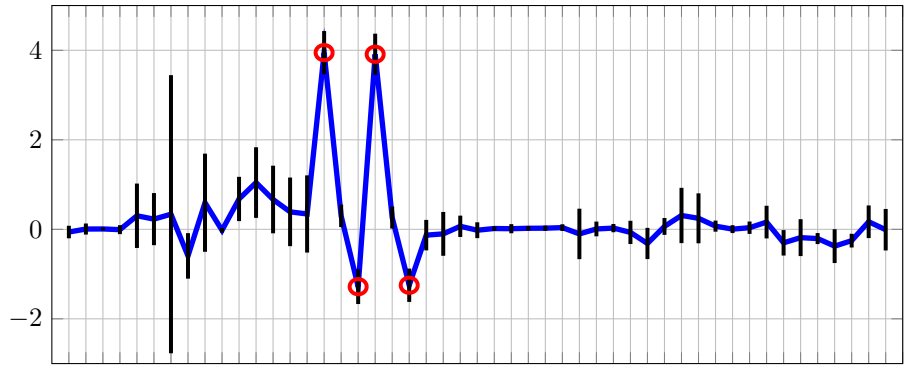
Results: biomarker discovery

Global feature F1



Results: biomarker discovery

Global feature F2



Indian buffet process (IBP)

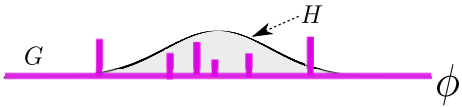
An alternative construction

- underlying block for infinite latent feature models

Indian buffet process (IBP)

An alternative construction

- underlying block for infinite latent feature models
- hierarchy of a Beta process (BP) with multiple Bernoulli processes (BeP)

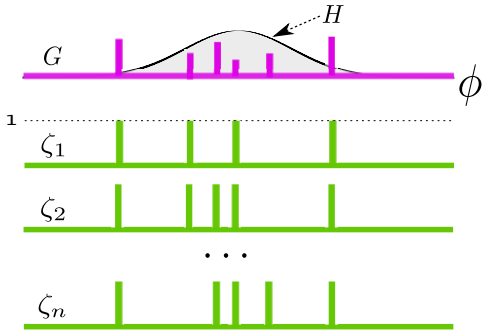


$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \sim \mathrm{BP}(c, \alpha, H)$$

Indian buffet process (IBP)

An alternative construction

- underlying block for infinite latent feature models
- hierarchy of a Beta process (BP) with multiple Bernoulli processes (BeP)



$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \sim \mathrm{BP}(c, \alpha, H)$$

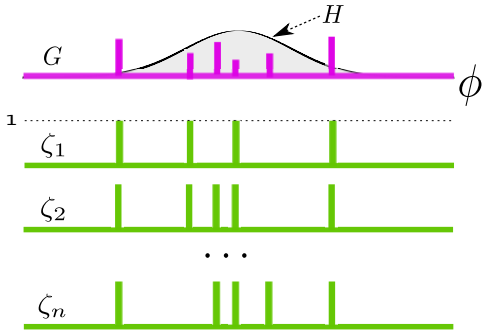
For $n = 1, \cdots, \infty$

$$\zeta_n = \sum_{k=1}^{\infty} z_{nk} \delta_{\phi_k} \sim \mathrm{BeP}(G)$$

Indian buffet process (IBP)

An alternative construction

- underlying block for infinite latent feature models
- hierarchy of a Beta process (BP) with multiple Bernoulli processes (BeP)



$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \sim \mathrm{BP}(c, \alpha, H)$$

For $n = 1, \cdots, \infty$

$$\zeta_n = \sum_{k=1}^{\infty} z_{nk} \delta_{\phi_k} \sim \mathrm{BeP}(G)$$

$$\mathbf{Z} \sim \mathrm{IBP}(\alpha)$$

Comparison RBFN versus BNN formulation (D=1)

$$f_{\boldsymbol{\theta}}(x) = B + \sum_{k=1}^{K} w_k \phi\big(s_k(x - c_k)\big) \qquad f_{\boldsymbol{\theta}}(x) = B + \sum_{k=1}^{K} w_k \phi\big(v_k x + b_k\big)$$

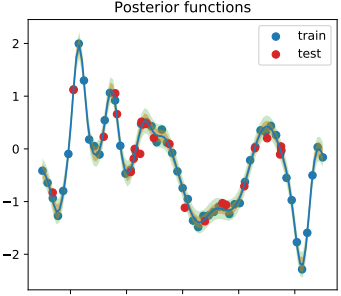
$$\begin{aligned} s_k^2 &\sim \text{Gamma}(\alpha_s, \beta_s) \\ c_k &\sim \mathcal{N}(0, \sigma_c^2) \\ w_k &\sim \mathcal{N}(0, \sigma_w^2) \\ b &\sim \mathcal{N}(0, \sigma_0^2) \end{aligned} \qquad \begin{aligned} v_k^2 &\sim \mathcal{N}(0, \sigma_v^2) \\ b_k &\sim \mathcal{N}(0, \sigma_b^2) \\ w_k &\sim \mathcal{N}(0, \sigma_w^2) \\ b &\sim \mathcal{N}(0, \sigma_0^2) \end{aligned}$$

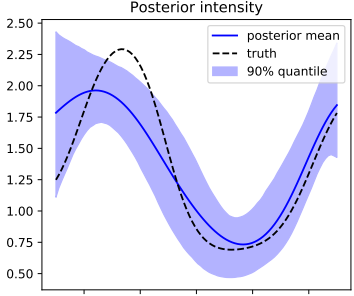
Take-away: priors on different random quantities, RBFN more intuitive

What if we don't know the intensity function?

Prior on Intensity Function of Poisson Process

$$\begin{aligned} h &\sim \mathrm{GP}(0, C(\cdot,\cdot)) \\ \lambda^* &\sim \mathrm{Gamma}(\alpha_\lambda, \beta_\lambda) \\ \lambda(c) &= \lambda^* \mathrm{sigmoid}(h(c)), \end{aligned}$$





Inference

1. Update network parameters $\boldsymbol{\theta}$ given fixed nr. of hidden units K via Hamiltonian Monte Carlo (HMC)

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{x}, K, \lambda) \propto \left(\prod_{n=1}^{N} \mathcal{N}(y_n; f(x_n; \boldsymbol{\theta}) \right) \mathcal{N}(b; 0, \sigma_b^2) \left(\prod_{k=1}^{K} \mathcal{N}(w_k; 0, \sigma_w^2)\ \lambda(c_k) \right)$$

2. Update network width K via birth/death moves
3. Update point-estimate for Poisson process intensity λ

$$\hat{\lambda}(c) \approx \frac{1}{S} \sum \lambda^{\star} \phi(h^{(s)}(c)),$$

where $h^{(s)} \sim p(h|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta})$.