Modular Gaussian Processes

Pablo **Moreno-Muñoz**

Section for Cognitive Systems
Technical University of Denmark (DTU)

DTU

*postdoc with Søren
and also in MLLS

Wednesday, 24 August 2021
@ MLLS Center, København

The **problem**

Complexity of probabilistic learning is typically dominated by the number of data points

The problem

Complexity of probabilistic learning is typically dominated by the number of data points

Examples

$$\hat{\boldsymbol{\theta}}_k = \frac{\sum_{i=1}^N r_{ik}\boldsymbol{x}_i}{\sum_{i=1}^N r_{ik}}$$

Mixture models

$$\Sigma^{-1}_{N \times N} \rightarrow \mathcal{O}(N^3)$$

Gaussian processes

$$\nabla_\theta \mathcal{L}_{1:N} = \sum_{i=1}^N \nabla_\theta \mathcal{L}_i$$

Gradient-based methods

The problem

Complexity of probabilistic learning is typically dominated by the number of data points

Examples

$$\hat{\boldsymbol{\theta}}_k = \frac{\sum_{i=1}^{N} r_{ik}\boldsymbol{x}_i}{\sum_{i=1}^{N} r_{ik}}$$

Mixture models

$$\Sigma_{N \times N}^{-1} \rightarrow \mathcal{O}(N^3)$$

Gaussian processes

$$\nabla_\theta \mathcal{L}_{1:N} = \sum_{i=1}^{N} \nabla_\theta \mathcal{L}_i$$

Gradient-based methods

The **problem**

Complexity of probabilistic learning is typically dominated by the number of data points

$$\Sigma^{-1}_{N\times N} \rightarrow \mathcal{O}(N^3)$$

Gaussian processes

There is **hope**

$$N = N_1 + N_2 + N_3 + \cdots + N_B$$

The **problem**

Complexity of probabilistic learning is typically dominated by the number of data points

$$\Sigma_{N\times N}^{-1} \to \mathcal{O}(N^3)$$

Gaussian processes

There is **hope**

$$N = N_1 + N_2 + N_3 + \cdots + N_B$$

$$(N_1)^3 + (N_2)^3 + (N_3)^3 + \cdots + (N_B)^3 \ll (N_1 + N_2 + N_3 + \cdots + N_B)^3$$

complexity given subsets is much smaller

The problem

Complexity of probabilistic learning is typically dominated by the number of data points

$$\Sigma^{-1}_{N\times N} \rightarrow \mathcal{O}(N^3)$$

Gaussian processes

There is **hope**

$$N = N_1 + N_2 + N_3 + \cdots + N_B$$

$$(1)^3 + (1)^3 + (1)^3 + \cdots + (1)^3 \ll (1000)^3$$

$N_b = 1$

complexity given subsets is much smaller

The problem

Complexity of probabilistic learning is typically dominated by the number of data points

$$\Sigma_{N\times N}^{-1} \rightarrow \mathcal{O}(N^3)$$

Gaussian processes

There is hope

$$N = N_1 + N_2 + N_3 + \cdots + N_B$$

$$1000 \ll (1000)^3$$

$$N_b = 1$$

complexity given subsets is much smaller

The problem

Complexity of probabilistic learning is typically dominated by the number of data points

$$\Sigma_{N\times N}^{-1} \rightarrow \mathcal{O}(N^3)$$

Gaussian processes

There is hope

$$N = N_1 + N_2 + N_3 + \cdots + N_B$$

$$\boxed{(2)^3 + (2)^3 + (2)^3 + \cdots + (2)^3 \ll (1000)^3} \quad N_b = 2$$

complexity given subsets is much smaller

The problem

Complexity of probabilistic learning is typically dominated by the number of data points

$$\Sigma^{-1}_{N \times N} \rightarrow \mathcal{O}(N^3)$$

Gaussian processes

There is **hope**

$$N = N_1 + N_2 + N_3 + \cdots + N_B$$

$$500 \cdot 8 \ll (1000)^3$$

$N_b = 2$

complexity given subsets is much smaller

The problem

Complexity of probabilistic learning is typically dominated by the number of data points

$$\Sigma_{N\times N}^{-1} \rightarrow \mathcal{O}(N^3)$$

Gaussian processes

There is hope $N = N_1 + N_2 + N_3 + \cdots + N_B$

$$4000 \ll (1000)^3$$

$N_b = 2$

complexity given subsets is much smaller

The **problem**

Complexity of probabilistic learning is typically dominated by the number of data points

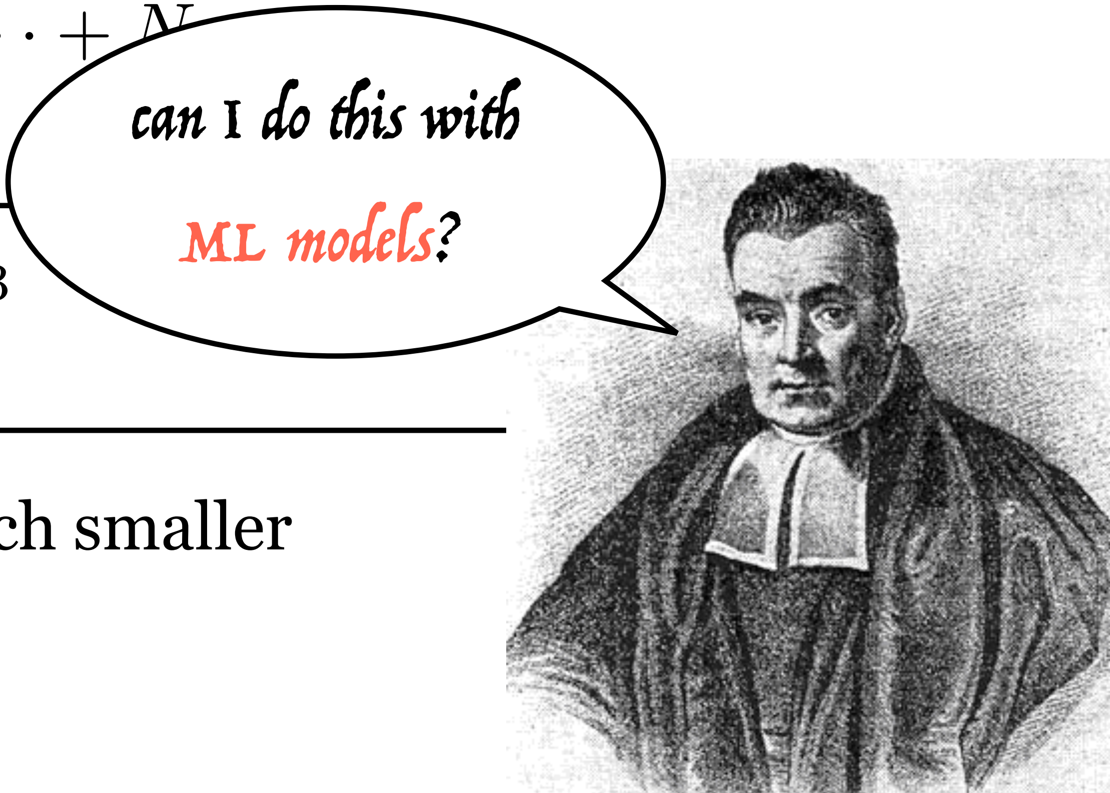
$$\Sigma^{-1}_{N\times N} \to \mathcal{O}(N^3)$$

Gaussian processes

There is **hope**

$$N = N_1 + N_2 + N_3 + \cdots + N$$

$$4000 \ll (1000)^3$$



complexity given subsets is much smaller

The **idea** (tourist metaphor)

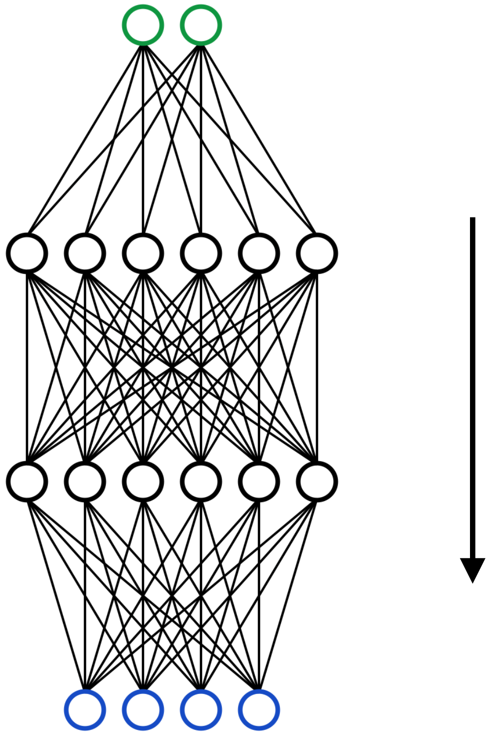
Nyhavn

The **idea** (tourist metaphor)

$N = 100$ observations

The **idea** (tourist metaphor)

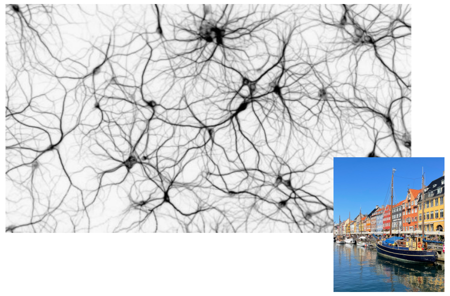
$N = 100$ observations



learning/inference process

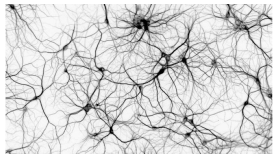
model expert on
Nyhavn data

$\mathcal{M}_{\boldsymbol{\theta}}$



The **idea** (tourist metaphor)

Nyhavn



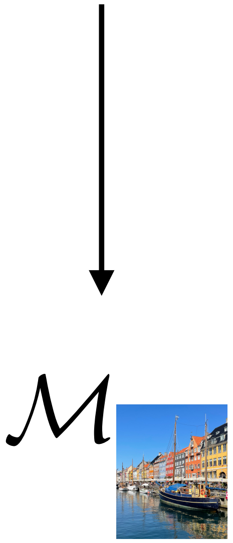
inference

$\mathcal{M}$

The **idea** (tourist metaphor)

Nyhavn

Eremitageslottet

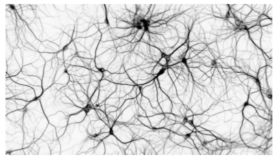


$\mathcal{M}$

The **idea** (tourist metaphor)

Nyhavn

Eremitageslottet



inference

$\mathcal{M}$

$\mathcal{M}$

The **idea** (tourist metaphor)

Nyhavn

Eremitageslottet

Amager strand

$\mathcal{M}$

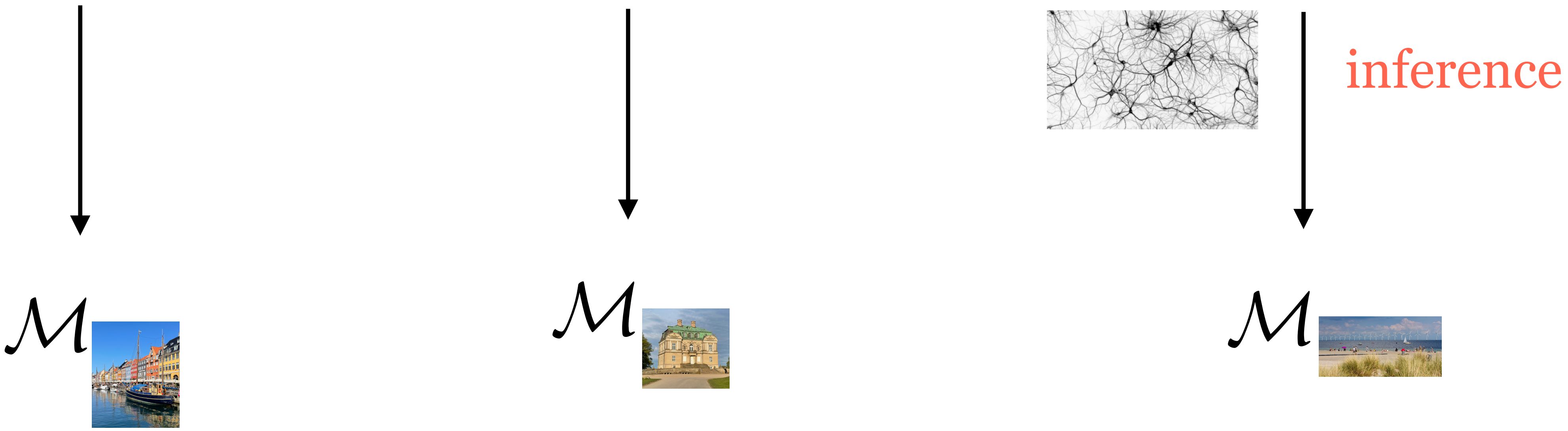
$\mathcal{M}$

The **idea** (tourist metaphor)

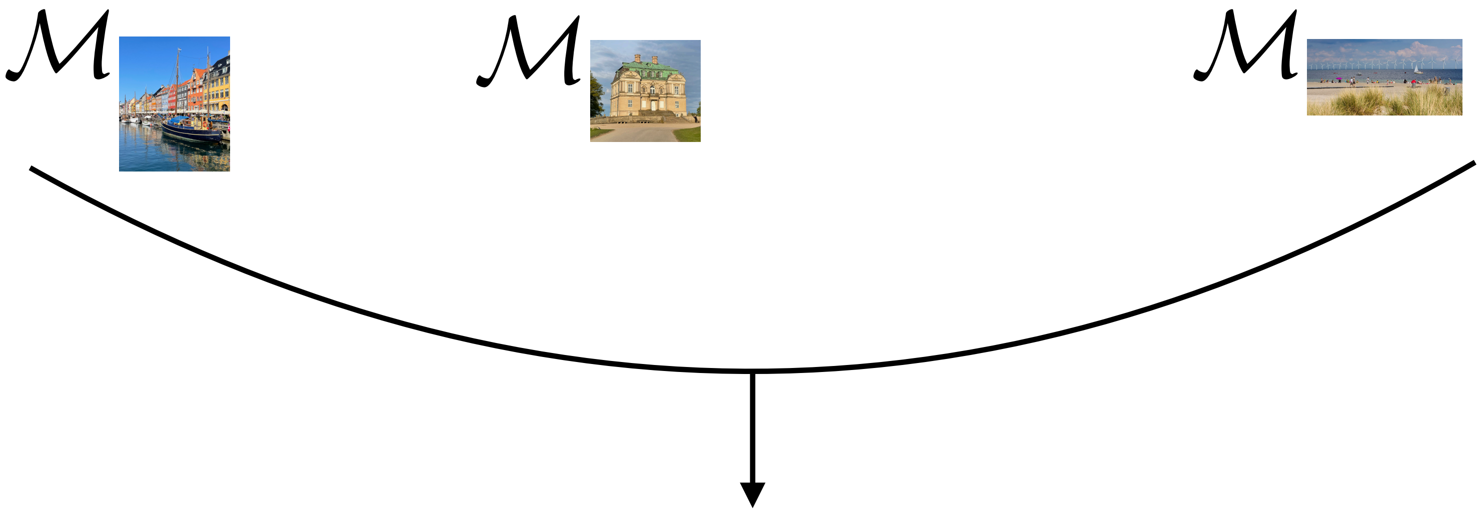
Nyhavn

Eremitageslottet

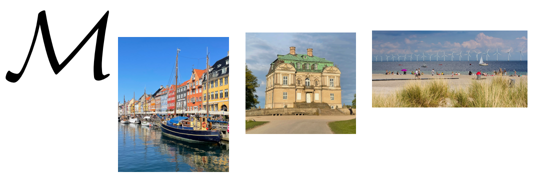
Amager strand



The **idea** (tourist metaphor)

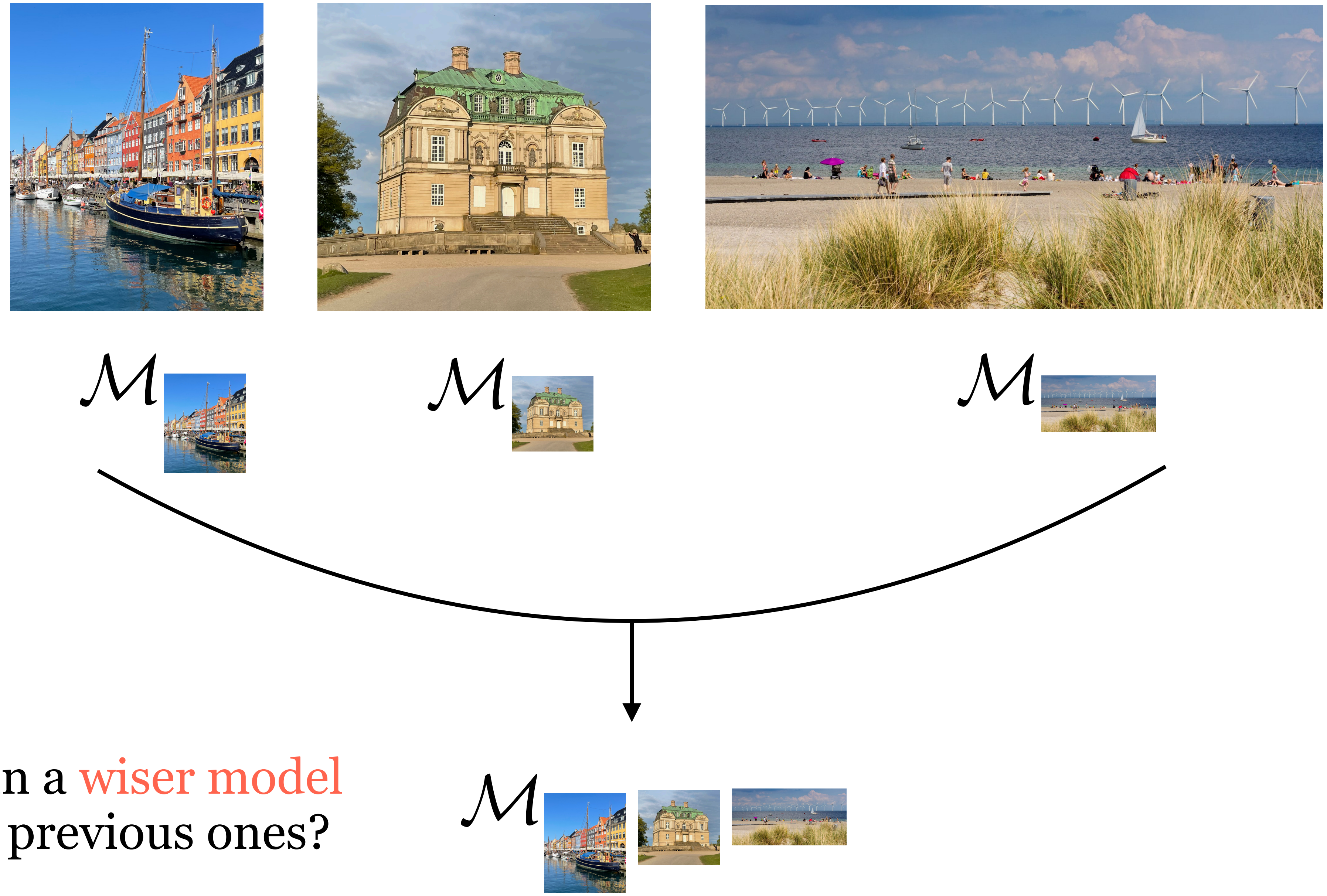


can we obtain a wiser model
from all the previous ones?



without revisiting data
(where complexity lies on)

The **idea** (tourist metaphor)

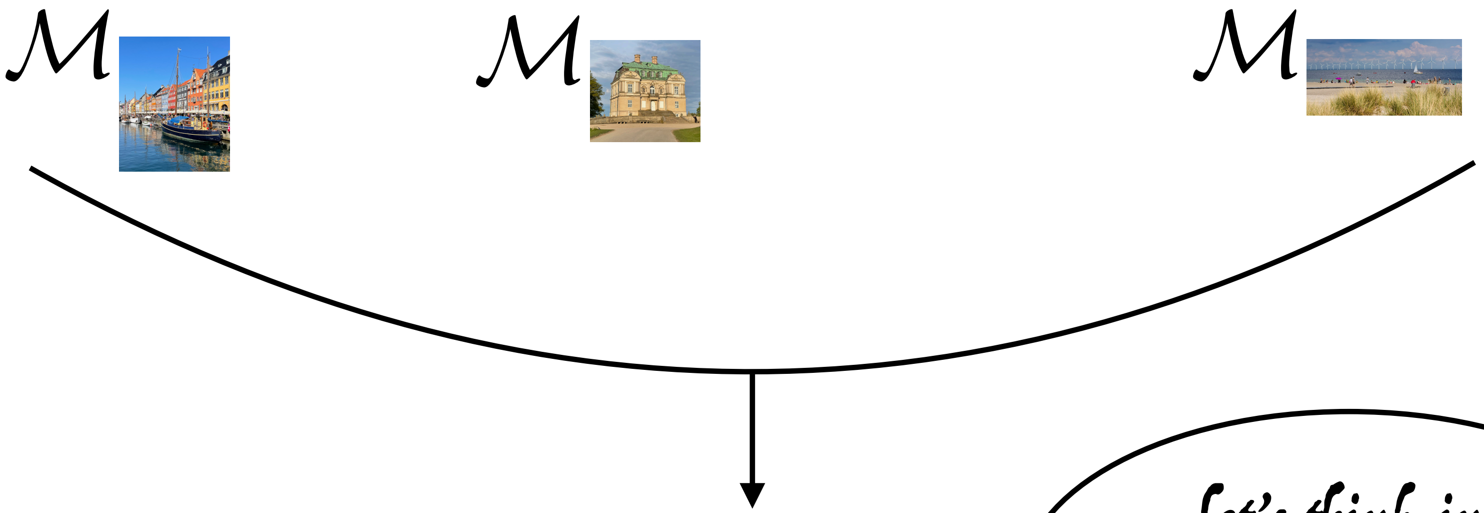


can we obtain a wiser model
from all the previous ones?

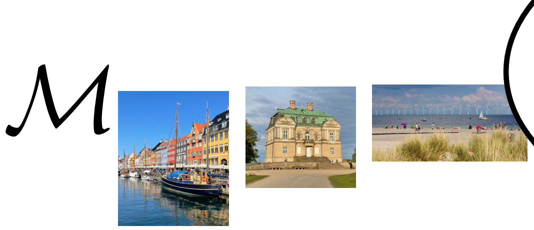
“Meta-model”

without revisiting data
(where complexity lies on)

The **idea** (tourist metaphor)



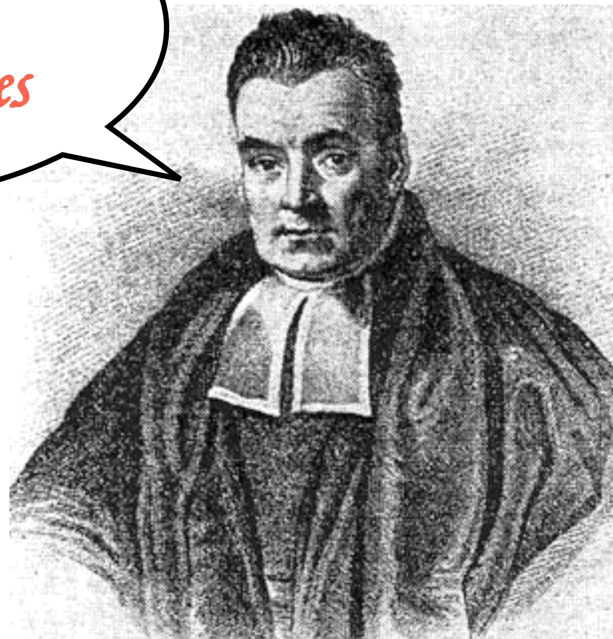
can we obtain a wiser model from all the previous ones?



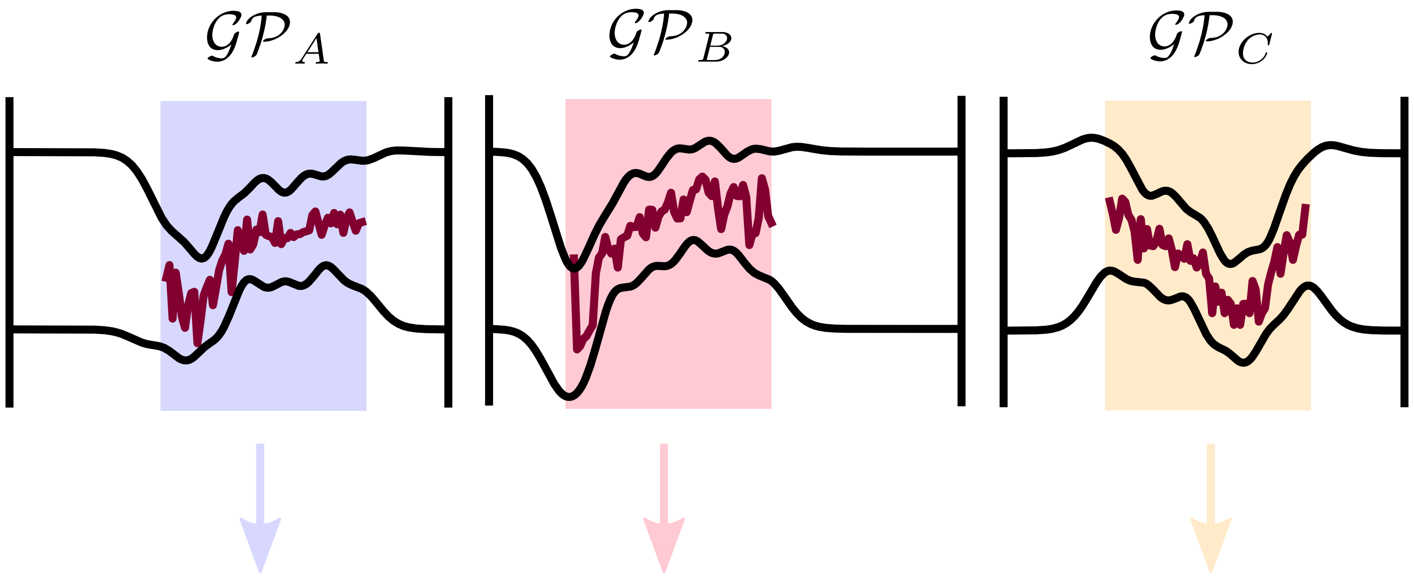
“Meta-model”

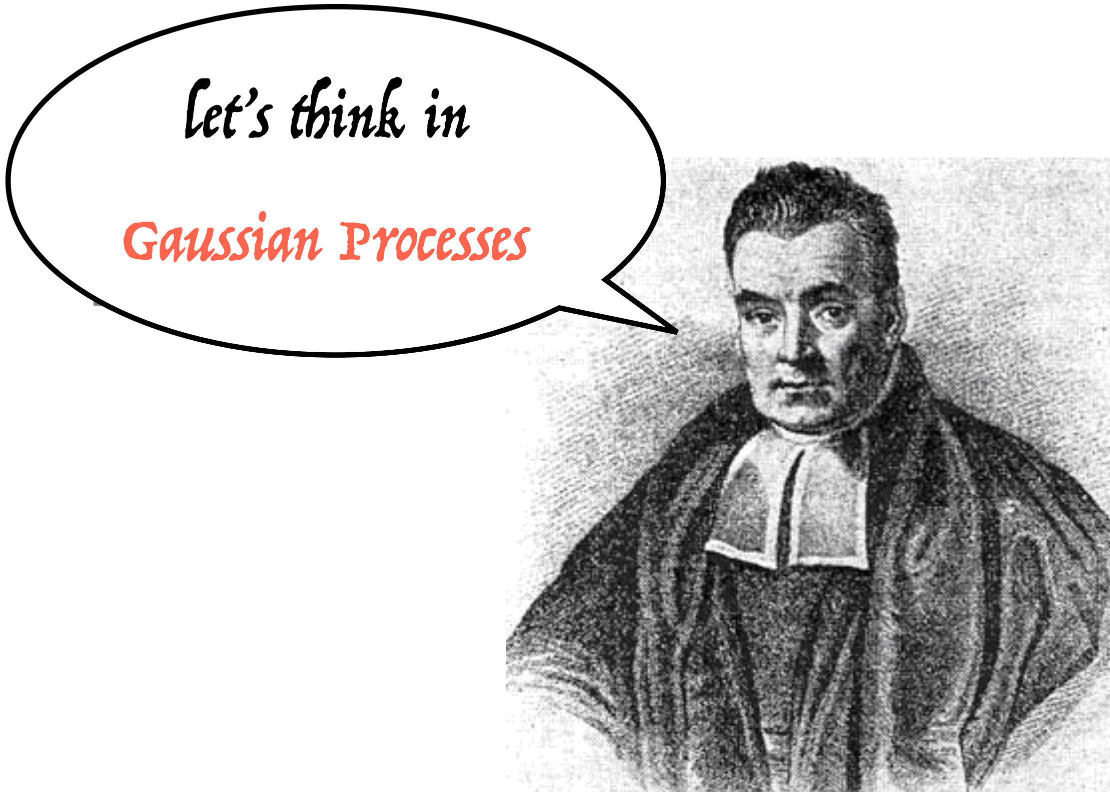
let's think in

Gaussian Processes

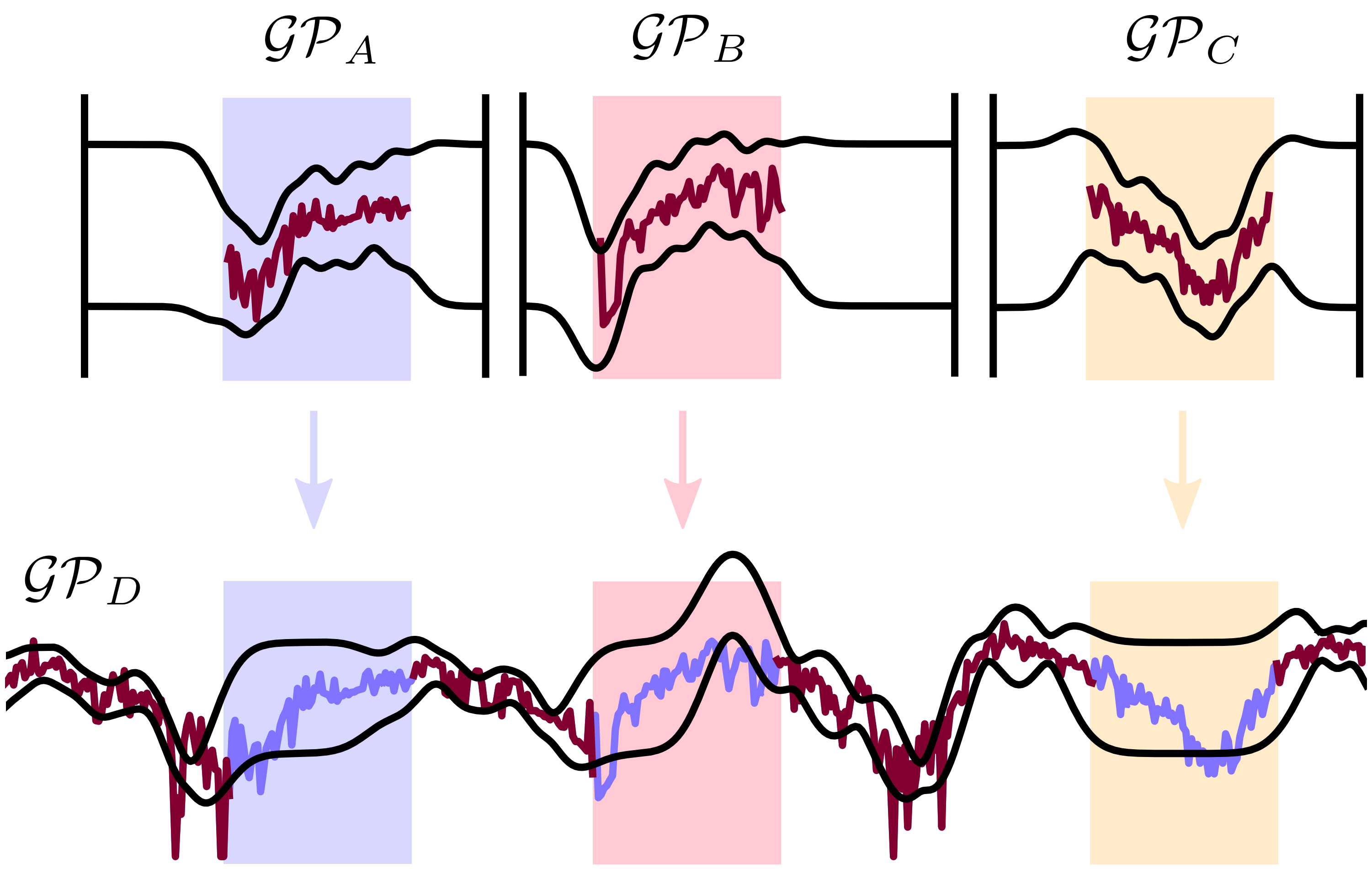


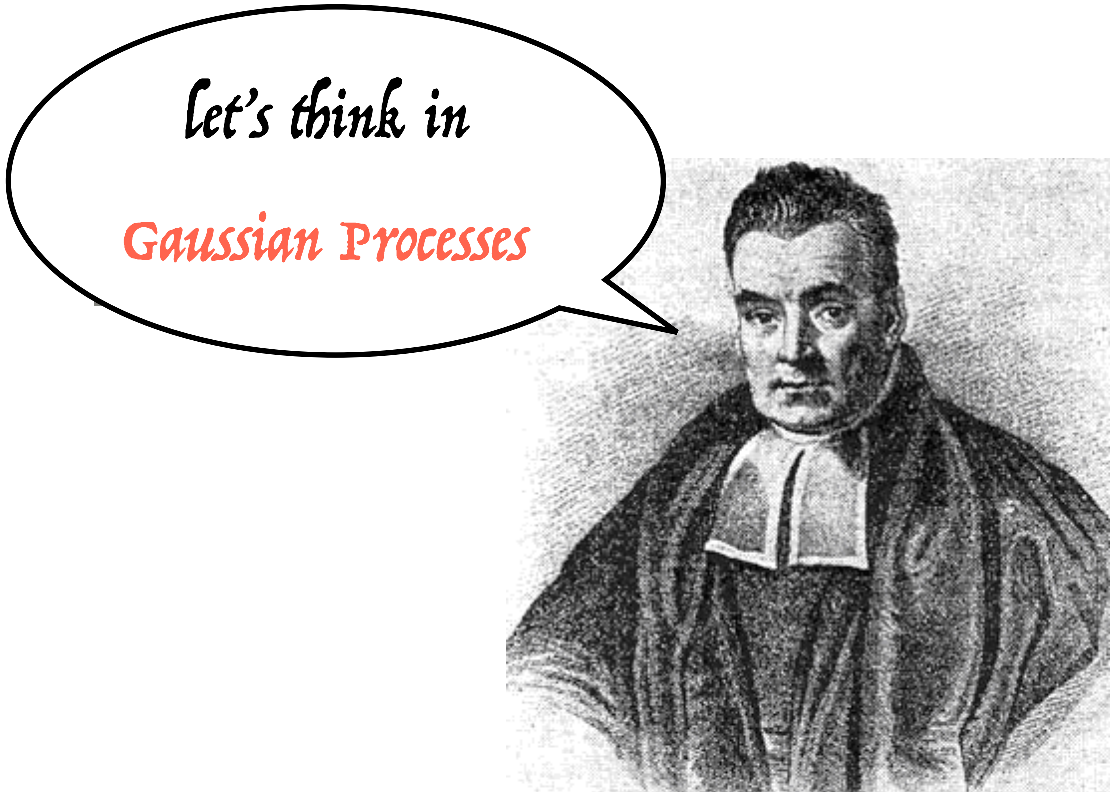
The **idea** (with GPs)



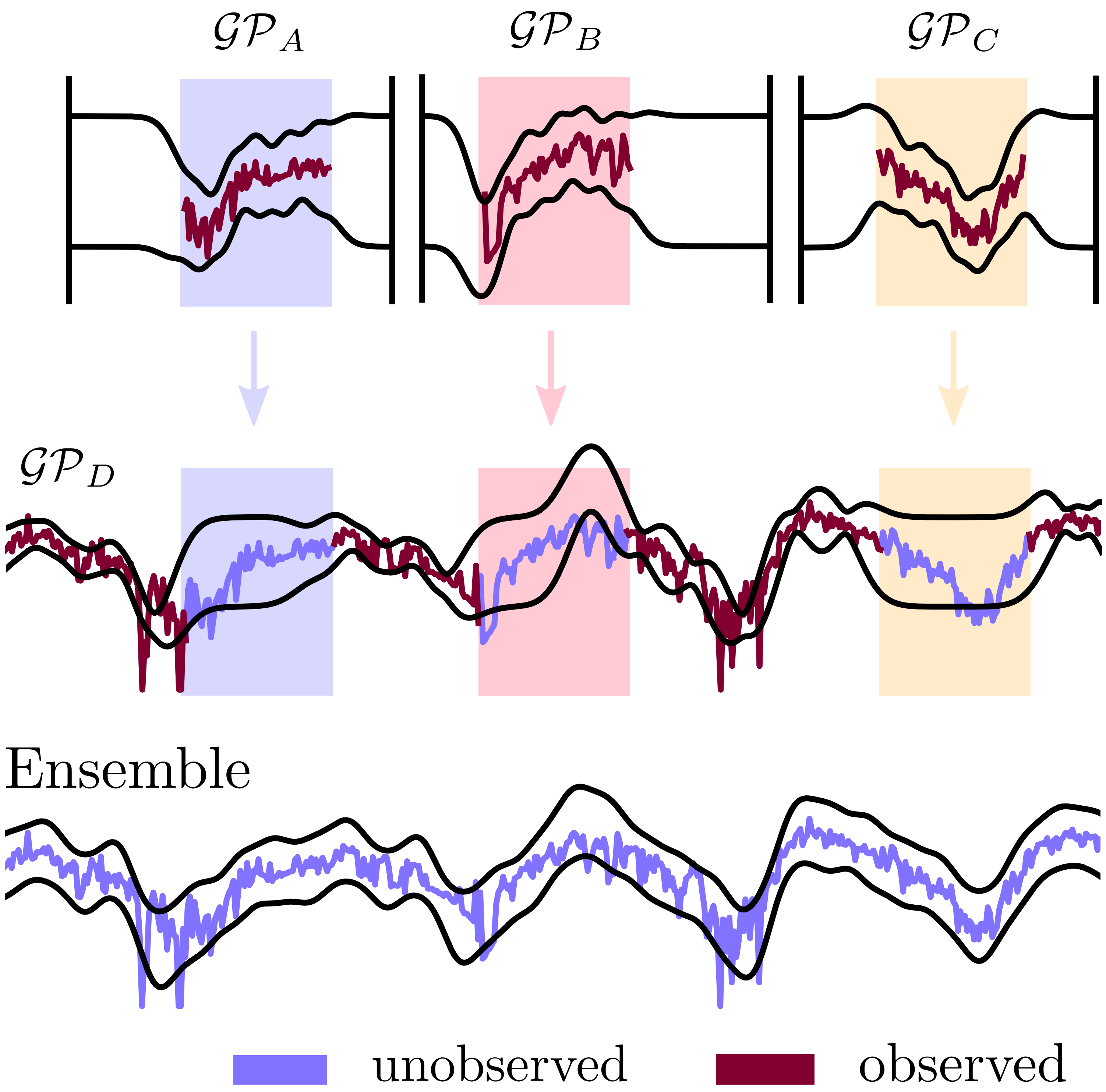


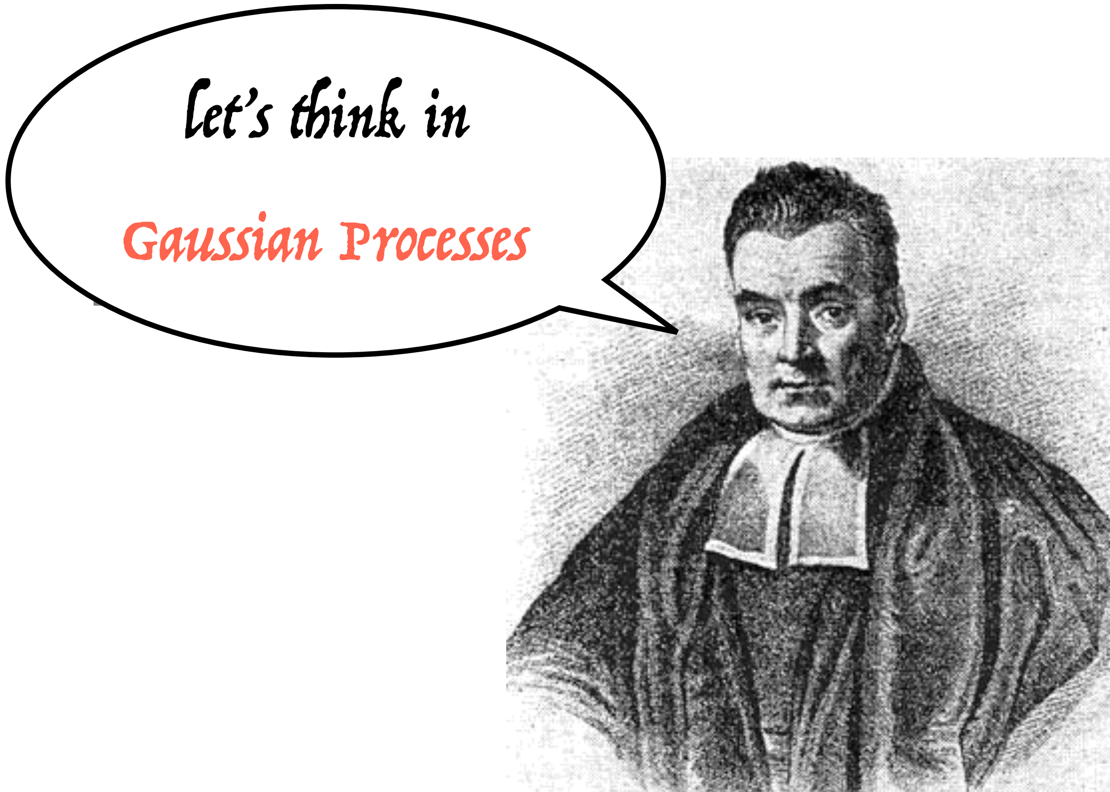
The **idea** (with GPs)





The **idea** (with GPs)





Summary index

I Gaussian processes (in a nutshell)

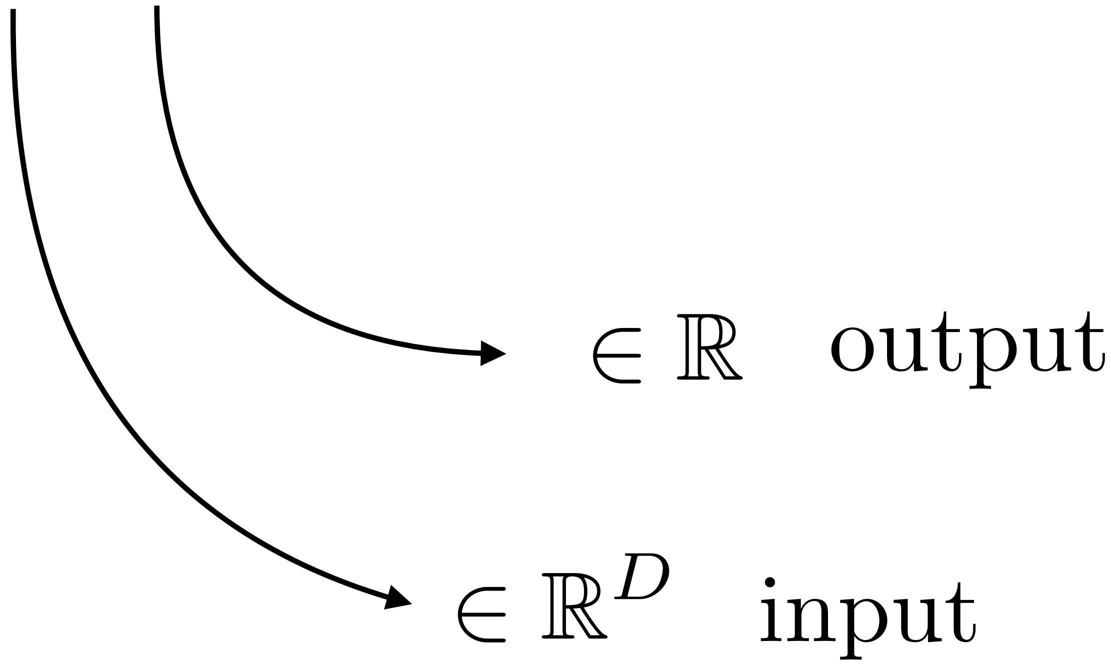
- gaussian likelihoods
- non-gaussian likelihoods
- sparse approximations

II Modular Gaussian processes

- factorisable (marginal) likelihoods
- Bayesian likelihood approximation
- lower ensemble bounds
- results

Gaussian Processes

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$



$\in \mathbb{R}$ output

$\in \mathbb{R}^D$ input

Gaussian Processes

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

Likelihood model

$$\mathbf{y}_i \sim p(\mathbf{y}_i | \theta)$$

Gaussian Processes

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$$

Likelihood model

$$\mathbf{y}_i \sim p(\mathbf{y}_i | \theta(\mathbf{x}_i))$$

Gaussian Processes

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

Classical GP model

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{y}_i | \mu, \sigma)$$

Gaussian Processes

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

Classical GP model

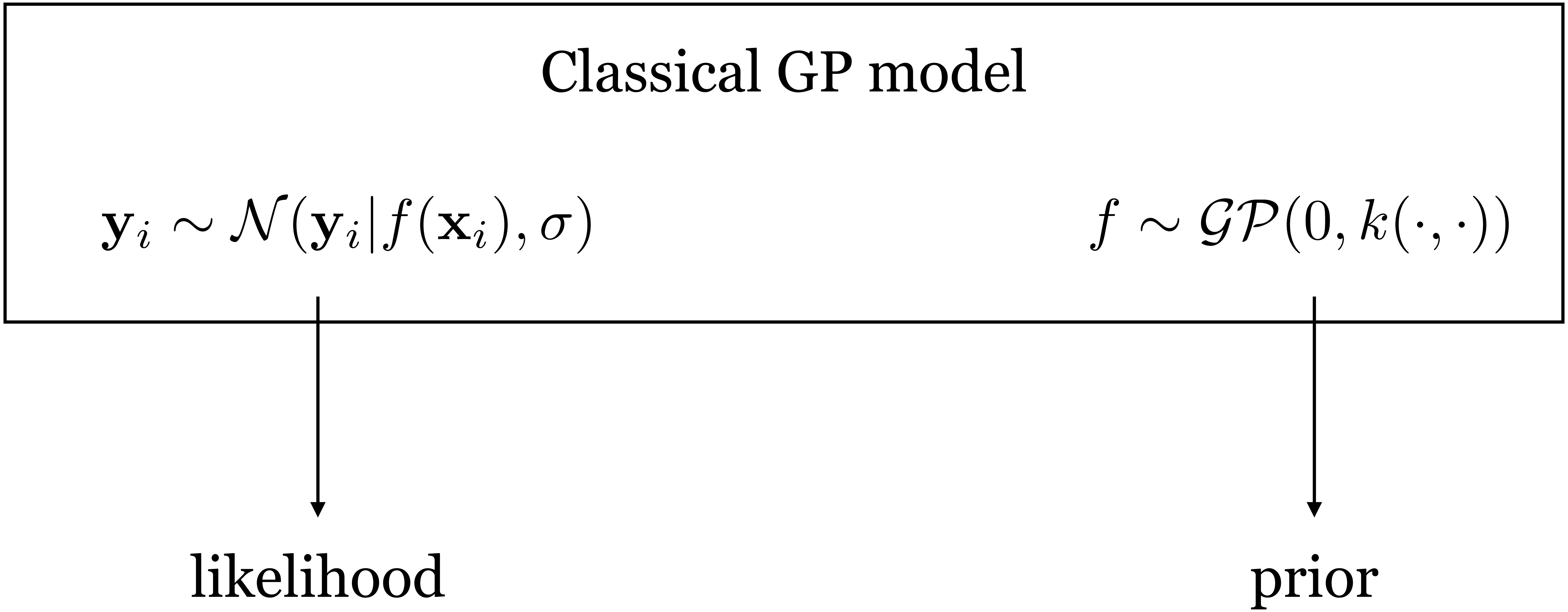
$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{y}_i | f(\mathbf{x}_i), \sigma)$$

$$\mu = f(\mathbf{x}_i)$$

non-linear function

Gaussian Processes

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$



Classical GP model

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{y}_i | f(\mathbf{x}_i), \sigma)$$

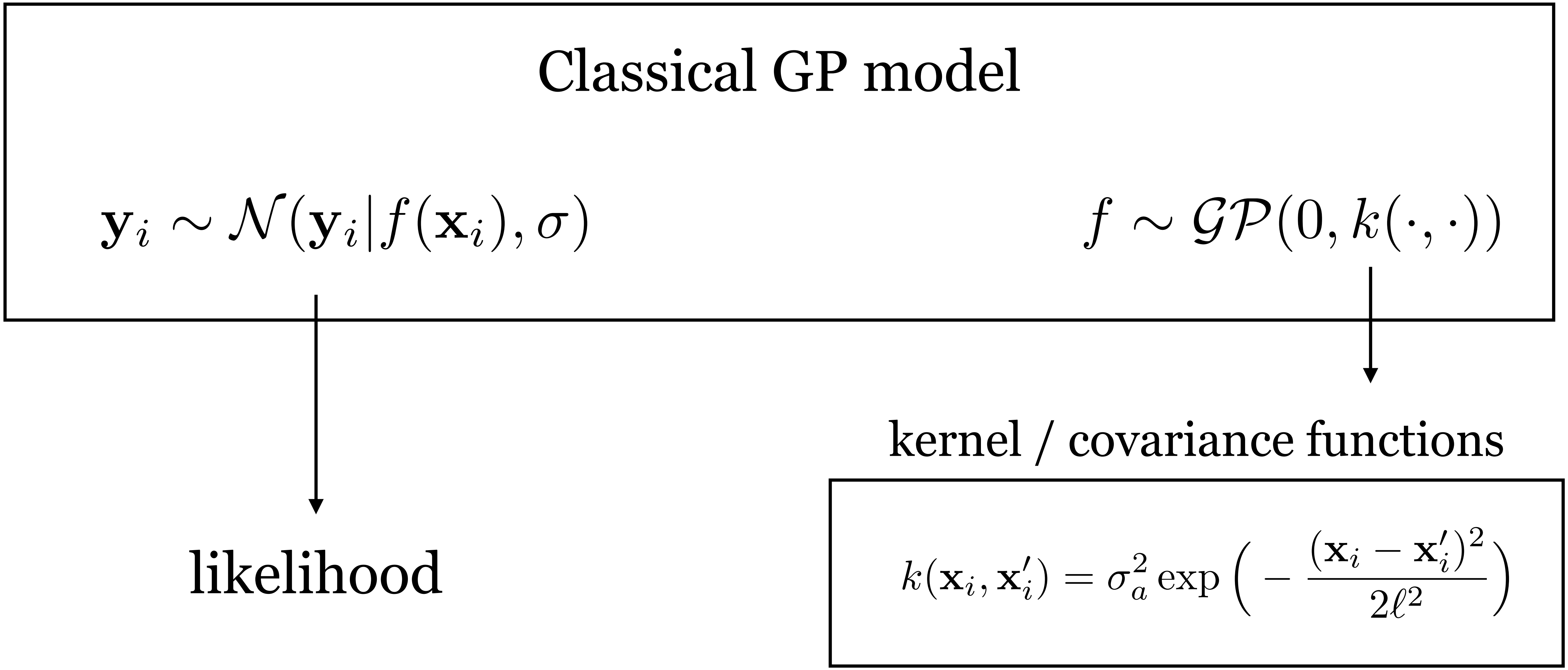
$$f \sim \mathcal{GP}(0, k(\cdot, \cdot))$$

likelihood

prior

Gaussian Processes

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$



Classical GP model

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{y}_i | f(\mathbf{x}_i), \sigma)$$

$$f \sim \mathcal{GP}(0, k(\cdot, \cdot))$$

likelihood

kernel / covariance functions

$$k(\mathbf{x}_i, \mathbf{x}_i') = \sigma_a^2 \exp\left(- \frac{(\mathbf{x}_i - \mathbf{x}_i')^2}{2\ell^2} \right)$$

Gaussian Processes

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

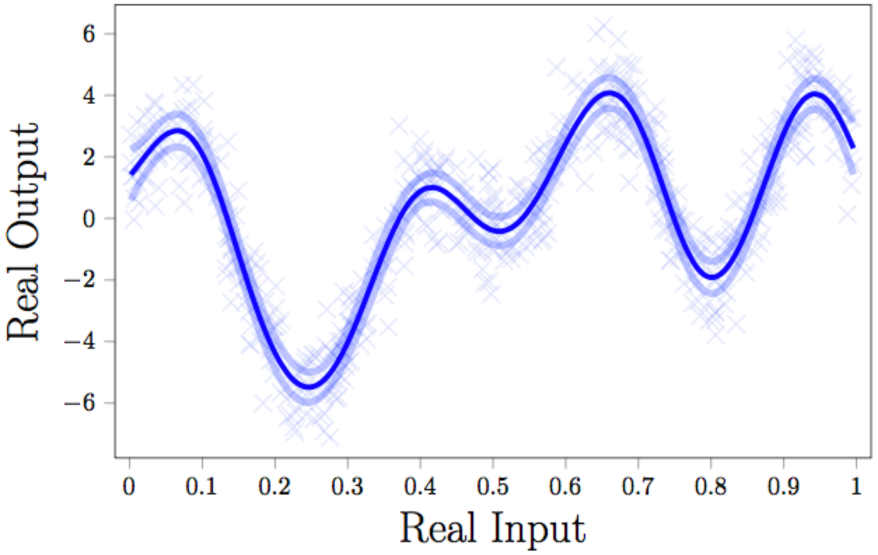
Classical GP model

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{y}_i | f(\mathbf{x}_i), \sigma) \qquad f \sim \mathcal{GP}(0, k(\cdot, \cdot))$$

$\downarrow$

kernel / covariance functions

$$k(\mathbf{x}_i, \mathbf{x}_i') = \sigma_a^2 \exp\left(-\frac{(\mathbf{x}_i - \mathbf{x}_i')^2}{2\ell^2}\right)$$



Summary index

I Gaussian processes (in a nutshell)

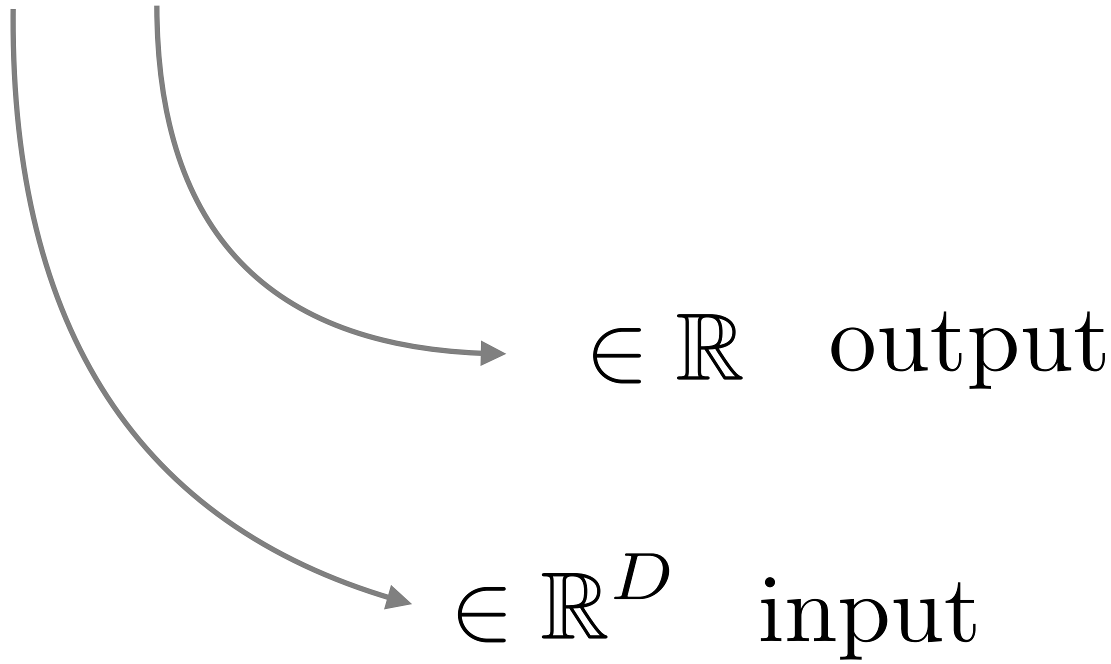
- gaussian likelihoods
- non-gaussian likelihoods
- sparse approximations

II Modular Gaussian processes

- factorisable (marginal) likelihoods
- Bayesian reconstruction “trick”
- lower ensemble bounds
- results

Non-Gaussian Likelihoods

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

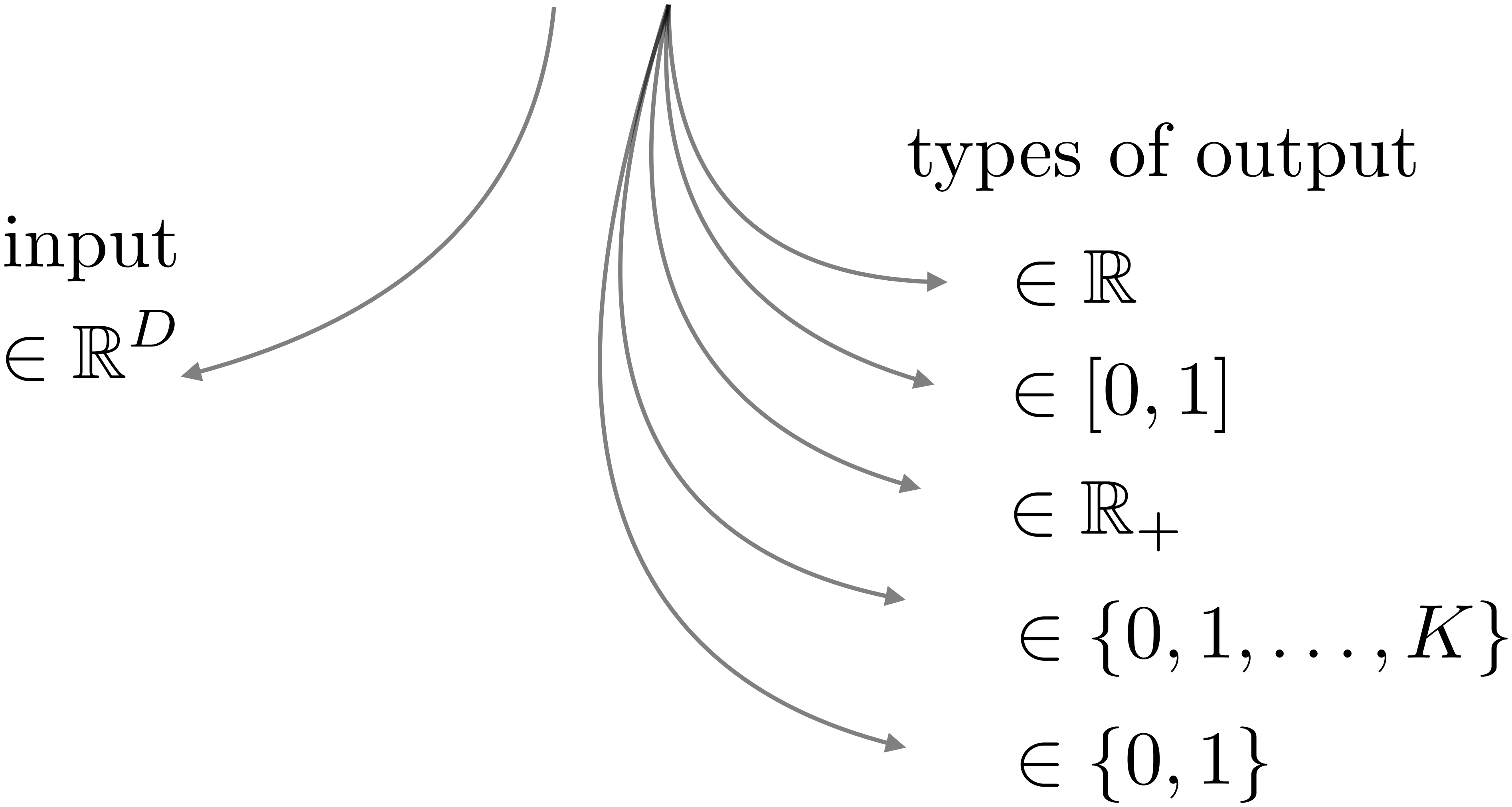


$\in \mathbb{R}$ output

$\in \mathbb{R}^D$ input

Non-Gaussian Likelihoods

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$



input

$\in \mathbb{R}^D$

types of output

$\in \mathbb{R}$

$\in [0, 1]$

$\in \mathbb{R}_+$

$\in \{0, 1, \ldots, K\}$

$\in \{0, 1\}$

Non-Gaussian Likelihoods

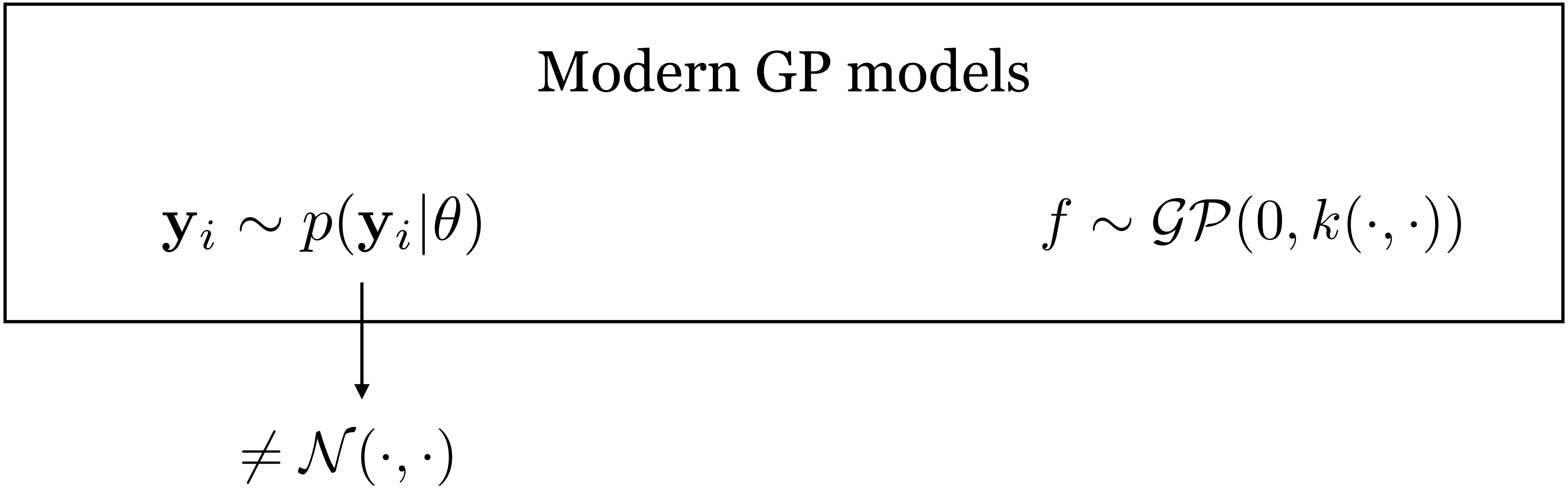
$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

Modern GP models

$$\mathbf{y}_i \sim p(\mathbf{y}_i|\theta) \qquad f \sim \mathcal{GP}(0, k(\cdot, \cdot))$$

Non-Gaussian Likelihoods

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$



Modern GP models

$$\mathbf{y}_i \sim p(\mathbf{y}_i | \theta)$$

$$f \sim \mathcal{GP}(0, k(\cdot, \cdot))$$

$$\neq \mathcal{N}(\cdot, \cdot)$$

Non-Gaussian Likelihoods

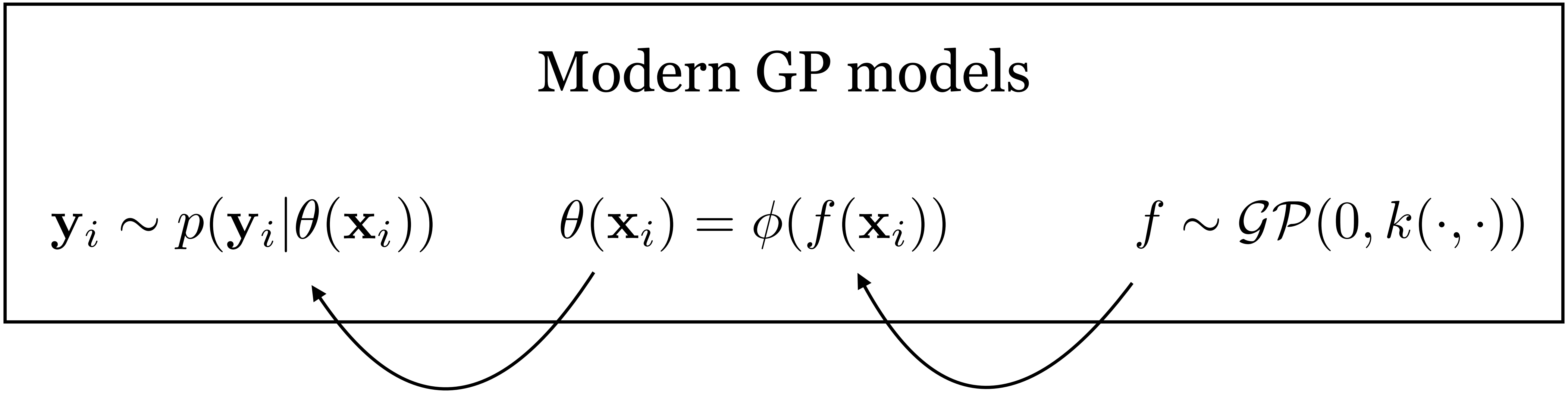
$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

Modern GP models

$$\begin{tikzcd}
\mathbf{y}_i \sim p(\mathbf{y}_i|\theta) & \theta = \phi(f) \arrow[l, bend left] & f \sim \mathcal{GP}(0, k(\cdot, \cdot)) \arrow[l, bend left]
\end{tikzcd}$$

Non-Gaussian Likelihoods

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$



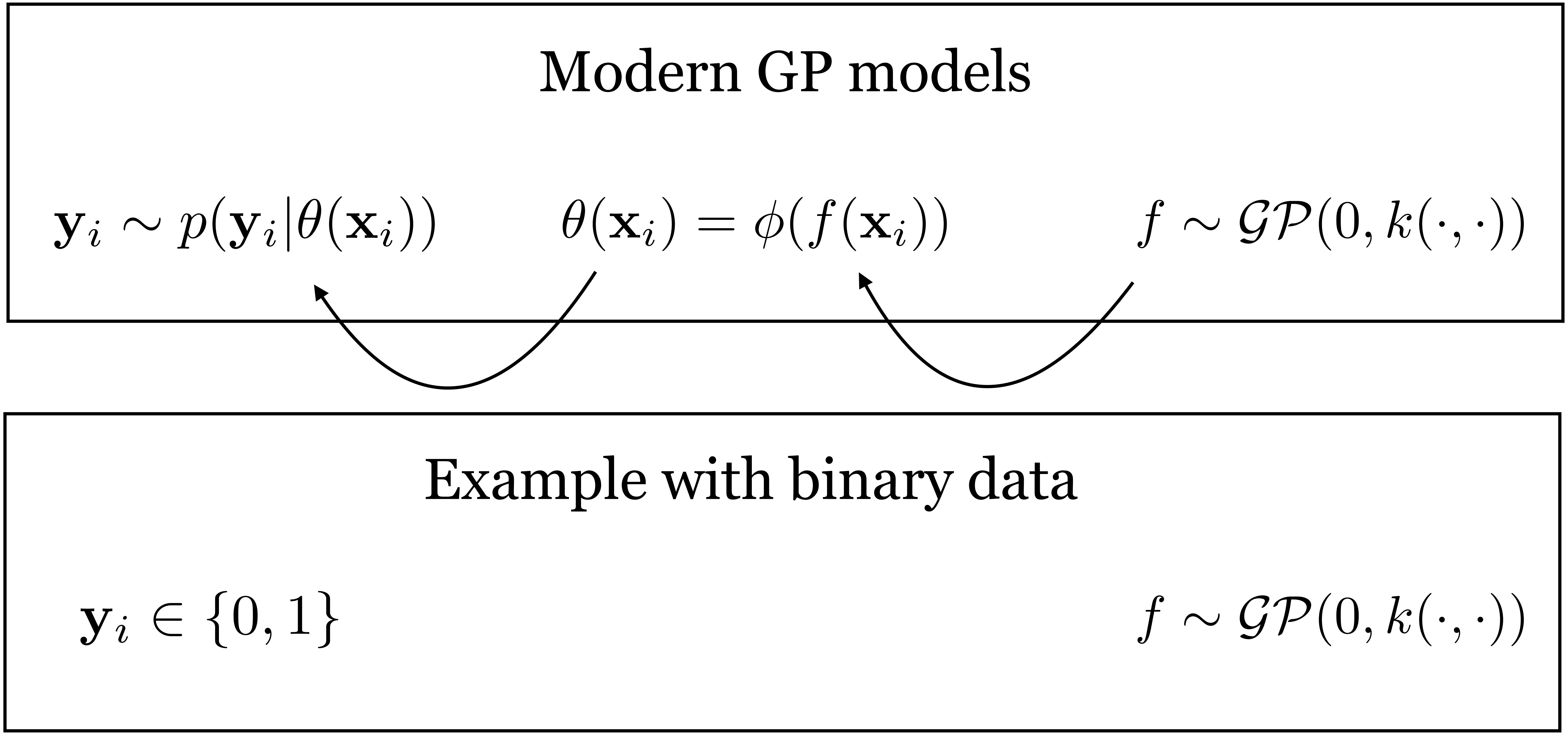
Modern GP models

$$\mathbf{y}_i \sim p(\mathbf{y}_i|\theta(\mathbf{x}_i)) \qquad \theta(\mathbf{x}_i) = \phi(f(\mathbf{x}_i)) \qquad f \sim \mathcal{GP}(0, k(\cdot,\cdot))$$

non-linear mappings
(linking functions)

Non-Gaussian Likelihoods

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$



Modern GP models

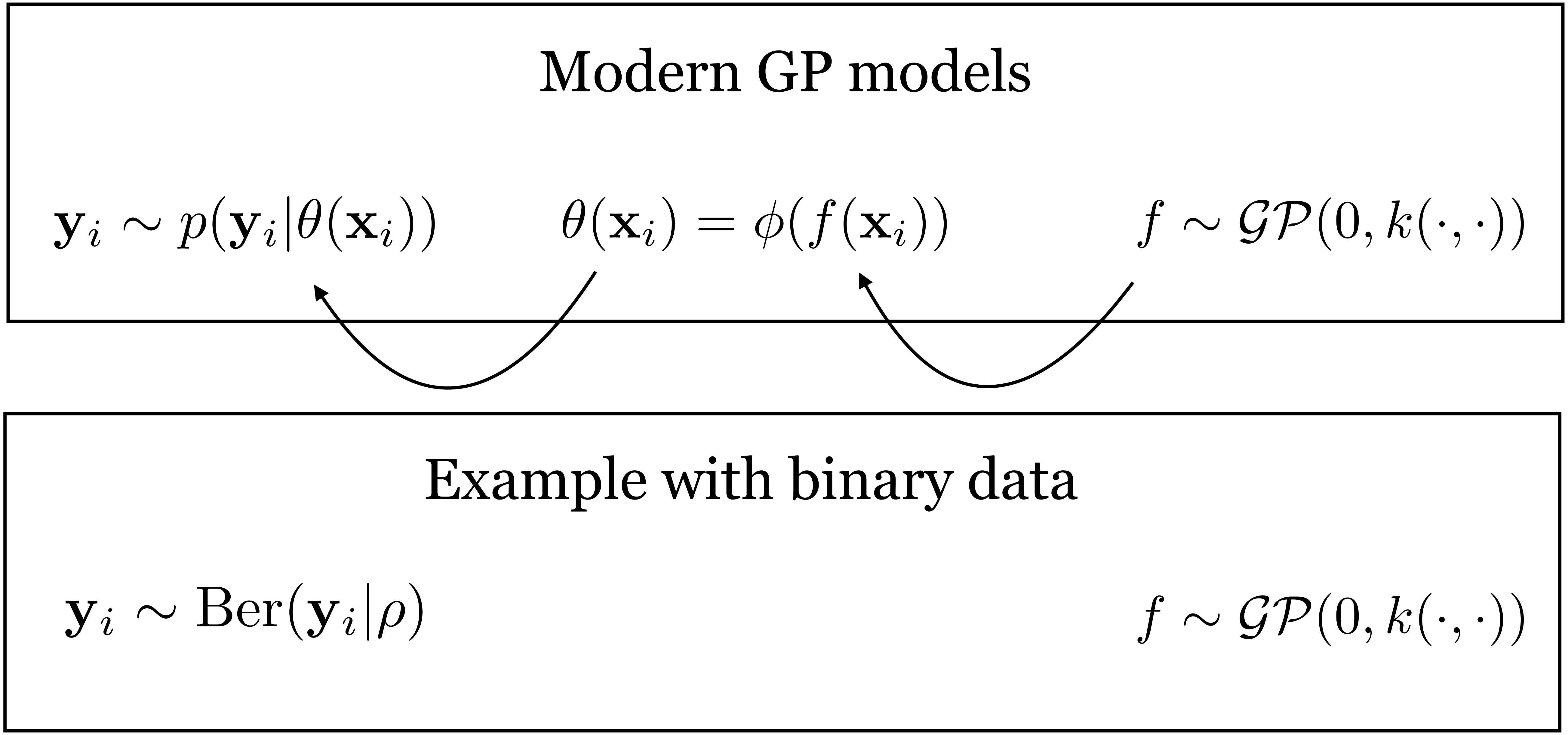
$\mathbf{y}_i \sim p(\mathbf{y}_i | \theta(\mathbf{x}_i))$ $\theta(\mathbf{x}_i) = \phi(f(\mathbf{x}_i))$ $f \sim \mathcal{GP}(0, k(\cdot, \cdot))$

Example with binary data

$\mathbf{y}_i \in \{0, 1\}$ $f \sim \mathcal{GP}(0, k(\cdot, \cdot))$

Non-Gaussian Likelihoods

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$$



Modern GP models

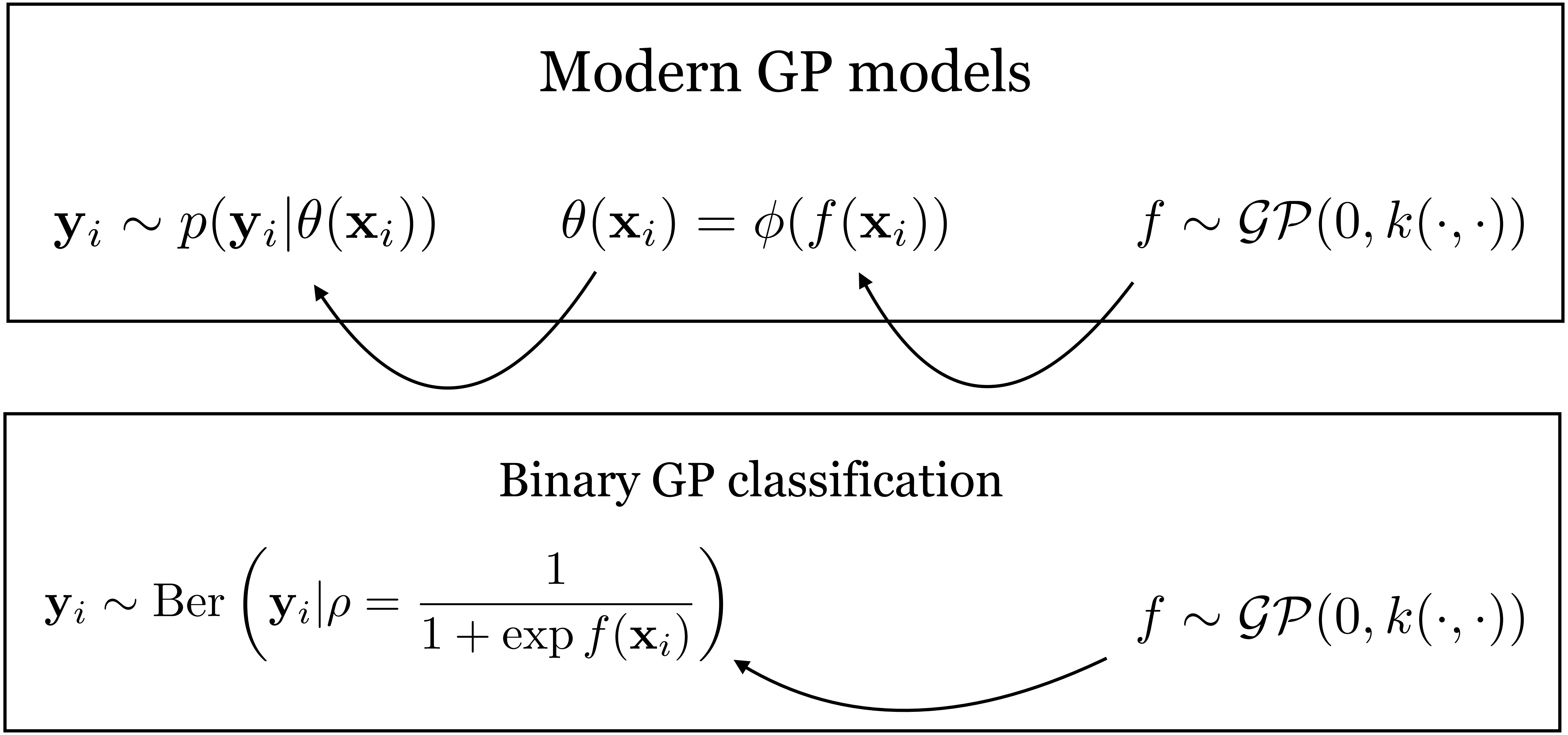
$$\mathbf{y}_i \sim p(\mathbf{y}_i | \theta(\mathbf{x}_i)) \qquad \theta(\mathbf{x}_i) = \phi(f(\mathbf{x}_i)) \qquad f \sim \mathcal{GP}(0, k(\cdot, \cdot))$$

Example with binary data

$$\mathbf{y}_i \sim \text{Ber}(\mathbf{y}_i | \rho) \qquad f \sim \mathcal{GP}(0, k(\cdot, \cdot))$$

Non-Gaussian Likelihoods

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$



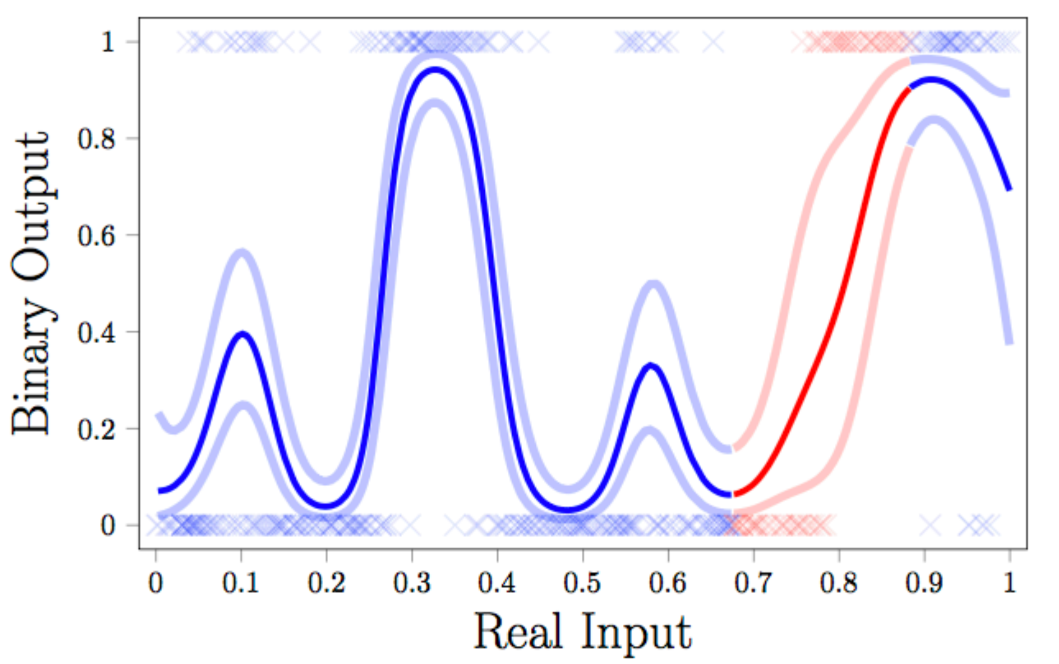
Modern GP models

$$\mathbf{y}_i \sim p(\mathbf{y}_i | \theta(\mathbf{x}_i)) \qquad \theta(\mathbf{x}_i) = \phi(f(\mathbf{x}_i)) \qquad f \sim \mathcal{GP}(0, k(\cdot,\cdot))$$

Binary GP classification

$$\mathbf{y}_i \sim \mathrm{Ber}\left(\mathbf{y}_i | \rho = \frac{1}{1 + \exp f(\mathbf{x}_i)}\right) \qquad f \sim \mathcal{GP}(0, k(\cdot,\cdot))$$

Non-Gaussian Likelihoods



$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

Modern GP models

$$\mathbf{y}_i \sim p(\mathbf{y}_i | \theta(\mathbf{x}_i)) \qquad \theta(\mathbf{x}_i) = \phi(f(\mathbf{x}_i)) \qquad f \sim \mathcal{GP}(0, k(\cdot,\cdot))$$

Binary GP classification

$$\mathbf{y}_i \sim \text{Ber}\left(\mathbf{y}_i | \rho = \frac{1}{1 + \exp f(\mathbf{x}_i)}\right) \qquad f \sim \mathcal{GP}(0, k(\cdot,\cdot))$$

Non-Gaussian Likelihoods

I

Three important contributions

M. Lázaro-Gredilla and M. K. Titsias
Variational Heteroscedastic Gaussian Process Regression
In International Conference in Machine Learning (ICML), 2011

$$\mathbf{y} \sim \mathcal{N}(\mathbf{y}|\mu = f(\mathbf{x}), \sigma = e^{g(\mathbf{x})})$$

J. Hensman, A. G. de G. Matthews and Z. Ghahramani
Scalable Variational Gaussian Process Classification
In Artificial Intelligence and Statistics (AISTATS), 2015

$$\mathbf{y} \sim \mathrm{Ber}(\mathbf{y}|\rho = \phi(f(\mathbf{x})))$$

A. D. Saul, J. Hensman, A. Vehtari and N. D. Lawrence
Chained Gaussian Processes
In Artificial Intelligence and Statistics (AISTATS), 2016

$$\mathbf{y} \sim \mathrm{Poisson}(\mathbf{y}|\lambda = \exp(f(\mathbf{x}) + g(\mathbf{x})))$$

Summary index

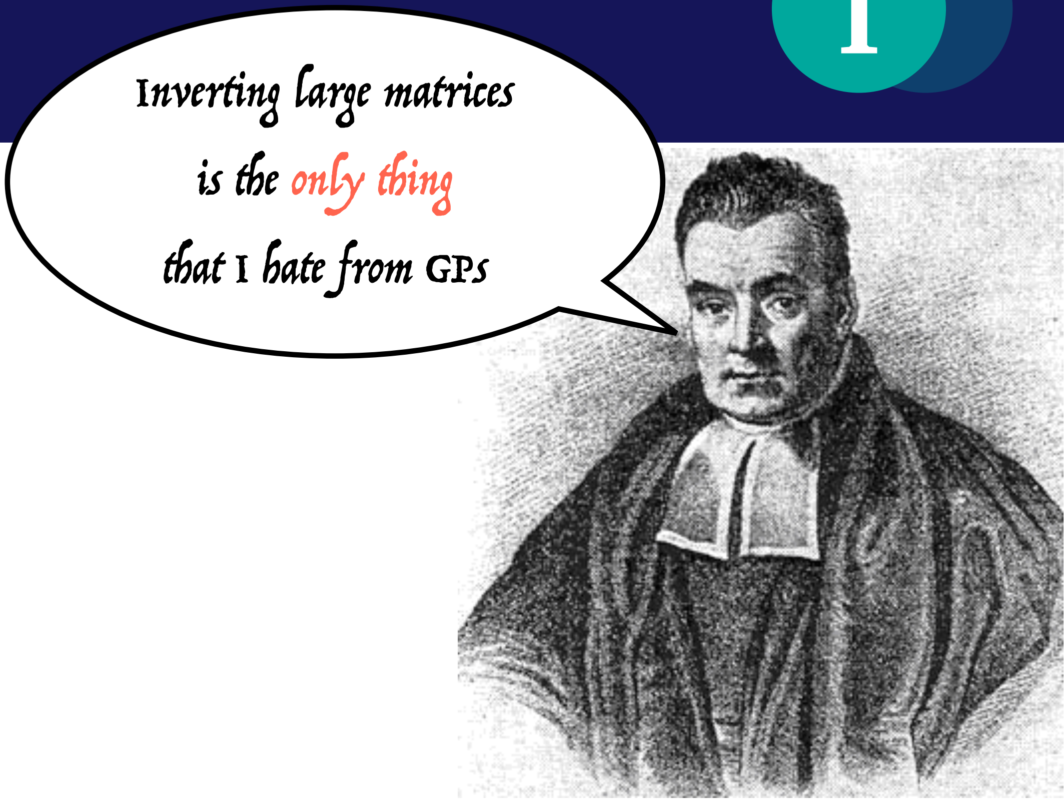
I Gaussian processes (in a nutshell)

- gaussian likelihoods
- non-gaussian likelihoods
- sparse approximations

II Modular Gaussian processes

- factorisable (marginal) likelihoods
- Bayesian likelihood approximation
- lower ensemble bounds
- results

Complexity problem



$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

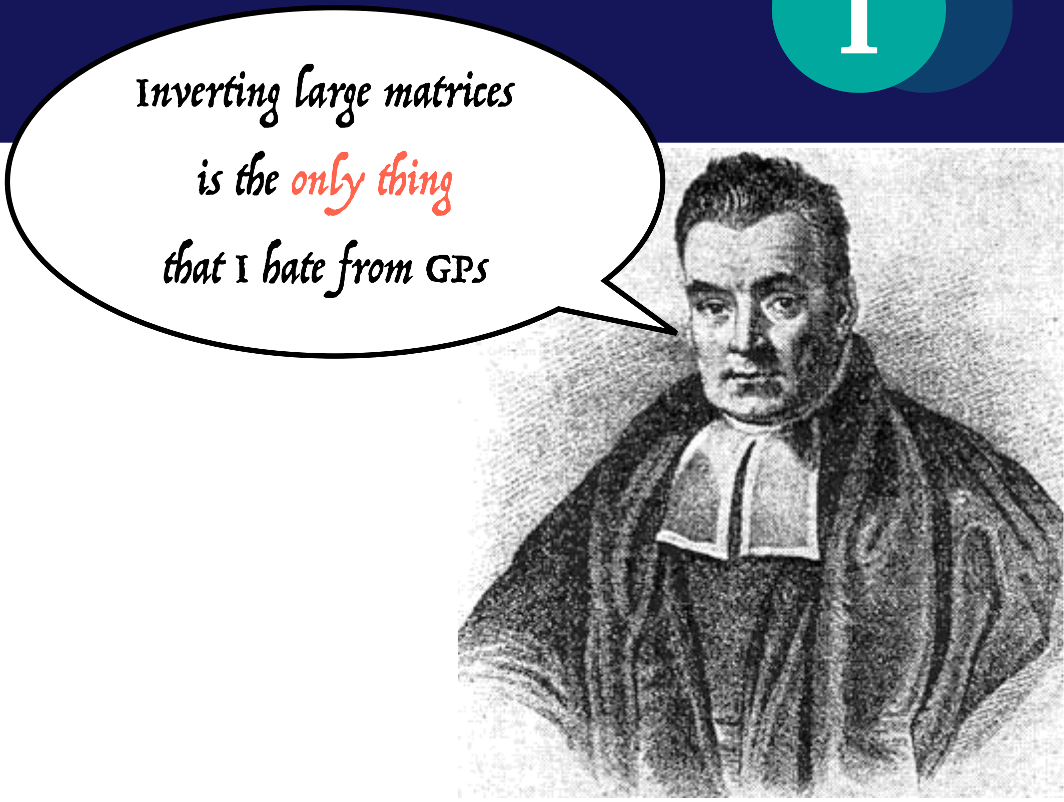
why?

$$\begin{tikzcd} p(f|\mathcal{D}) & \int p(\mathbf{y}_i|f(\mathbf{x}_i))p(f(\mathbf{x}_i))df(\mathbf{x}_i) \arrow[l, bend left, "\Sigma^{-1}"', "\mathcal{O}(N^3)"] \end{tikzcd}$$

marginal likelihood integral

posterior inference of the underlying GP function

Complexity problem


Inverting large matrices is the only thing that I hate from GPs

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

why?

$$p(f|\mathcal{D}) \xleftarrow[\mathcal{O}(N^3)]{\cancel{\Sigma^{-1}}} \int p(\mathbf{y}_i|f(\mathbf{x}_i))p(f(\mathbf{x}_i))df(\mathbf{x}_i)$$

marginal likelihood integral

posterior inference of the underlying GP function

Sparse Gaussian Processes

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

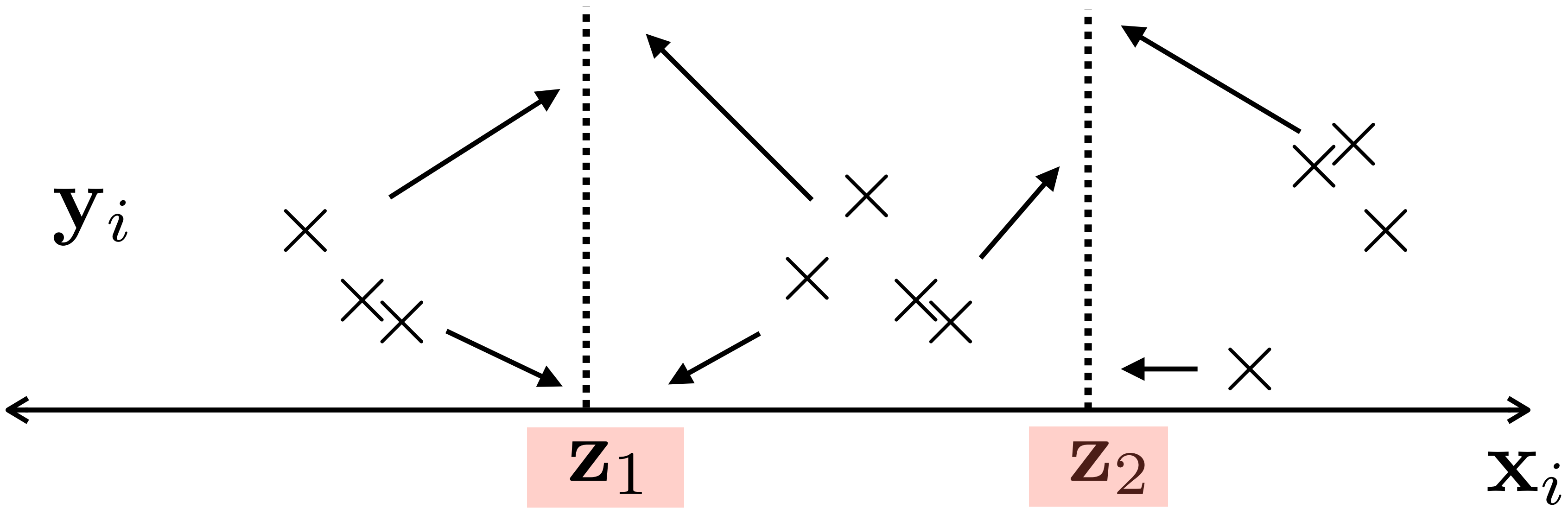
Modern GP model

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{y}_i | f(\mathbf{x}_i), \sigma) \qquad f \sim \mathcal{GP}(0, k(\cdot, \cdot))$$

seems equal but..

Sparse Gaussian Processes

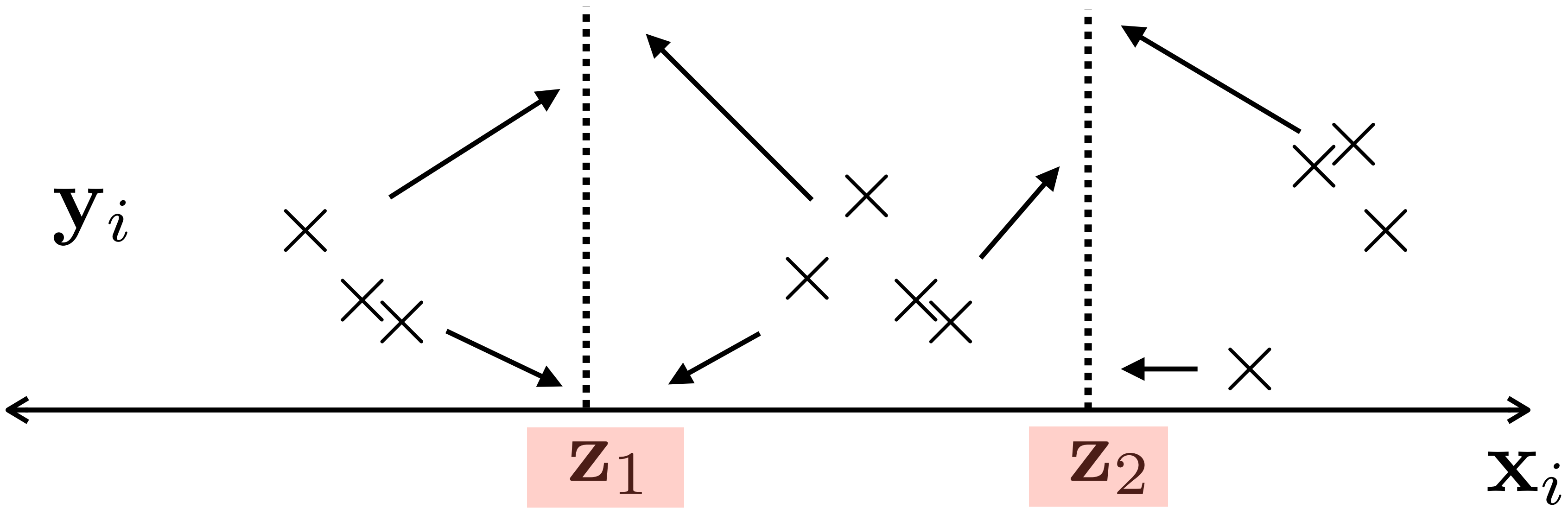
$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$



conditioning is power!

Sparse Gaussian Processes

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$



Notation

$$\mathbf{u} = f(\mathbf{z})$$

$$\mathbf{f} = f(\mathbf{x})$$

Sparse Gaussian Processes

Before

$$\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})d\mathbf{f}$$

marginal likelihood integral

Sparse Gaussian Processes

Now

$$\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{f}d\mathbf{u}$$

marginal likelihood integral

$$p(\mathbf{f}|\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K}_{\mathbf{fu}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{u}, \mathbf{K}_{\mathbf{ff}} - \mathbf{K}_{\mathbf{fu}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{uf}}^{\top})$$

Gaussian conditional

$$\mathcal{O}(NM^2)$$

$$M \ll N$$

Sparse Gaussian Processes

Variational inference

Our (new) goal

$$q(f,u) \approx p(f,u|\mathcal{D})$$ ✓

Now

$$\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{u})p(\mathbf{u})d\mathbf{f}d\mathbf{u}$$

$$p(\mathbf{f}|\mathbf{u}) = \mathcal{N}(\mathbf{f}|\mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{u}, \mathbf{K_{ff}} - \mathbf{K_{fu}}\mathbf{K_{uu}^{-1}}\mathbf{K_{uf}^{\top}})$$

Gaussian conditional

$$\mathcal{O}(NM^2)$$

$$M \ll N$$

Sparse Gaussian Processes

Data

$$\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{N}$$

Model

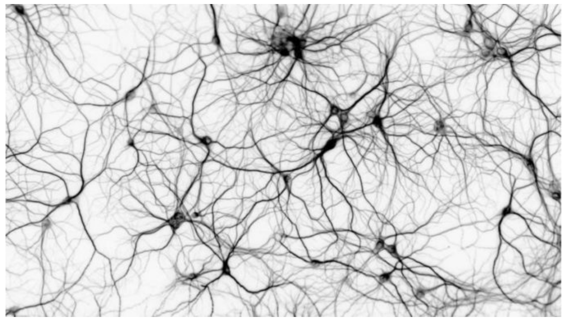
$\mathcal{M}$

$$\mathbf{y}_i \sim \mathcal{N}(\mathbf{y}_i | f(\mathbf{x}_i), \sigma)$$

$$f \sim \mathcal{GP}(0, k(\cdot, \cdot))$$

Inference

$$q(f, u) \approx p(f, u | \mathcal{D})$$



Summary index

I Gaussian processes (in a nutshell)

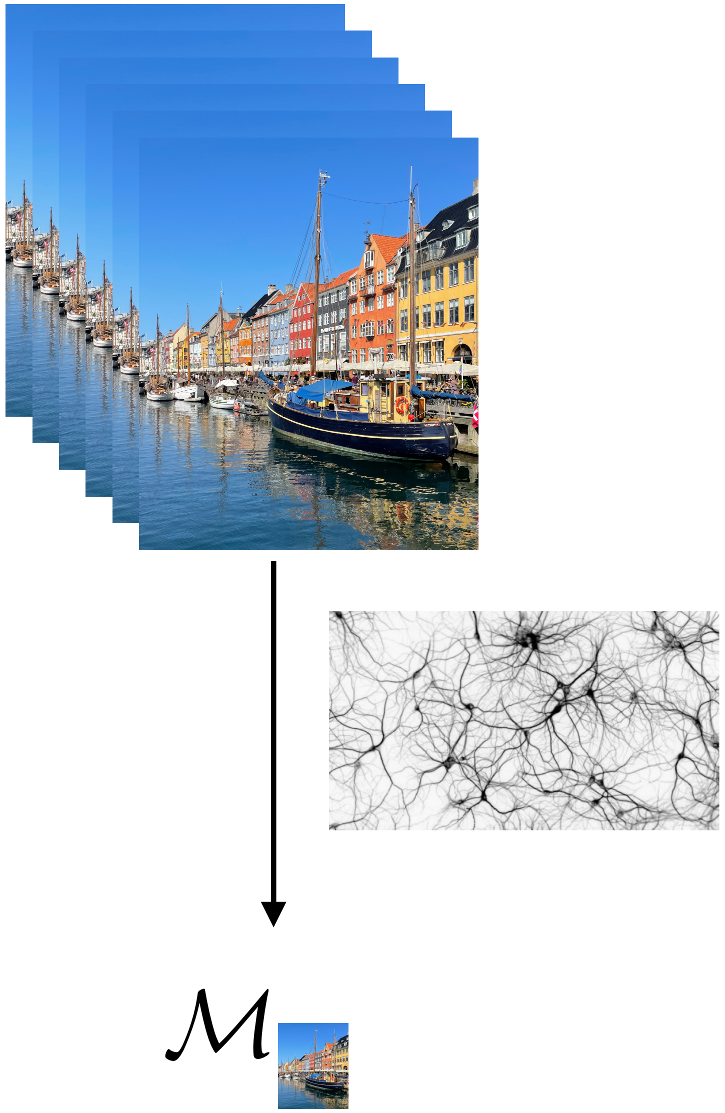
- gaussian likelihoods
- non-gaussian likelihoods
- sparse approximations

II Modular Gaussian processes

- factorisable (marginal) likelihoods
- Bayesian likelihood approximation
- lower ensemble bounds
- results

Modular Gaussian Processes

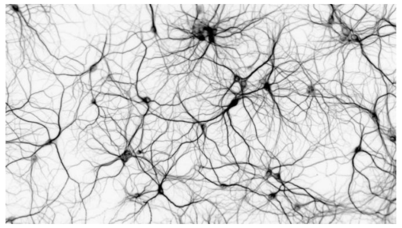
coming back to the metaphor

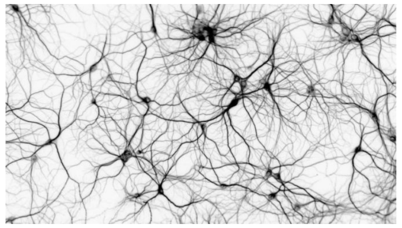


Modular Gaussian Processes

coming back to the metaphor

$$\mathcal{D}_k = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_k}$$





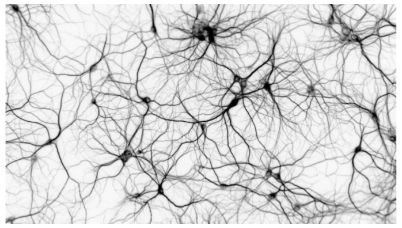
$\mathcal{M}$

$$\mathcal{M}_k = \{\boldsymbol{\phi}_k, \boldsymbol{\psi}_k, \boldsymbol{Z}_k\}$$

parameters

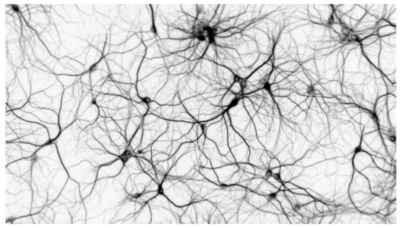
Modular Gaussian Processes

coming back to the metaphor



$\mathcal{M}$

$$\mathcal{D}_k = \{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1}^{N_k}$$



$$\mathcal{M}_k = \{\boldsymbol{\phi}_k, \boldsymbol{\psi}_k, \boldsymbol{Z}_k\}$$

"module"

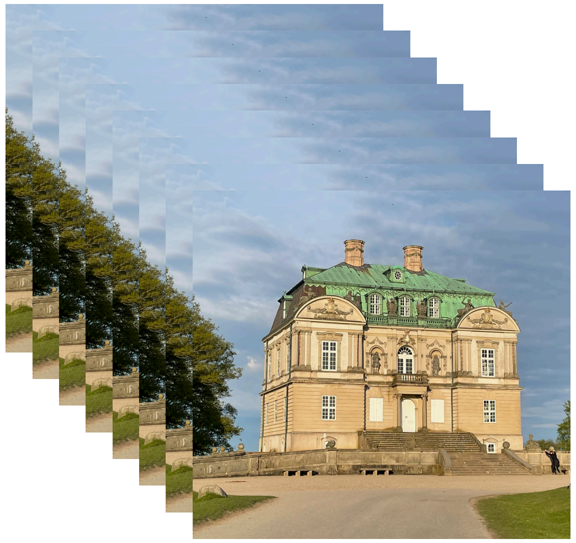
$\boldsymbol{\phi}_k$ — variational parameters

$\boldsymbol{\psi}_k$ — kernel hyperparameters

$\boldsymbol{u}_k, \boldsymbol{Z}_k$ — inducing points

Modular Gaussian Processes

doing these learning processes independently



$\mathcal{M}_1 = \{\boldsymbol{\phi}_1, \boldsymbol{\psi}_1, \boldsymbol{Z}_1\}$ $\mathcal{M}_2 = \{\boldsymbol{\phi}_2, \boldsymbol{\psi}_2, \boldsymbol{Z}_2\}$ $\mathcal{M}_3 = \{\boldsymbol{\phi}_3, \boldsymbol{\psi}_3, \boldsymbol{Z}_3\}$

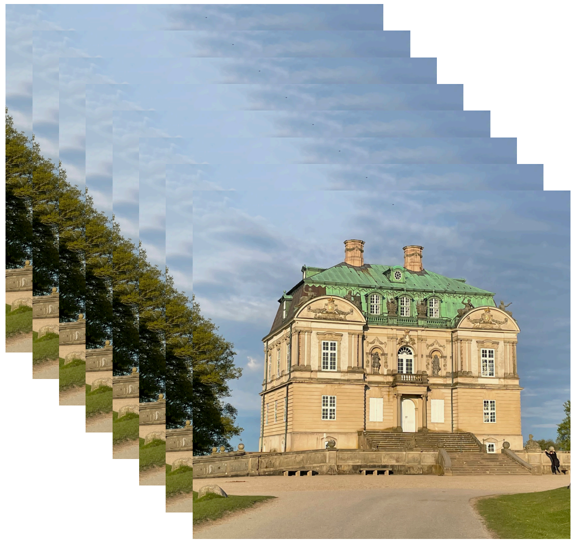
we obtain different objects with parameters
where data is no longer needed

Modular Gaussian Processes

doing these learning processes independently

$\mathcal{M}_1 = \{\phi_1, \psi_1, \boldsymbol{Z}_1\}$

module 1



$\mathcal{M}_2 = \{\phi_2, \psi_2, \boldsymbol{Z}_2\}$

module 2

$\mathcal{M}_3 = \{\phi_3, \psi_3, \boldsymbol{Z}_3\}$

module 3

meta-module
meta-GP

$\mathcal{M}_* = \{\phi_*, \psi_*, \boldsymbol{Z}_*\}$

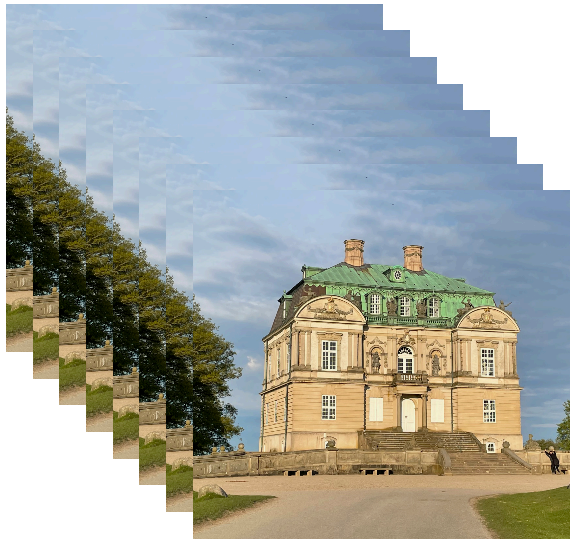
Modular Gaussian Processes

II

doing these learning processes independently

$\mathcal{M}_1 = \{\phi_1, \psi_1, \boldsymbol{Z}_1\}$

module 1



$\mathcal{M}_2 = \{\phi_2, \psi_2, \boldsymbol{Z}_2\}$

module 2

$\mathcal{M}_3 = \{\phi_3, \psi_3, \boldsymbol{Z}_3\}$

module 3

meta-module
meta-GP

$\mathcal{M}_* = \{\phi_*, \psi_*, \boldsymbol{Z}_*\}$

$\boldsymbol{\phi}_*$ — new variational parameters

$\boldsymbol{\psi}_*$ — new kernel hyperparameters

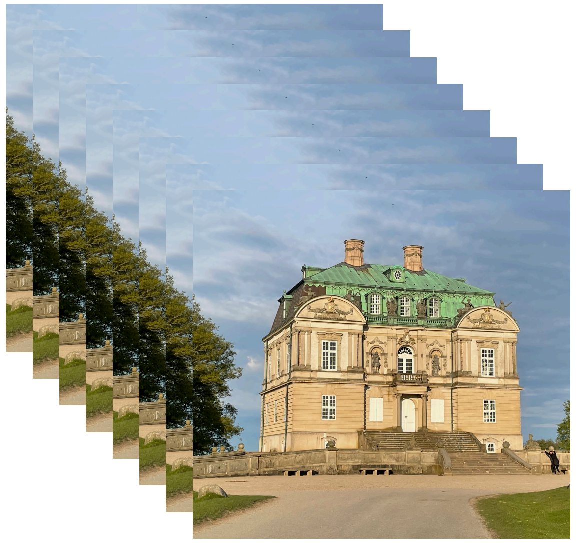
$\boldsymbol{u}_*, \boldsymbol{Z}_*$ — new inducing points

Modular Gaussian Processes

doing these learning processes independently

$\mathcal{M}_1 = \{\phi_1, \psi_1, \boldsymbol{Z}_1\}$

module 1



$\mathcal{M}_2 = \{\phi_2, \psi_2, \boldsymbol{Z}_2\}$

module 2

$\mathcal{M}_3 = \{\phi_3, \psi_3, \boldsymbol{Z}_3\}$

We need the log-marginal likelihood!

meta-module
meta-GP

$\mathcal{M}_* = \{\phi_*, \psi_*, \boldsymbol{Z}_*\}$

ϕ_* — new

ψ_* — new

$\boldsymbol{u}_*, \boldsymbol{Z}_*$ — new

Factorisable (marginal) likelihood

first step — data divided in K subsets

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N} \qquad \mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$$

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_K) = \log \int p(\boldsymbol{y}, f_+) f_+$$

$$\log p(\boldsymbol{y}) = \log \iint q(\boldsymbol{u}_*) p(f_{+\neq \boldsymbol{u}_*} | \boldsymbol{u}_*) p(\boldsymbol{y}|f_+) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} df_{+\neq \boldsymbol{u}_*} d\boldsymbol{u}_*$$

$$\geq \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*} | \boldsymbol{u}_*)} [\log p(\boldsymbol{y}|f_+)] + \log \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} \right]$$

Factorisable (marginal) likelihood

II

first step — data divided in K subsets

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N} \qquad \mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K\}$$

second step — augmentation + large-dimensional integrals

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_K) = \log \int p(\boldsymbol{y}, f_+) f_+$$

$$\begin{aligned}\log p(\boldsymbol{y}) &= \log \iint q(\boldsymbol{u}_*) p(f_{+\neq \boldsymbol{u}_*} | \boldsymbol{u}_*) p(\boldsymbol{y} | f_+) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} df_{+\neq \boldsymbol{u}_*} d\boldsymbol{u}_* \\ &\geq \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*} | \boldsymbol{u}_*)} [\log p(\boldsymbol{y} | f_+)] + \log \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} \right]\end{aligned}$$

Factorisable (marginal) likelihood

II

first step — data divided in K subsets

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N} \qquad \mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$$

second step — augmentation + large-dimensional integrals

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_K) = \log \int p(\boldsymbol{y}, f_+) f_+$$

third step — conditioning on new inducing points

$$\log p(\boldsymbol{y}) = \log \iint q(\boldsymbol{u}_*) p(f_{+\neq\boldsymbol{u}_*}|\boldsymbol{u}_*) p(\boldsymbol{y}|f_+) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} df_{+\neq\boldsymbol{u}_*} d\boldsymbol{u}_*$$

$$\geq \mathbb{E}_{q(\boldsymbol{u}_*)}\left[\mathbb{E}_{p(f_{+\neq\boldsymbol{u}_*}|\boldsymbol{u}_*)}[\log p(\boldsymbol{y}|f_+)] + \log \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)}\right]$$

Factorisable (marginal) likelihood

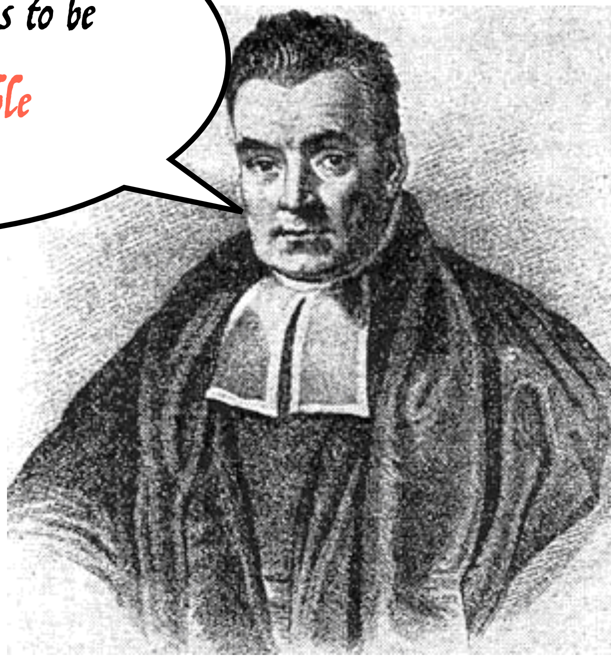
II

first step — data divided in K subsets

$$\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N} \qquad \mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \dots, \mathcal{D}_K\}$$

second step — augmentation + large-dimen tegrals

$$\log p(\boldsymbol{y}) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_K) =$$



third step — conditioning on new inducing points

$$\log p(\boldsymbol{y}) = \log \iint q(\boldsymbol{u}_*) p(f_{+\neq \boldsymbol{u}_*} | \boldsymbol{u}_*) p(\boldsymbol{y}|f_+) \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} df_{+\neq \boldsymbol{u}_*} d\boldsymbol{u}_*$$

$$\geq \mathbb{E}_{q(\boldsymbol{u}_*)} \left[\mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*)} [\log p(\boldsymbol{y}|f_+)] + \log \frac{p(\boldsymbol{u}_*)}{q(\boldsymbol{u}_*)} \right]$$

Summary index

I Gaussian processes (in a nutshell)

- gaussian likelihoods
- non-gaussian likelihoods
- sparse approximations

II Modular Gaussian processes

- factorisable (marginal) likelihoods
- Bayesian likelihood approximation
- module-driven lower bounds
- results

Bayesian likelihood approximation

$$\boxed{\mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}[\log p(\boldsymbol{y}|f_+)]}$$

some manipulations are in order

Bayesian likelihood approximation

II

$$\boxed{\mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}[\log p(\boldsymbol{y}|f_+)]}$$

$$\log p(\boldsymbol{y}|f_+) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_K|f_+)$$

expanding the likelihood wrt modules

Bayesian likelihood approximation

II

$$\boxed{\mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}[\log p(\boldsymbol{y}|f_+)]}$$

$$\log p(\boldsymbol{y}|f_+) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_K|f_+)$$

expanding the likelihood wrt modules

$$= \log \prod_{k=1}^{K} p(\boldsymbol{y}_k|f_+)$$

applying conditional indep. (CI)

Bayesian likelihood approximation

II

$$\boxed{\mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}[\log p(\boldsymbol{y}|f_+)]}$$

$$\log p(\boldsymbol{y}|f_+) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_K|f_+)$$

expanding the likelihood wrt modules

$$= \log \prod_{k=1}^{K} p(\boldsymbol{y}_k|f_+)$$

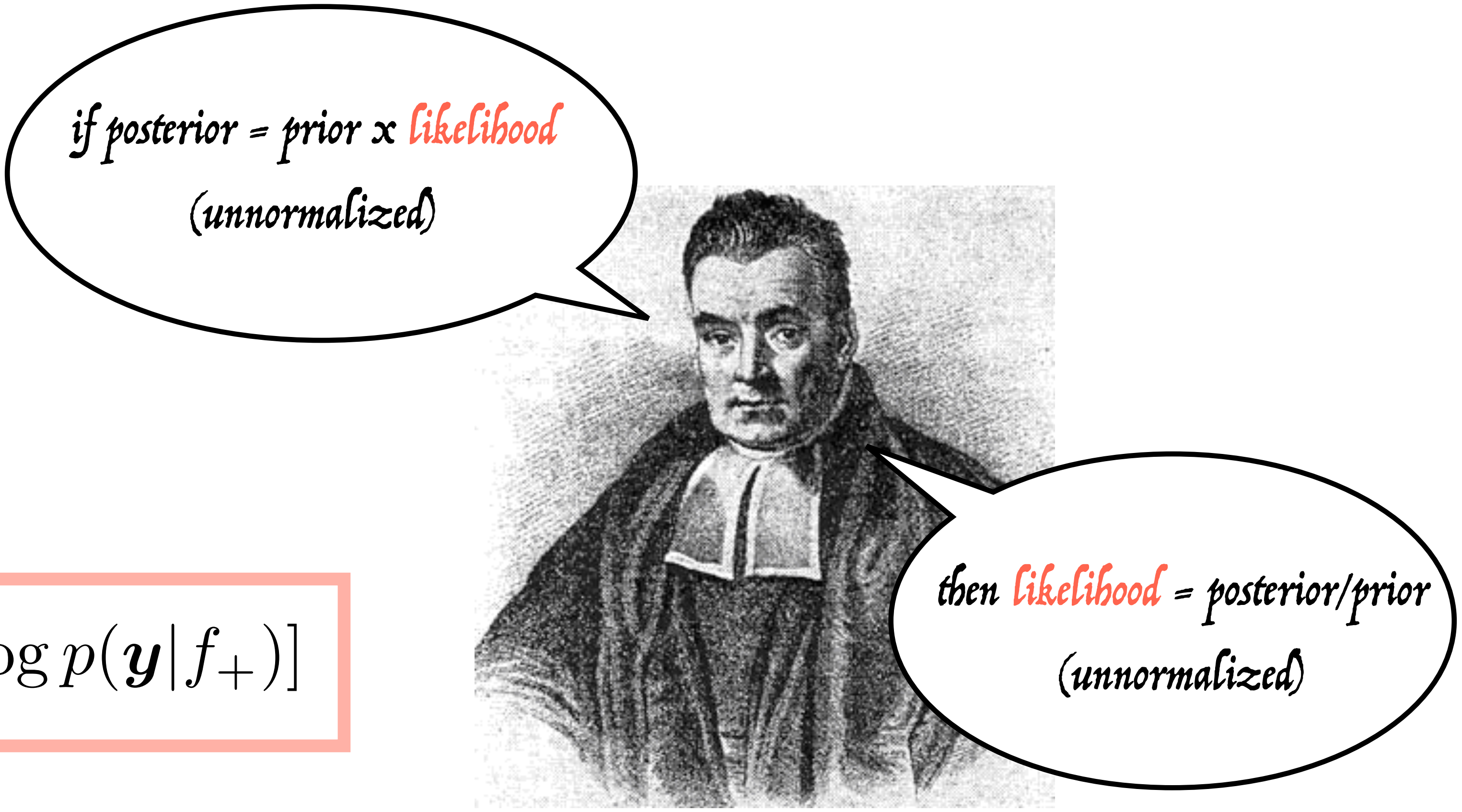
applying conditional indep. (CI)

$$= \sum_{k=1}^{K} \log p(\boldsymbol{y}_k|f_+) \quad ✗$$

observations are still there!

Bayesian likelihood approximation

II


if posterior = prior x likelihood (unnormalized)
then likelihood = posterior/prior (unnormalized)

$$\boxed{\mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}[\log p(\boldsymbol{y}|f_+)]}$$

$$\log p(\boldsymbol{y}|f_+) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_K|f_+)$$

expanding the likelihood wrt modules

$$= \log \prod_{k=1}^{K} p(\boldsymbol{y}_k|f_+)$$

applying conditional indep. (CI)

$$= \sum_{k=1}^{K} \log p(\boldsymbol{y}_k|f_+) \approx \sum_{k=1}^{K} \log Z_k \frac{q_k(f_+)}{p_k(f_+)}$$ ✓

Bayesian likelihood approximation

II

✓

$$\mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}[\log p(\boldsymbol{y}|f_+)] \approx \sum_{k=1}^{K} \mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}\left[\log Z_k \frac{q_k(f_+)}{p_k(f_+)}\right]$$

no more data-dependency!

Bayesian likelihood approximation

II

expectation integrals got reduced

$$\mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*)}[\log p(\boldsymbol{y}|f_+)] \approx \sum_{k=1}^{K} \mathbb{E}_{p(f_{+\neq \boldsymbol{u}_*}|\boldsymbol{u}_*)} \left[\log Z_k \frac{q_k(f_+)}{p_k(f_+)}\right] = \sum_{k=1}^{K} \mathbb{E}_{p(\boldsymbol{u}_k|\boldsymbol{u}_*)} \left[\log Z_k \frac{q_k(\boldsymbol{u}_k)}{p_k(\boldsymbol{u}_k)}\right]$$

thanks to Gaussian marginal properties

Summary index

I Gaussian processes (in a nutshell)

- gaussian likelihoods
- non-gaussian likelihoods
- sparse approximations

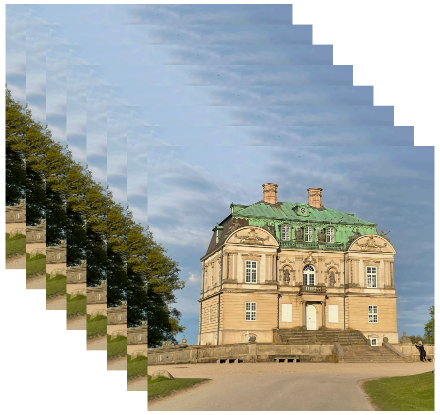
II Modular Gaussian processes

- factorisable (marginal) likelihoods
- Bayesian likelihood approximation
- module-driven lower bounds
- results

Module-driven bound

II

$\mathcal{M}_1 = \{\phi_1, \psi_1, \boldsymbol{Z}_1\}$



$\mathcal{M}_2 = \{\phi_2, \psi_2, \boldsymbol{Z}_2\}$

$\mathcal{M}_3 = \{\phi_3, \psi_3, \boldsymbol{Z}_3\}$

$\cdots$

$\mathcal{M}_K = \{\phi_K, \psi_K, \boldsymbol{Z}_K\}$

A **bound** without data!

$$\boxed{\mathcal{L}_{\mathcal{E}} = \sum_{k=1}^{K} \mathbb{E}_{q_{\mathcal{C}}(\boldsymbol{u}_k)} \left[\log q_k(\boldsymbol{u}_k) - \log p(\boldsymbol{u}_k)\right] - \textbf{KL}\left[q(\boldsymbol{u}_*)||p(\boldsymbol{u}_*)\right]}$$

new complexity: $\boxed{\mathcal{O}((\sum_k M_k)M^2)}$

Summary index

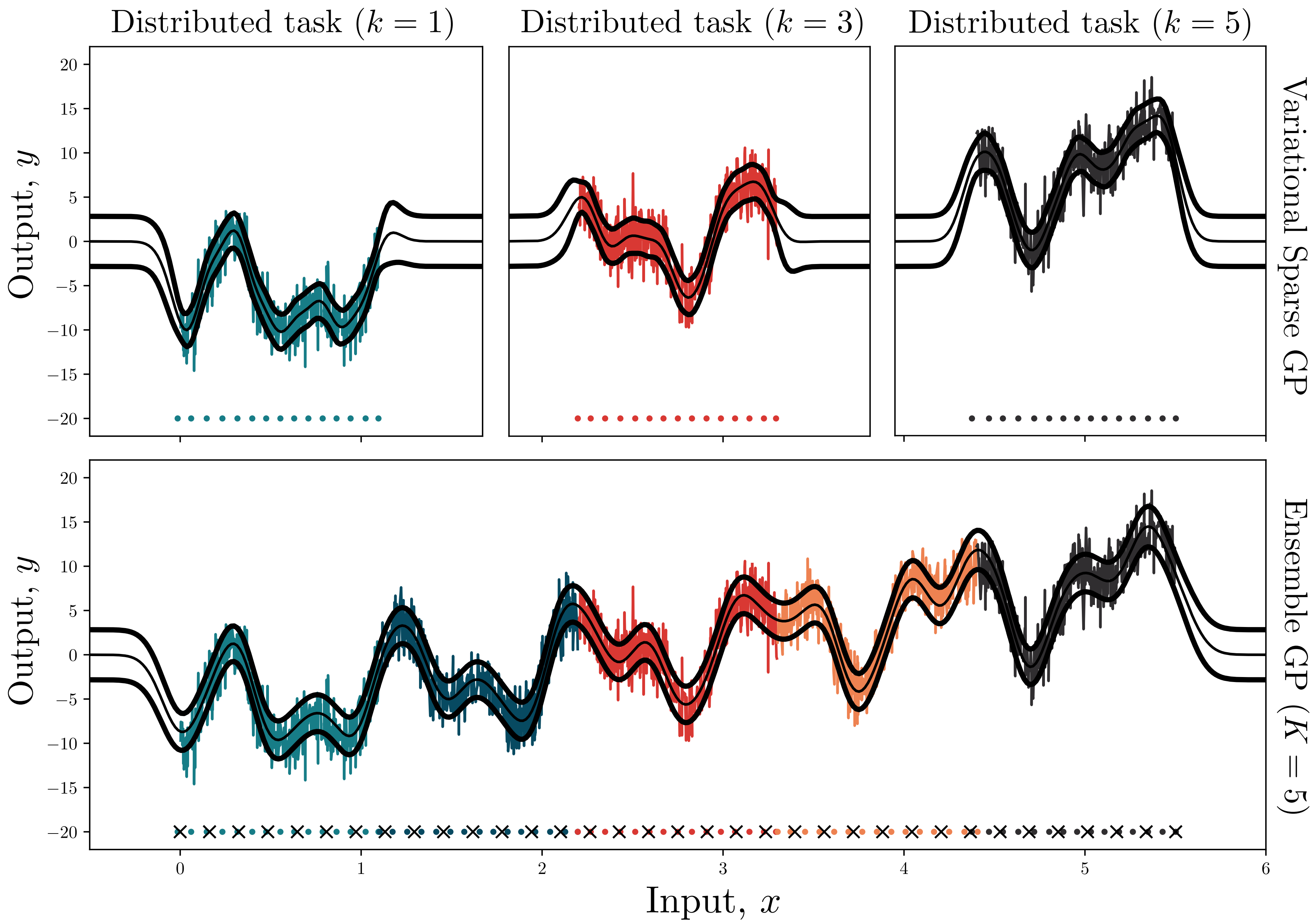
I Gaussian processes (in a nutshell)

- gaussian likelihoods
- non-gaussian likelihoods
- sparse approximations

II Modular Gaussian processes

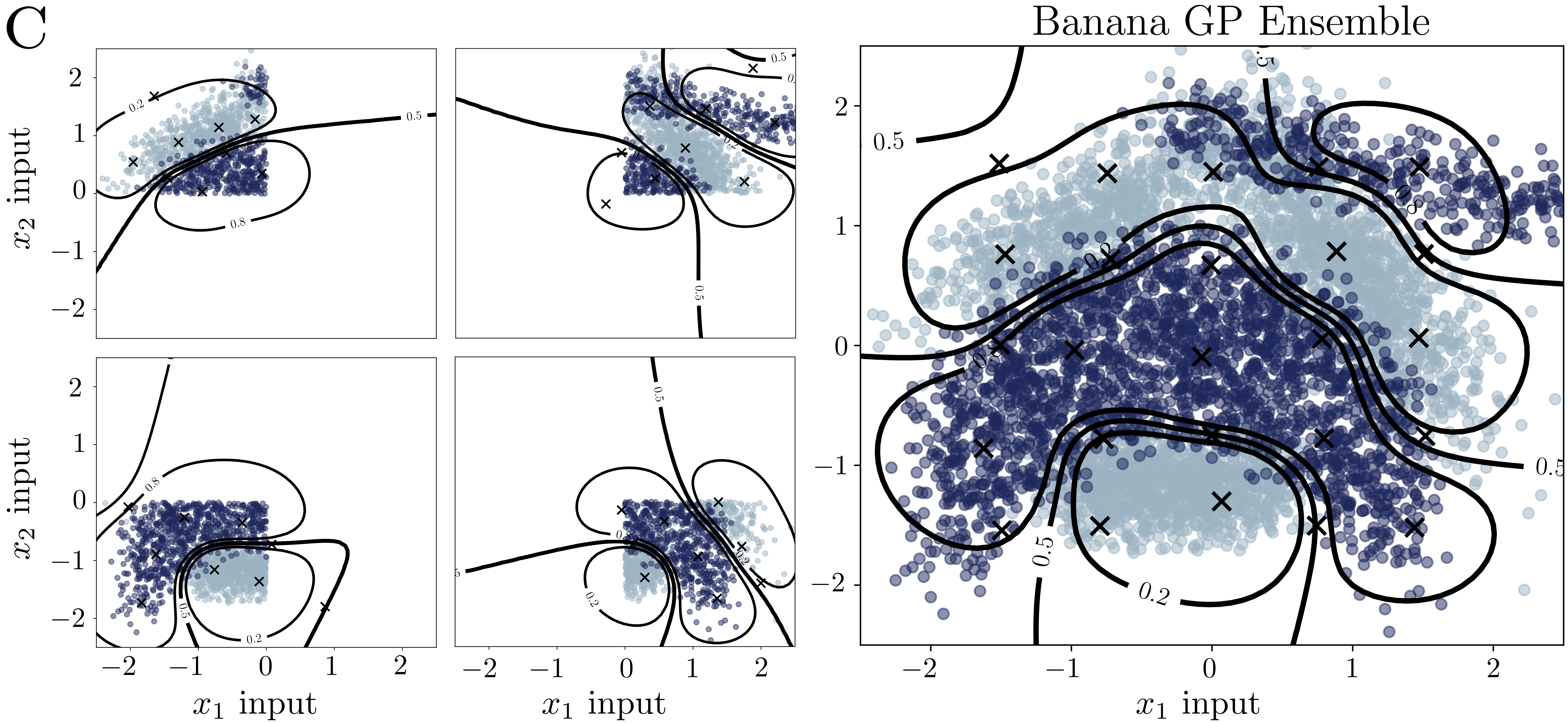
- factorisable (marginal) likelihoods
- Bayesian likelihood approximation
- module-driven lower bounds
- results

Results / parallel inference



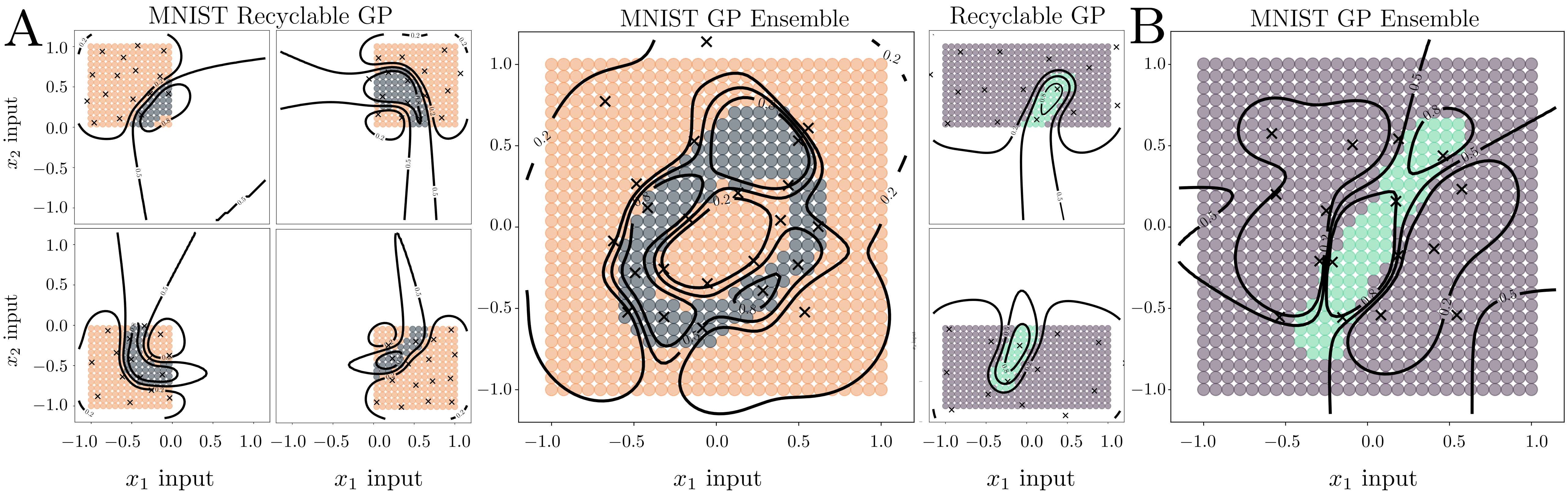
Regression w. 5 independent tasks

Results / banana classification



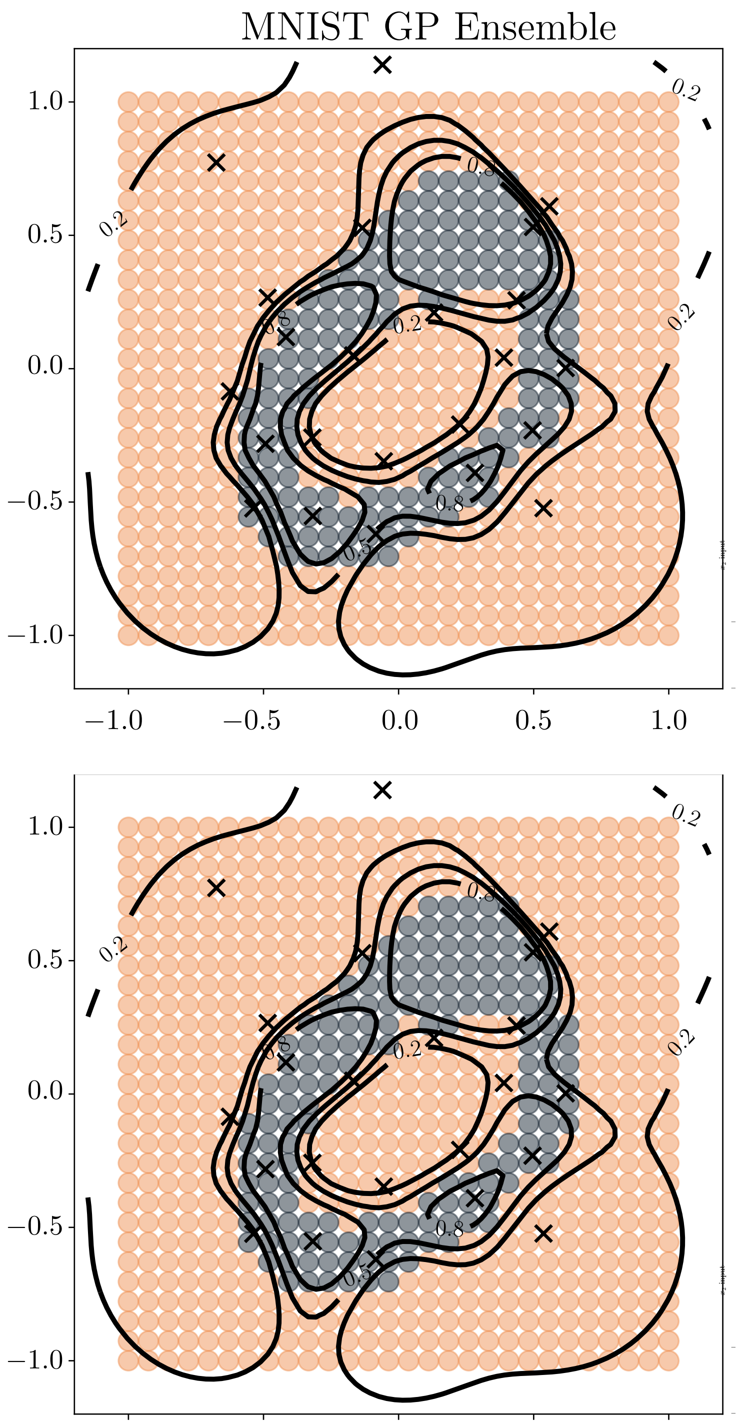
Classification in $\mathbb{R}^2$

Results / image recognition

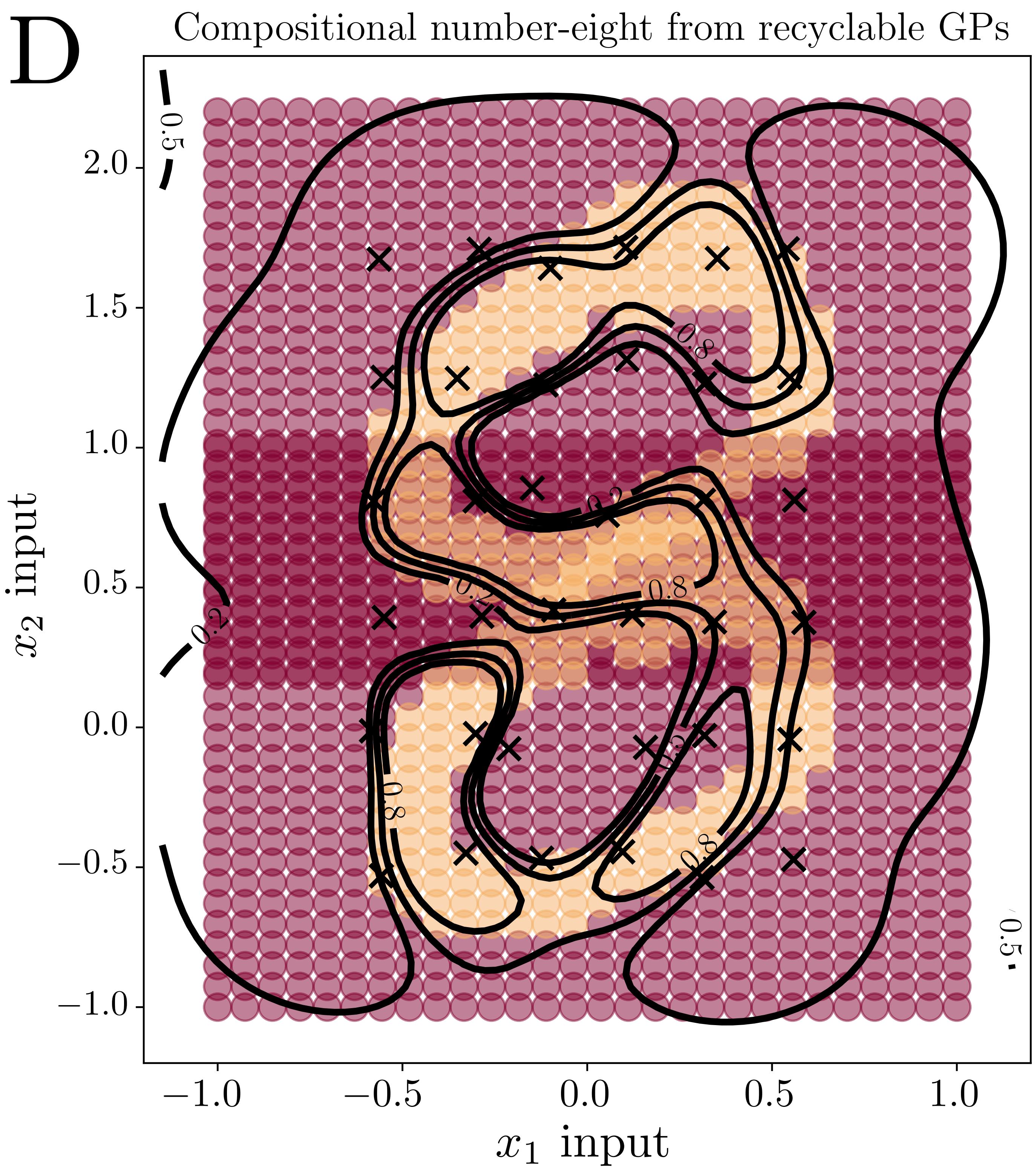


Recognition of $\{0, 1\}$ digits from pieces

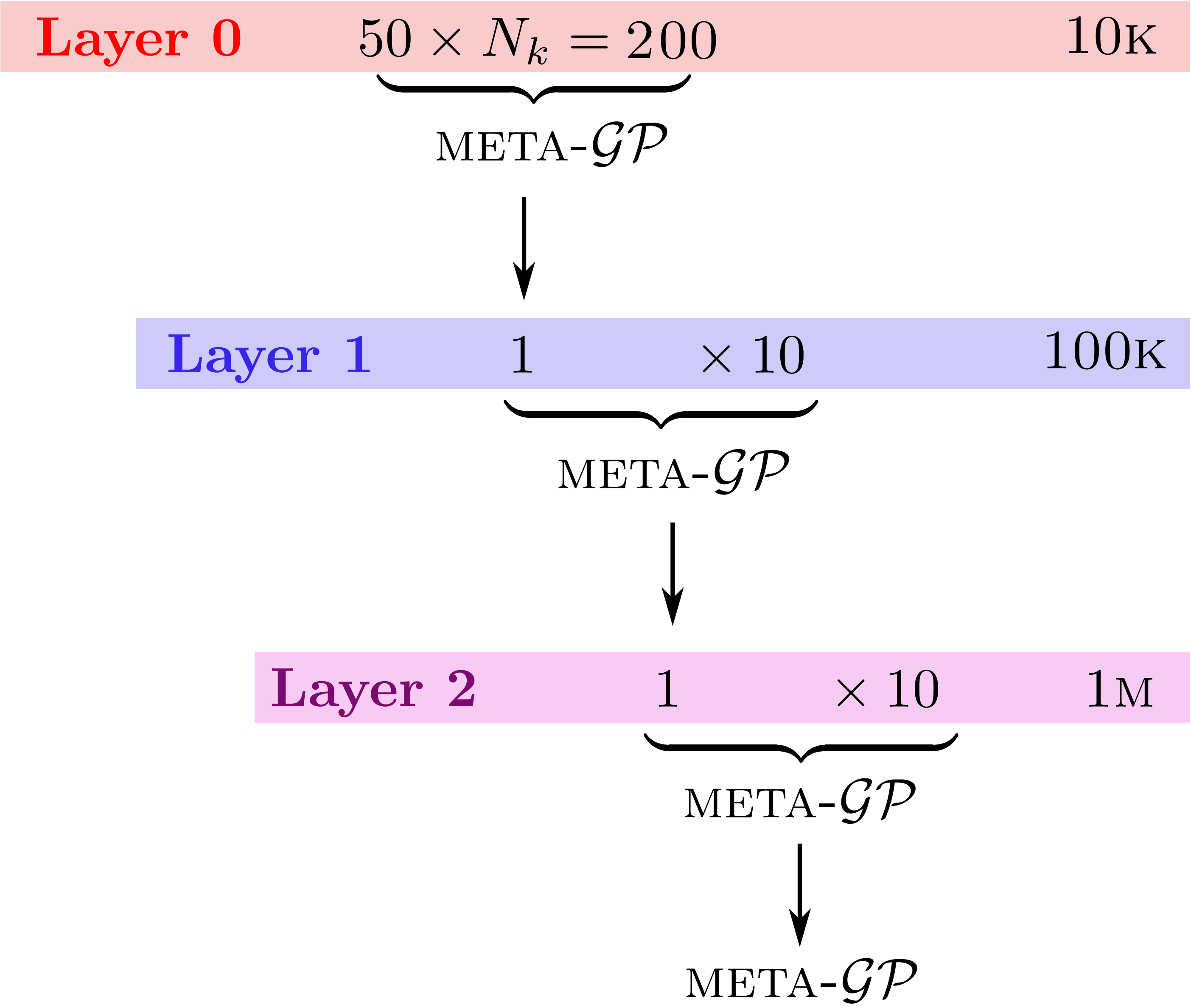
Results / compositional prediction



from two ensembles of zeros



Results / meta-models from meta-models II



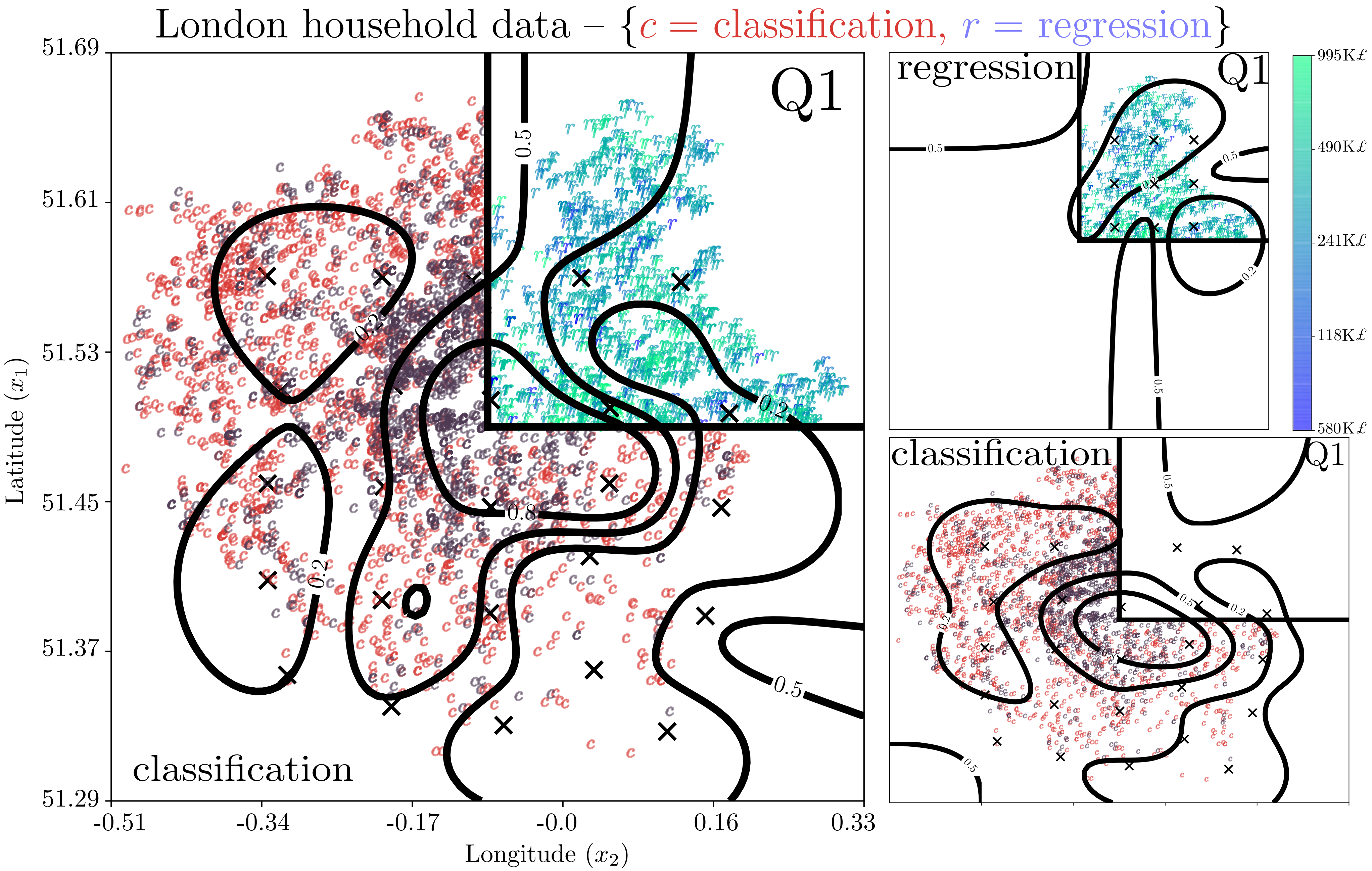
Layer 0 $\underbrace{50 \times N_k = 200}_{\text{META-}\mathcal{GP}}$ 10K

Layer 1 $\underbrace{1 \qquad \times 10}_{\text{META-}\mathcal{GP}}$ 100K

Layer 2 $\underbrace{1 \qquad \times 10}_{\text{META-}\mathcal{GP}}$ 1M

META-$\mathcal{GP}$

Results / heterogeneous



we can also mix binary + real-valued data

Machine Learning + **Life Sciences**

Why is this project interesting for life sciences?

Machine Learning + **Life Sciences**

II

Why is this project **interesting** for life sciences?

- personalized models for patients as **modules**
- population studies without **data-centralisation**
- post-learning **correlation** analysis
- **transfer** learning
- **parallel inference** and computational cost

Collaboration/authors

Pablo **Moreno-Muñoz**

@pablorenoz

Antonio **Artés-Rodríguez**

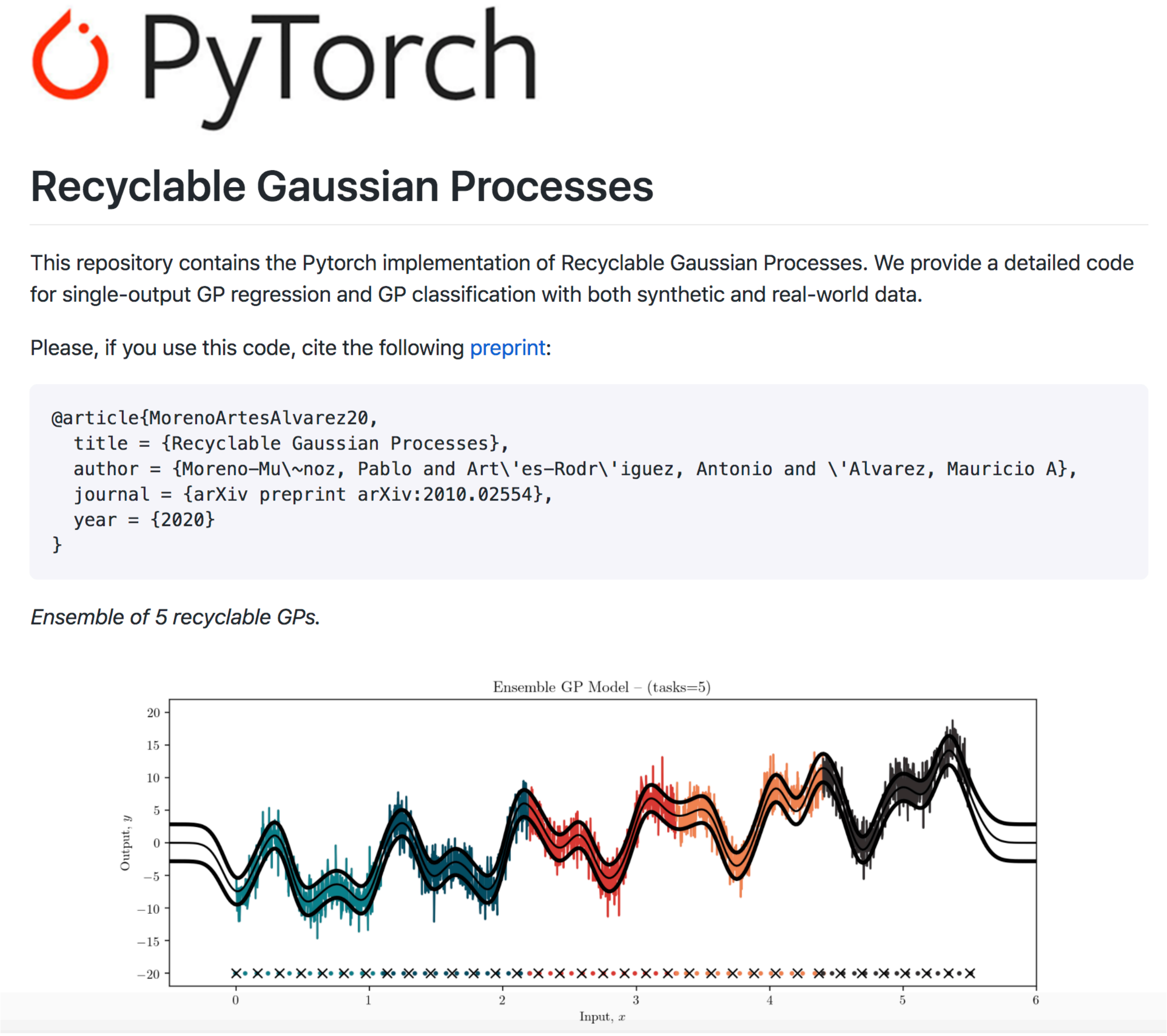
Universidad Carlos III de Madrid, Spain

Mauricio A. **Alvárez**

University of Sheffield
United Kingdom

Find the paper & code!

Already submitted



Recyclable Gaussian Processes

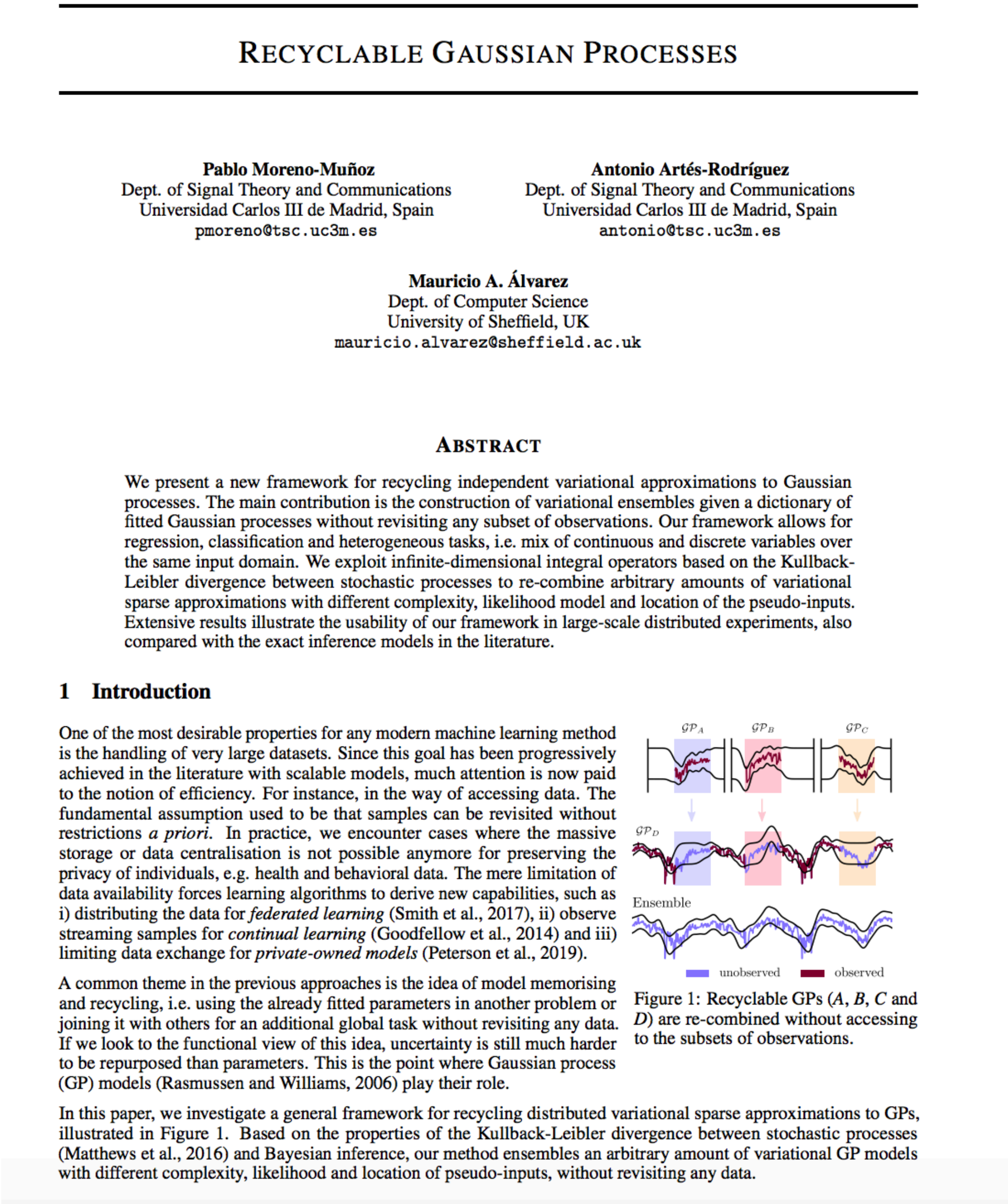
This repository contains the Pytorch implementation of Recyclable Gaussian Processes. We provide a detailed code for single-output GP regression and GP classification with both synthetic and real-world data.

Please, if you use this code, cite the following preprint:

```
@article{MorenoArtesAlvarez20,
  title = {Recyclable Gaussian Processes},
  author = {Moreno-Mu\~noz, Pablo and Art\'es-Rodr\'iguez, Antonio and \'Alvarez, Mauricio A},
  journal = {arXiv preprint arXiv:2010.02554},
  year = {2020}
}
```

Ensemble of 5 recyclable GPs.

RecyclableGP GitHub repo



Recyclable Gaussian Processes

Pablo Moreno-Muñoz
Dept. of Signal Theory and Communications
Universidad Carlos III de Madrid, Spain
pmoreno@tsc.uc3m.es

Antonio Artés-Rodríguez
Dept. of Signal Theory and Communications
Universidad Carlos III de Madrid, Spain
antonio@tsc.uc3m.es

Mauricio A. Álvarez
Dept. of Computer Science
University of Sheffield, UK
mauricio.alvarez@sheffield.ac.uk

Abstract

We present a new framework for recycling independent variational approximations to Gaussian processes. The main contribution is the construction of variational ensembles given a dictionary of fitted Gaussian processes without revisiting any subset of observations. Our framework allows for regression, classification and heterogeneous tasks, i.e. mix of continuous and discrete variables over the same input domain. We exploit infinite-dimensional integral operators based on the Kullback-Leibler divergence between stochastic processes to re-combine arbitrary amounts of variational sparse approximations with different complexity, likelihood model and location of the pseudo-inputs. Extensive results illustrate the usability of our framework in large-scale distributed experiments, also compared with the exact inference models in the literature.

1 Introduction

One of the most desirable properties for any modern machine learning method is the handling of very large datasets. Since this goal has been progressively achieved in the literature with scalable models, much attention is now paid to the notion of efficiency. For instance, in the way of accessing data. The fundamental assumption used to be that samples can be revisited without restrictions *a priori*. In practice, we encounter cases where the massive storage or data centralisation is not possible anymore for preserving the privacy of individuals, e.g. health and behavioral data. The mere limitation of data availability forces learning algorithms to derive new capabilities, such as i) distributing the data for *federated learning* (Smith et al., 2017), ii) observe streaming samples for *continual learning* (Goodfellow et al., 2014) and iii) limiting data exchange for *private-owned models* (Peterson et al., 2019).

A common theme in the previous approaches is the idea of model memorising and recycling, i.e. using the already fitted parameters in another problem or joining it with others for an additional global task without revisiting any data. If we look to the functional view of this idea, uncertainty is still much harder to be repurposed than parameters. This is the point where Gaussian process (GP) models (Rasmussen and Williams, 2006) play their role.

Figure 1: Recyclable GPs (*A*, *B*, *C* and *D*) are re-combined without accessing to the subsets of observations.

In this paper, we investigate a general framework for recycling distributed variational sparse approximations to GPs, illustrated in Figure 1. Based on the properties of the Kullback-Leibler divergence between stochastic processes (Matthews et al., 2016) and Bayesian inference, our method ensembles an arbitrary amount of variational GP models with different complexity, likelihood and location of pseudo-inputs, without revisiting any data.

arXiv:2010.02554 preprint

The (very) **end**

thanks!