Unsupervised Learning with Gaussian Processes

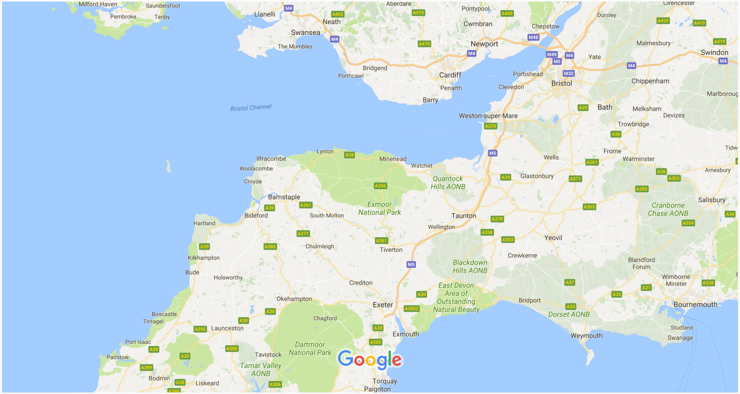
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

September 11, 2019

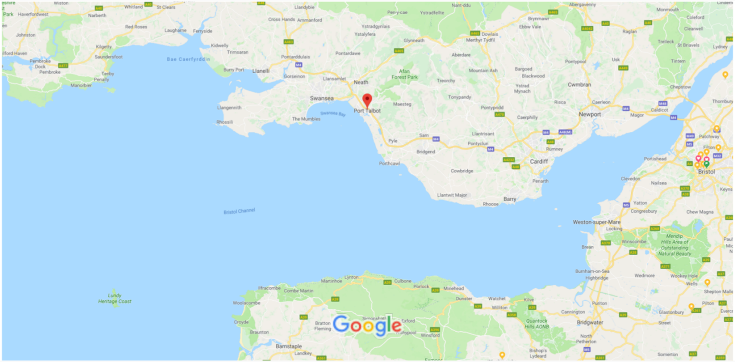
http://www.carlhenrik.com

Introductions

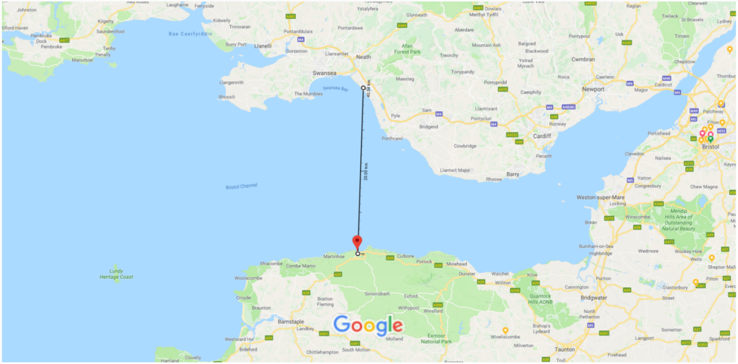
This where I live



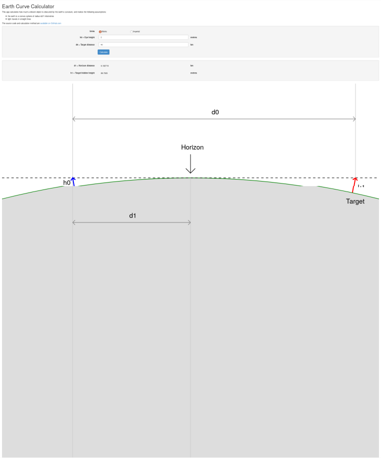
This is what I do


Haverfordwest
Narberth
Whitland
St Clears
Red Roses
Laugharne
Ferryside
Kidwelly
Trimsaran
Johnston
Kilgetty
Saundersfoot
Pembroke Dock
Pembroke
Tenby
Penally
Manorbier
Bae Caerfyrddin
Burry Port
Llanelli
Llandybie
Cross Hands
Ammanford
Pontarddulais
Gorseinon
Pontardawe
Llansamlet
Neath
Swansea
Port Talbot
Llangennith
Rhossili
The Mumbles
Swansea Bay
Ystradgynlais
Ystalyfera
Glynneath
Pen-y-cae
Ystradfellte
Nant-ddu
Afan Forest Park
Maesteg
Aberdare
Treorchy
Tonypandy
Mountain Ash
Merthyr Tydfil
Dowlais
Ebbw Vale
Brynmawr
Abergavenny
Bargoed
Blackwood
Ystrad Mynach
Pontypridd
Caerphilly
Pontypool
Cwmbran
Usk
Raglan
Monmouth
Coleford
Risca
Caerleon
Newport
Magor
Caldicot
Chepstow
Tintern
Lydney
Thornbury
Pyle
Sarn
Bridgend
Porthcawl
Llantrisant
Pontyclun
Cowbridge
Cardiff
Penarth
Llantwit Major
Barry
Rhoose
Portishead
Patchway
Filton
Bristol
Clevedon
Nailsea
Yatton
Congresbury
Chew Magna
Weston-super-Mare
Locking
Winscombe
Brean
Mendip Hills Area of Outstanding Natural Beauty
Burnham-on-Sea
Highbridge
Wedmore
Wookey Hole
Wells
Shepton Mallet
Bristol Channel
Lundy Heritage Coast
Ilfracombe
Combe Martin
Martinhoe
Lynton
Culbone
Porlock
Minehead
Dunster
Watchet
Williton
Kilve
Quantock Hills AONB
Cannington
Bridgwater
North Petherton
Puriton
Glastonbury
Street
Pilton
Somerton
Woolacombe
Croyde
Braunton
Barnstaple
Landkey
Bratton Fleming
Exford
Winsford
Bishop's Lydeard
Google


Bae Caerfyrddin
Llanelli
Burry Port
Neath
Swansea
Swansea Bay
The Mumbles
Port Talbot
Afan Forest Park
Maesteg
Pyle
Porthcawl
Bridgend
Cowbridge
Llantwit Major
Barry
Rhoose
Penarth
Cardiff
Caerphilly
Pontypridd
Newport
Cwmbran
Caerleon
Magor
Caldicot
Chepstow
Portishead
Clevedon
Nailsea
Bristol
Patchway
Filton
Yatton
Congresbury
Chew Magna
Weston-super-Mare
Locking
Winscombe
Mendip Hills Area of Outstanding Natural Beauty
Brean
Burnham-on-Sea
Highbridge
Wedmore
Wells
Glastonbury
Street
Bridgwater
Cannington
North Petherton
Taunton
Bishop's Lydeard
Quantock Hills AONB
Kilve
Watchet
Williton
Wiveliscombe
Milverton
Dunster
Minehead
Porlock
Culbone
Lynton
Martinhoe
Combe Martin
Ilfracombe
Woolacombe
Croyde
Braunton
Barnstaple
Bratton Fleming
Simonsbath
Exford
Withypool
Winsford
Exmoor National Park
Molland
South Molton
Chittlehampton
Landkey
Westward Ho!
Bideford
Clovelly
Hartland
Bristol Channel
Lundy Heritage Coast
Tenby
Penally
Manorbier
Saundersfoot
Kilgetty
Pembroke
Pembroke Dock
Laugharne
Ferryside
Kidwelly
Trimsaran
Pontarddulais
Pontardawe
Gorseinon
Llansamlet
Llangennith
Rhossili
Glynneath
Aberdare
Merthyr Tydfil
Mountain Ash
Treorchy
Tonypandy
Llantrisant
Pontyclun
Sarn
Blackwood
Bargoed
Risca
Pontypool
Usk
Raglan
Tintern
Lydney

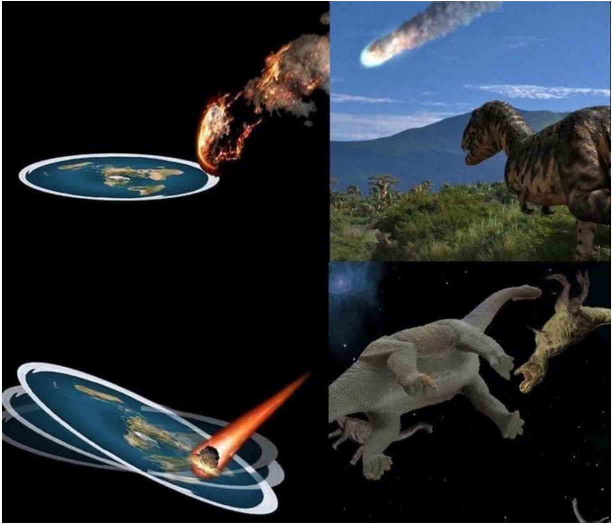


Swansea
Neath
Llanelli
Bae Caerfyrddin
Bristol Channel
Lundy
Heritage Coast
Cardiff
Newport
Cwmbran
Barry
Penarth
Bridgend
Porthcawl
Cowbridge
Llantwit Major
Weston-super-Mare
Bristol
Portishead
Clevedon
Minehead
Porlock
Martinhoe
Ilfracombe
Combe Martin
Woolacombe
Barnstaple
Bideford
Bridgwater
Taunton
Exmoor
National Park
Quantock
Hills AONB
Mendip
Hills Area of
Outstanding
Natural Beauty
20.00 km
Google



Distance to horizon $6.2km$

Hidden height $125.6m$




Lord Nelson
PALE ALE

Learning Theory

- $\mathcal{F}$ space of functions
- $\mathcal{A}$ learning algorithm
- $\mathcal{S} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$
- $\mathcal{S} \sim P(\mathcal{X} \times \mathcal{Y})$
- $\ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y)$ loss function

Statistical Learning

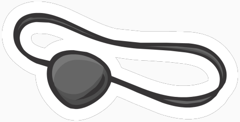
$$e(\mathcal{S}, \mathcal{A}, \mathcal{F}) = \mathbb{E}_{P(\{\mathcal{X},\mathcal{Y}\})}\left[\ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y)\right]$$

$$
\begin{aligned}
e(\mathcal{S}, \mathcal{A}, \mathcal{F}) &= \mathbb{E}_{P(\{\mathcal{X}, \mathcal{Y}\})}\left[\ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x, y)\right] \\
&\approx \frac{1}{M} \sum_{n=1}^{M} \ell(\mathcal{A}_{\mathcal{F}}(\mathcal{S}), x_n, y_n)
\end{aligned}
$$

No Free Lunch

We can come up with a combination of $\{\mathcal{S}, \mathcal{A}, \mathcal{F}\}$ that makes $e(\mathcal{S}, \mathcal{A}, \mathcal{F})$ take an arbitary value

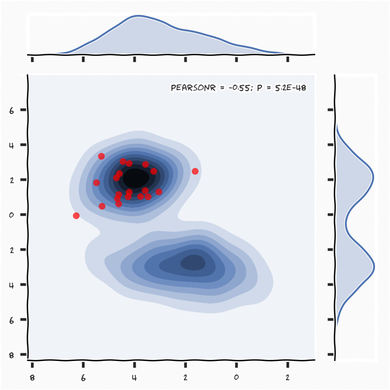
Assumptions: Algorithms



Statistical Learning

$$\mathcal{A}_{\mathcal{F}}(\mathcal{S})$$

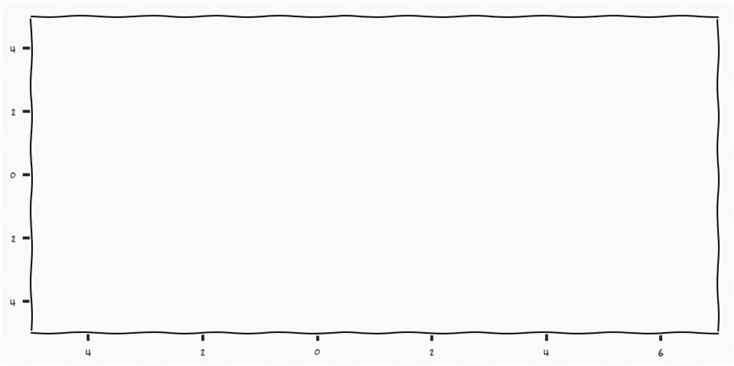
Assumptions: Biased Sample



Statistical Learning

$$\mathcal{A}_{\mathcal{F}}(\mathcal{S})$$

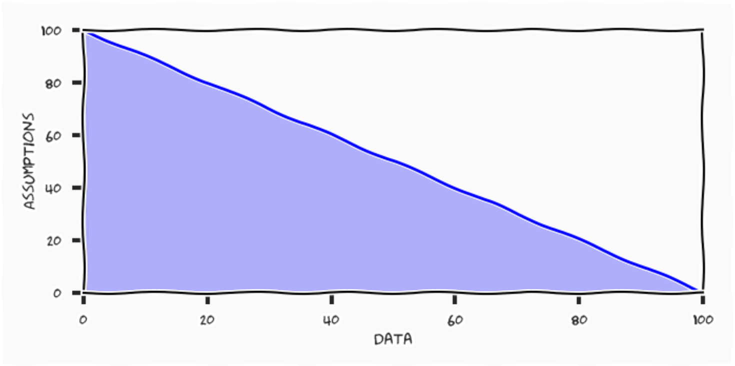
Assumptions: Hypothesis space



Statistical Learning

$$\mathcal{A}_{\mathcal{F}}(\mathcal{S})$$

Data and Knowledge



IUDICIUM POSTERIUM DISCIPULUS EST PRIORIS[1]

[1] The posterior is the student of the prior

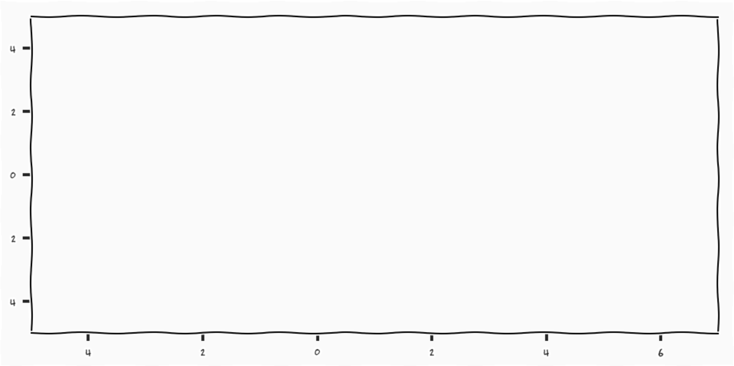
Today

September 11, 2019

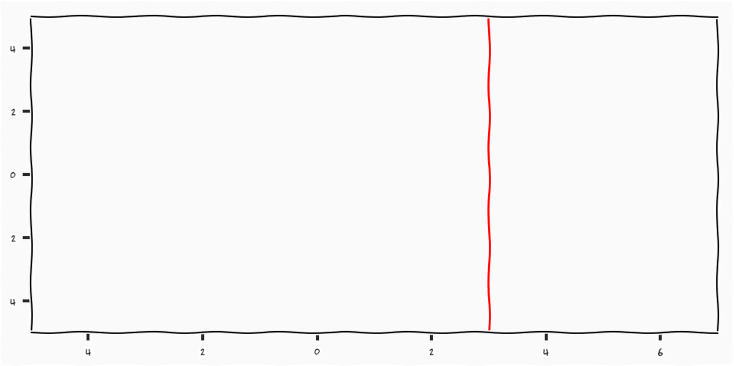
Neill Campbell, Carl Henrik Ek, David Fernandes, Ivan Ustyuzhaninov, Aidan Scannell, Emelie Barman, Erik Bodin, Andrew Lawrence, Markus Kaiser, Alessandro di Martino, Ieva Kazlauskaite, Akshaya Thippur

Gaussian Processes

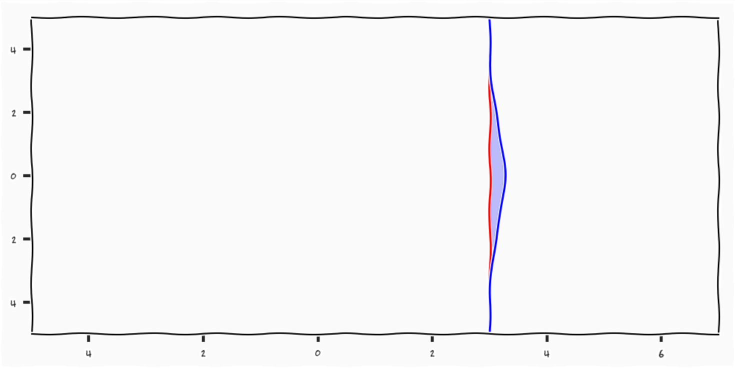
Gaussian Processes



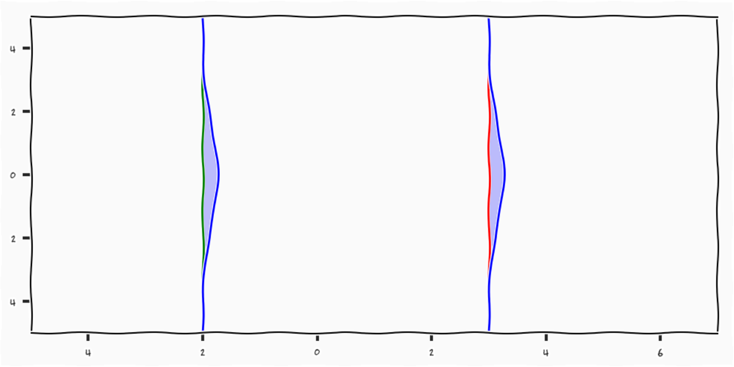
Gaussian Processes



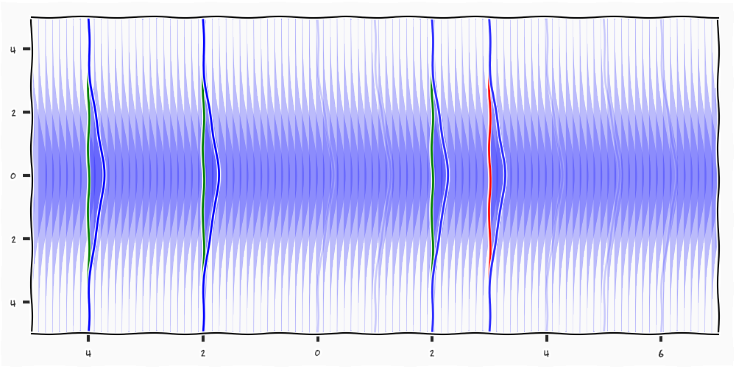
Gaussian Processes



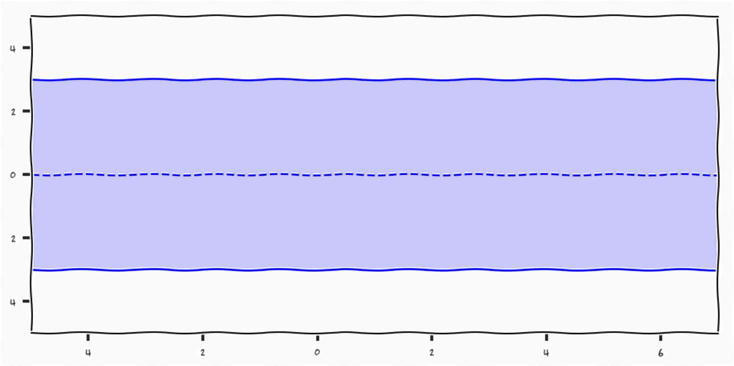
Gaussian Processes



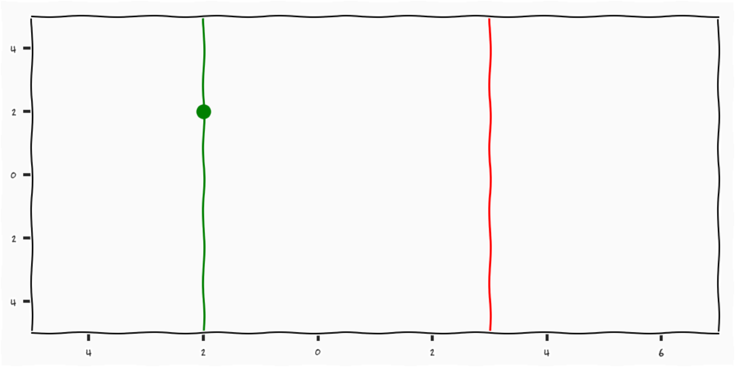
Gaussian Processes



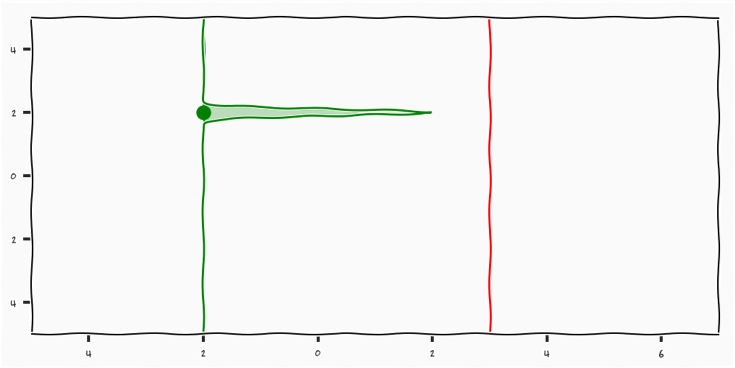
Gaussian Processes



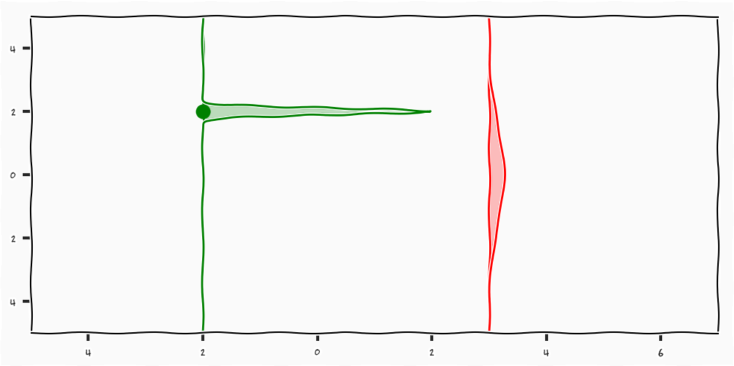
Gaussian Processes

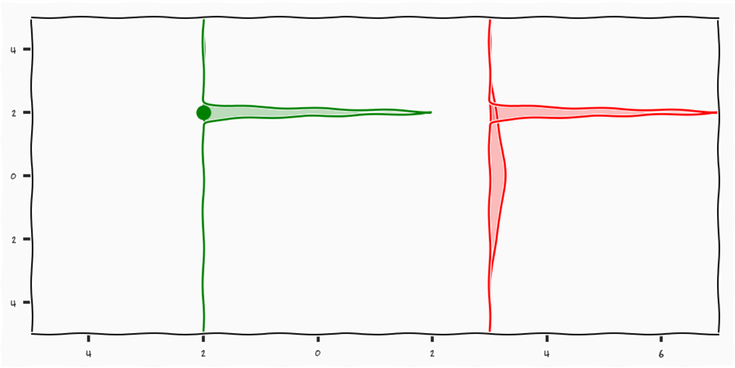


Gaussian Processes

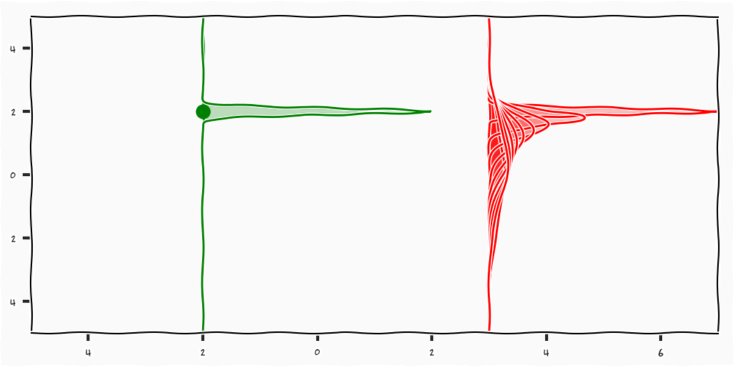


Gaussian Processes

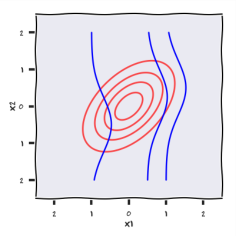



4
2
0
2
4
4
2
0
2
4
6

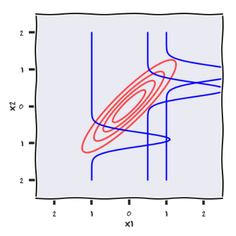
Gaussian Processes



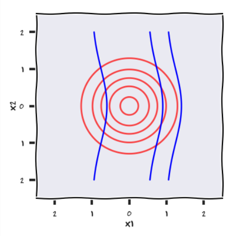
Conditional Gaussians



$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}\right)$$

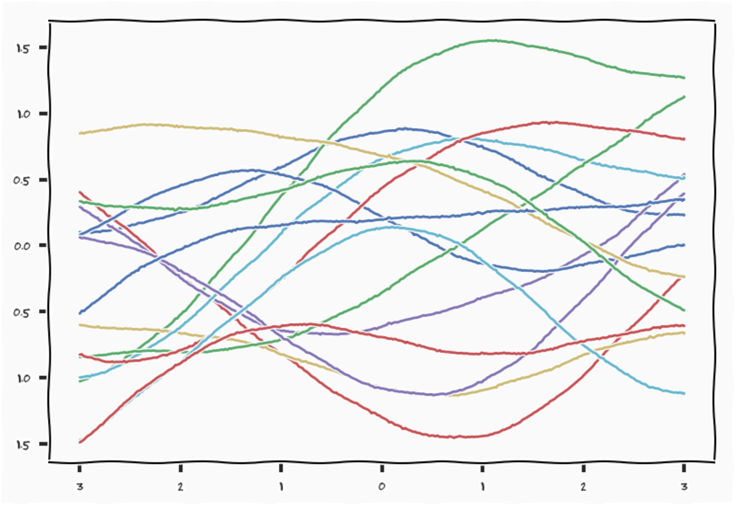


$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.9 \\ 0.9 & 1 \end{bmatrix}\right)$$

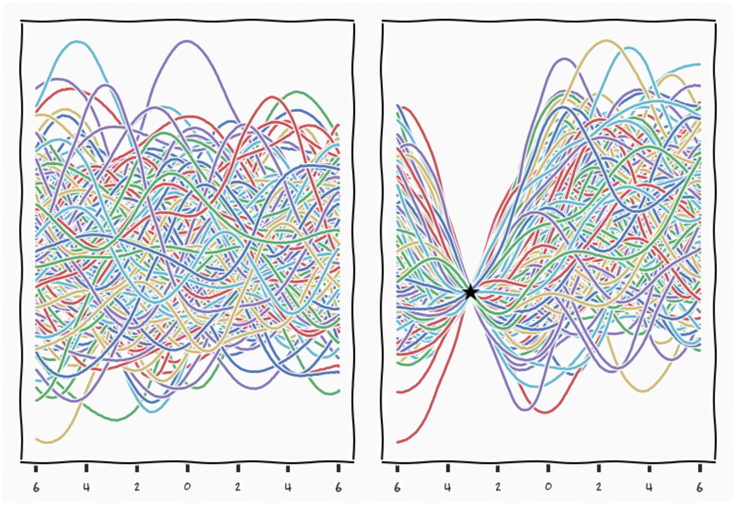


$$N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

Gaussian Processes



Gaussian Processes



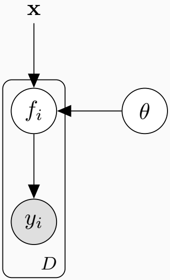
The Gaussian Identities

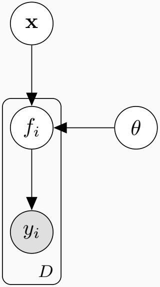
$$p(x_1, x_2) \qquad p(x_1) = \int p(x_1, x_2)\mathrm{d}x \qquad p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)}$$

Gaussian Identities

Unsupervised Learning with GPs

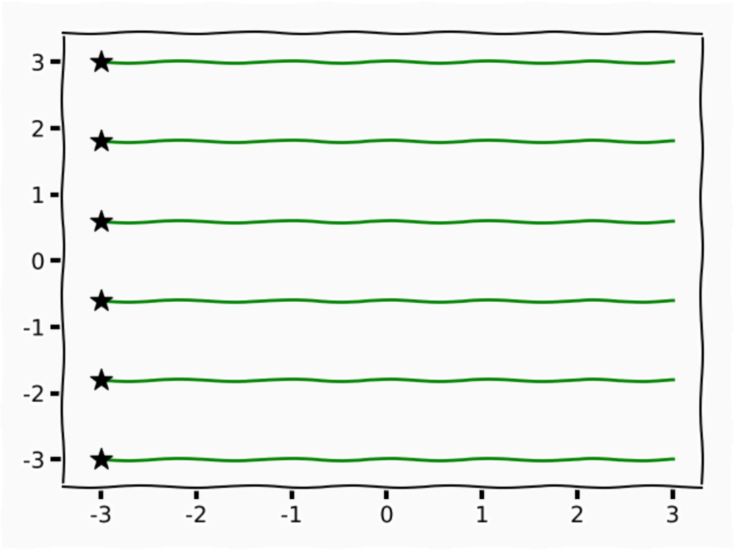
Unsupervised Learning



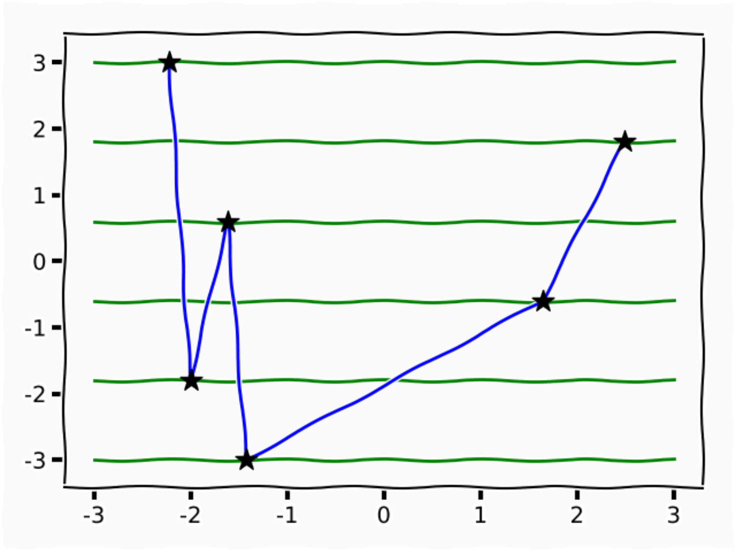


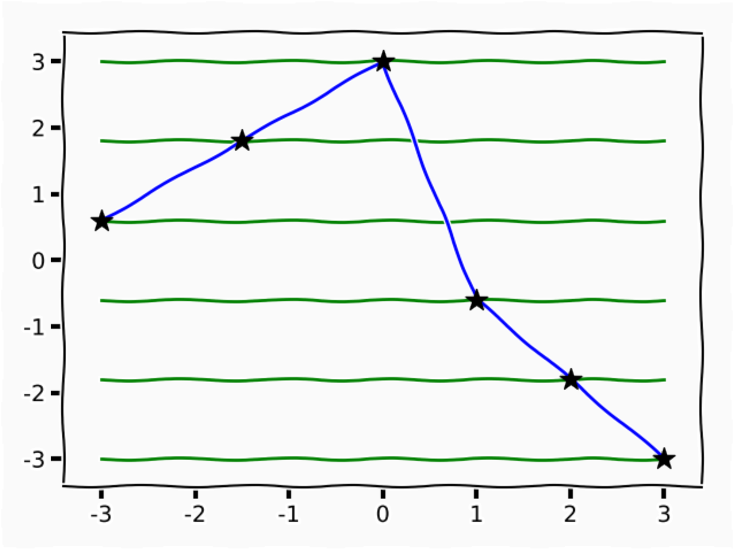
$p(y|x)$ $p(y)$

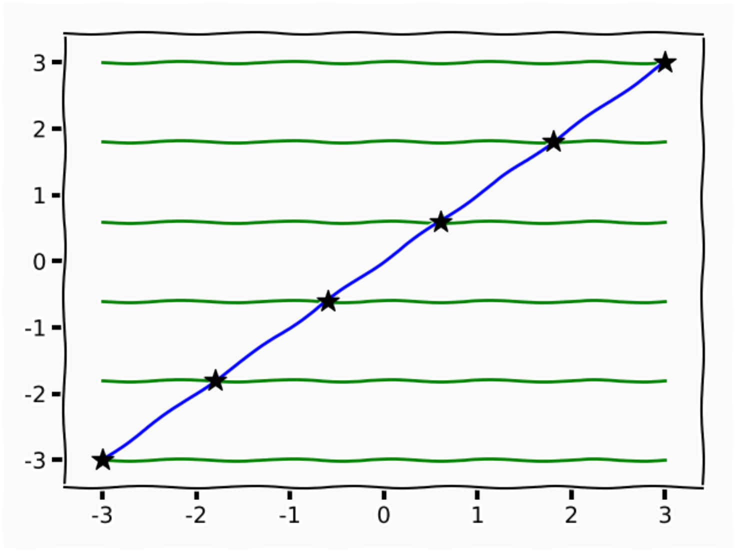
Unsupervised Learning



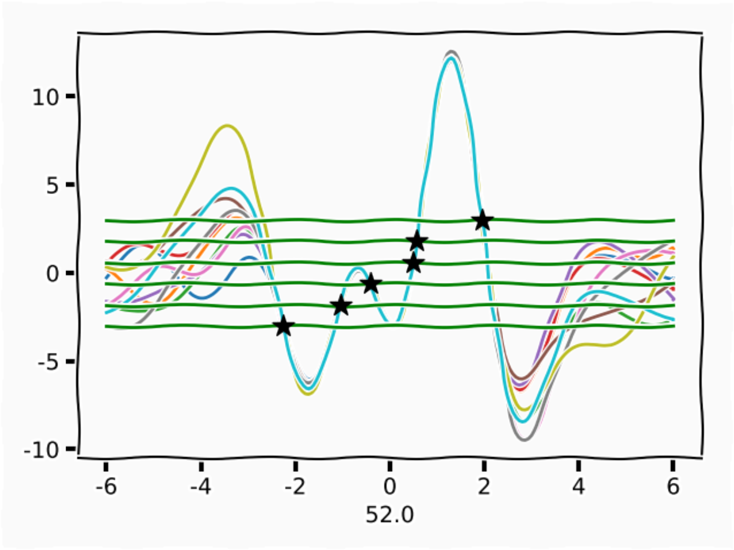
Unsupervised Learning



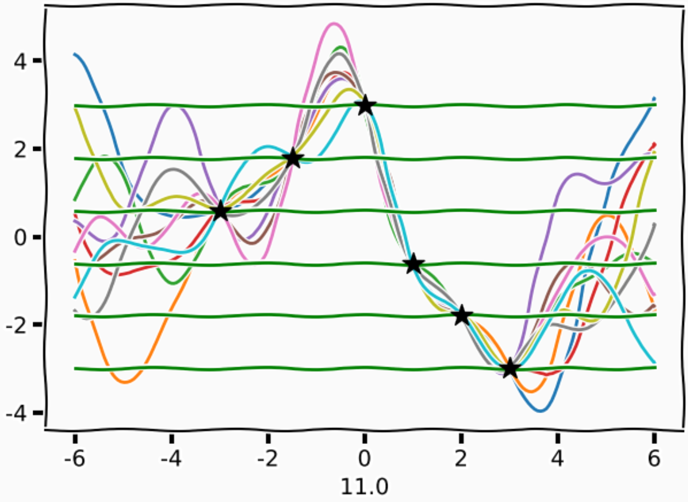

3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3


3
2
1
0
-1
-2
-3
-3
-2
-1
0
1
2
3

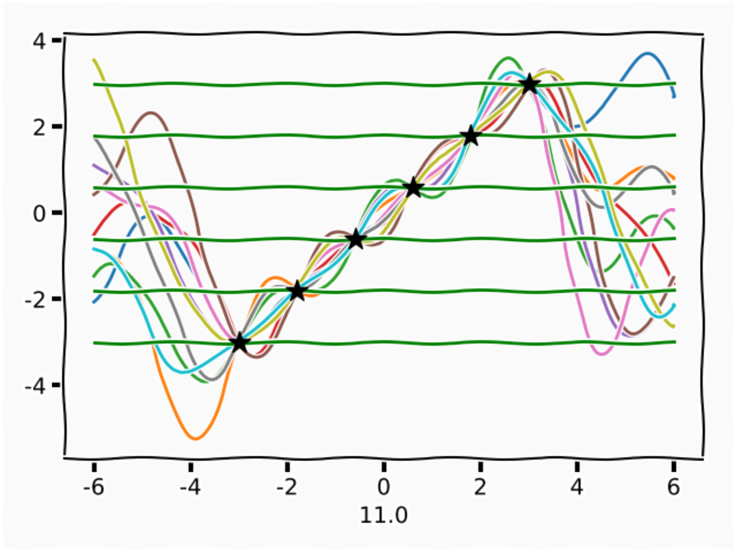
Unsupervised Learning

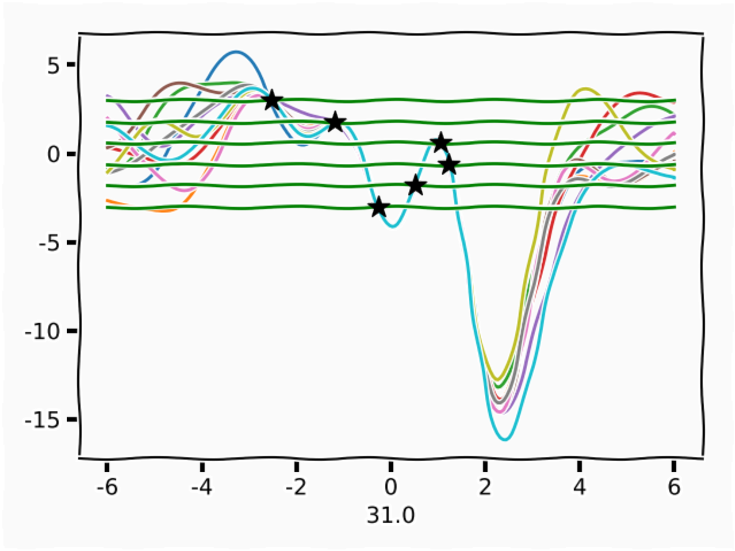


Unsupervised Learning

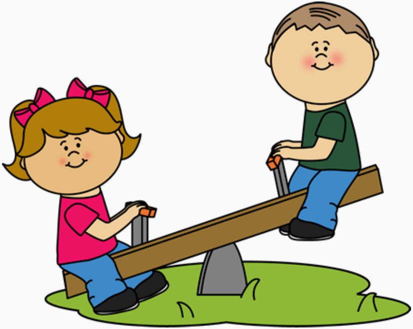


Unsupervised Learning




5
0
-5
-10
-15
-6
-4
-2
0
2
4
6
31.0

Priors



Priors

$$p(y) = \int p(y|f)p(f|x)p(x)\mathrm{d}f\mathrm{d}x$$

$$p(x|y) = p(y|x)\frac{p(x)}{p(y)}$$

1. Priors that makes sense

   p(f) describes our belief/assumptions and defines our notion of complexity in the function

   p(x) expresses our belief/assumptions and defines our notion of complexity in the latent space

2. Now lets churn the handle

Relationship between x and data

$$p(y) = \int p(y|f)p(f|x)p(x)\mathrm{d}f\,\mathrm{d}x$$

- GP prior

$$p(f|x) \sim \mathcal{N}(0, K) \propto e^{-\frac{1}{2}(f^{\mathrm{T}}K^{-1}f)}$$

$$K_{ij} = e^{-(x_i - x_j)^{\mathrm{T}}M^{\mathrm{T}}M(x_i - x_j)}$$

Relationship between x and data

$$p(y) = \int p(y|f)p(f|x)p(x)\mathrm{d}f\mathrm{d}x$$

- GP prior

$$p(f|x) \sim \mathcal{N}(0, K) \propto e^{-\frac{1}{2}(f^{\mathrm{T}}K^{-1}f)}$$

$$K_{ij} = e^{-(x_i - x_j)^{\mathrm{T}}M^{\mathrm{T}}M(x_i - x_j)}$$

- Likelihood

$$p(y|f) \sim N(y|f, \beta) \propto e^{-\frac{1}{2\beta}\mathrm{tr}(y-f)^{\mathrm{T}}(y-f)}$$

Relationship between x and data

$$p(y) = \int p(y|f)p(f|x)p(x)\mathrm{d}f\mathrm{d}x$$

- GP prior

$$p(f|x) \sim \mathcal{N}(0, K) \propto e^{-\frac{1}{2}(f^{\mathrm{T}}K^{-1}f)}$$

$$K_{ij} = e^{-(x_i - x_j)^{\mathrm{T}}M^{\mathrm{T}}M(x_i - x_j)}$$

- Likelihood

$$p(y|f) \sim N(y|f, \beta) \propto e^{-\frac{1}{2\beta}\mathrm{tr}(y-f)^{\mathrm{T}}(y-f)}$$

- Analytically intractable (Non Elementary Integral) and infinitely differientiable

Laplace Integration

"Nature laughs at the difficulties of integrations"
– Simon Laplace

Approximate Inference

ML/MAP[2]

$$\hat{x} = \text{argmax}_x \int p(y|f)p(f|x)\mathrm{d}f p(x)$$

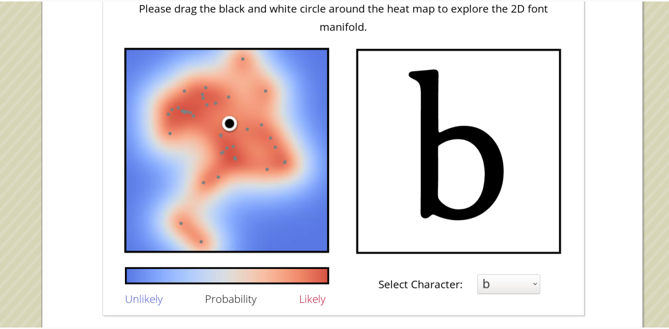
$$= \text{argmin}_x \frac{1}{2}y^{\mathrm{T}}\mathbf{K}^{-1}y + \frac{1}{2}|\mathbf{K}| - \log\, p(x)$$

[2]Lawrence, N. D. (2005). Probabilistic non-linear principal component analysis with Gaussian process latent variable models.

GP-LVM

- Li, W., Viola, F., Starck, J., Brostow, G. J., & Campbell, N. D. (2016). Roto++: accelerating professional rotoscoping using shape manifolds. (In proceeding of ACM SIGGRAPH'16)
- Grochow, K., Martin, S. L., Hertzmann, A., & Popovi\'c, Zoran (2004). Style-based inverse kinematics. SIGGRAPH '04: SIGGRAPH 2004
- Urtasun, R., Fleet, D. J., & Fua, P. (2006). 3D people tracking with Gaussian process dynamical models. Computer Vision and Pattern Recognition, 2006

Font Demo



URL

Bayesian GP-LVM[4]

- Challenges with ML estimation
  - How to initialise x?
  - What is the dimensionality q?
- *Our assumption on the latent space does not reach the data*

[3]Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes.
[4]Titsias, M., & Lawrence, N. D. (2010). Bayesian Gaussian Process Latent Variable Model

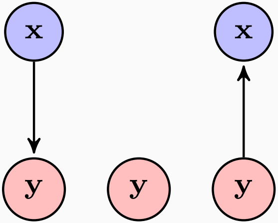
Bayesian GP-LVM[4]

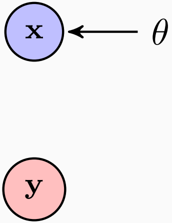
- Challenges with ML estimation
  - How to initialise x?
  - What is the dimensionality q?
- *Our assumption on the latent space does not reach the data*
- Approximate integration![3]

[3] Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes.

[4] Titsias, M., & Lawrence, N. D. (2010). Bayesian Gaussian Process Latent Variable Model

ELBO





$$p(\mathbf{y}) = \int_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{x}|\mathbf{y})}$$

$$q_\theta(\mathbf{x}) \approx p(\mathbf{x}|\mathbf{y})$$

Variational Bayes

$$p(y)$$

Variational Bayes

$$\log p(y)$$

Variational Bayes

$$\log\, p(y) = \log\, p(y) + \int \log \frac{p(x|y)}{p(x|y)}$$

$$
\begin{aligned}
\log\, p(y) &= \log\, p(y) + \int \log\, \frac{p(x|y)}{p(x|y)} \\
&= \int q(x) \log\, p(y) \mathrm{d}x + \int q(x) \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x
\end{aligned}
$$

$$\log\, p(y) = \log\, p(y) + \int \log\, \frac{p(x|y)}{p(x|y)}$$
$$= \int q(x) \log\, p(y) \mathrm{d}x + \int q(x) \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x$$
$$= \int q(x) \log\, \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x$$

Variational Bayes

$$
\begin{aligned}
\log\, p(y) &= \log\, p(y) + \int \log\, \frac{p(x|y)}{p(x|y)} \\
&= \int q(x) \log\, p(y) \mathrm{d}x + \int q(x) \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log\, \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log\, \frac{q(x)}{q(x)} \mathrm{d}x + \int q(x) \log\, p(x, y) \mathrm{d}x + \int q(x)\, \log\, \frac{1}{p(x|y)} \mathrm{d}x
\end{aligned}
$$

Variational Bayes

$$
\begin{aligned}
\log\, p(y) &= \log\, p(y) + \int \log\, \frac{p(x|y)}{p(x|y)} \\
&= \int q(x) \log\, p(y) \mathrm{d}x + \int q(x) \log\, \frac{p(x|y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log\, \frac{p(x|y)p(y)}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log\, \frac{q(x)}{q(x)} \mathrm{d}x + \int q(x) \log\, p(x,y) \mathrm{d}x + \int q(x)\, \log\, \frac{1}{p(x|y)} \mathrm{d}x \\
&= \int q(x) \log\, q(x) \mathrm{d}x + \int q(x) \log\, p(x,y) \mathrm{d}x + \int q(x)\, \log\, \frac{q(x)}{p(x|y)} \mathrm{d}x
\end{aligned}
$$

The $\log$ term

$$KL(q(x)||q(x|y)) = \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x$$

The log term

$$
\begin{aligned}
{\color{red}KL(q(x)||q(x|y))} &= \int q(x)\,\log\frac{q(x)}{p(x|y)}\mathrm{d}x \\
&= -\int q(x)\,\log\frac{p(x|y)}{q(x)}\mathrm{d}x
\end{aligned}
$$

The log term

$$
\begin{aligned}
KL(q(x)||q(x|y)) &= \int q(x) \log \frac{q(x)}{p(x|y)} \mathrm{d}x \\
&= - \int q(x) \log \frac{p(x|y)}{q(x)} \mathrm{d}x \\
&\geq -\log \int p(x|y) \mathrm{d}x = -\log 1 = 0
\end{aligned}
$$

ELBO

$$\begin{aligned} log\;\; p(y) &= \mathrm{KL}(q(x)||p(x|y)) + \underbrace{\mathbb{E}_{q(x)}\left[log\;\; p(x,y)\right] - H(q(x))}_{\mathrm{ELBO}} \\ &\geq \mathbb{E}_{q(x)}\left[log\;\; p(x,y)\right] - H(q(x)) = \mathcal{L}(q(x)) \end{aligned}$$

- if we maximise the ELBO we,
    - find an approximate posterior
    - get an approximation to the marginal likelihood
- *maximising* $p(\mathbf{Y})$ is learning
- finding $p(\mathbf{X}|\mathbf{Y}) \approx q(\mathbf{X})$ is prediction

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do

– Ryan Adams[5]

[5] Talking Machines Season 2, Episode 5

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do
- Taking the expectation of a log is usually easier than the expectation

– Ryan Adams[5]

[5] Talking Machines Season 2, Episode 5

Why is this useful?

Why is this a sensible thing to do?

- If we can't formulate the joint distribution there isn't much we can do
- Taking the expectation of a log is usually easier than the expectation
- We are allowed to choose the distribution to take the expectation over

– Ryan Adams[5]

[5] `Talking Machines Season 2, Episode 5`

Lower Bound[6]

$$\mathcal{L} = \int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}, \mathbf{F}, \mathbf{X})}{q(\mathbf{X})} \right)$$

[6] Damianou, A. C. (2015). Deep Gaussian Processes and Variational Propagation of Uncertainty (Doctoral dissertation)

Lower Bound[6]

$$\mathcal{L} = \int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}, \mathbf{F}, \mathbf{X})}{q(\mathbf{X})} \right)$$

$$\int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}|\mathbf{F}) p(\mathbf{F}|\mathbf{X}) p(\mathbf{X})}{q(\mathbf{X})} \right)$$

[6] Damianou, A. C. (2015). Deep Gaussian Processes and Variational Propagation of Uncertainty (Doctoral dissertation)

Lower Bound[6]

$$
\begin{aligned}
\mathcal{L} = &\int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}, \mathbf{F}, \mathbf{X})}{q(\mathbf{X})} \right) \\
&\int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}|\mathbf{F}) p(\mathbf{F}|\mathbf{X}) p(\mathbf{X})}{q(\mathbf{X})} \right) \\
&= \int_{\mathbf{F},\mathbf{X}} q(\mathbf{X}) \log p(\mathbf{Y}|\mathbf{F}) p(\mathbf{F}|\mathbf{X}) - \int_{\mathbf{X}} q(\mathbf{X}) \log \frac{q(\mathbf{X})}{p(\mathbf{X})}
\end{aligned}
$$

[6] Damianou, A. C. (2015). Deep Gaussian Processes and Variational Propagation of Uncertainty (Doctoral dissertation)

Lower Bound[6]

$$
\begin{aligned}
\mathcal{L} = \int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y},\mathbf{F},\mathbf{X})}{q(\mathbf{X})} \right) \\
\int_{\mathbf{X},\mathbf{F}} q(\mathbf{X}) \log \left(\frac{p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{X})p(\mathbf{X})}{q(\mathbf{X})} \right) \\
= \int_{\mathbf{F},\mathbf{X}} q(\mathbf{X}) \log p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{X}) - \int_{\mathbf{X}} q(\mathbf{X}) \log \frac{q(\mathbf{X})}{p(\mathbf{X})} \\
= \tilde{\mathcal{L}} - \mathrm{KL}\left(q(\mathbf{X}) \,\|\, p(\mathbf{X})\right)
\end{aligned}
$$

[6] Damianou, A. C. (2015). Deep Gaussian Processes and Variational Propagation of Uncertainty (Doctoral dissertation)

Lower Bound

$$\tilde{\mathcal{L}} = \int_{\mathbf{F},\mathbf{X}} q(\mathbf{X}) \log p(\mathbf{Y}|\mathbf{F})p(\mathbf{F}|\mathbf{X})$$

- Has not eliviate the problem at all, X still needs to go through F to reach the data
- Idea of sparse approximations[7]

[7] Quinonero-Candela, Joaquin, & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression & Snelson, E., & Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs

Lower Bound

- Add another set of samples from the same prior

$$p(\mathbf{U}|\mathbf{Z}) = \prod_{j=1}^{d} \mathcal{N}(\mathbf{u}_{:,j}|\mathbf{0}, \mathbf{K})$$

- Add another set of samples from the same prior

$$p(\mathbf{U}|\mathbf{Z}) = \prod_{j=1}^{d} \mathcal{N}(\mathbf{u}_{:,j}|\mathbf{0}, \mathbf{K})$$

- Conditional distribution

$$p(\mathbf{f}_{:,j}, \mathbf{u}_{:,j}|\mathbf{X}, \mathbf{Z}) = p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z})p(\mathbf{u}_{:,j}|\mathbf{Z})$$

$$= \mathcal{N}\left(\mathbf{f}_{:,j}|\mathbf{K}_{fu}(\mathbf{K}_{uu})^{-1}\mathbf{u}_{:,j}, \mathbf{K}_{ff} - \mathbf{K}_{fu}(\mathbf{K}_{uu})^{-1}\mathbf{K}_{uf}\right)\mathcal{N}\left(\mathbf{u}_{:,j}|\mathbf{0}, \mathbf{K}_{uu}\right),$$

Lower Bound

$$p(\mathbf{Y}, \mathbf{F}, \mathbf{U}, \mathbf{X}|\mathbf{Z}) = p(\mathbf{X}) \prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}) p(\mathbf{u}_{:,j}|\mathbf{Z})$$

- we have done nothing to the model, just added *halucinated* observations

Lower Bound

$$p(\mathbf{Y}, \mathbf{F}, \mathbf{U}, \mathbf{X}|\mathbf{Z}) = p(\mathbf{X}) \prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}) p(\mathbf{u}_{:,j}|\mathbf{Z})$$

- we have done nothing to the model, just added *halucinated* observations
- however, we will now interpret $\mathbf{U}$ and $\mathbf{X}_u$ not as random variables but variational parameters

Lower Bound

$$p(\mathbf{Y},\mathbf{F},\mathbf{U},\mathbf{X}|\mathbf{Z}) = p(\mathbf{X})\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j})p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X})p(\mathbf{u}_{:,j}|\mathbf{Z})$$

- we have done nothing to the model, just added *halucinated* observations
- however, we will now interpret $\mathbf{U}$ and $\mathbf{X}_u$ not as random variables but variational parameters
- i.e. parametrise approximate posterior using these parameters (remember sparse motivation)

Lower Bound

- Variational distributions are approximations to intractable posteriors,

$$
\begin{aligned}
q(\mathbf{U}) &\approx p(\mathbf{U}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{F}) \\
q(\mathbf{F}) &\approx p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{Y}) \\
q(\mathbf{X}) &\approx p(\mathbf{X}|\mathbf{Y})
\end{aligned}
$$

Lower Bound

- Variational distributions are approximations to intractable posteriors,

$$q(\mathbf{U}) \approx p(\mathbf{U}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{F})$$
$$q(\mathbf{F}) \approx p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{Y})$$
$$q(\mathbf{X}) \approx p(\mathbf{X}|\mathbf{Y})$$

- Assume that we can *find* $\mathbf{U}$ that completely represents $\mathbf{F}$, i.e. $\mathbf{U}$ is sufficient statistics of $\mathbf{F}$,

$$q(\mathbf{F}) \approx p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z}, \mathbf{Y}) = p(\mathbf{F}|\mathbf{U}, \mathbf{X}, \mathbf{Z})$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{F},\mathbf{U}|\mathbf{X},\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})}$$

Lower Bound

$$
\begin{aligned}
\tilde{\mathcal{L}} &= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{F},\mathbf{U}|\mathbf{X},\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})} \\
&= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j})p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X},\mathbf{Z})p(\mathbf{u}_{:,j}|\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})}
\end{aligned}
$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{p(\mathbf{Y},\mathbf{F},\mathbf{U}|\mathbf{X},\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})}$$

$$= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) \log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j})p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j},\mathbf{X},\mathbf{Z})p(\mathbf{u}_{:,j}|\mathbf{Z})}{q(\mathbf{F})q(\mathbf{U})}$$

- Assume that $\mathbf{U}$ is sufficient statistics for $\mathbf{F}$

$$q(\mathbf{F})q(\mathbf{U})q(\mathbf{X}) = p(\mathbf{F}|\mathbf{U},\mathbf{X},\mathbf{Z})q(\mathbf{U})q(\mathbf{X})$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{\mathbf{X},\mathbf{F},\mathbf{U}} \prod_{j=1}^{d} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) q(\mathbf{u}_{:,j}) q(\mathbf{X})$$

$$\log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) p(\mathbf{u}_{:,j}|\mathbf{Z})}{\prod_{j=1}^{d} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) q(\mathbf{u}_{:,j})} =$$

Lower Bound

$$\tilde{\mathcal{L}} = \int_{\mathbf{X},\mathbf{F},\mathbf{U}} \prod_{j=1}^{d} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) q(\mathbf{u}_{:,j}) q(\mathbf{X})$$

$$\log \frac{\prod_{j=1}^{d} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) p(\mathbf{u}_{:,j}|\mathbf{Z})}{\prod_{j=1}^{d} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) q(\mathbf{u}_{:,j})} =$$

$$= \int_{\mathbf{X},\mathbf{F},\mathbf{U}} \prod_{j=1}^{p} p(\mathbf{f}_{:,j}|\mathbf{u}_{:,j}, \mathbf{X}, \mathbf{Z}) q(\mathbf{u}_{:,j}) q(\mathbf{X}) \log \frac{\prod_{j=1}^{p} p(\mathbf{y}_{:,j}|\mathbf{f}_{:,j}) p(\mathbf{u}_{:,j}|\mathbf{Z})}{\prod_{j=1}^{p} q(\mathbf{u}_{:,j})}$$

$$= \mathbb{E}_{q(\mathbf{F}),q(\mathbf{X}),q(\mathbf{U})} \left[p(\mathbf{Y}|\mathbf{F}) \right] - \mathrm{KL} \left(q(\mathbf{U}) || p(\mathbf{U}|\mathbf{Z}) \right)$$

Summary

$$\mathbb{E}_{q(\mathbf{F}),q(\mathbf{X}),q(\mathbf{U})}\left[p(\mathbf{Y}|\mathbf{F})\right] - \mathrm{KL}\left(q(\mathbf{U})||p(\mathbf{U}|\mathbf{Z})\right) - \mathrm{KL}\left(q(\mathbf{X})||p(\mathbf{X})\right)$$

- Expectation tractable (for some co-variances)
- Reduces to expectations over co-variance functions know as Ψ statistics
- Allows us to place priors and not "regularisers" over the latent representation

Latent space priors

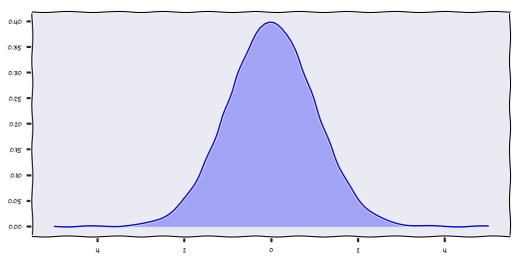
Latent space priors[8]

$$\mathbb{E}_{q(\mathbf{F}),q(\mathbf{X}),q(\mathbf{U})}\left[p(\mathbf{Y}|\mathbf{F})\right] - \mathrm{KL}\left(q(\mathbf{U})||p(\mathbf{U}|\mathbf{Z})\right) - \mathrm{KL}\left(q(\mathbf{X})||p(\mathbf{X})\right)$$

- Importantly $p(\mathbf{X})$ appears only in KL term
- Allows us to express stronger assumptions about the model

[8]Damianou, A. C., Titsias, M., & Lawrence, Neil D, Variational Inference for Uncertainty on the Inputs of Gaussian Process Models (2014)

The Gaussian blob



$$p(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$$

Automatic Relevance Determination

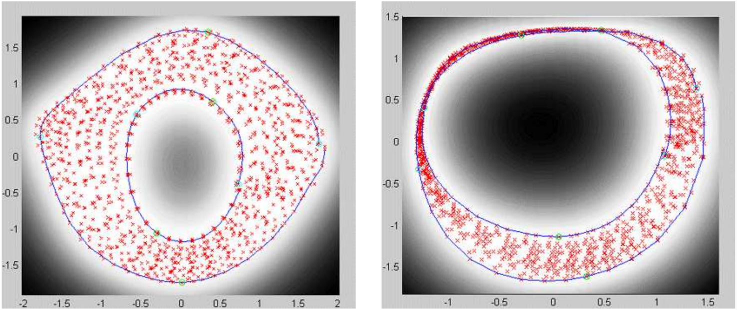
$$k(\mathbf{x}_i, \mathbf{x}_j) = \sigma e^{-\sum_d^D \alpha_d \cdot (x_{i,d} - x_{j,d})^2}$$

GPy

Code

```
RBF(...,ARD=True)
Matern32(...,ARD=True)
```

Dynamic Gaussian Processes[9, 10]



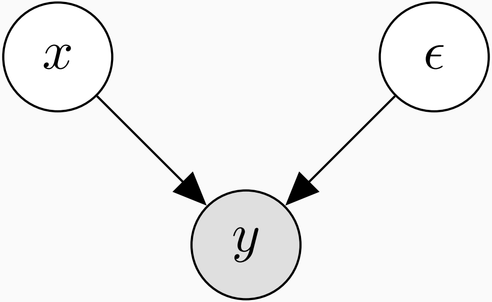
$$p(y, f, x|t) = p(y|f)p(f|x) \underbrace{p(x|t)}_{\sim\mathcal{N}(\mathbf{0},\sqcup)}$$

[9] Urtasun, R., Fleet, D. J., & Fua, P., 3d people tracking with gaussian process dynamical models, CVPR(2006)

[10] Damianou, A. C., Titsias, M., & Lawrence, N. D., Variational Gaussian Process Dynamical Systems, 2011

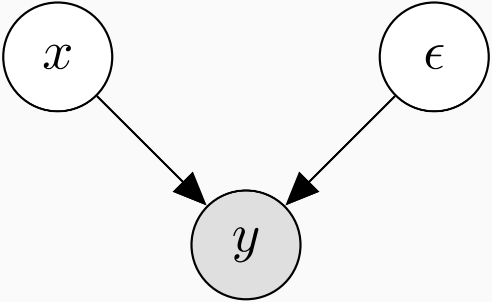
Latent space structures

Explaining Away



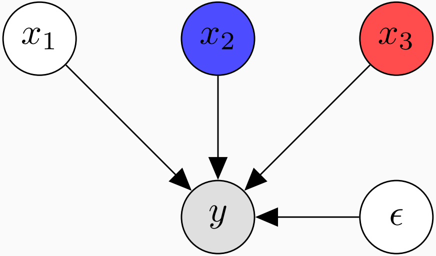
$$y = f(x) + \epsilon$$

Explaining Away



$$y - \epsilon = f(x)$$

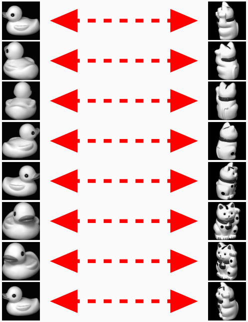
Factor Analysis



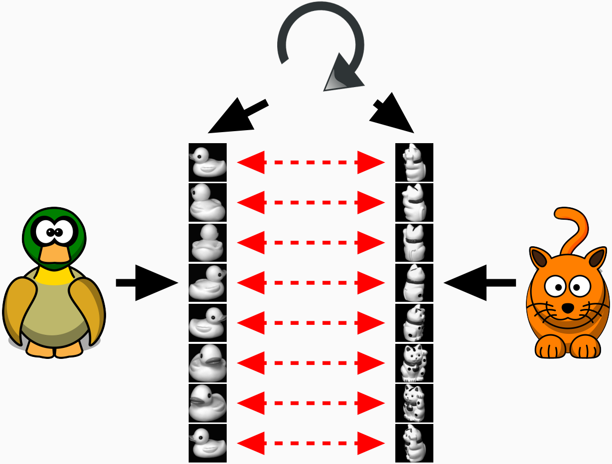
$$y = f(x_1, x_2, x_3) + \epsilon$$

Alignments

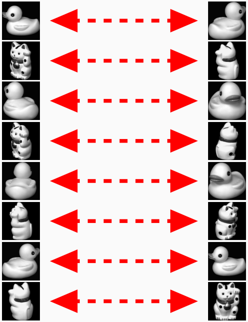
Alignments



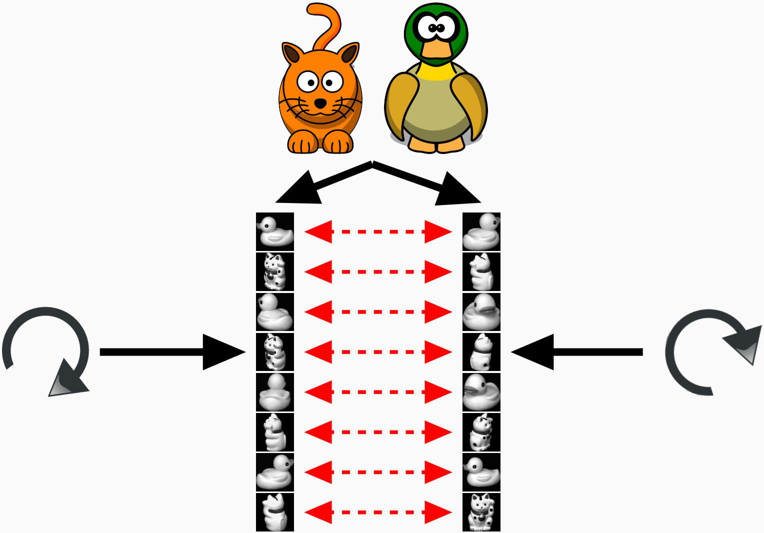
Alignments



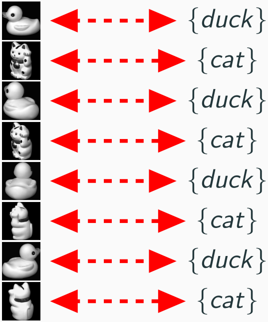
Alignments



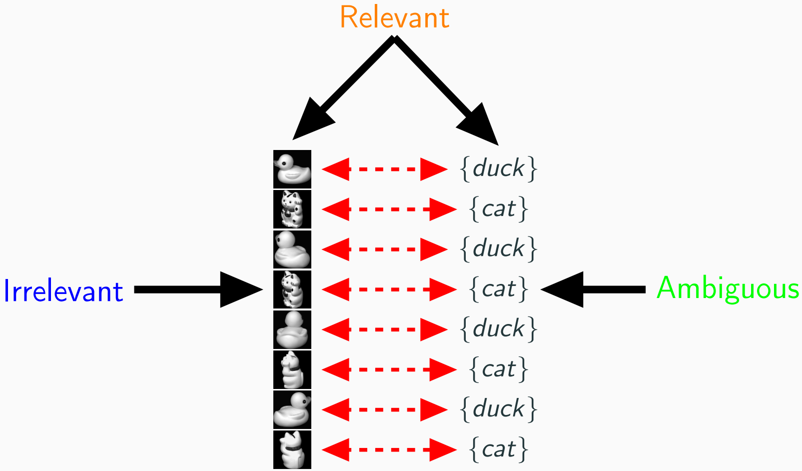
Alignments



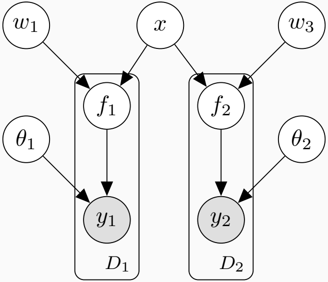
Alignments



Alignments


Relevant
Irrelevant
Ambiguous
{duck}
{cat}
{duck}
{cat}
{duck}
{cat}
{duck}
{cat}

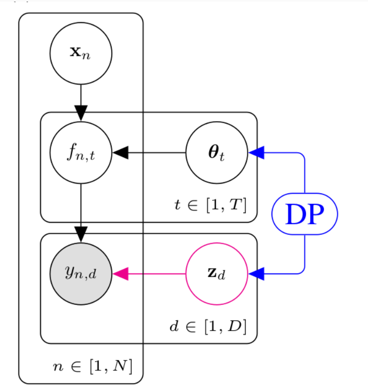
IBFA with GP-LVM[11]

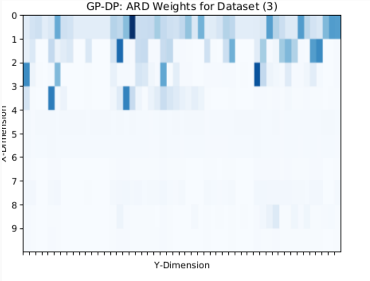


$$y_1 = f(w_1^{\mathrm{T}} x) \quad y_2 = f(w_2^{\mathrm{T}} x)$$

[11] Damianou, A., Lawrence, N. D., & Ek, C. H. (2016). Multi-view learning as a nonparametric nonlinear inter-battery factor analysis

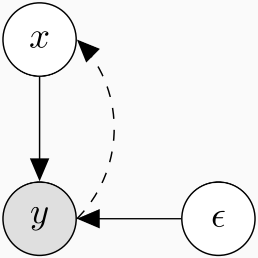
GP-DP[12]





[12] Lawrence, A. R., Ek, C. H., & Campbell, N. W., DP-GP-LVM: A bayesian non-parametric model for learning multivariate dependency structures, ICML (2019)

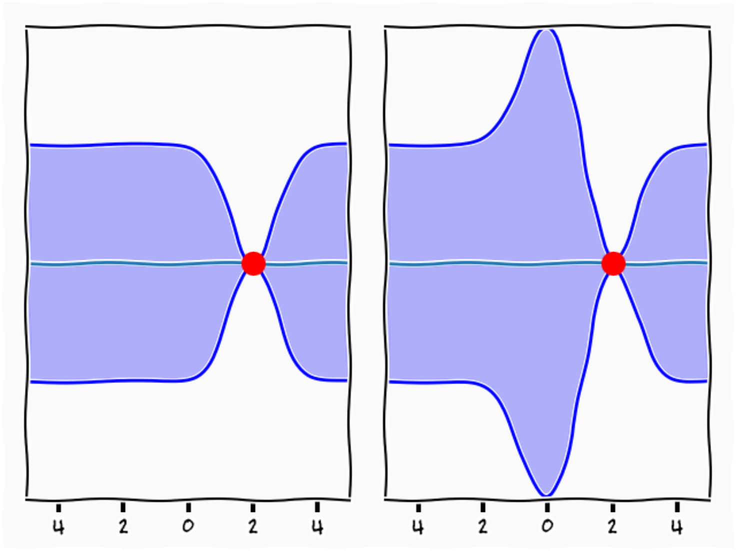
Constrained Latent Space[13]



$$y = f(g(y)) + \epsilon$$

[13] Lawrence, N. D., & Quinonero-Candela, Joaquin, Local distance preservation in the gp-lvm through back constraints, ICML, 2006

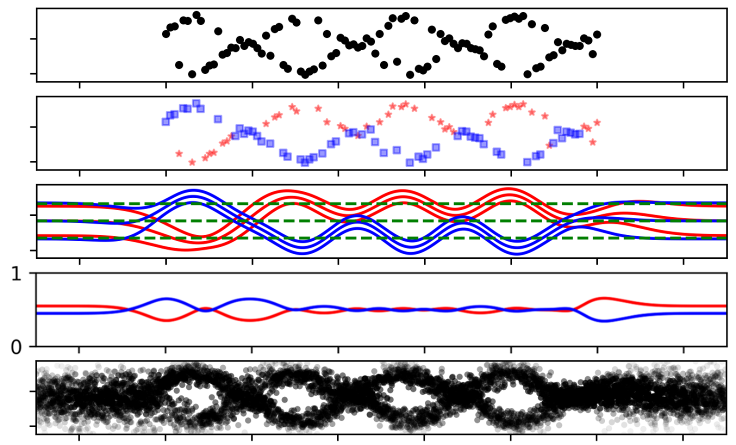
Geometry



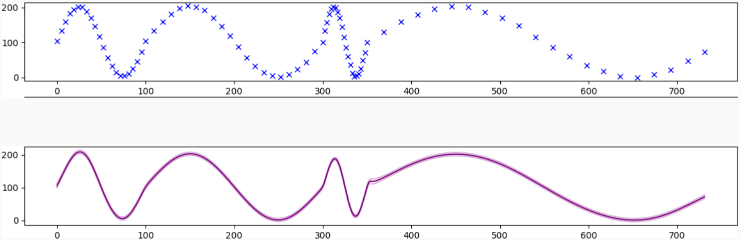
Latent GP-Regression[14]

$$p(\mathbf{Y}|\mathbf{X}) = \int p(\mathbf{Y}|\mathbf{F})\, p(\mathbf{F}|\mathbf{X}, \mathbf{X}^{(C)})\, p(\mathbf{X}^{(C)})\, \mathrm{d}\mathbf{F}\, \mathrm{d}\mathbf{X}^{(C)}.$$

[14] Bodin, E., Campbell, N. D. F., & Ek, C. H., Latent Gaussian Process Regression (2017).


1
0

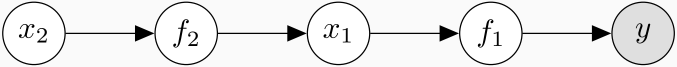
Continous



200
100
0
0
100
200
300
400
500
600
700
200
100
0
0
100
200
300
400
500
600
700

Composite Functions

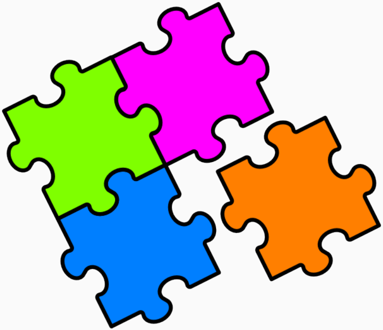
Deep Gaussian Processes[15]



- Place a GP as a warping function, that is warped, ...

[15] Damianou, A. C., & Lawrence, N. D. (2013). Deep Gaussian Processes

Composite Functions

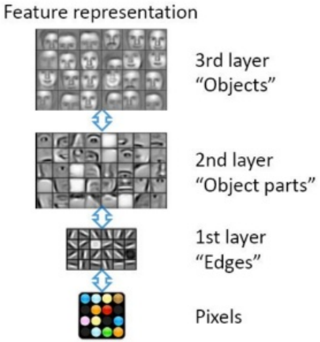


$$y = f_k(f_{k-1}(\dots f_0(x))) = f_k \circ f_{k-1} \circ \dots \circ f_1(x)$$

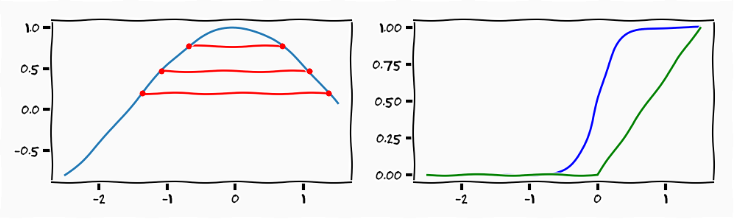
Composite Functions

Diff Levels of Abstraction

- Hierarchical Learning
  - Natural progression from low level to high level structure as seen in natural complexity
  - Easier to monitor what is being learnt and to guide the machine to better subspaces
  - A good lower level representation can be used for many distinct tasks



Composite functions



$$y = f_k(f_{k-1}(\dots f_0(x))) = f_k \circ f_{k-1} \circ \dots \circ f_1(x)$$

$$\mathrm{Kern}(f_1) \subseteq \mathrm{Kern}(f_{k-1} \circ \dots \circ f_2 \circ f_1) \subseteq \mathrm{Kern}(f_k \circ f_{k-1} \circ \dots \circ f_2 \circ f_1)$$

$$\mathrm{Im}(f_k \circ f_{k-1} \circ \dots \circ f_2 \circ f_1) \subseteq \mathrm{Im}(f_k \circ f_{k-1} \circ \dots \circ f_2) \subseteq \dots \subseteq \mathrm{Im}(f_k)$$

Change of Variables

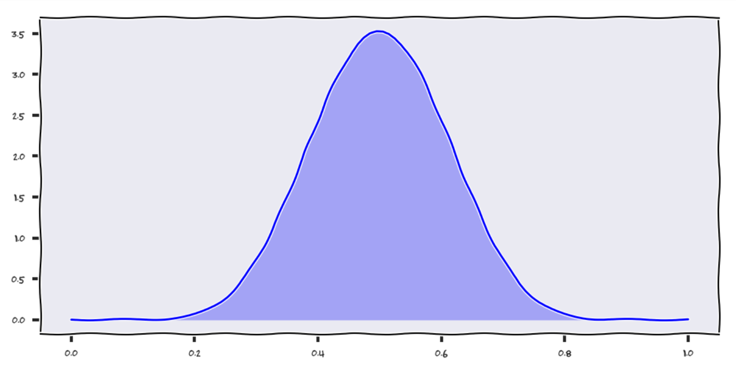
Theorem (Change of Variable)

Let $x \in \mathcal{X} \subseteq \mathbb{R}^n$ be a random vector with a probability density function given by $p_x(x)$, and let $y \in \mathcal{Y} \subseteq \mathbb{R}^n$ be a random vector such that $\psi(y) = x$, where the function $\psi : \mathcal{Y} \to \mathcal{X}$ is bijective of class of $\mathcal{C}^1$ and $| \bigtriangledown \psi(y)| > 0, \forall y \in \mathcal{Y}$. Then, the probability density function $p_y(\cdot)$ induced in $\mathcal{Y}$ is given by

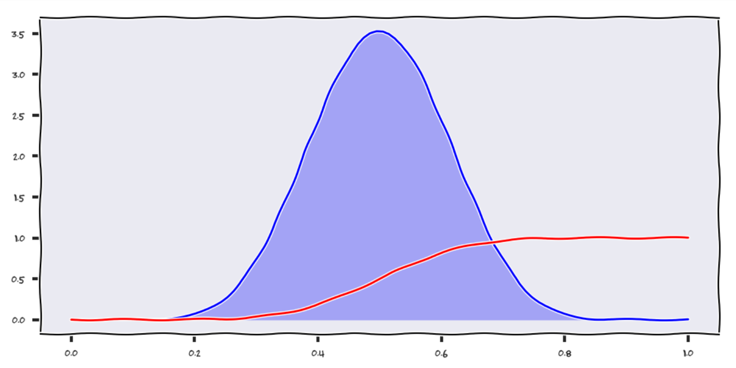
$$p_y(y) = p_x(\psi(y))| \bigtriangledown \psi(y)|$$

where $\bigtriangledown\psi(\cdot)$ denotes the Jacobian of $\psi(\cdot)$, and $| \cdot |$ denotes the determinant operator.

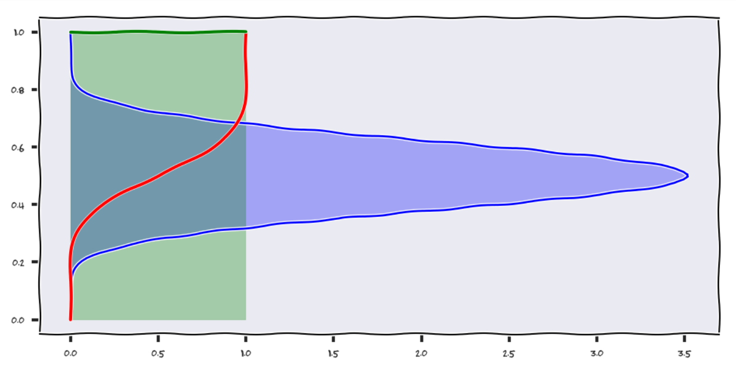
Sampling



Sampling



Sampling



Change of Variables

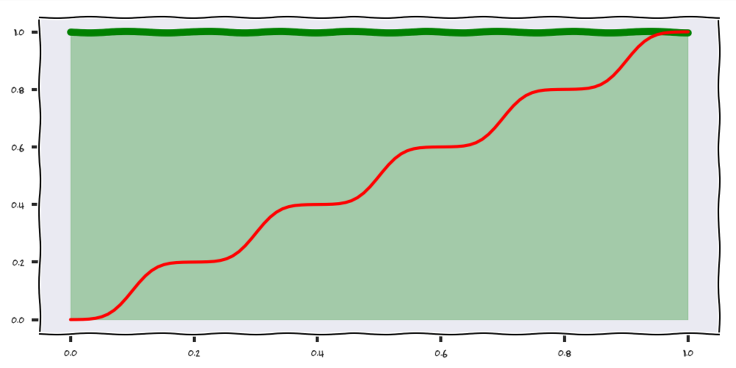
Theorem (Change of Variable)

Let $x \in \mathcal{X} \subseteq \mathbb{R}^n$ be a random vector with a probability density function given by $p_x(x)$, and let $y \in \mathcal{Y} \subseteq \mathbb{R}^n$ be a random vector such that $\psi(y) = x$, where the function $\psi : \mathcal{Y} \to \mathcal{X}$ is bijective of class of $\mathcal{C}^1$ and $| \bigtriangledown \psi(y)| > 0, \forall y \in \mathcal{Y}$. Then, the probability density function $p_y(\cdot)$ induced in $\mathcal{Y}$ is given by

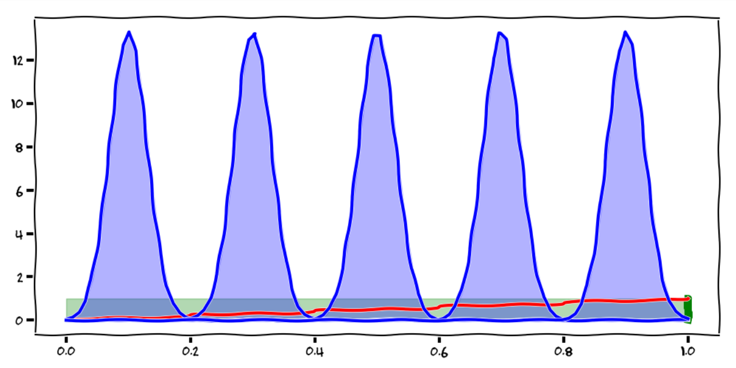
$$p_y(y) = p_x(\psi(y))| \bigtriangledown \psi(y)|$$

where $\bigtriangledown\psi(\cdot)$ denotes the Jacobian of $\psi(\cdot)$, and $| \cdot |$ denotes the determinant operator.

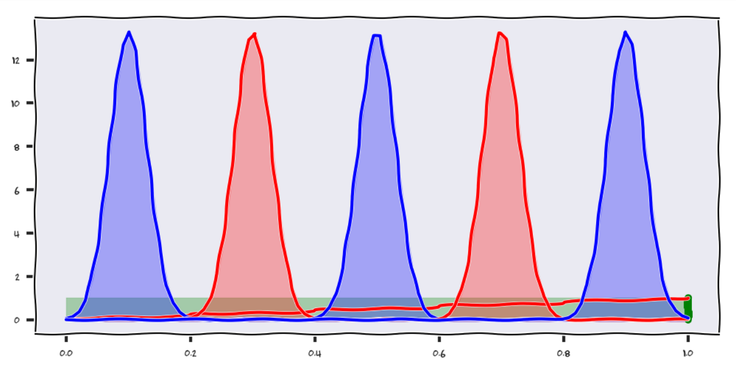
Change of Variables



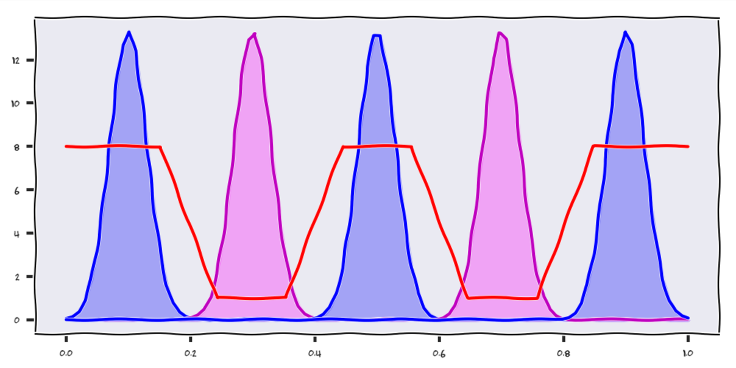
Change of Variables



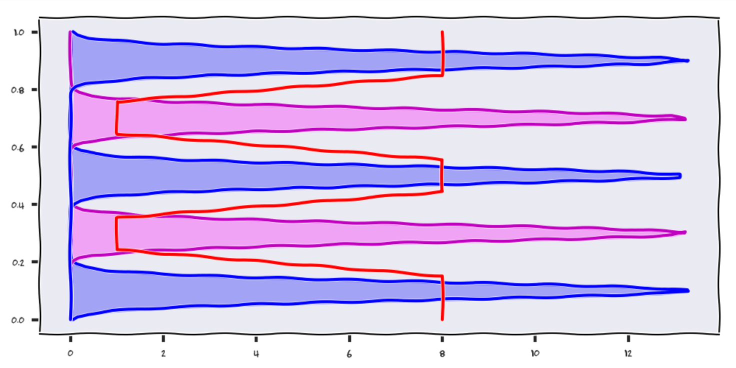
Change of Variables



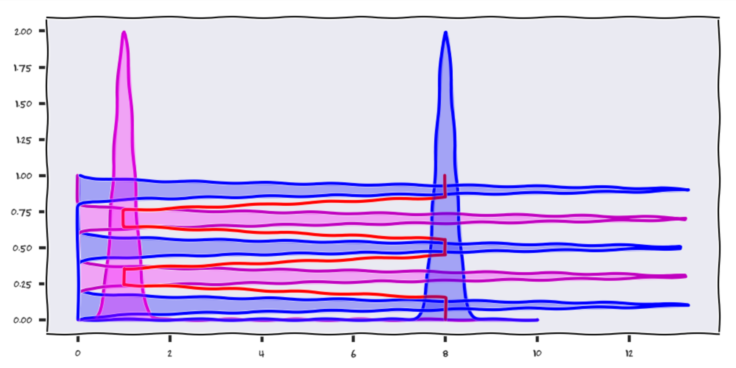
Change of Variables



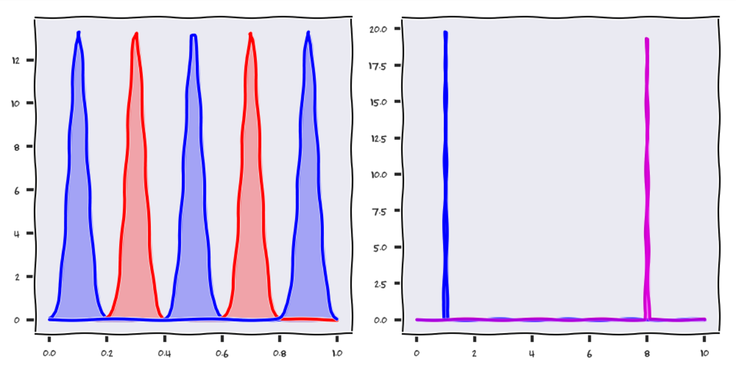
Change of Variables



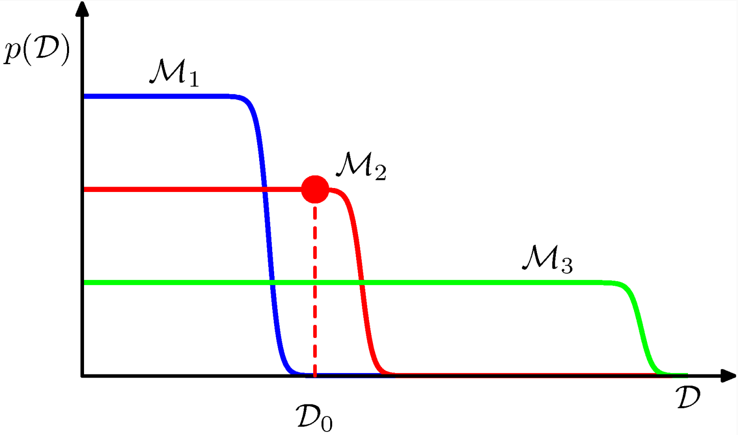
Change of Variables



Change of Variables



MacKay plot



$p(\mathcal{D})$
$\mathcal{M}_1$
$\mathcal{M}_2$
$\mathcal{M}_3$
$\mathcal{D}$
$\mathcal{D}_0$

When do I want Composite Functions

$$y = f_k \circ f_{k-1} \circ \cdots \circ f_1(x)$$

1. My generative process is composite
   - my prior knowledge is composite
2. I want to "re-parametrise" my kernel in a learning setting
   - i have knowledge of the re-parametrisation

Windfarms

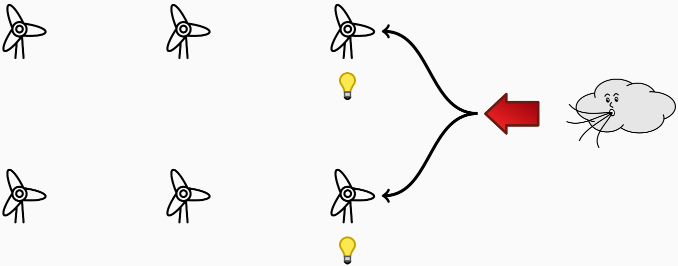
Problem

- Effectiveness of modern windfarm
    - 25-60% (of Betz Limit)
- Turbine has several parameters
    - angle and direction of blades
    - gear
    - etc.

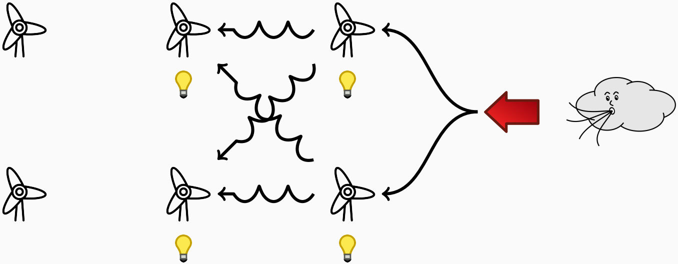
Problem

- Effectiveness of modern windfarm
    - 25-60% (of Betz Limit)
- Turbine has several parameters
    - angle and direction of blades
    - gear
    - etc.
- *How can we maximise the efficiency of a windfarm?*

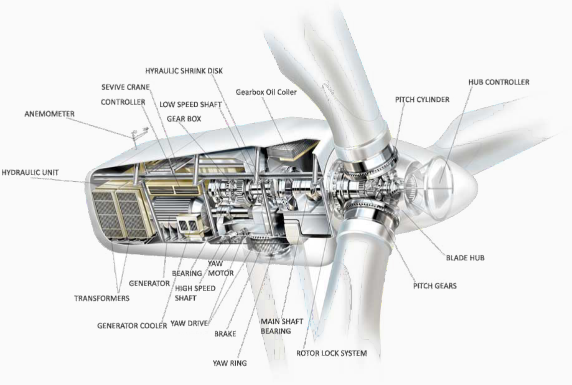
Windfarm



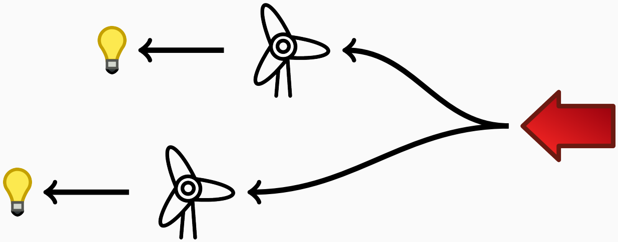
Windfarm



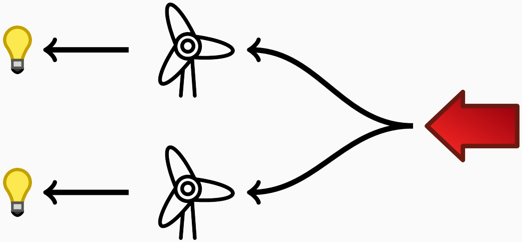
The Wind Turbine



Model

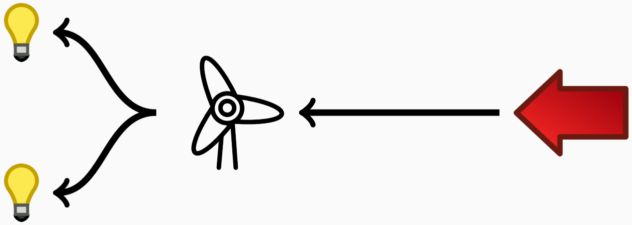


Model: Alignment



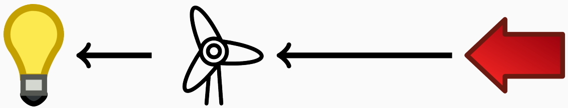
$$w_1(t) = w_2(a(t))$$

Model: Windfront



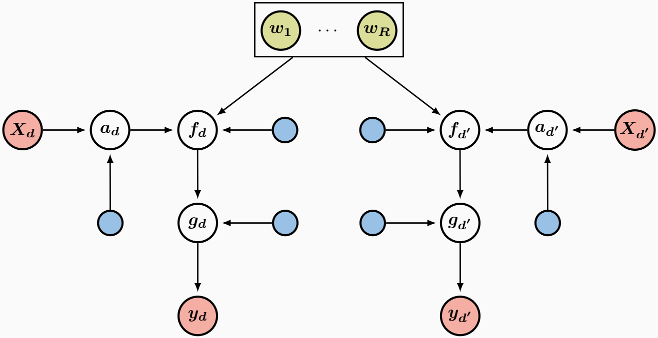
$$f_d(x) = \sum_{r=1}^{R} \int T_{d,r}(x-z) \cdot w_r(z) \frac{\mathrm{d}}{\mathrm{d}z}$$

Model: Windturbine



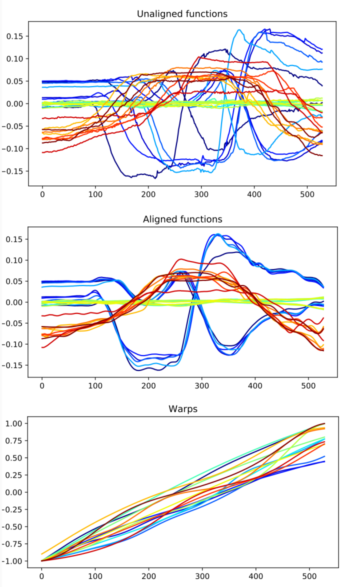
$$y_d = g_d(f_d)$$

Model: Graphical Model [16]

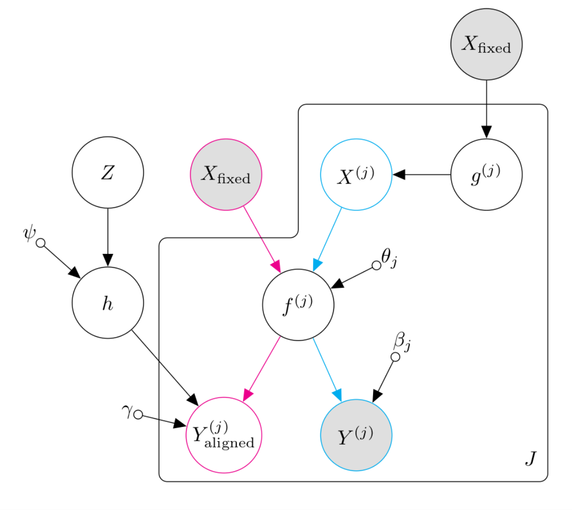


[16]Kaiser, M., Otte, C., Runkler, T., & Ek, C.~H., Bayesian alignments of warped multi-output gaussian processes, NIPS, 2018

Alignment Learning

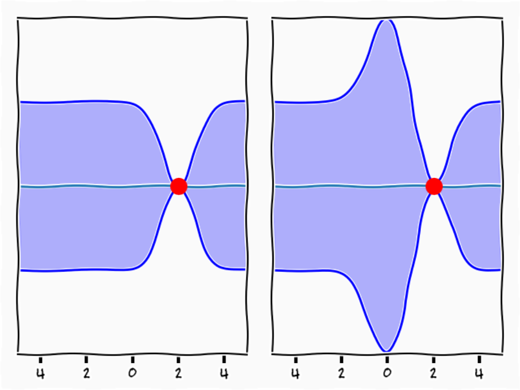


Alignment Learning[17]



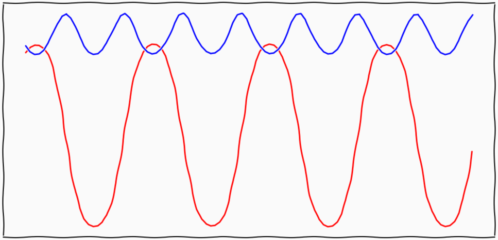
[17] Kazlauskaite, I., Ek, C. H., & Campbell, N. D. F., Gaussian Process Latent Variable Alignment Learning, AISTATS 2019

Kernel Re-Parametrisation

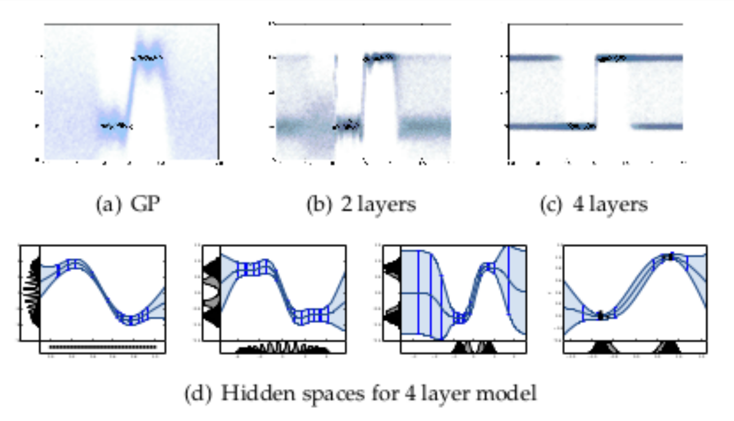


$$k(x_1', x_2') = k(f(x_1), f(x_2)) = k([x_1, z_1], [x_2, z_2])$$

Composition: priors

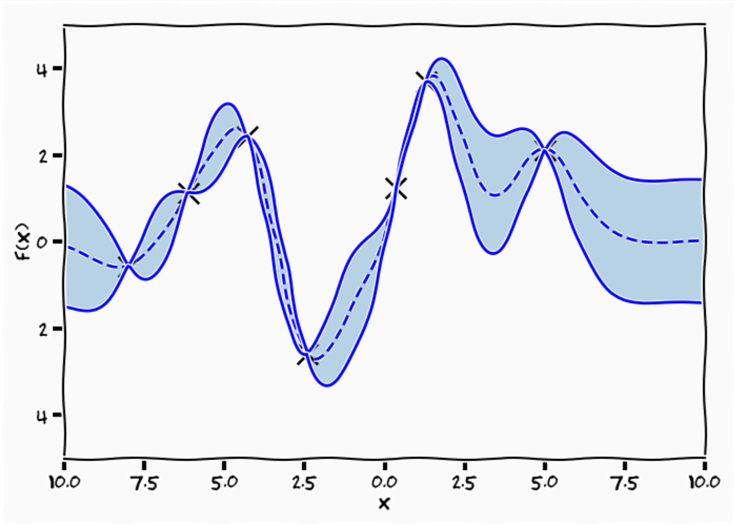


Composition: priors[18]

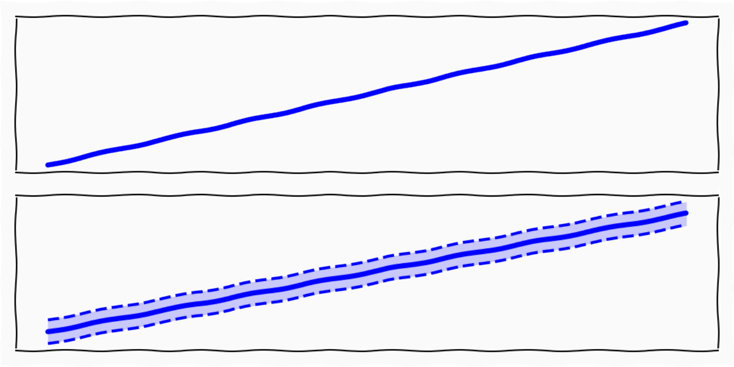


[18] Slides by James Hensman

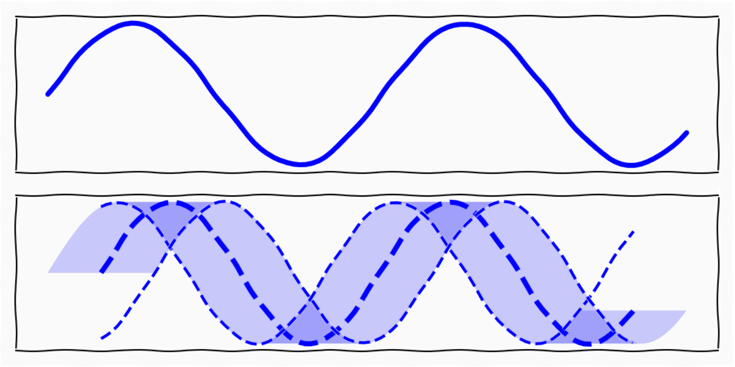
Propagation of Uncertainty



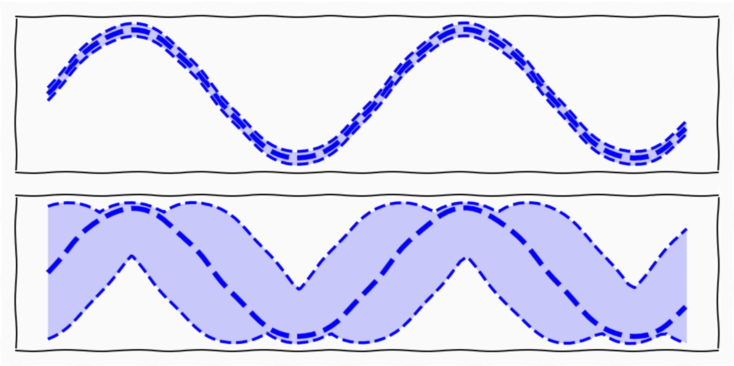
Composition: uncertainty



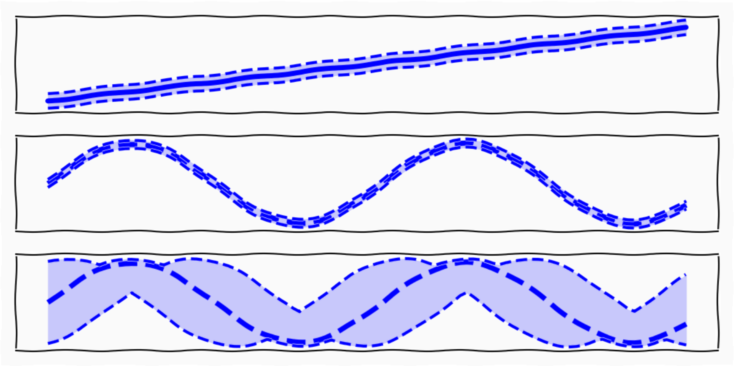
Composition: uncertainty



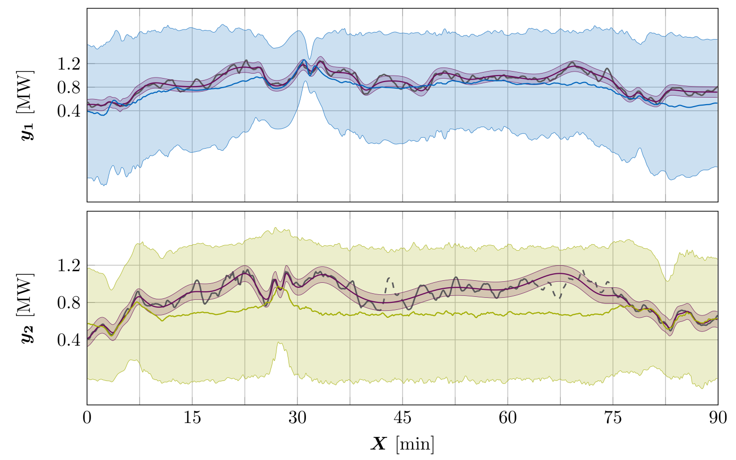
Composition: uncertainty



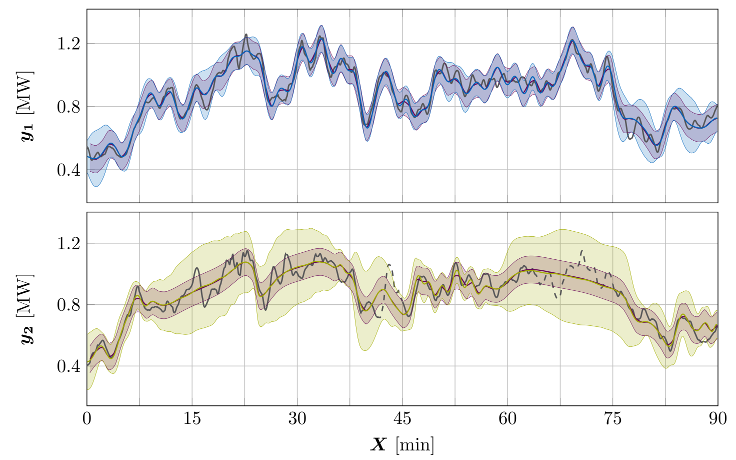
Composition: uncertainty



Composition: uncertainty



Composition: uncertainty



Summary

Summary

- Unsupervised learning is very hard[19]

[19] I would argue that there is no such thing

Summary

- Unsupervised learning is very hard[19]
    - *Its actually not, its really really easy.*

[19] I would argue that there is no such thing

Summary

- Unsupervised learning is very hard[19]
  - *Its actually not, its really really easy.*
- Relevant assumptions needed to learn anything useful

[19] I would argue that there is no such thing

Summary

- Unsupervised learning is very hard[19]
  - *Its actually not, its really really easy.*
- Relevant assumptions needed to learn anything useful
- Strong assumptions needed to learn anything from "sensible" amounts of data

[19] I would argue that there is no such thing

Summary

- Unsupervised learning is very hard[19]
    - *Its actually not, its really really easy.*
- Relevant assumptions needed to learn anything useful
- Strong assumptions needed to learn anything from "sensible" amounts of data
- Stochastic processes such as GPs provide strong, interpretative assumptions that aligns well to our intuitions allowing us to make relevant assumptions

[19] I would argue that there is no such thing

Summary II

- Composite functions cannot model more things

Summary II

- Composite functions cannot model more things
- However, they can easily warp the input space to model less things

Summary II

- Composite functions cannot model more things
- However, they can easily warp the input space to model less things
- This leads to high requirments on data

Summary II

- Composite functions cannot model more things
- However, they can easily warp the input space to model less things
- This leads to high requirments on data
- Even bigger need for uncertainty propagation, we cannot assume noiseless data

Summary II

- Composite functions cannot model more things
- However, they can easily warp the input space to model less things
- This leads to high requirments on data
- Even bigger need for uncertainty propagation, we cannot assume noiseless data
- Intuitions needs to change, we need to think of priors over hierarchies

Summary II

- Composite functions cannot model more things
- However, they can easily warp the input space to model less things
- This leads to high requirments on data
- Even bigger need for uncertainty propagation, we cannot assume noiseless data
- Intuitions needs to change, we need to think of priors over hierarchies
- We need to think about correlated uncertainty, not marginals

eof