DS-GA 3001.008 Modelling time series data

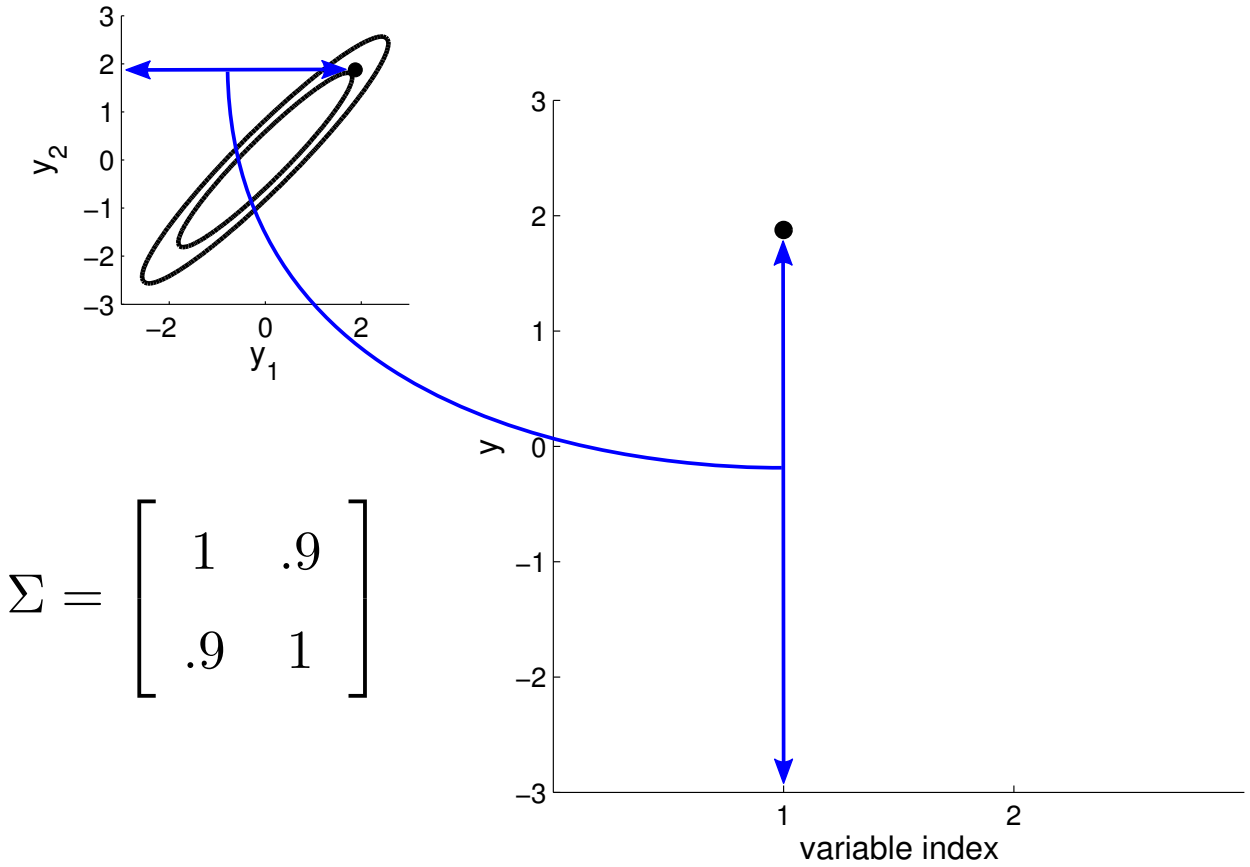
L10. Gaussian Processes for Time Series

Instructor: Cristina Savin

NYU, CNS & CDS

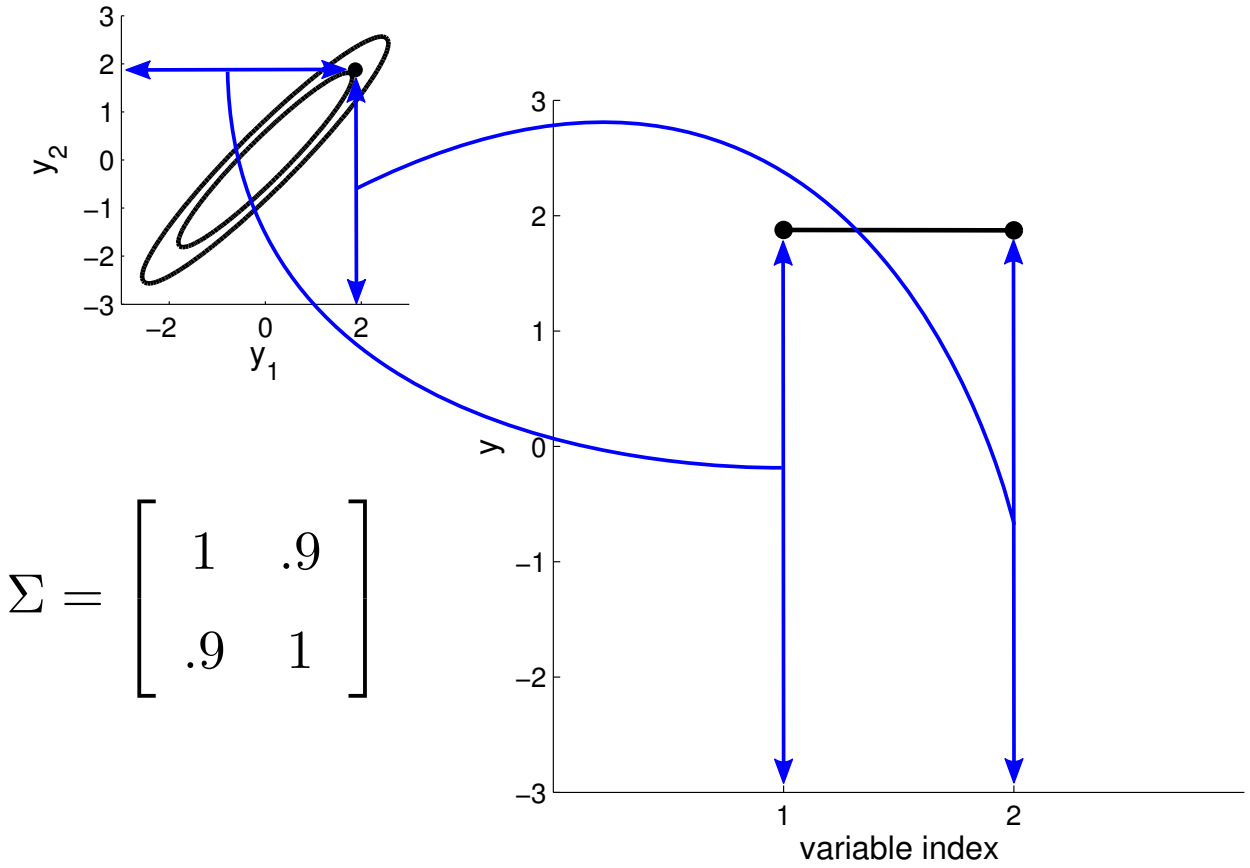
GP intro reminder

New visualisation



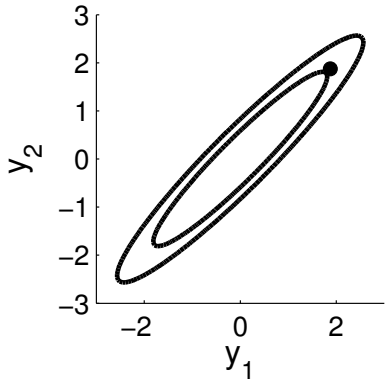
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

New visualisation

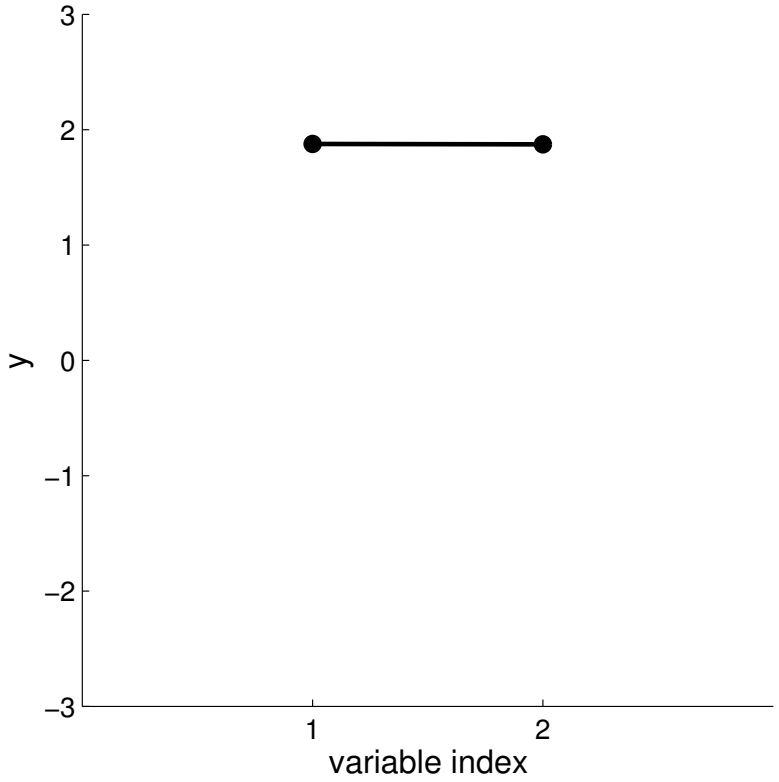


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

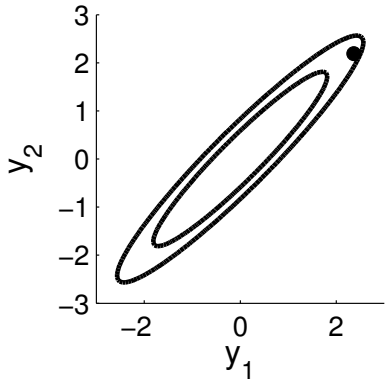
New visualisation



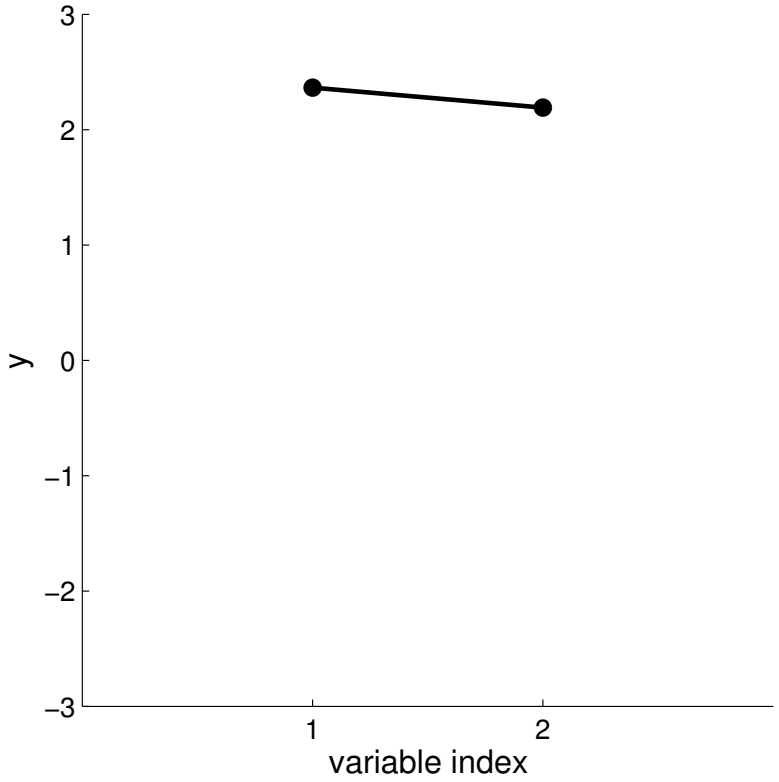
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



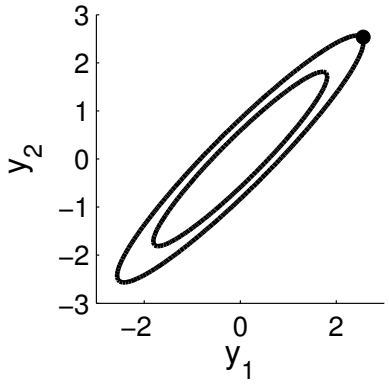
New visualisation



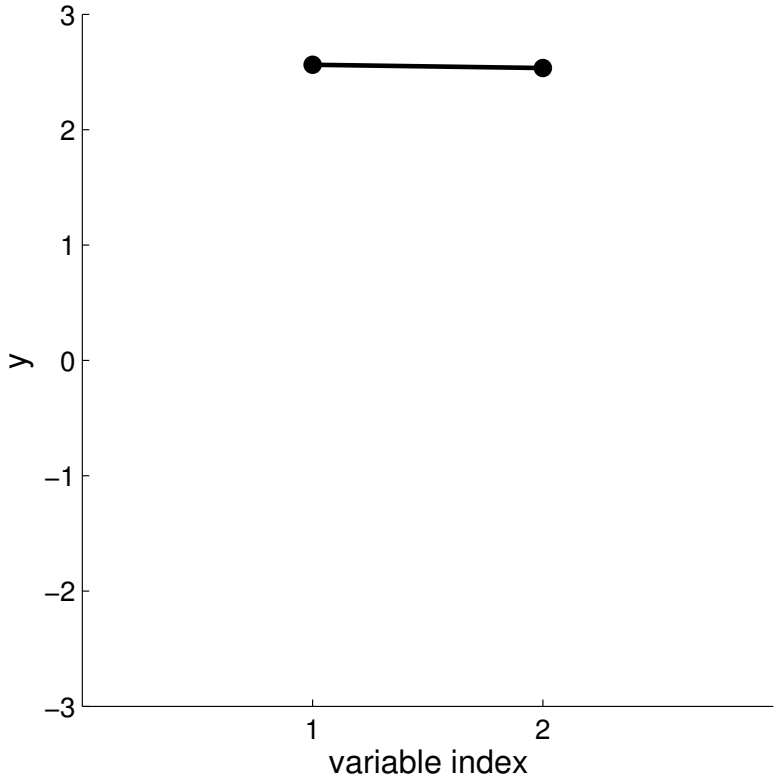
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



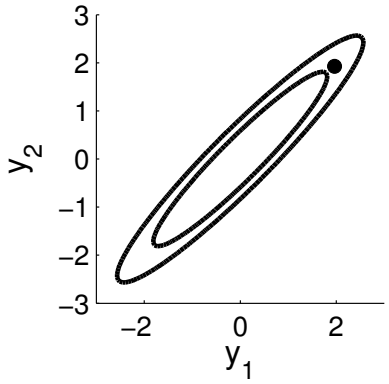
New visualisation



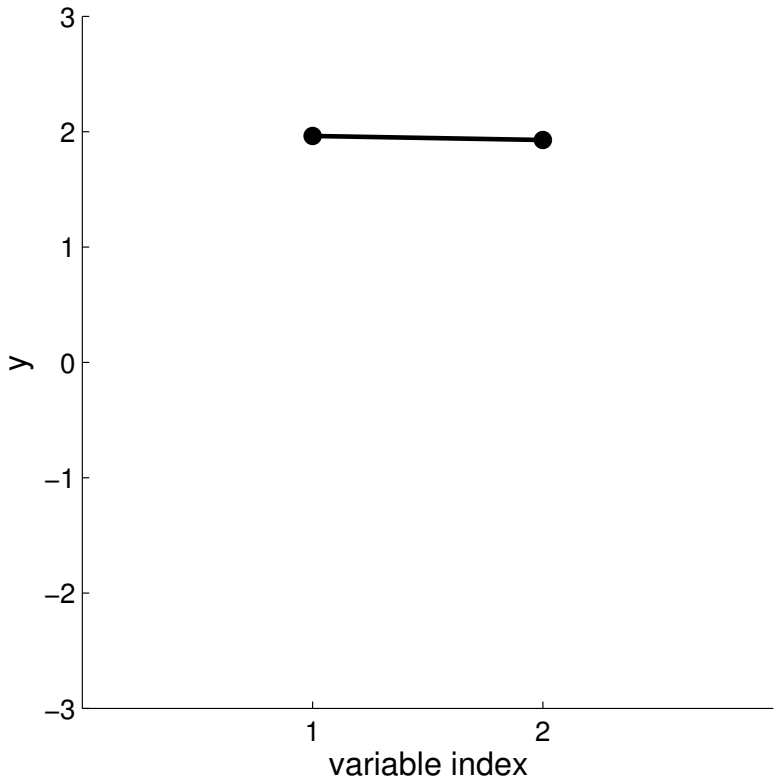
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



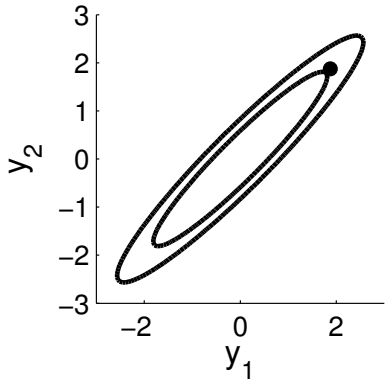
New visualisation



$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

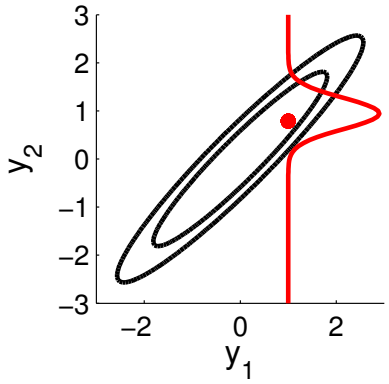


New visualisation

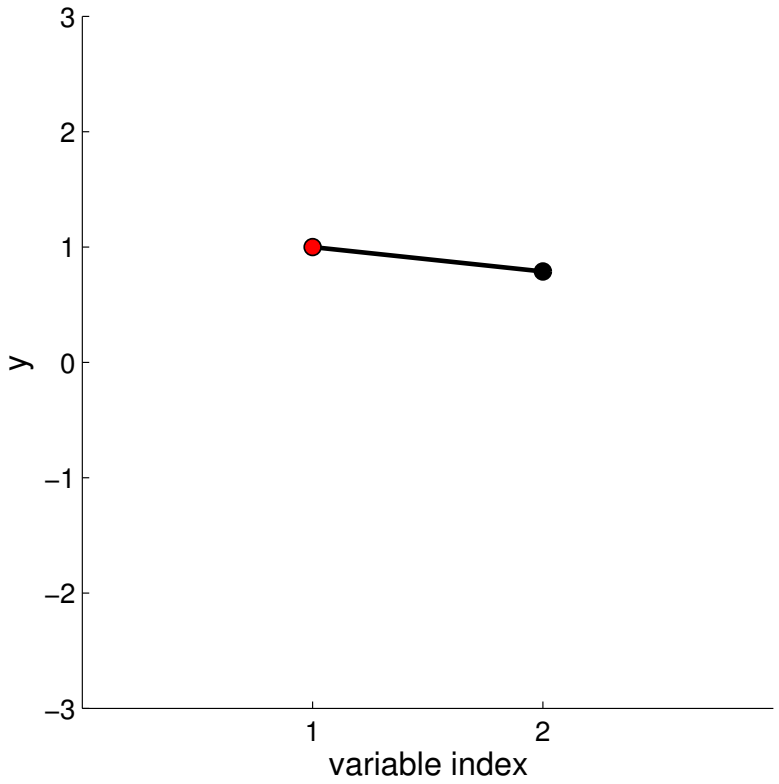


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

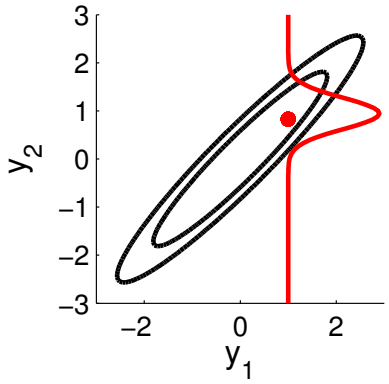
New visualisation



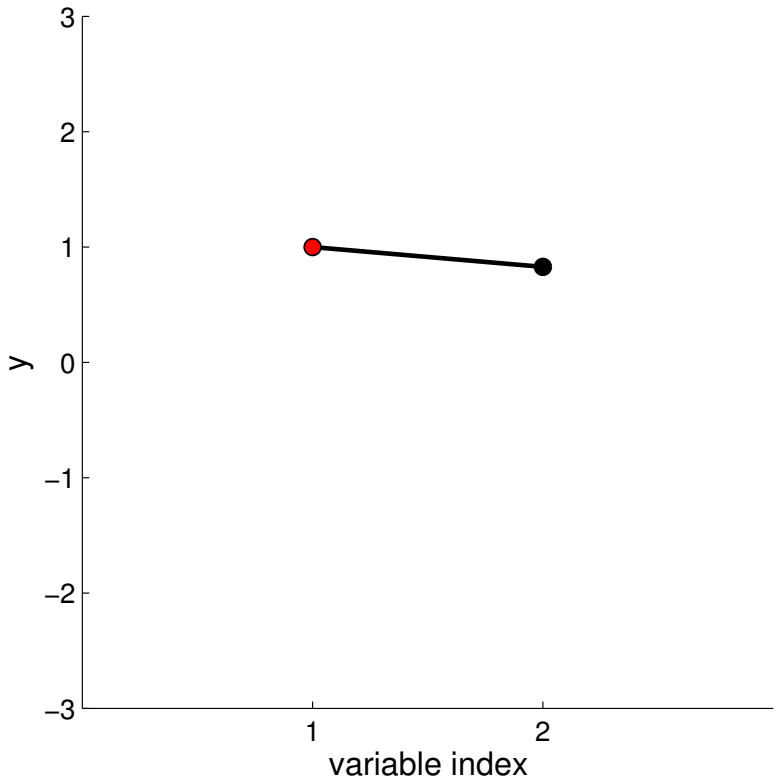
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



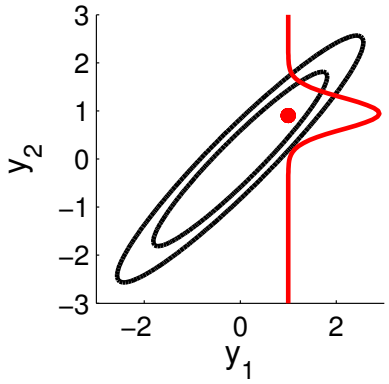
New visualisation



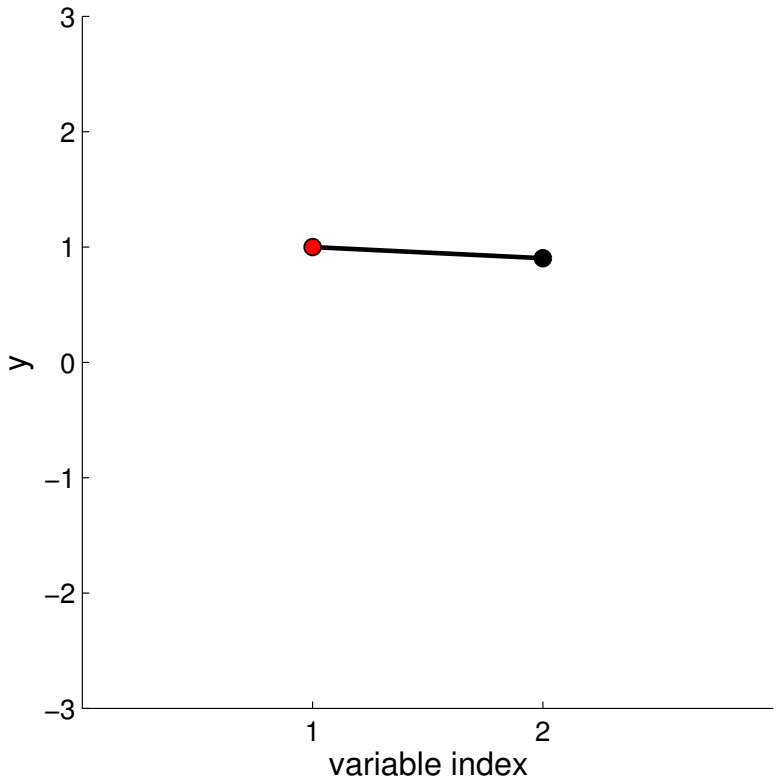
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



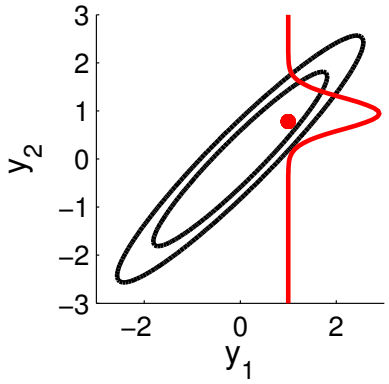
New visualisation



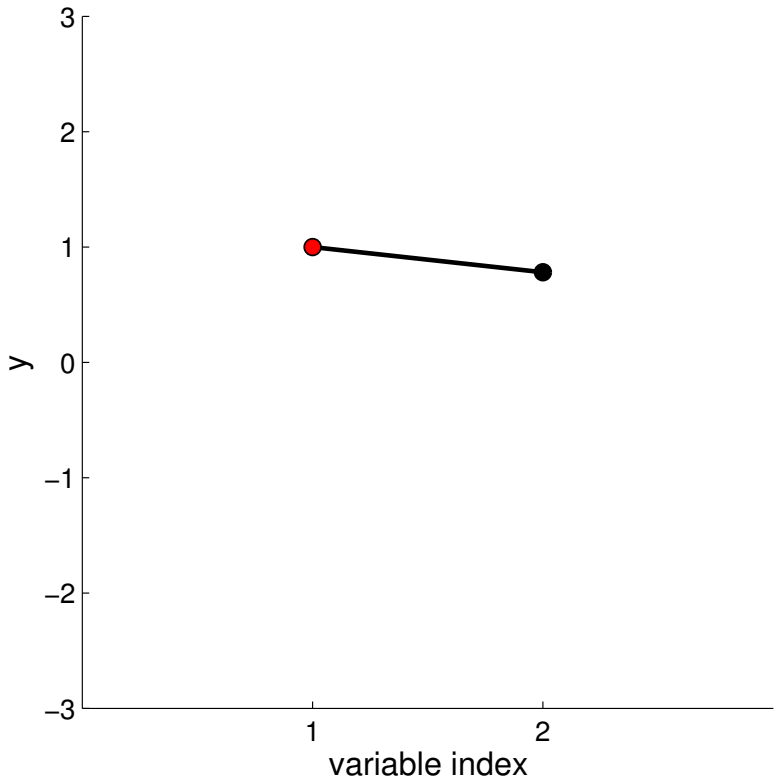
$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$



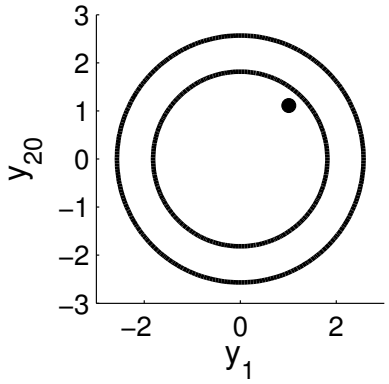
New visualisation

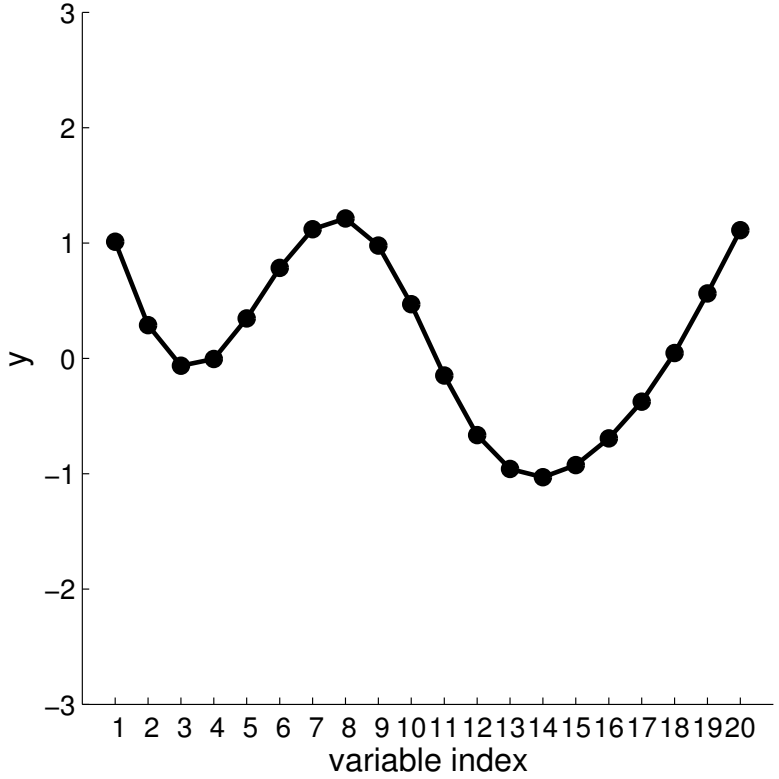


$$\Sigma = \begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$$

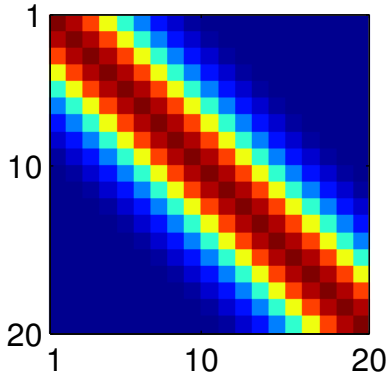


New visualisation

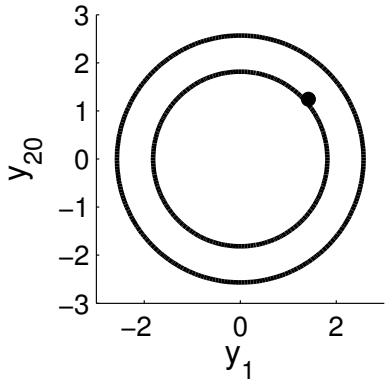


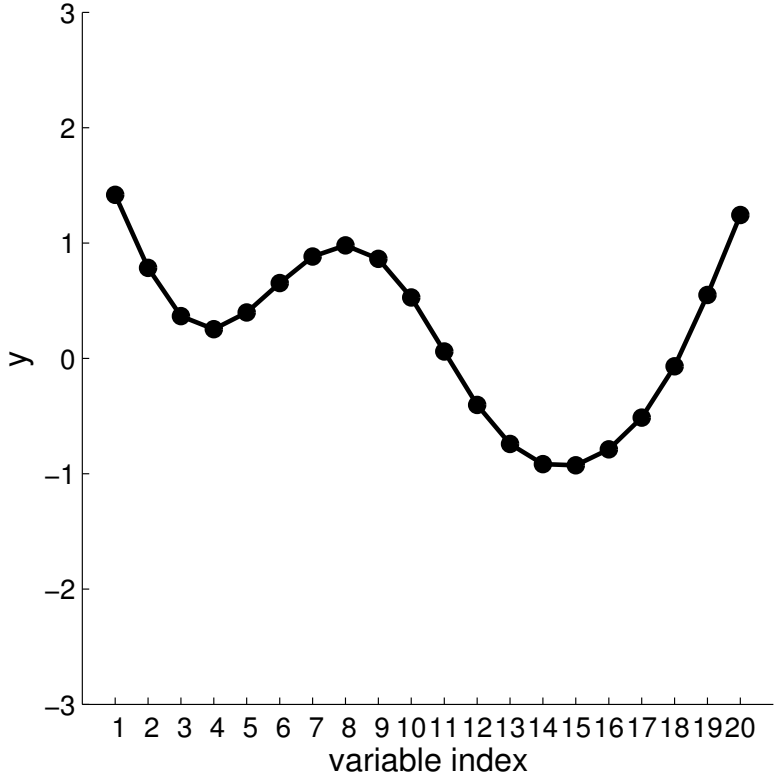


$\Sigma =$

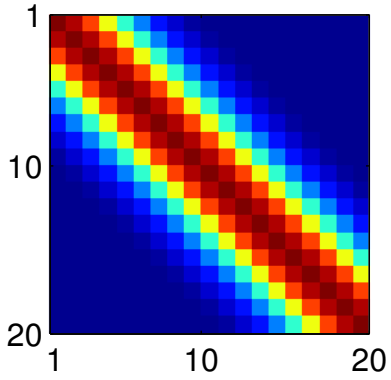


New visualisation

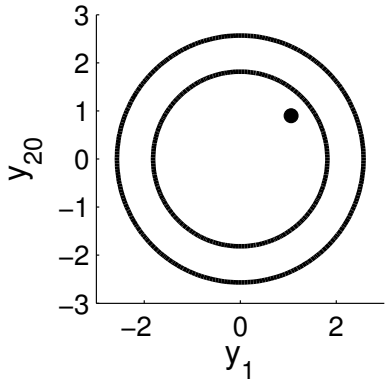


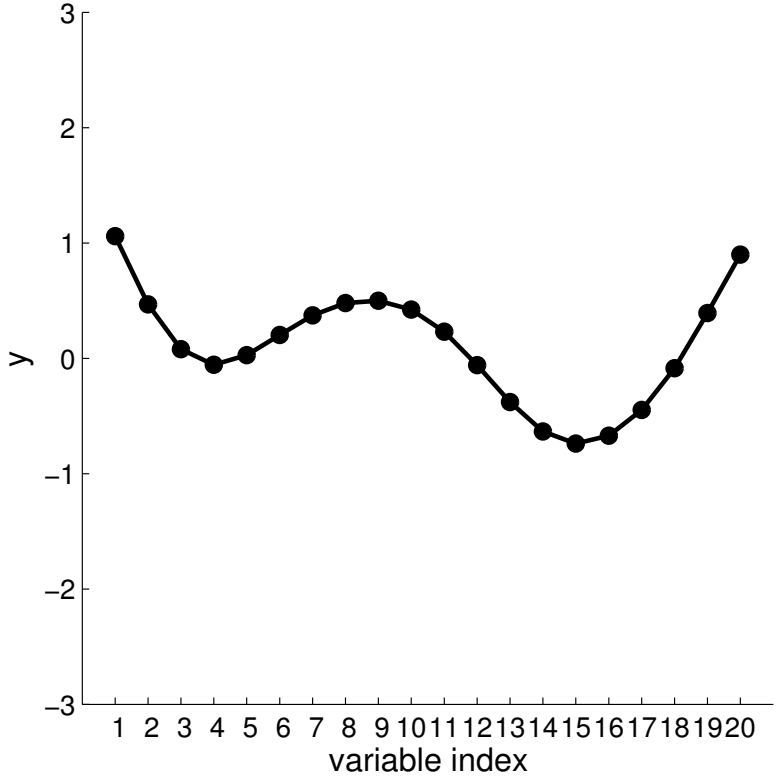


$\Sigma =$

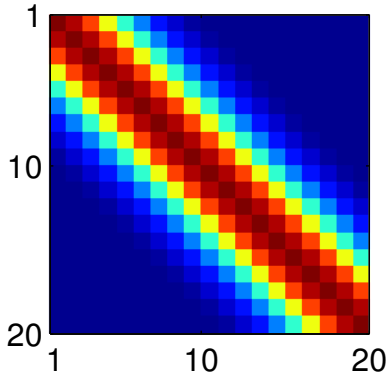


New visualisation

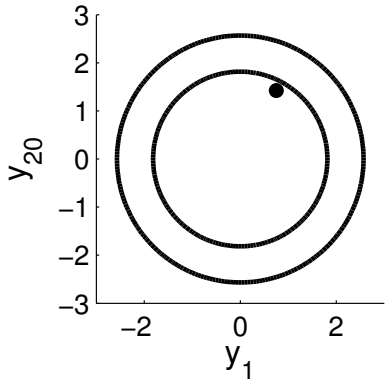


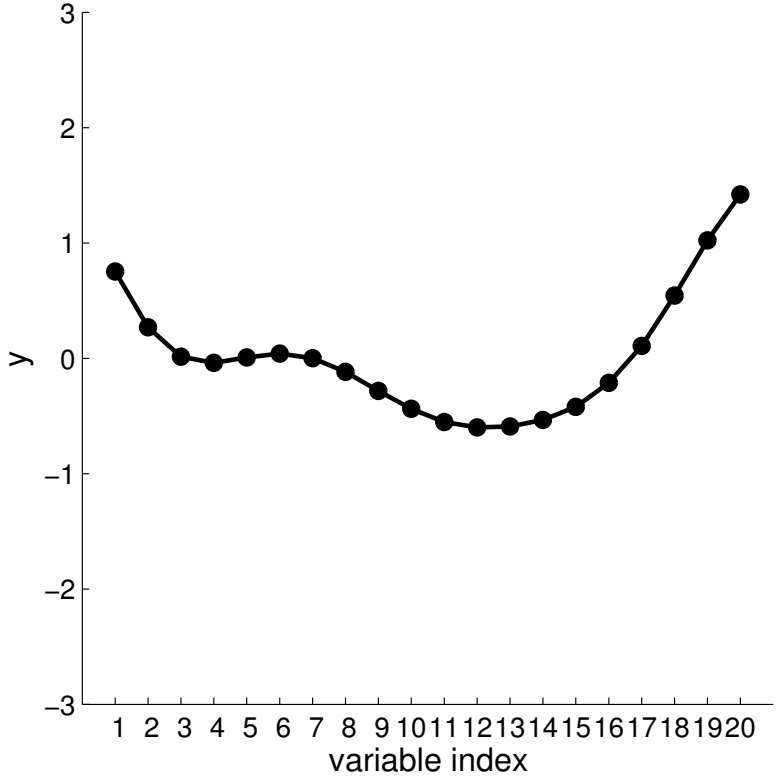


$\Sigma =$

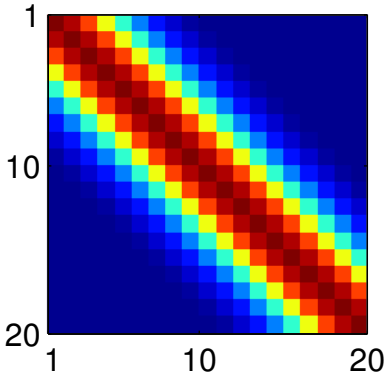


New visualisation

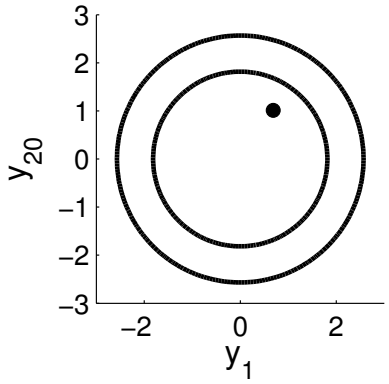


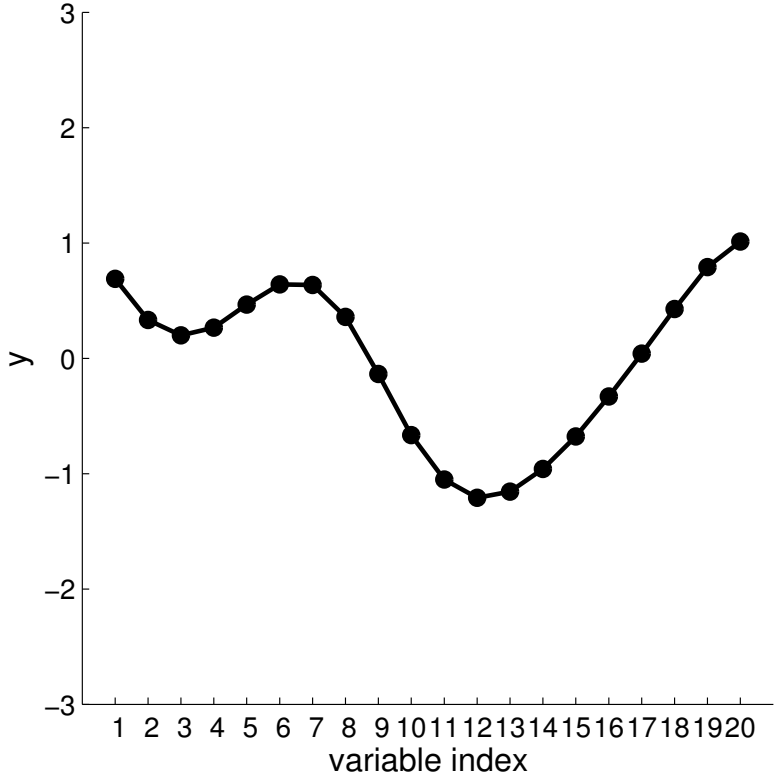


$\Sigma =$

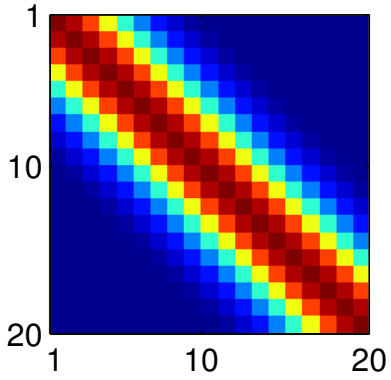


New visualisation

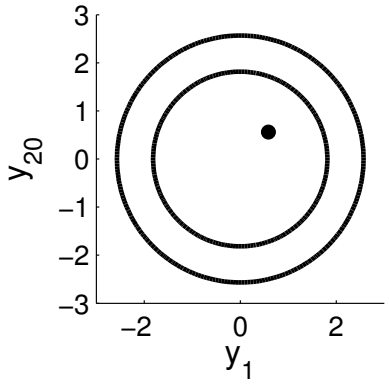


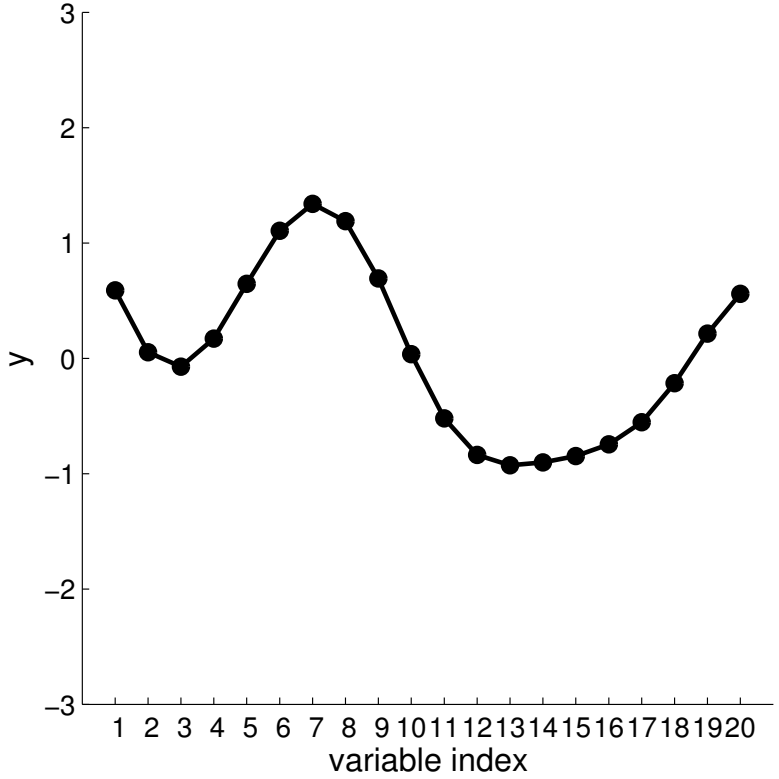


$\Sigma =$

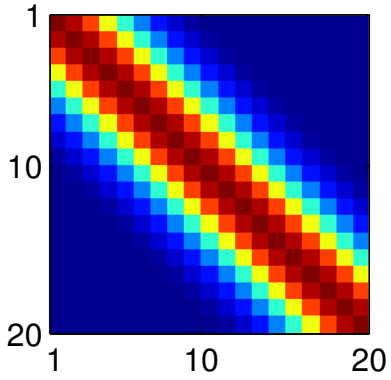


New visualisation



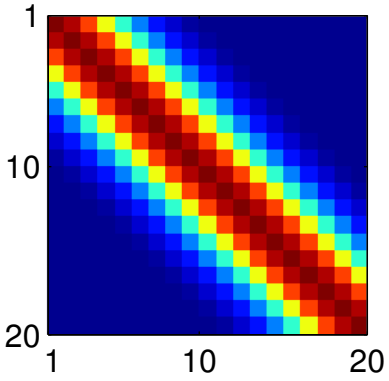


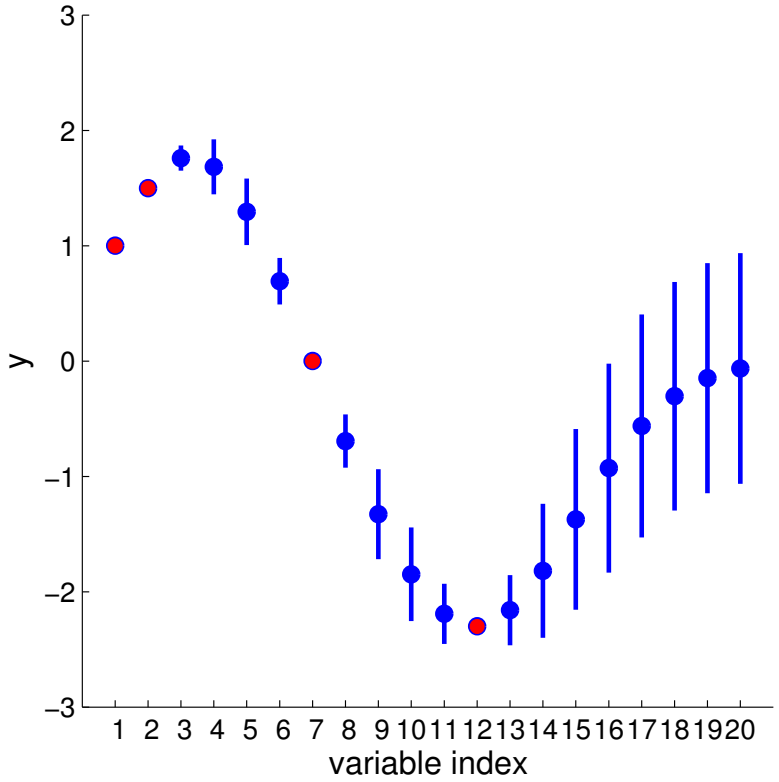
$\Sigma =$



Regression using Gaussians

$\Sigma =$



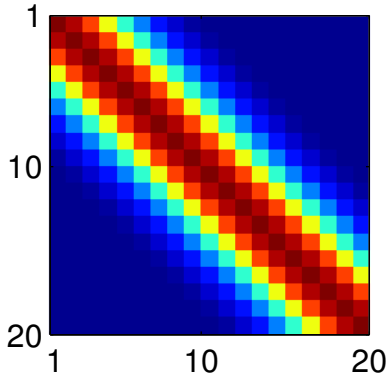


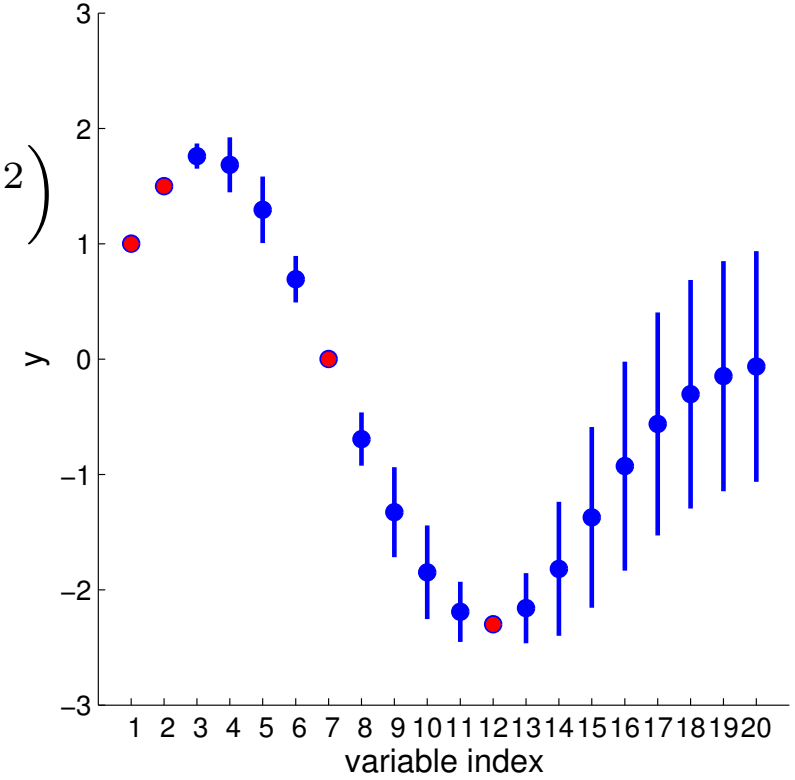
Regression using Gaussians

$$\Sigma(\mathsf{x}_1, \mathsf{x}_2) = \mathsf{K}(\mathsf{x}_1, \mathsf{x}_2) + \mathsf{I}\sigma_{\mathsf{y}}^2$$

$$\mathsf{K}(\mathsf{x}_1, \mathsf{x}_2) = \sigma^2 \exp\left(-\frac{1}{2\mathsf{l}^2}(\mathsf{x}_1 - \mathsf{x}_2)^2\right)$$

$\Sigma =$



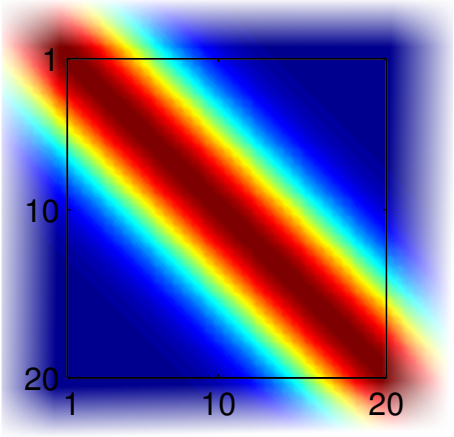


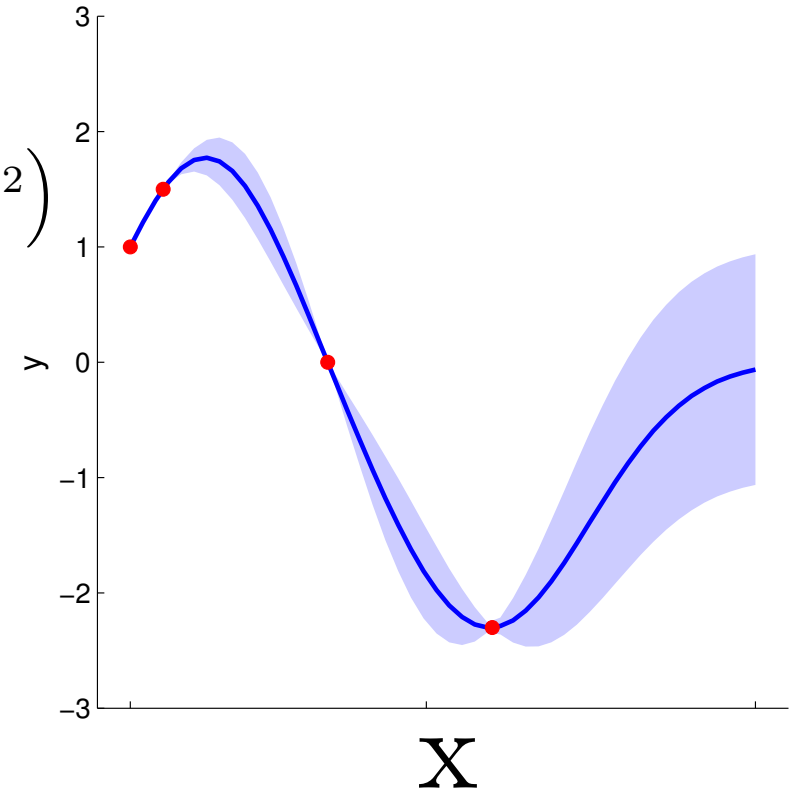
Regression: probabilistic inference in function space

$$\Sigma(x_1, x_2) = K(x_1, x_2) + I\sigma_y^2$$

$$K(x_1, x_2) = \sigma^2 \exp\left(-\frac{1}{2l^2}(x_1 - x_2)^2\right)$$

$\Sigma =$





Regression: probabilistic inference in function space

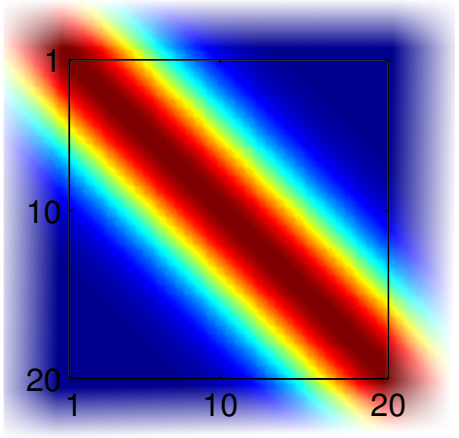
Non-parametric (∞-parametric)

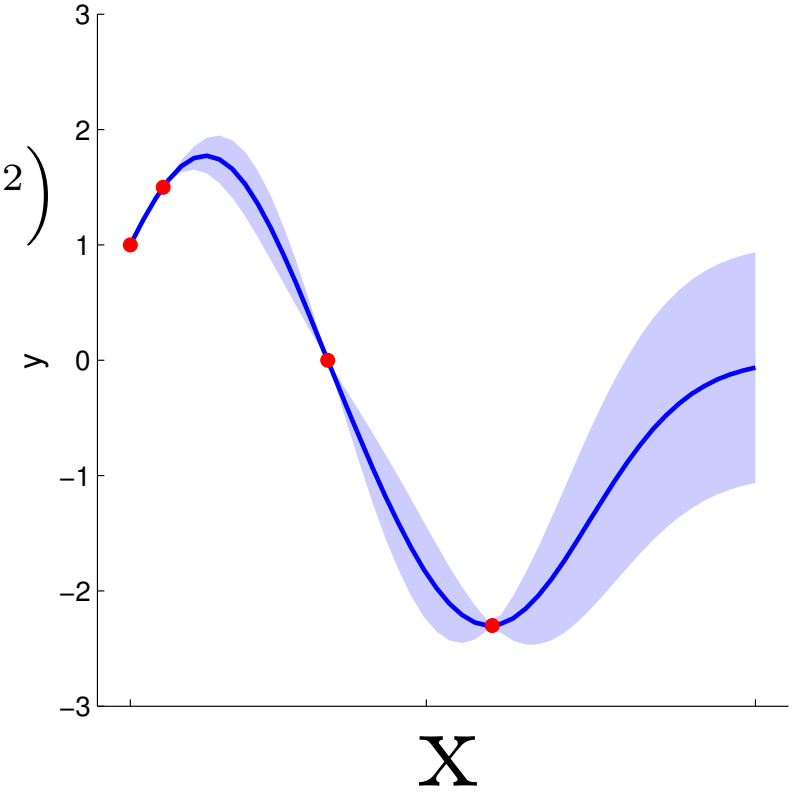
$$p(\mathsf{y}|\theta) = \mathcal{N}(0, \Sigma)$$

$$\Sigma(\mathsf{x}_1, \mathsf{x}_2) = \mathsf{K}(\mathsf{x}_1, \mathsf{x}_2) + \mathsf{I}\sigma_{\mathsf{y}}^2$$

$$\mathsf{K}(\mathsf{x}_1, \mathsf{x}_2) = \sigma^2 \exp\left(-\frac{1}{2\mathsf{l}^2}(\mathsf{x}_1 - \mathsf{x}_2)^2\right)$$

$\Sigma =$



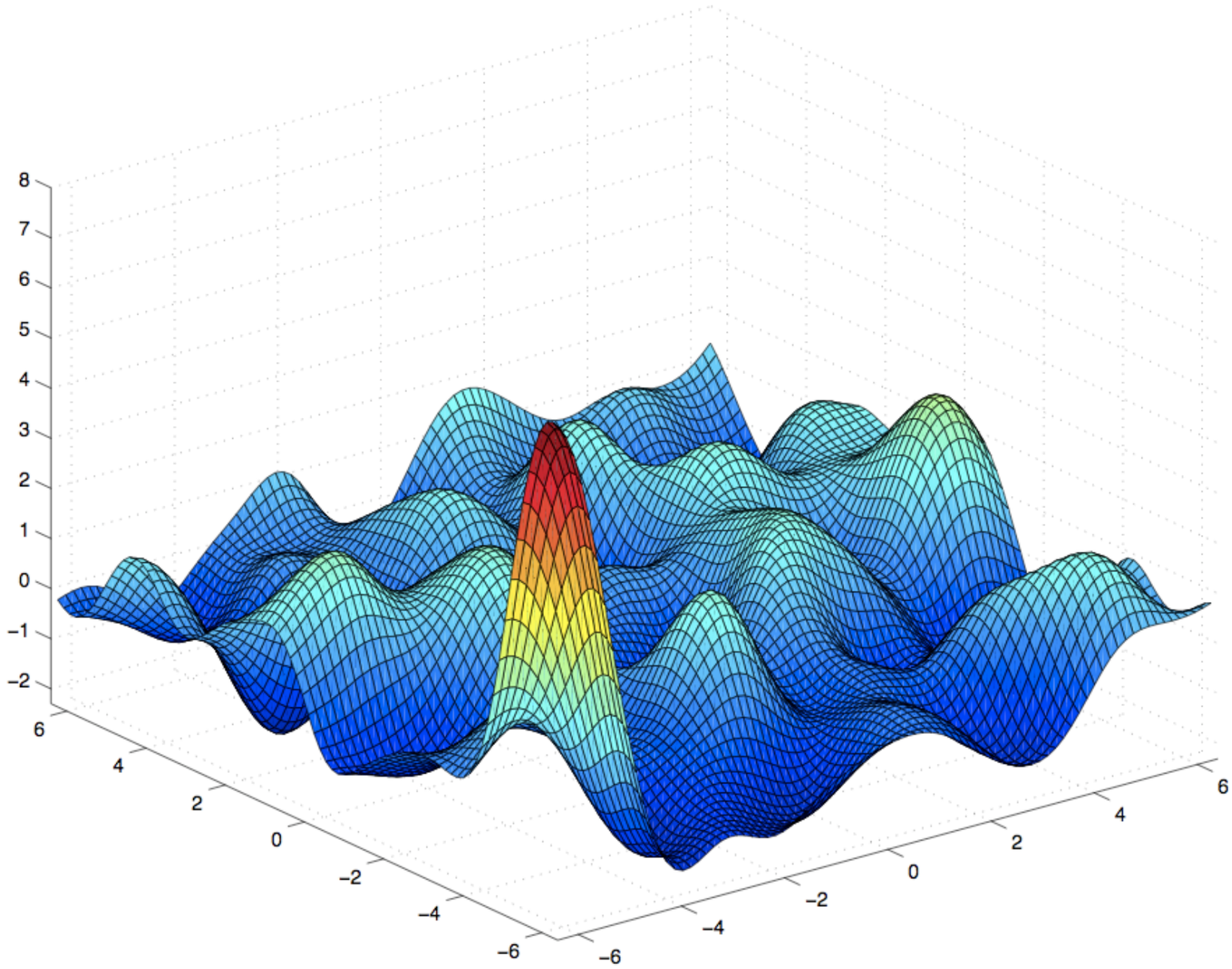


Parametric model

$$\mathsf{y}(\mathsf{x}) = f(\mathsf{x}; \theta) + \sigma_{\mathsf{y}}\epsilon$$

$$\epsilon \sim \mathcal{N}(0, 1)$$

Sample from a 2D Gaussian process



Mathematical Foundations: Definition

Gaussian process = generalization of multivariate Gaussian distribution to infinitely many variables.

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

A Gaussian distribution is fully specified by a mean vector, $\boldsymbol{\mu}$, and covariance matrix Σ:

$$\mathbf{f} = (\mathsf{f}_1, \ldots, \mathsf{f}_n) \sim \mathcal{N}(\mu, \Sigma), \quad \text{indices} \quad i = 1, \ldots, n$$

A Gaussian process is fully specified by a mean function $m(\mathbf{x})$ and covariance function $\mathsf{K}(\mathbf{x}, \mathbf{x}')$:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), \mathsf{K}(\mathbf{x}, \mathbf{x}')\right), \quad \text{indices} \quad \mathbf{x}$$

We seek to learn a function that maps inputs $\mathbf{x}_1, \ldots, \mathbf{x}_N$ to outputs $y_1, \ldots, y_N$.
Instead of assuming a specific form, consider a random function drawn from a GP prior:

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

Any function is possible (no restriction on support) but some are (much) more likely *a priori*.
Observations y_i are taken to be noisy versions of the (almost surely continuous) latent $f(\mathbf{x}_i)$:

$$y_i|\mathbf{x}_i, f(\cdot) \sim \mathcal{N}\left(f(\mathbf{x}_i), \sigma^2\right) \qquad \text{[so } Y \sim \mathcal{N}\left(0, \widetilde{K}_{XX}\right) \text{ with } \widetilde{K}_{XX} = K_{XX} + \sigma^2 I\text{]}$$

Evidence: given by the multivariate Gaussian likelihood:

$$P(Y|X) = |2\pi(K_{XX} + \sigma^2 I)|^{-\frac{1}{2}} e^{-\frac{1}{2} Y (K_{XX} + \sigma^2 I)^{-1} Y^{\mathsf{T}}}$$

Posterior: on latent f is also a GP:

$$f(\cdot)|X, Y \sim \mathcal{GP}(K_{\cdot X}(K_{XX} + \sigma^2 I)^{-1} Y^{\mathsf{T}}, K(\cdot, \cdot) - K_{\cdot X}(K_{XX} + \sigma^2 I)^{-1} K_{X \cdot}))$$

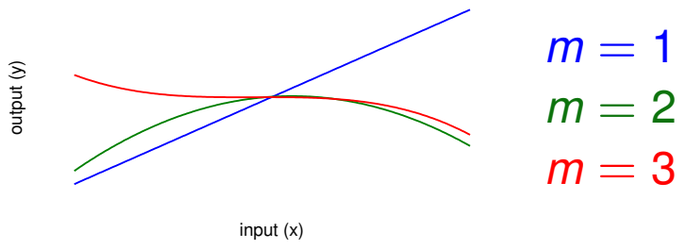
Predictions: posterior on f, plus observation noise:

$$y|X, Y, \mathbf{x} \sim \mathcal{N}\left(E[f(\mathbf{x})|X, Y], \text{Var}[f(\mathbf{x})|X, Y] + \sigma^2\right) = \mathcal{N}\left(K_{\mathbf{x}X} \widetilde{K}_{XX}^{-1} Y, K_{\mathbf{x}X} \widetilde{K}_{XX}^{-1} K_{X\mathbf{x}} + \sigma^2\right)$$

- Polynomial:

$$K(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^{\mathsf{T}}\mathbf{x}')^m \qquad m = 1, 2, \ldots$$

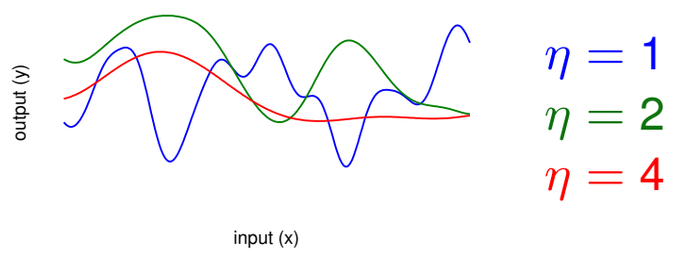
f is inhomogeneous polynomial of degree m



- Squared-exponential (or exponentiated-quadratic):

$$K(\mathbf{x}, \mathbf{x}') = \theta^2 e^{-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\eta^2}}$$

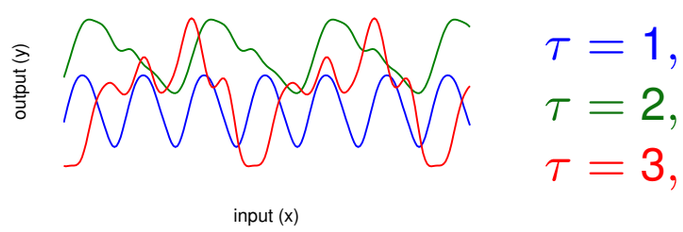
f is smooth (C^∞ almost surely) on length scale η



- Periodic (exp-sine):

$$K(x, x') = \theta^2 e^{-\frac{2\sin^2(\pi(x-x')/\tau)}{\eta^2}}$$

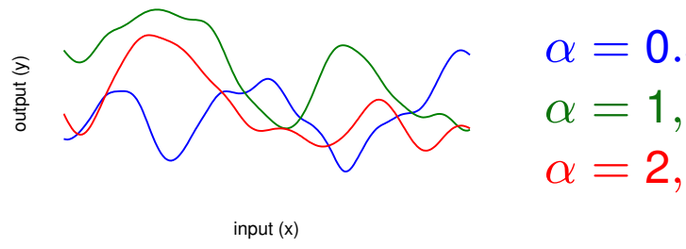
f is smooth and periodic



- Rational Quadratic:

$$K(\mathbf{x}, \mathbf{x}') = \left(1 + \frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\alpha\eta^2}\right)^{-\alpha} \qquad \alpha > 0$$

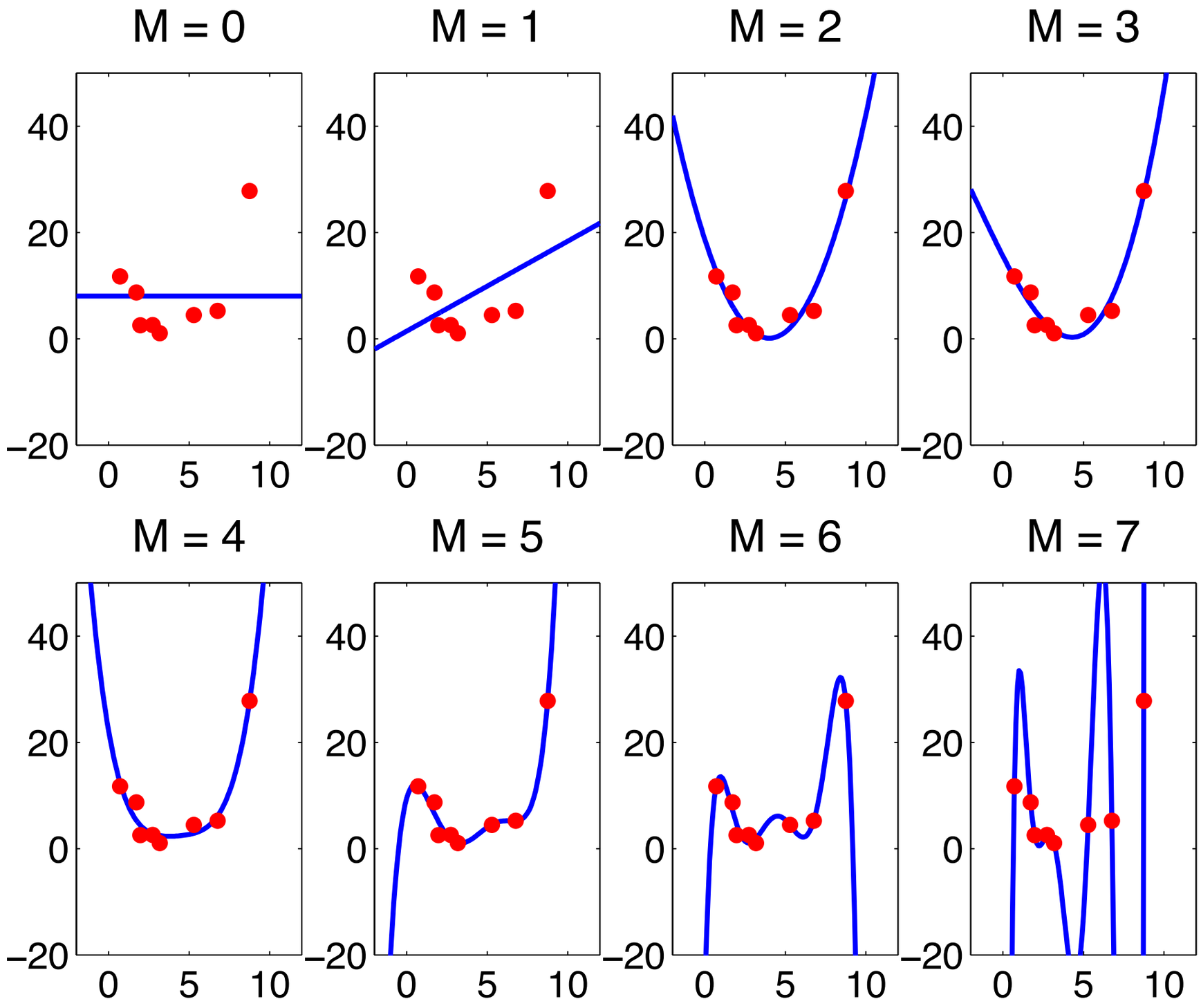
f is smooth over multiple scales



Why is this a smart thing to do:

flexible learning

Model selection



Occam's razor

The simplest model consistent with the data, but not simpler.

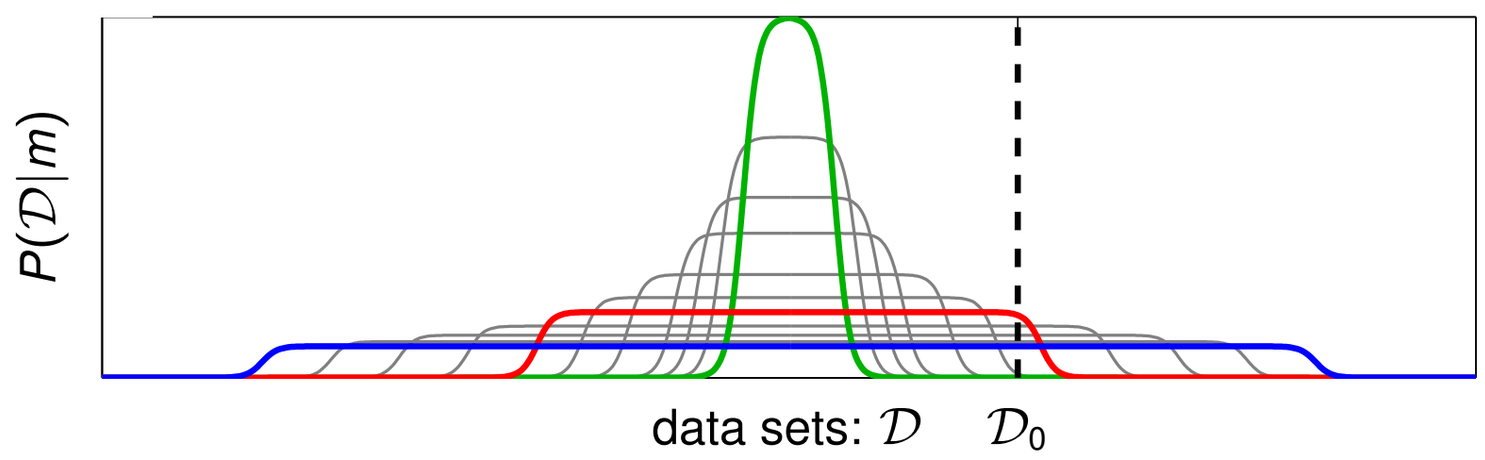
Compare model classes m using their posterior probability given the data:

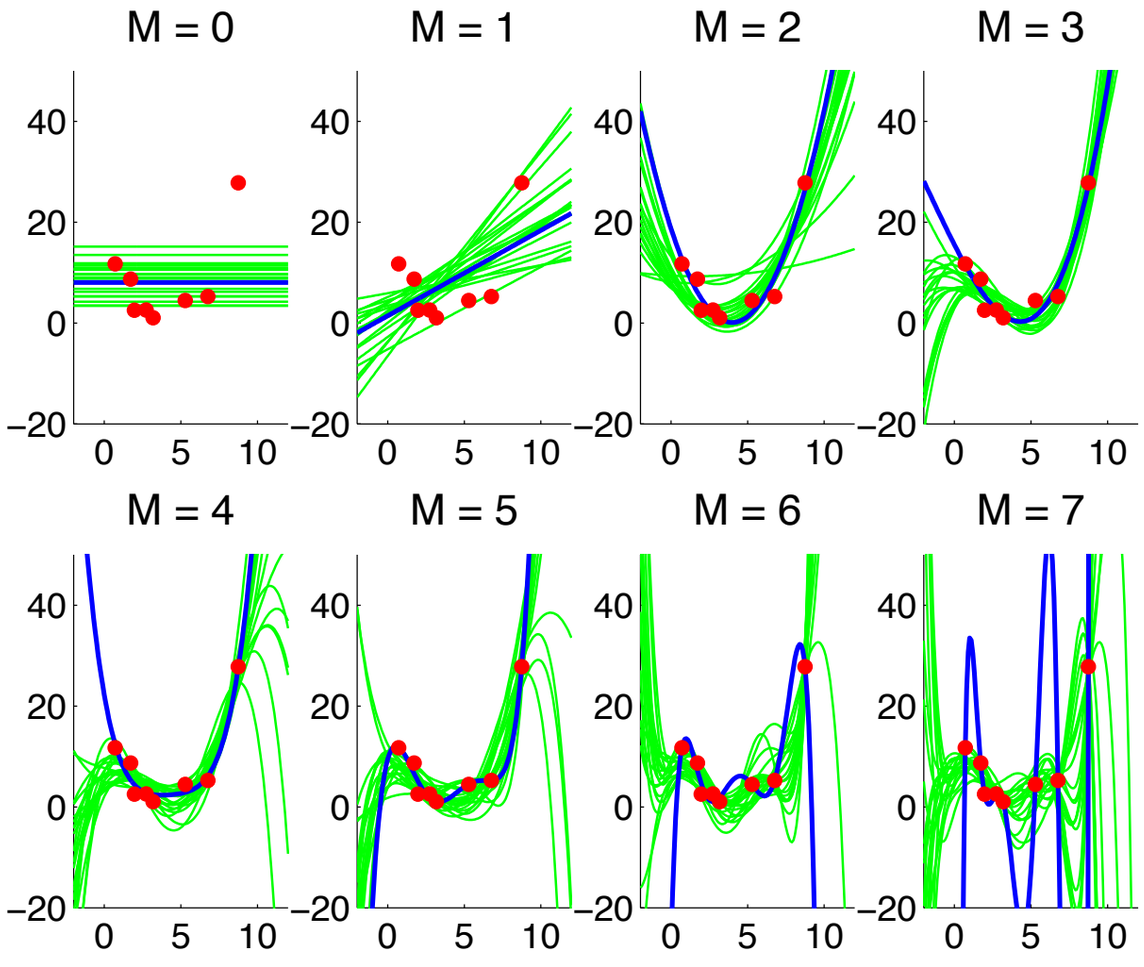
$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}, \qquad P(\mathcal{D}|m) = \int_{\Theta_m} P(\mathcal{D}|\boldsymbol{\theta}_m, m)P(\boldsymbol{\theta}_m|m)\, d\boldsymbol{\theta}_m$$

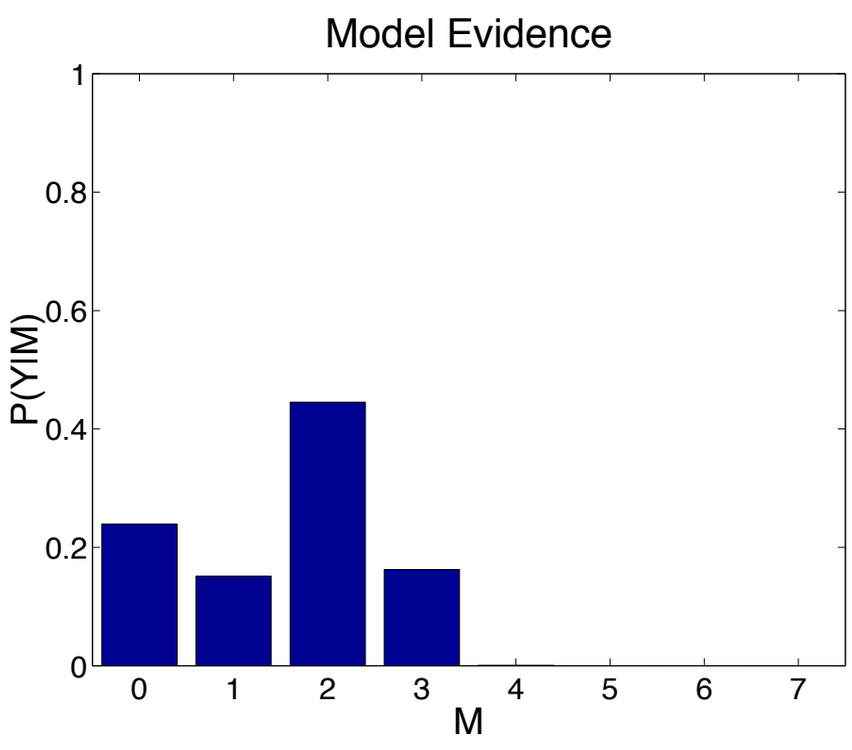
$P(\mathcal{D}|m)$: The probability that *randomly selected* parameter values from the model class would generate data set $\mathcal{D}$.

Model classes that are too simple are unlikely to generate the observed data set.
Model classes that are too complex can generate many possible data sets, so again, they are unlikely to generate that particular data set at random.

favour a model that is just right.




M = 0
M = 1
M = 2
M = 3
M = 4
M = 5
M = 6
M = 7
40
20
0
−20
0
5
10


Model Evidence
P(Y|M)
M
1
0.8
0.6
0.4
0.2
0
0
1
2
3
4
5
6
7

Nonparametric Bayesian Models and Occam's Razor

Overparameterised models can overfit. In the GP, the "parameter" is the function $f(\mathbf{x})$ (or "weights" in non-linear feature space) which can be infinite-dimensional.

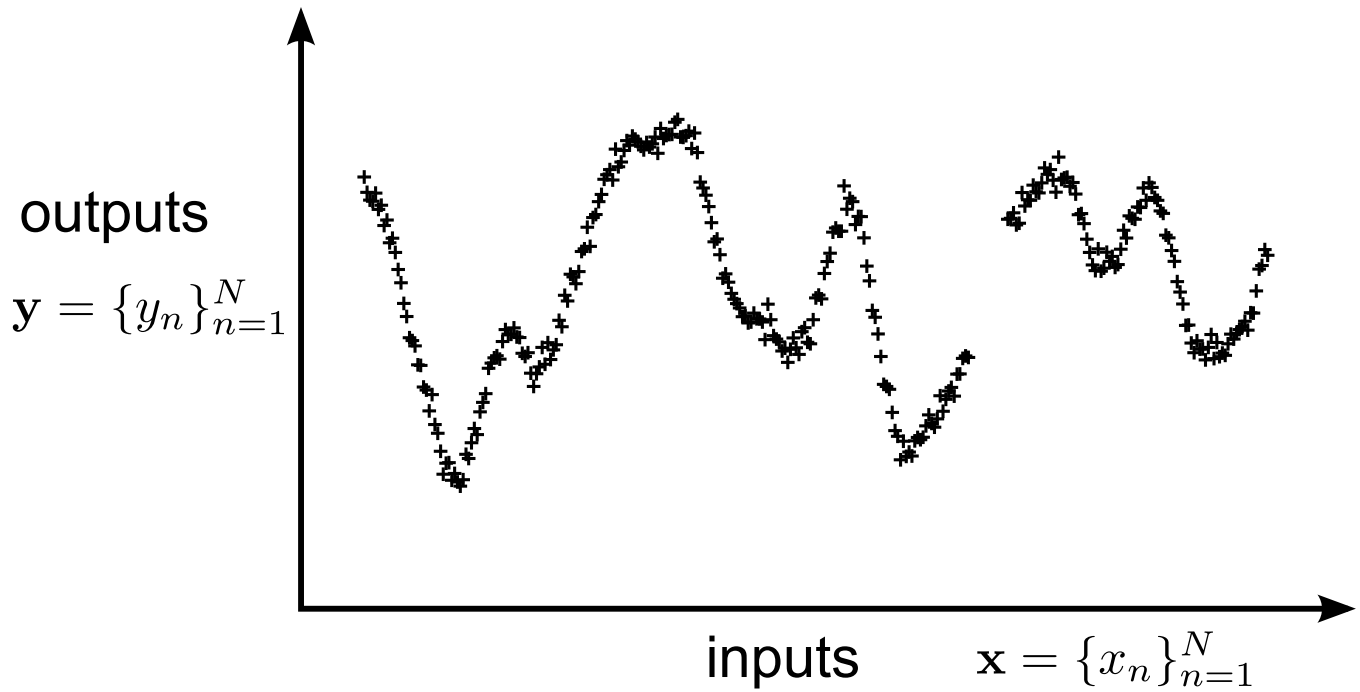
However, the Bayesian treatment integrates over these parameters: we never identify a single "best fit" f, just a posterior (and posterior mean). So f cannot be adjusted to overfit the data.

The GP is an example of the larger class of **nonparametric Bayesian models**.

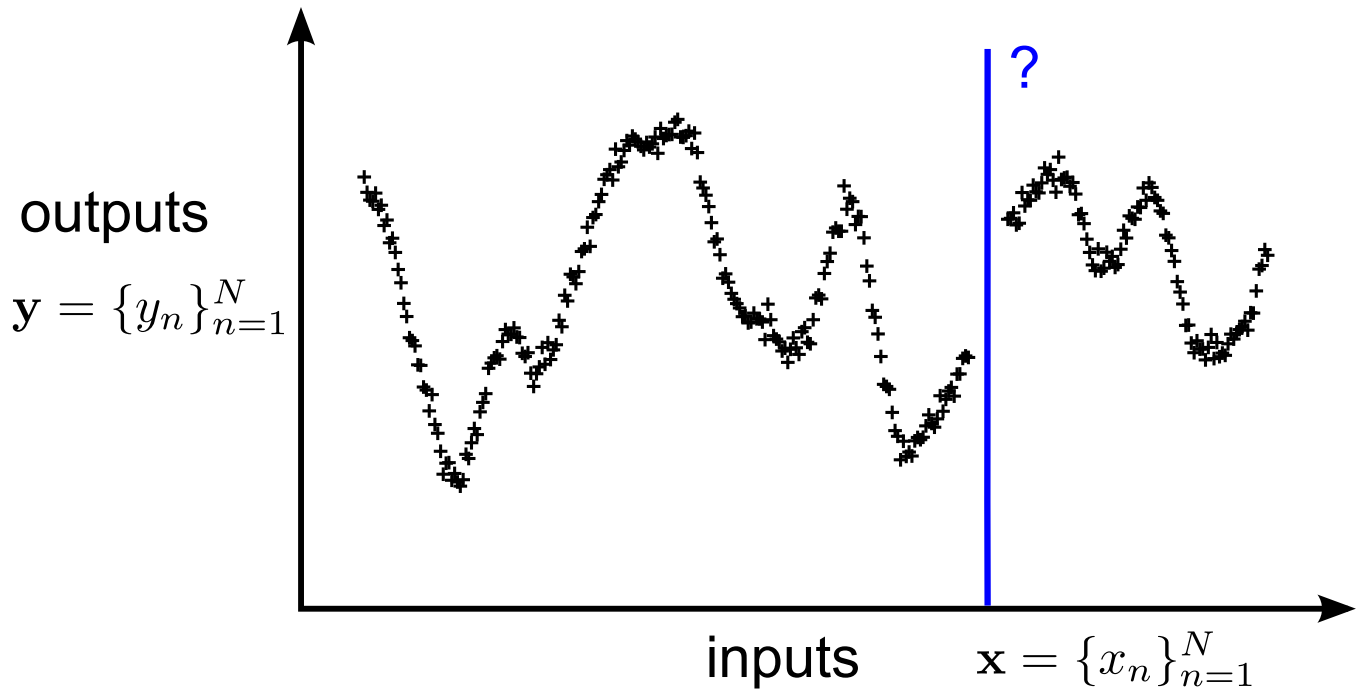
- Infinite number of parameters.
- Often constructed as the infinite limit of a nested family of finite models (sometimes equivalent to infinite model averaging).
- Parameters integrated out, so effective number of parameters to overfit is zero or small (hyperparameters).
- No need for model selection. Bayesian posterior on parameters will concentrate on "submodel" with largest integral automatically.
- No explicit need for Occam's razor, validation or added regularisation penalty.
- Examples include the Dirichlet process (infinite mixtures), Infinite Binary Prior (infinite binary factor models), Infinite HMM ...

Main challenge: computational cost

Motivation: Gaussian Process Regression



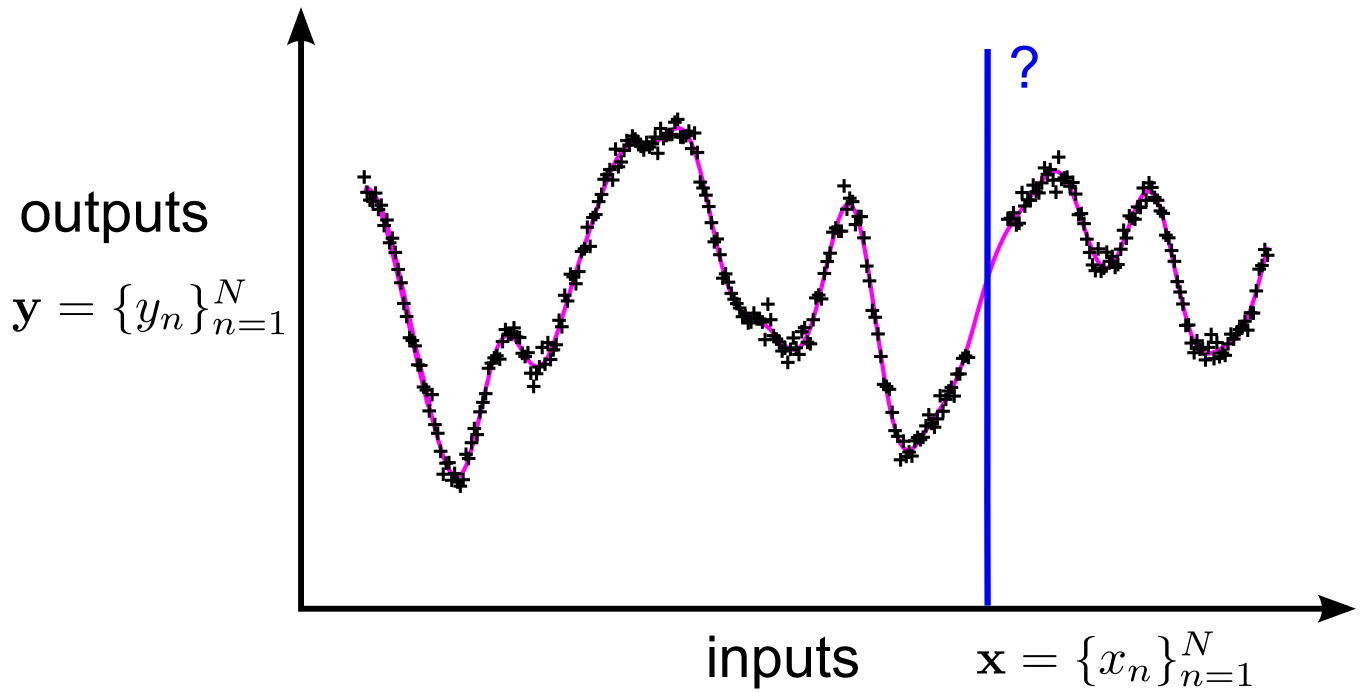
Motivation: Gaussian Process Regression



Motivation: Gaussian Process Regression

$$p(f|\theta) = \mathcal{GP}(f; 0, \mathrm{K}_\theta)$$

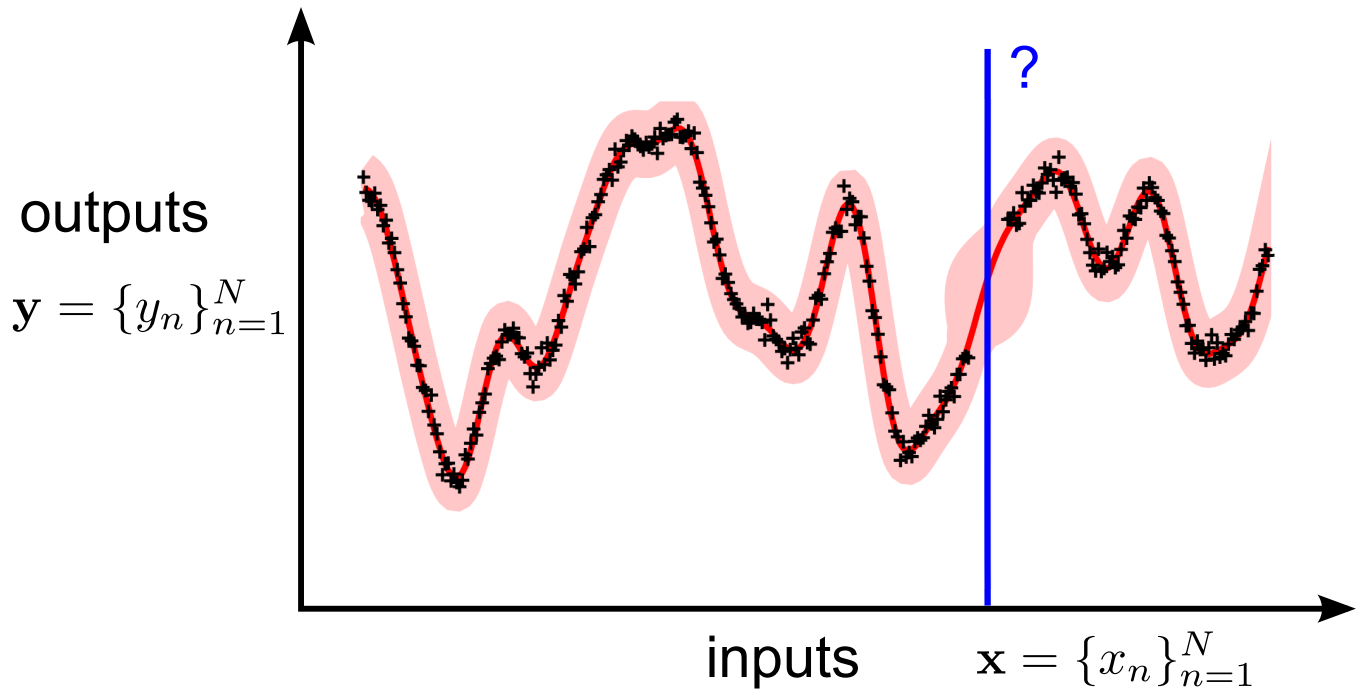
$$p(y_n|f, x_n, \theta)$$



Motivation: Gaussian Process Regression

$$p(f|\theta) = \mathcal{GP}(f; 0, \mathrm{K}_\theta) \qquad \text{inference \& learning} \qquad p(f|\mathbf{y}, \mathbf{x}, \theta)$$

$$p(y_n|f, x_n, \theta) \qquad\qquad p(\mathbf{y}|\mathbf{x}, \theta)$$



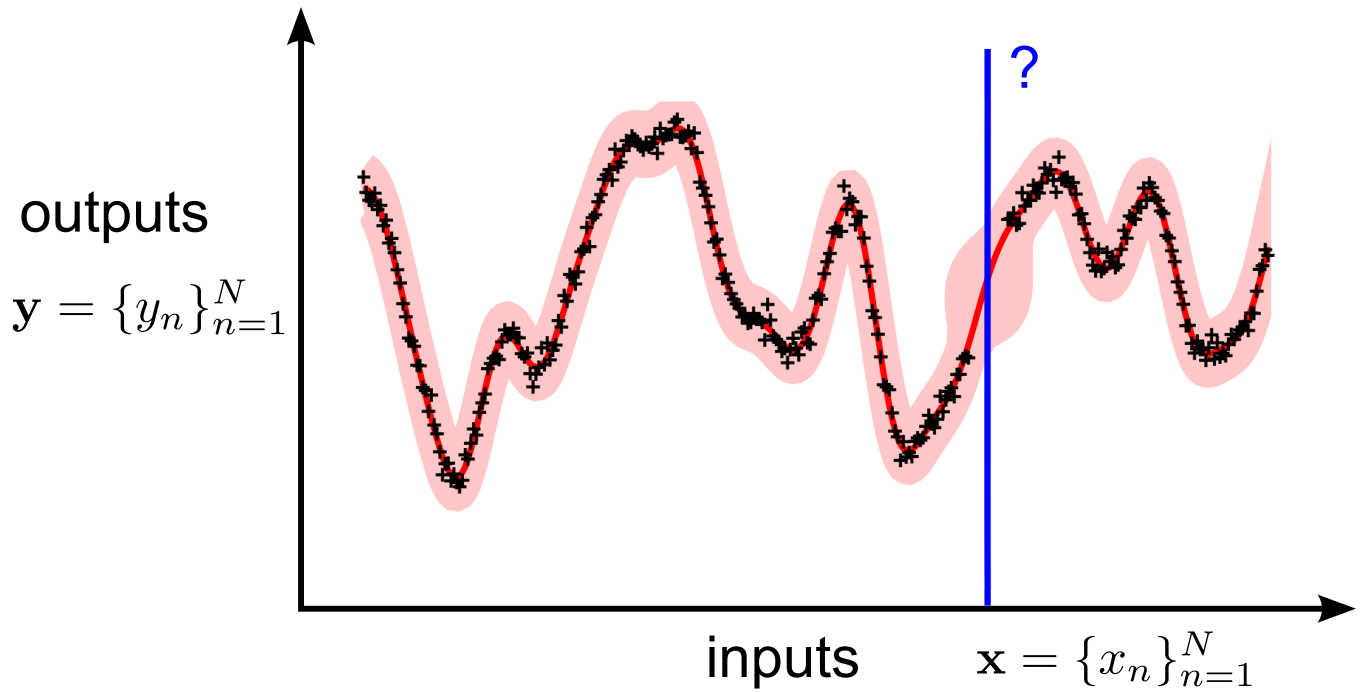
Motivation: Gaussian Process Regression

$$p(f|\theta) = \mathcal{GP}(f; 0, \mathrm{K}_\theta) \qquad p(f|\mathbf{y}, \mathbf{x}, \theta)$$

$$p(y_n|f, x_n, \theta) \qquad p(\mathbf{y}|\mathbf{x}, \theta)$$

inference & learning

intractabilities
computational $\mathcal{O}(N^3)$
analytic



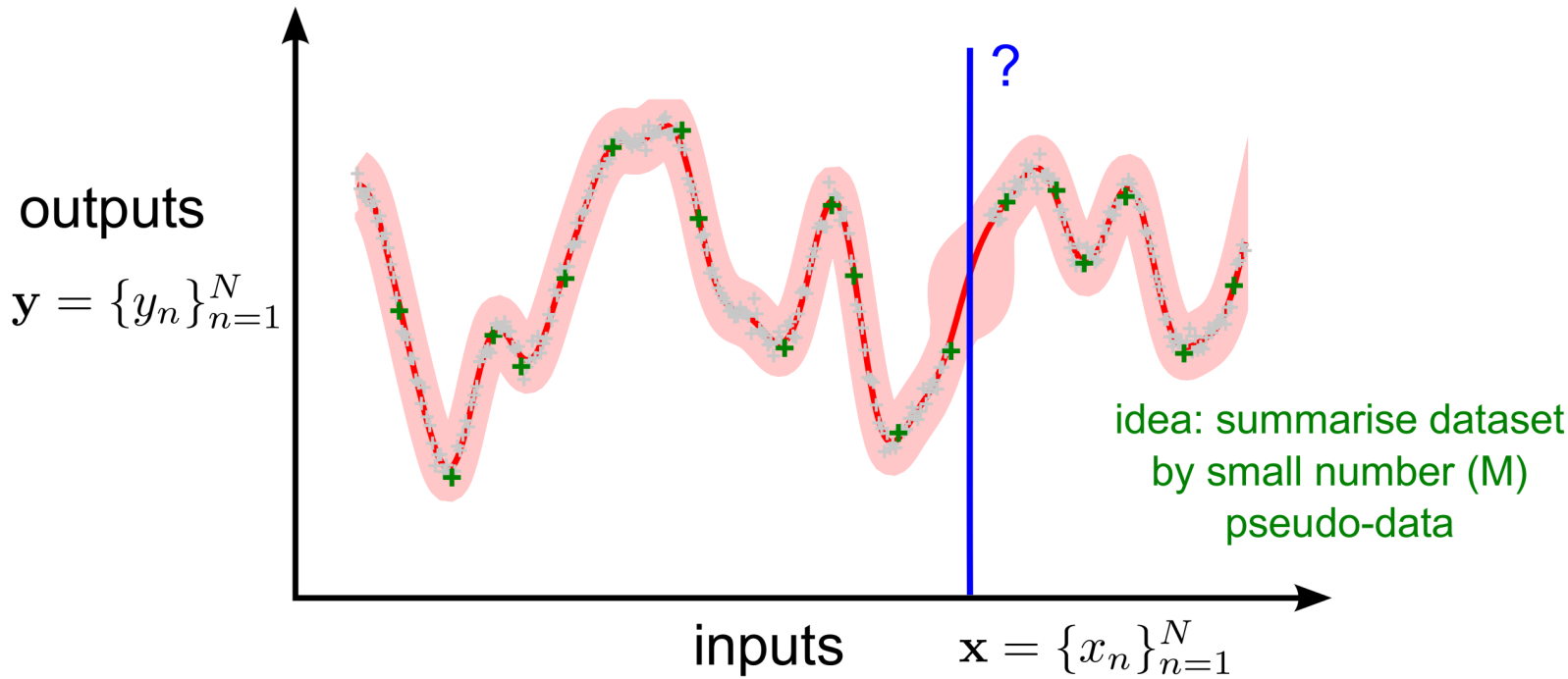
Motivation: Gaussian Process Regression

$$p(f|\theta) = \mathcal{GP}(f; 0, \mathrm{K}_\theta) \qquad p(y_n|f, x_n, \theta)$$

inference & learning

intractabilities
computational $\mathcal{O}(N^3)$
analytic

$$p(f|\mathbf{y}, \mathbf{x}, \theta) \qquad p(\mathbf{y}|\mathbf{x}, \theta)$$



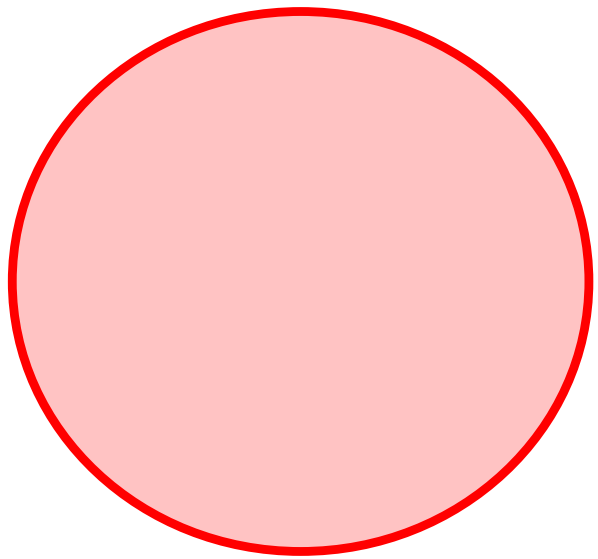
A Brief History of Gaussian Process Approximations

FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

$\text{div}[p(\mathbf{f}, \mathbf{y})||q(\mathbf{f}, \mathbf{y})]$



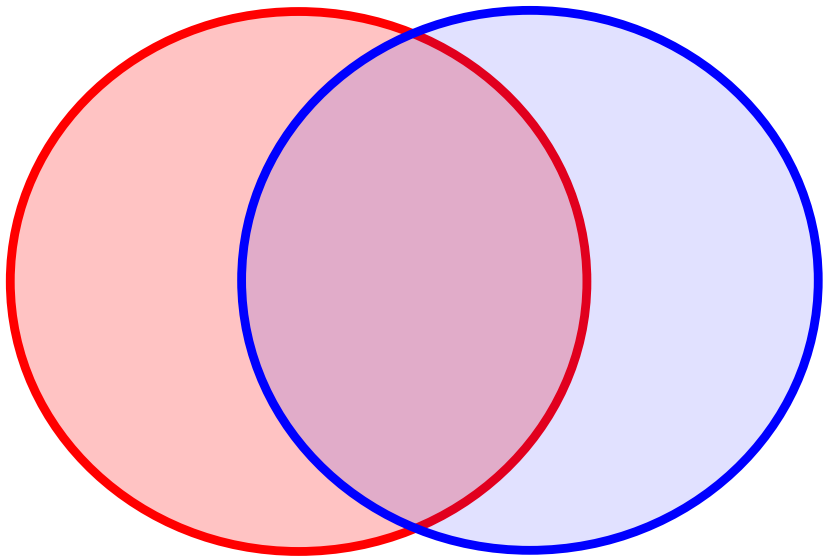
FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

methods employing
pseudo-data

$\text{div}[p(\mathbf{f}, \mathbf{y})||q(\mathbf{f}, \mathbf{y})]$



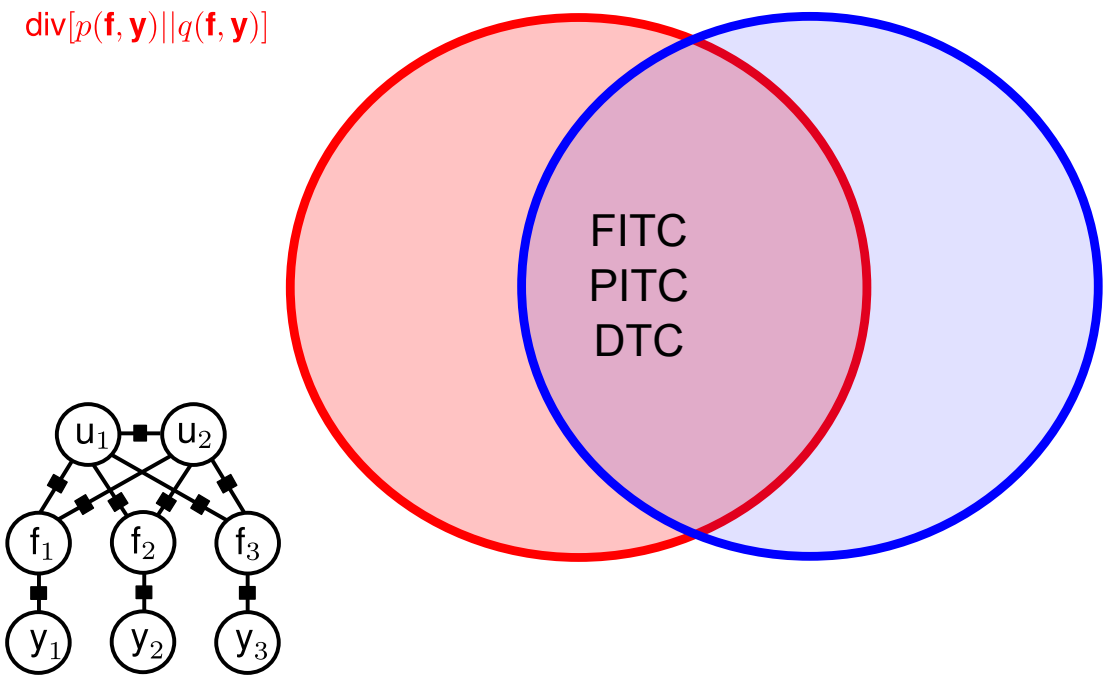
FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

methods employing
pseudo-data

$\text{div}[p(\mathbf{f},\mathbf{y})||q(\mathbf{f},\mathbf{y})]$



FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

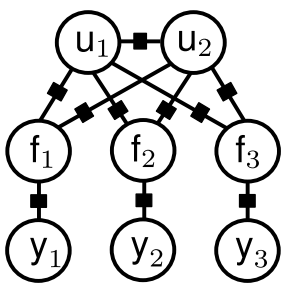
A Brief History of Gaussian Process Approximations

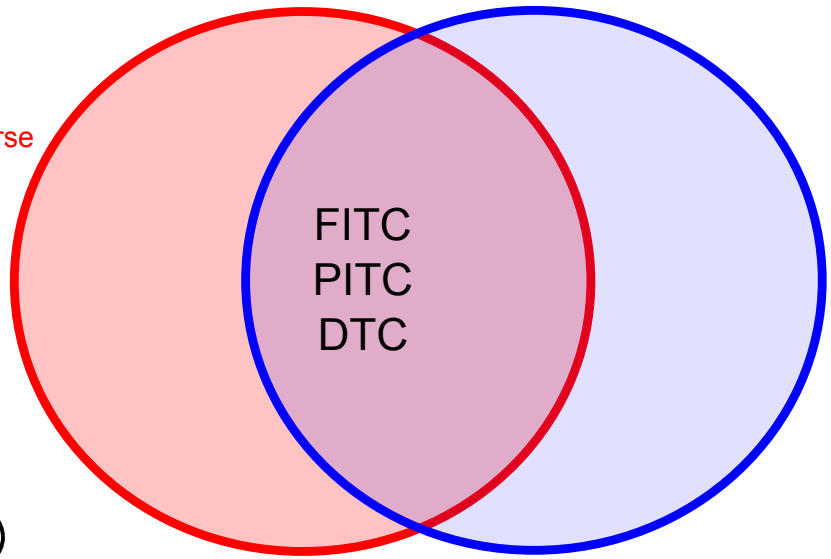
approximate generative model
exact inference

methods employing
pseudo-data

$\text{div}[p(\mathbf{f},\mathbf{y})||q(\mathbf{f},\mathbf{y})]$

A Unifying View of Sparse Approximate Gaussian Process Regression
Quinonero-Candela & Rasmussen, 2005
(FITC, PITC, DTC)





FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

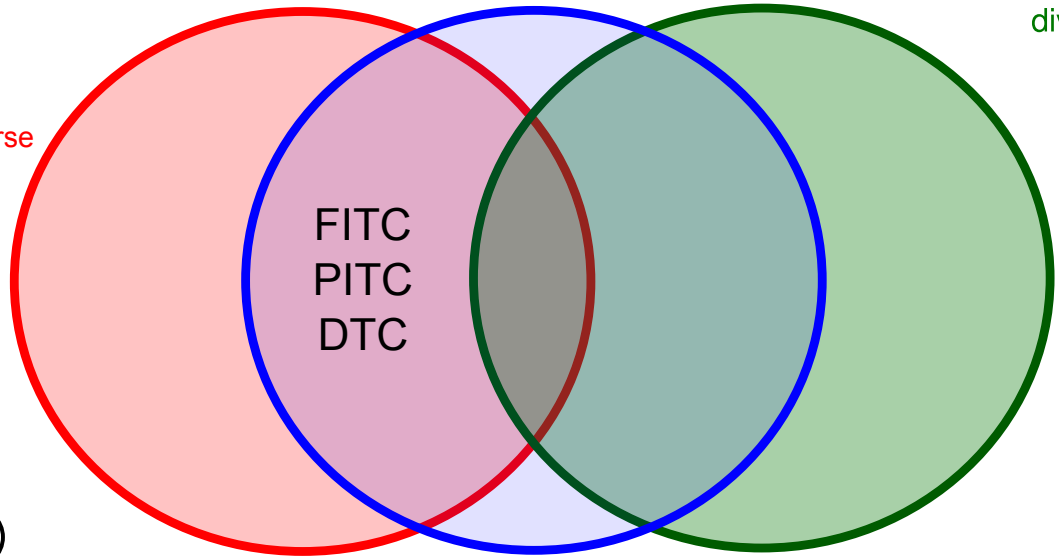
methods employing
pseudo-data

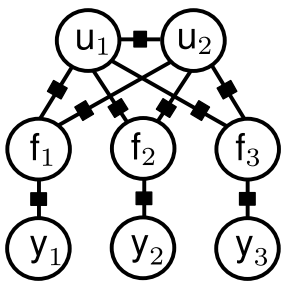
exact generative model
approximate inference

$\text{div}[p(\mathbf{f},\mathbf{y})||q(\mathbf{f},\mathbf{y})]$

$\text{div}[p(\mathbf{f}|\mathbf{y})||q(\mathbf{f})]$

A Unifying View of Sparse Approximate Gaussian Process Regression
Quinonero-Candela & Rasmussen, 2005
(FITC, PITC, DTC)





FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

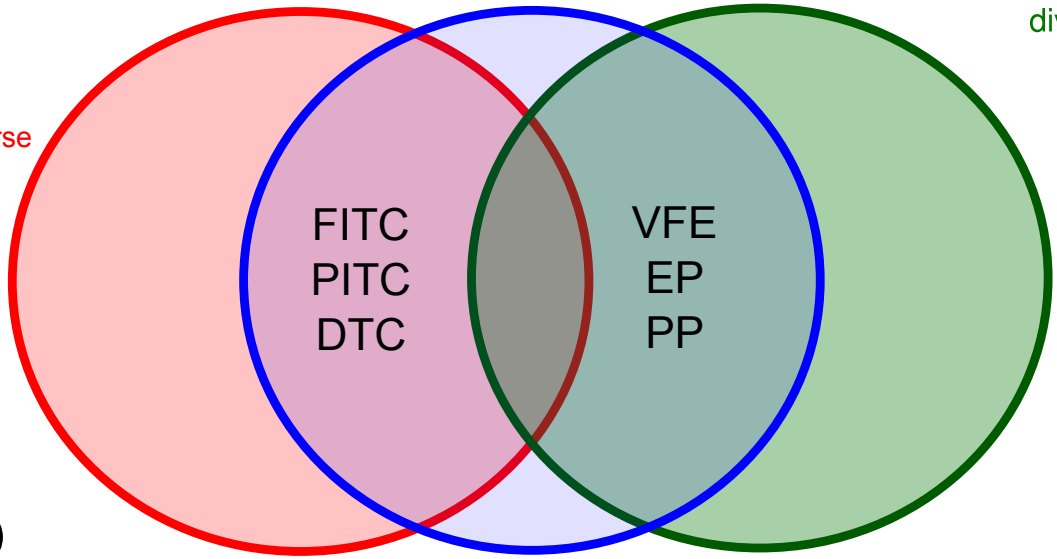
methods employing
pseudo-data

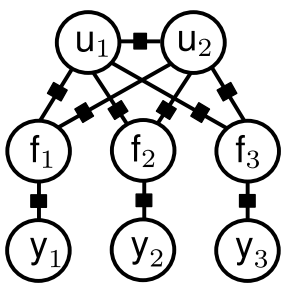
exact generative model
approximate inference

$\text{div}[p(\mathbf{f},\mathbf{y})||q(\mathbf{f},\mathbf{y})]$

$\text{div}[p(\mathbf{f}|\mathbf{y})||q(\mathbf{f})]$

A Unifying View of Sparse Approximate Gaussian Process Regression
Quinonero-Candela & Rasmussen, 2005
(FITC, PITC, DTC)





FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

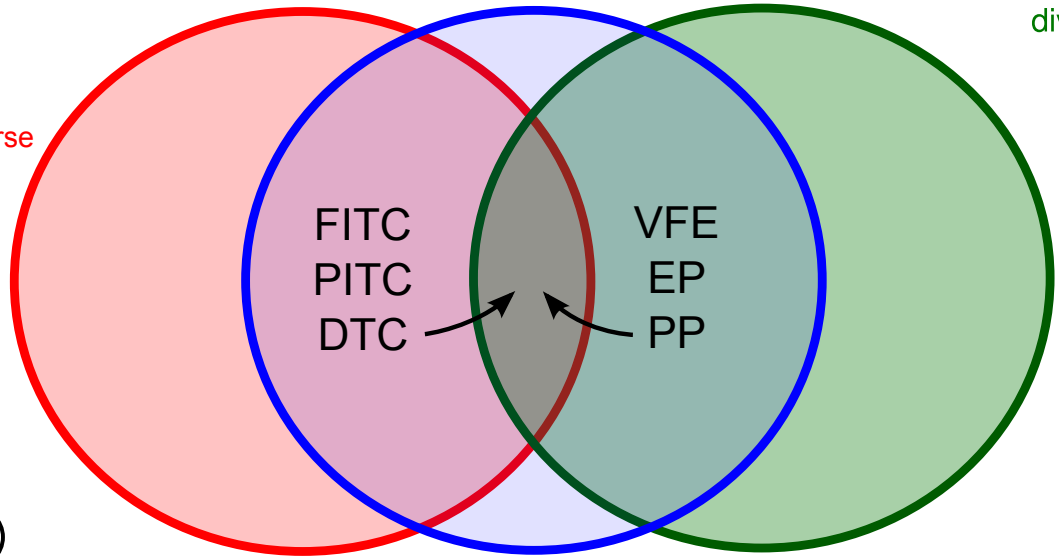
methods employing
pseudo-data

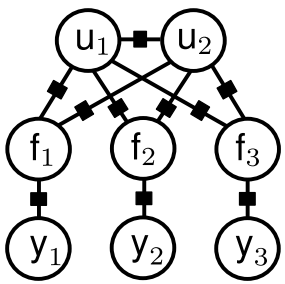
exact generative model
approximate inference

$\text{div}[p(\mathbf{f},\mathbf{y})||q(\mathbf{f},\mathbf{y})]$

$\text{div}[p(\mathbf{f}|\mathbf{y})||q(\mathbf{f})]$

A Unifying View of Sparse Approximate Gaussian Process Regression
Quinonero-Candela & Rasmussen, 2005
(FITC, PITC, DTC)





FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

A Brief History of Gaussian Process Approximations

approximate generative model
exact inference

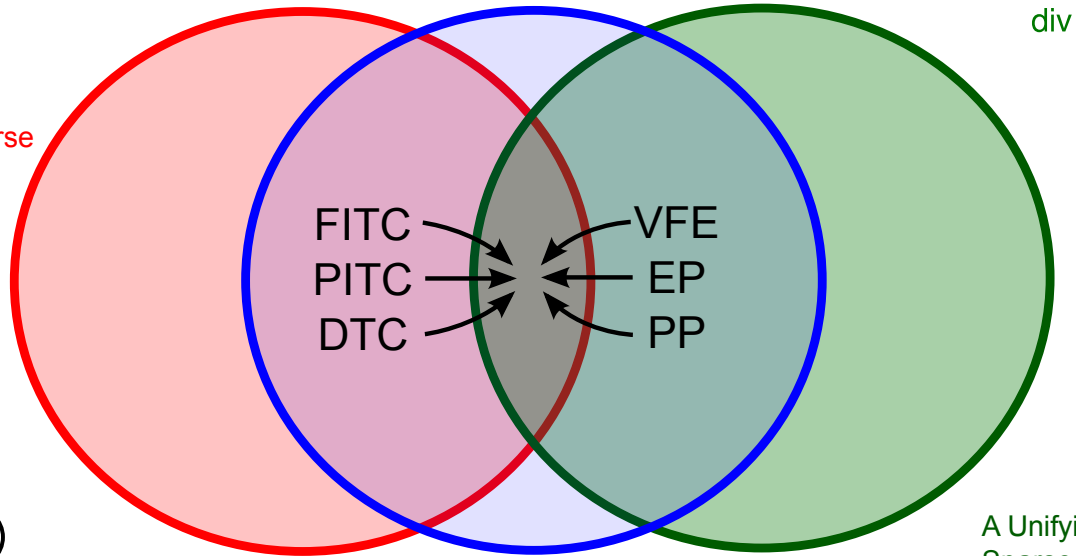
methods employing
pseudo-data

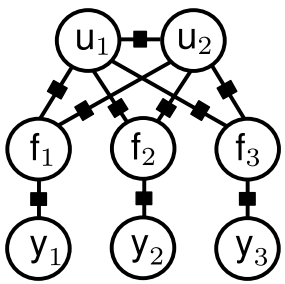
exact generative model
approximate inference

$\text{div}[p(\mathbf{f},\mathbf{y})||q(\mathbf{f},\mathbf{y})]$

$\text{div}[p(\mathbf{f}|\mathbf{y})||q(\mathbf{f})]$

A Unifying View of Sparse Approximate Gaussian Process Regression
Quinonero-Candela & Rasmussen, 2005
(FITC, PITC, DTC)





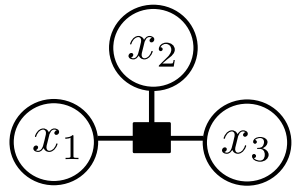
A Unifying Framework for Sparse Gaussian Process Approximation using Power Expectation Propagation
Bui, Yan and Turner, 2016
(VFE, EP, FITC, PITC ...)

FITC: Snelson et al. “Sparse Gaussian Processes using Pseudo-inputs”
PITC: Snelson et al. “Local and global sparse Gaussian process approximations”
EP: Csato and Opper 2002 / Qi et al. "Sparse-posterior Gaussian Processes for general likelihoods.”
VFE: Titsias “Variational Learning of Inducing Variables in Sparse Gaussian Processes”
DTC / PP: Seeger et al. “Fast Forward Selection to Speed Up Sparse Gaussian Process Regression”

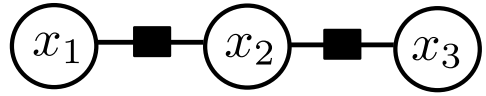
Factor Graphs: introduction / reminder

factor graph examples

$$p(x_1, x_2, x_3) = g(x_1, x_2, x_3)$$



$$p(x_1, x_2, x_3) = g_1(x_1, x_2) g_2(x_2, x_3)$$

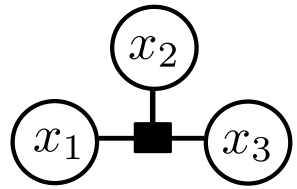


$$x_3 \perp x_1 | x_2$$

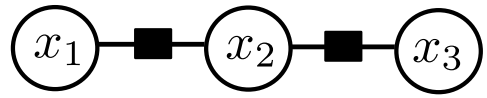
Factor Graphs: introduction / reminder

factor graph examples

$$p(x_1, x_2, x_3) = g(x_1, x_2, x_3)$$



$$p(x_1, x_2, x_3) = g_1(x_1, x_2) g_2(x_2, x_3)$$



$x_3 \perp x_1 | x_2$

what is the minimal factor graph for this multivariate Gaussian?

$$p(\mathbf{x}|\mu, \Sigma) = \mathcal{N}(\mathbf{x}; \mu, \Sigma)$$

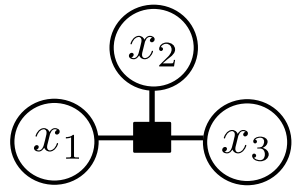
4 dimensional

$$\Sigma = \begin{bmatrix} 1 & 1/2 & 1/2 & 1/4 \\ 1/2 & 5/4 & 1/4 & 1/8 \\ 1/2 & 1/4 & 5/4 & 5/8 \\ 1/4 & 1/8 & 5/8 & 21/16 \end{bmatrix} \quad \Sigma^{-1} = \begin{bmatrix} 1.5 & -1/2 & -1/2 & 0 \\ -1/2 & 1 & 0 & 0 \\ -1/2 & 0 & 5/4 & -1/2 \\ 0 & 0 & -1/2 & 1 \end{bmatrix}$$

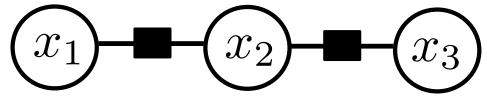
Factor Graphs: introduction / reminder

factor graph examples

$$p(x_1, x_2, x_3) = g(x_1, x_2, x_3)$$



$$p(x_1, x_2, x_3) = g_1(x_1, x_2) g_2(x_2, x_3)$$



$$x_3 \perp x_1 | x_2$$

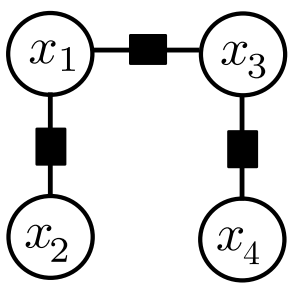
what is the minimal factor graph for this multivariate Gaussian?

$$p(\mathbf{x}|\mu, \Sigma) = \mathcal{N}(\mathbf{x}; \mu, \Sigma)$$

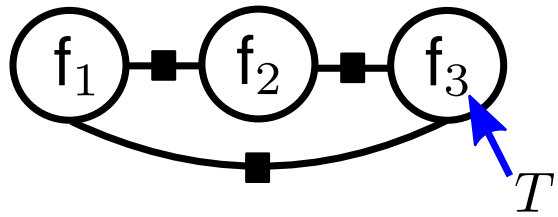
4 dimensional

$$\Sigma = \begin{bmatrix} 1 & 1/2 & 1/2 & 1/4 \\ 1/2 & 5/4 & 1/4 & 1/8 \\ 1/2 & 1/4 & 5/4 & 5/8 \\ 1/4 & 1/8 & 5/8 & 21/16 \end{bmatrix} \quad \Sigma^{-1} = \begin{bmatrix} 1.5 & -1/2 & -1/2 & 0 \\ -1/2 & 1 & 0 & 0 \\ -1/2 & 0 & 5/4 & -1/2 \\ 0 & 0 & -1/2 & 1 \end{bmatrix}$$

solution:



Fully independent training conditional (FITC) approximation

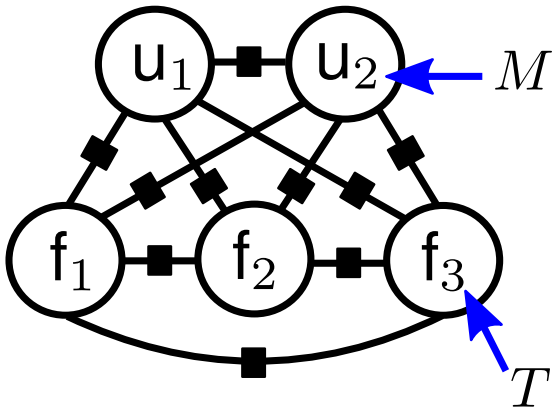


construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

Fully independent training conditional (FITC) approximation

1. augment model with M<T pseudo data

$$p(\mathbf{f}, \mathbf{u}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{u} \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathsf{K}_{\mathbf{ff}} & \mathsf{K}_{\mathbf{fu}} \\ \mathsf{K}_{\mathbf{uf}} & \mathsf{K}_{\mathbf{uu}} \end{bmatrix}\right)$$

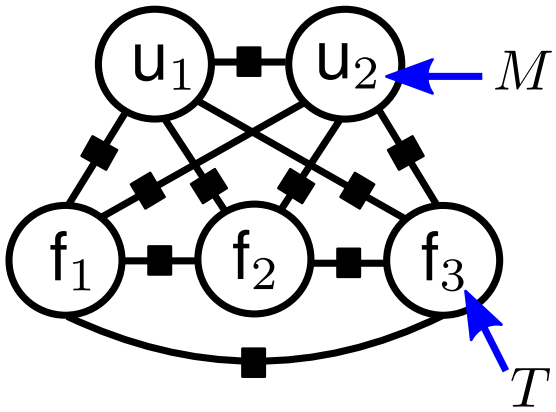


construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

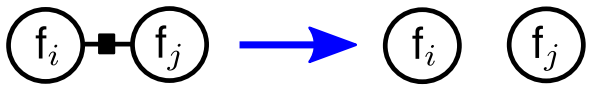
Fully independent training conditional (FITC) approximation

1. augment model with M<T pseudo data

$$p(\mathbf{f}, \mathbf{u}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{u} \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{K_{ff}} & \mathbf{K_{fu}} \\ \mathbf{K_{uf}} & \mathbf{K_{uu}} \end{bmatrix}\right)$$



2. remove some of the dependencies
   (results in simpler model)



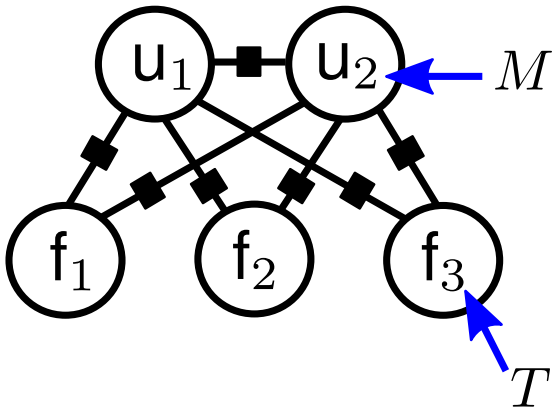
all factors

construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

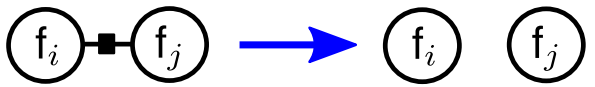
Fully independent training conditional (FITC) approximation

1. augment model with M<T pseudo data

$$p(\mathbf{f}, \mathbf{u}) = \mathcal{N}\left(\begin{bmatrix} \mathbf{f} \\ \mathbf{u} \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{K_{ff}} & \mathbf{K_{fu}} \\ \mathbf{K_{uf}} & \mathbf{K_{uu}} \end{bmatrix}\right)$$



2. remove some of the dependencies
   (results in simpler model)



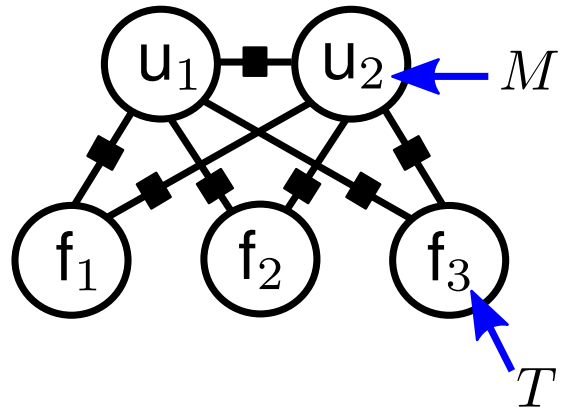
all factors

construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

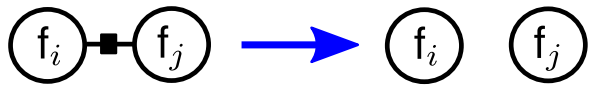
Fully independent training conditional (FITC) approximation

1. augment model with M<T pseudo data

$$p(\mathbf{f},\mathbf{u}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{u}\end{bmatrix}; \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\mathbf{K_{ff}} & \mathbf{K_{fu}}\\ \mathbf{K_{uf}} & \mathbf{K_{uu}}\end{bmatrix}\right)$$



2. remove some of the dependencies
   (results in simpler model)



all factors

3. calibrate model
   (e.g. using KL divergence, many choices)

$$\underset{q(\mathbf{u}),\{q(\mathsf{f}_t|\mathbf{u})\}_{t=1}^{T}}{\arg\min} \mathrm{KL}(p(\mathbf{f},\mathbf{u})||q(\mathbf{u})\prod_{t=1}^{T} q(\mathsf{f}_t|\mathbf{u})) \implies \begin{aligned} q(\mathbf{u}) &= p(\mathbf{u}) \\ q(\mathsf{f}_t|\mathbf{u}) &= p(\mathsf{f}_t|\mathbf{u}) \end{aligned}$$

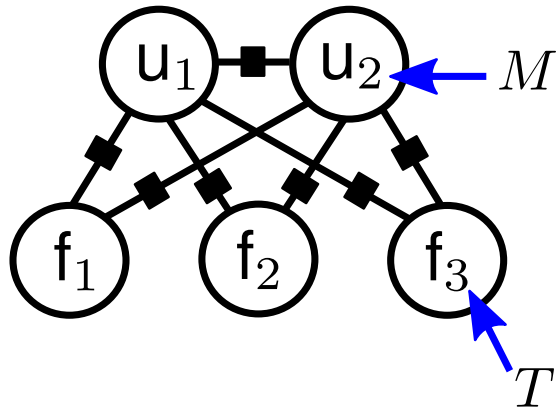
equal to exact conditionals

construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

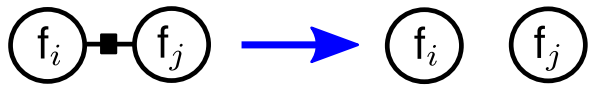
Fully independent training conditional (FITC) approximation

1. augment model with M<T pseudo data

$$p(\mathbf{f},\mathbf{u}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{f}\\ \mathbf{u}\end{bmatrix}; \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\mathbf{K_{ff}} & \mathbf{K_{fu}}\\ \mathbf{K_{uf}} & \mathbf{K_{uu}}\end{bmatrix}\right)$$



2. remove some of the dependencies
   (results in simpler model)



all factors

3. calibrate model
   (e.g. using KL divergence, many choices)

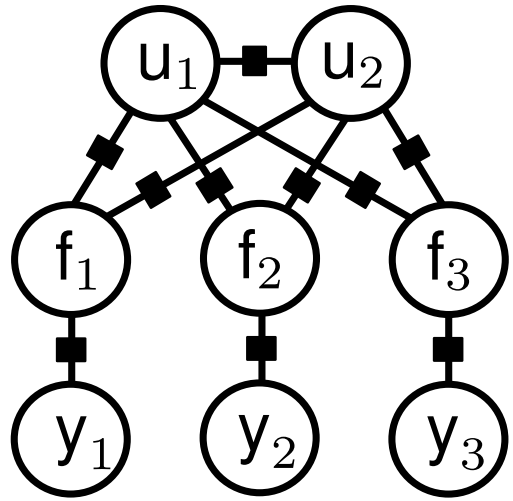
$$\underset{q(\mathbf{u}),\{q(\mathrm{f}_t|\mathbf{u})\}_{t=1}^{T}}{\arg\min} \mathrm{KL}(p(\mathbf{f},\mathbf{u})||q(\mathbf{u})\prod_{t=1}^{T} q(\mathrm{f}_t|\mathbf{u})) \implies \begin{aligned} q(\mathbf{u}) &= p(\mathbf{u}) \\ q(\mathrm{f}_t|\mathbf{u}) &= p(\mathrm{f}_t|\mathbf{u}) \end{aligned}$$

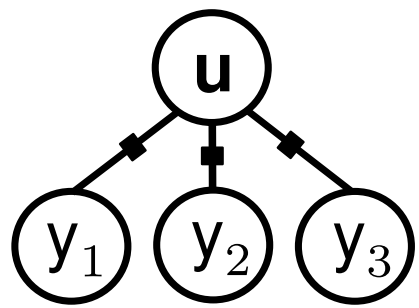
equal to exact conditionals

construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

indirect posterior approximation

Fully independent training conditional (FITC) approximation



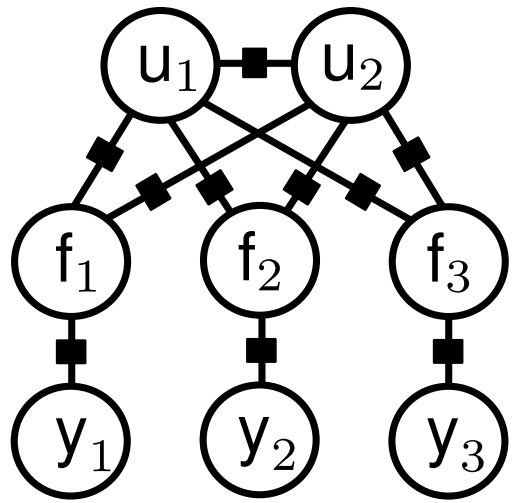


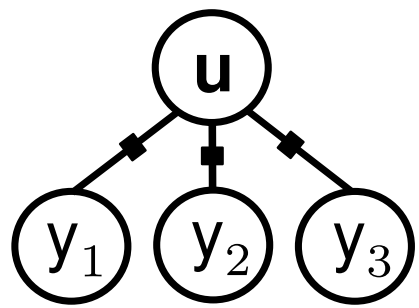
construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

indirect posterior approximation

Fully independent training conditional (FITC) approximation

$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$





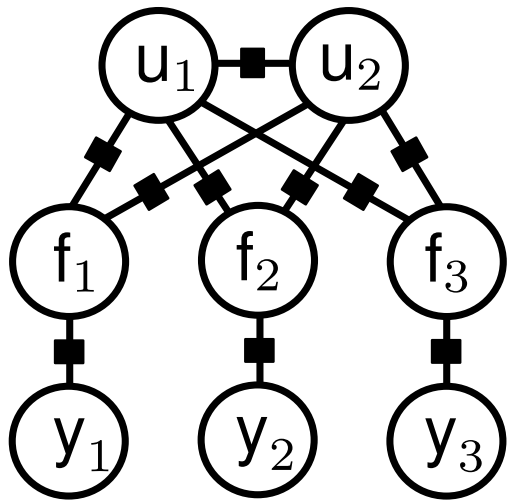
construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

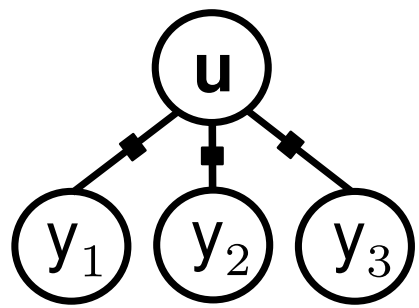
indirect posterior approximation

Fully independent training conditional (FITC) approximation

$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$

$$q(\mathsf{f}_t|\mathbf{u}) = p(\mathsf{f}_t|\mathbf{u})$$





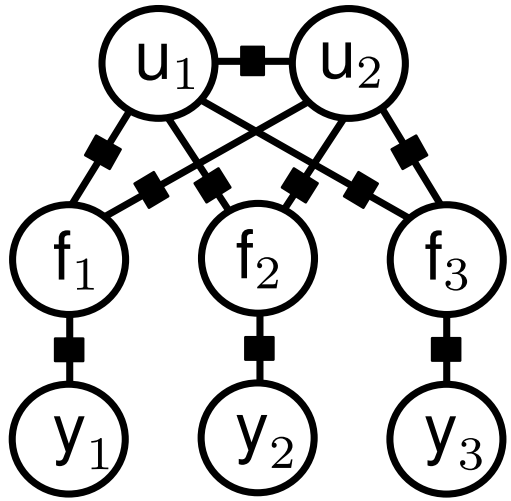
construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

indirect posterior approximation

Fully independent training conditional (FITC) approximation

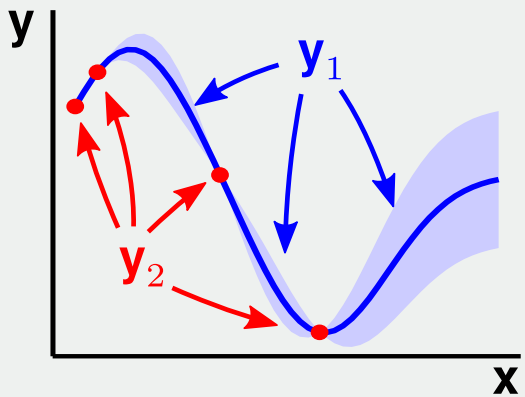
$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$

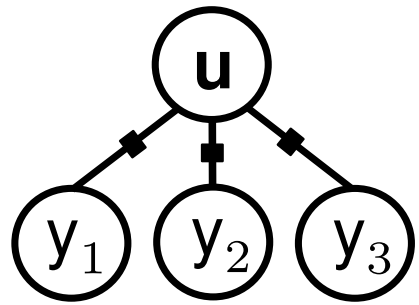
$$q(\mathsf{f}_t|\mathbf{u}) = p(\mathsf{f}_t|\mathbf{u})$$



How do we make predictions?

$$p(\mathbf{y}_1|\mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1; \Sigma_{12}\Sigma_{22}^{-1}\mathbf{y}_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\mathsf{T}})$$





construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

indirect posterior approximation

Fully independent training conditional (FITC) approximation

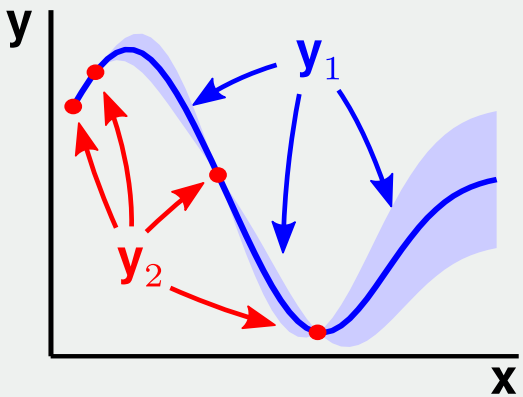
$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$

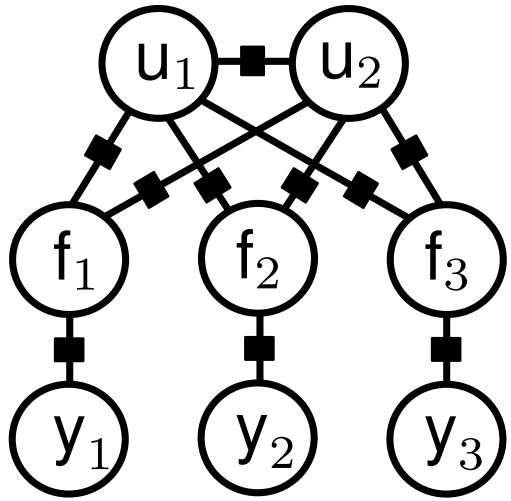
$$q(\mathsf{f}_t|\mathbf{u}) = p(\mathsf{f}_t|\mathbf{u})$$

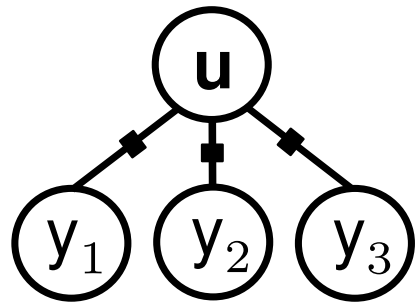
$$= \mathcal{N}(\mathsf{f}_t; \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathbf{u}, \mathsf{K}_{\mathsf{f}_t\mathsf{f}_t} - \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}_t})$$

How do we make predictions?

$$p(\mathbf{y}_1|\mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1; \Sigma_{12}\Sigma_{22}^{-1}\mathbf{y}_2, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\mathsf{T}})$$







construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

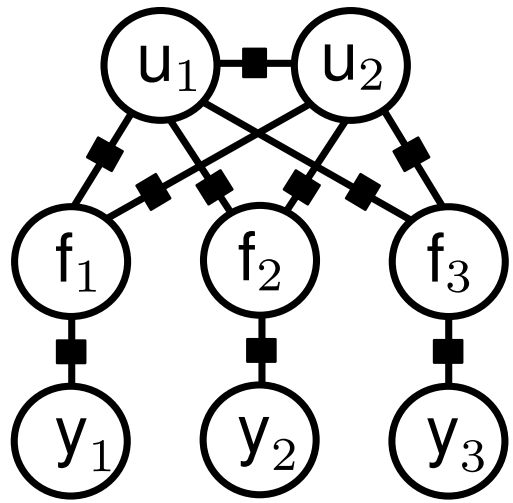
indirect posterior approximation

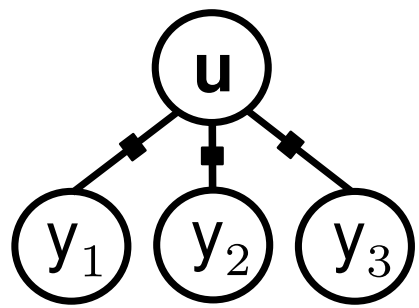
Fully independent training conditional (FITC) approximation

$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$

$$q(\mathsf{f}_t|\mathbf{u}) = p(\mathsf{f}_t|\mathbf{u})$$

$$= \mathcal{N}(\mathsf{f}_t; \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathbf{u}, \mathsf{K}_{\mathsf{f}_t\mathsf{f}_t} - \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}_t})$$





construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

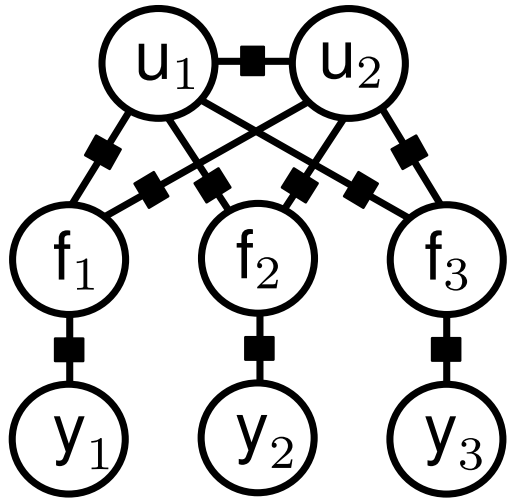
indirect posterior approximation

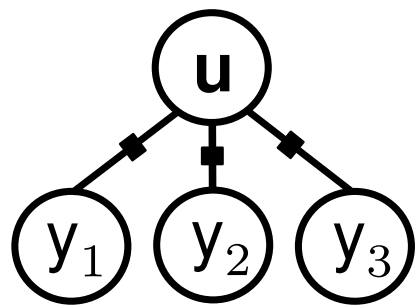
Fully independent training conditional (FITC) approximation

$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$

$$q(\mathsf{f}_t|\mathbf{u}) = p(\mathsf{f}_t|\mathbf{u})$$

$$= \mathcal{N}(\mathsf{f}_t; \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathbf{u}, \underbrace{\mathsf{K}_{\mathsf{f}_t\mathsf{f}_t} - \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}_t}}_{\mathsf{D}_{tt}})$$





construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

indirect posterior approximation

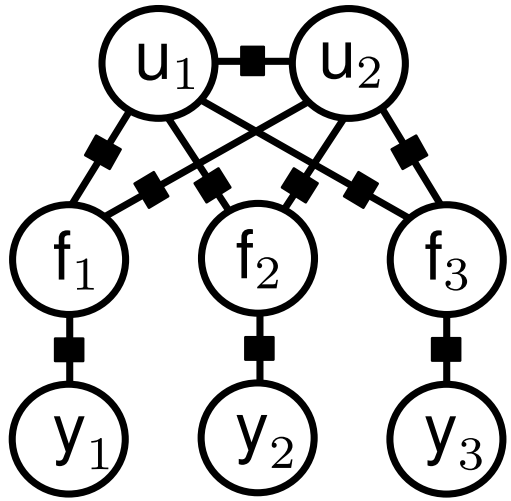
Fully independent training conditional (FITC) approximation

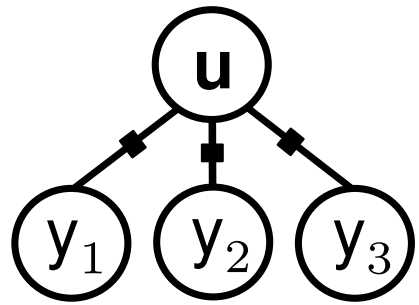
$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$

$$q(\mathsf{f}_t|\mathbf{u}) = p(\mathsf{f}_t|\mathbf{u})$$

$$= \mathcal{N}(\mathsf{f}_t; \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathbf{u}, \underbrace{\mathsf{K}_{\mathsf{f}_t\mathsf{f}_t} - \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}_t}}_{\mathsf{D}_{tt}})$$

$$q(\mathsf{y}_t|\mathsf{f}_t) = p(\mathsf{y}_t|\mathsf{f}_t) = \mathcal{N}(\mathsf{y}_t; \mathsf{f}_t, \sigma_{\mathsf{y}}^2)$$





construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

indirect posterior approximation

Fully independent training conditional (FITC) approximation

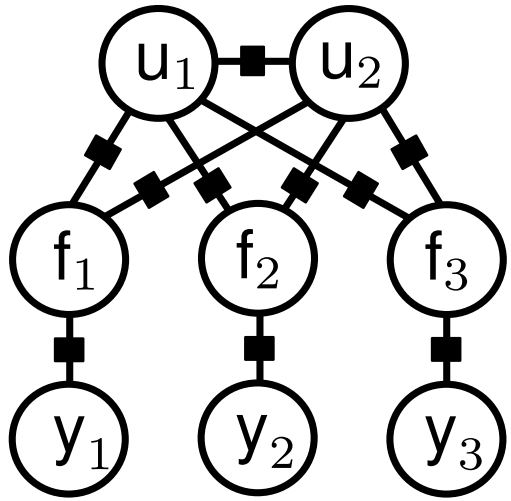
$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$

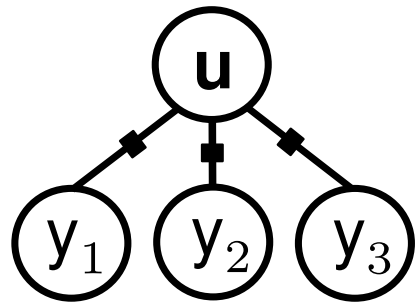
$$q(\mathsf{f}_t|\mathbf{u}) = p(\mathsf{f}_t|\mathbf{u})$$

$$= \mathcal{N}(\mathsf{f}_t; \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathbf{u}, \underbrace{\mathsf{K}_{\mathsf{f}_t\mathsf{f}_t} - \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}_t}}_{\mathsf{D}_{tt}})$$

$$q(\mathsf{y}_t|\mathsf{f}_t) = p(\mathsf{y}_t|\mathsf{f}_t) = \mathcal{N}(\mathsf{y}_t; \mathsf{f}_t, \sigma_{\mathsf{y}}^2)$$

cost of computing likelihood is $\mathcal{O}(TM^2)$





construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

indirect posterior approximation

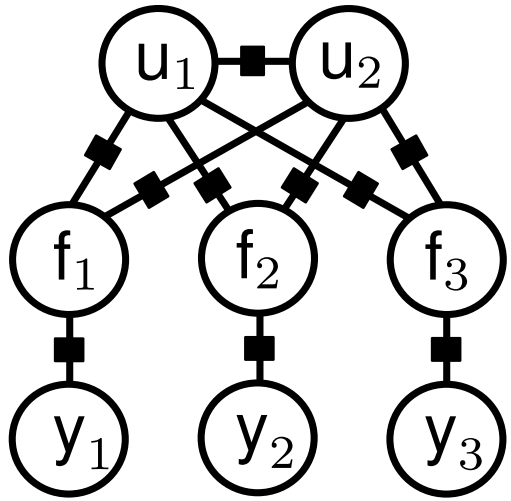
Fully independent training conditional (FITC) approximation

$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$

$$q(\mathsf{f}_t|\mathbf{u}) = p(\mathsf{f}_t|\mathbf{u})$$

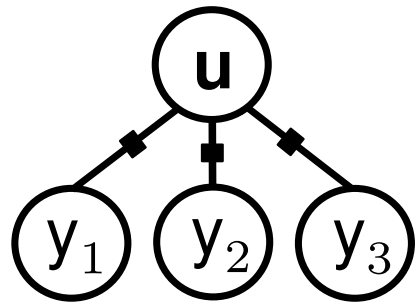
$$= \mathcal{N}(\mathsf{f}_t; \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathbf{u}, \underbrace{\mathsf{K}_{\mathsf{f}_t\mathsf{f}_t} - \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}_t}}_{\mathsf{D}_{tt}})$$

$$q(\mathsf{y}_t|\mathsf{f}_t) = p(\mathsf{y}_t|\mathsf{f}_t) = \mathcal{N}(\mathsf{y}_t; \mathsf{f}_t, \sigma_{\mathsf{y}}^2)$$



cost of computing likelihood is $\mathcal{O}(TM^2)$

$$p(\mathbf{y}_t|\theta) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathsf{K}_{\mathsf{fu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}} + \mathsf{D} + \sigma_{\mathsf{y}}^2\mathsf{I})$$



construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

indirect posterior approximation

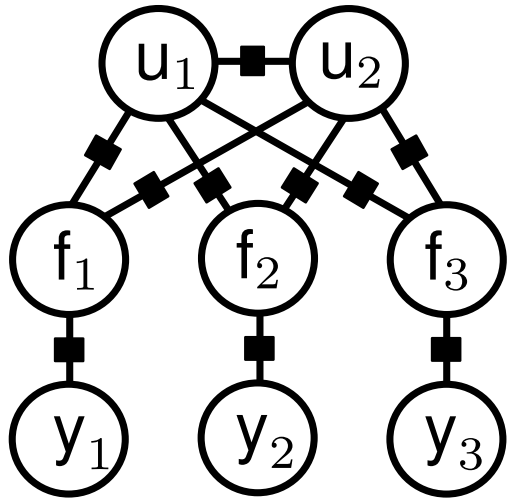
Fully independent training conditional (FITC) approximation

$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$

$$q(\mathsf{f}_t|\mathbf{u}) = p(\mathsf{f}_t|\mathbf{u})$$

$$= \mathcal{N}(\mathsf{f}_t; \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathbf{u}, \underbrace{\mathsf{K}_{\mathsf{f}_t\mathsf{f}_t} - \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}_t}}_{\mathsf{D}_{tt}})$$

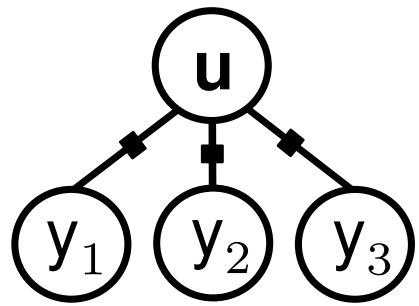
$$q(\mathsf{y}_t|\mathsf{f}_t) = p(\mathsf{y}_t|\mathsf{f}_t) = \mathcal{N}(\mathsf{y}_t; \mathsf{f}_t, \sigma_{\mathsf{y}}^2)$$



cost of computing likelihood is $\mathcal{O}(TM^2)$

$$p(\mathbf{y}_t|\theta) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathsf{K}_{\mathsf{fu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}} + \mathsf{D} + \sigma_{\mathsf{y}}^2\mathsf{I})$$

$$= \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathsf{K}_{\mathsf{fu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}} + \mathsf{D} + \sigma_{\mathsf{y}}^2\mathsf{I})$$



construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

indirect posterior approximation

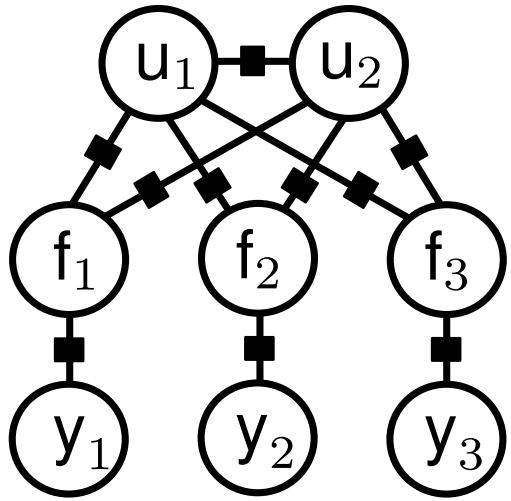
Fully independent training conditional (FITC) approximation

$$q(\mathbf{u}) = p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathsf{K}_{\mathsf{uu}})$$

$$q(\mathsf{f}_t|\mathbf{u}) = p(\mathsf{f}_t|\mathbf{u})$$

$$= \mathcal{N}(\mathsf{f}_t; \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathbf{u}, \underbrace{\mathsf{K}_{\mathsf{f}_t\mathsf{f}_t} - \mathsf{K}_{\mathsf{f}_t\mathsf{u}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}_t}}_{\mathsf{D}_{tt}})$$

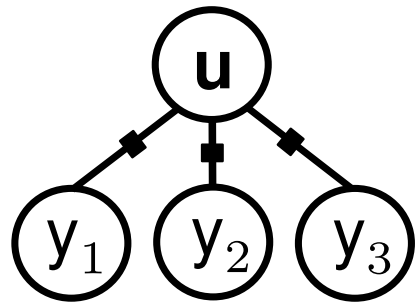
$$q(\mathsf{y}_t|\mathsf{f}_t) = p(\mathsf{y}_t|\mathsf{f}_t) = \mathcal{N}(\mathsf{y}_t; \mathsf{f}_t, \sigma_{\mathsf{y}}^2)$$



cost of computing likelihood is $\mathcal{O}(TM^2)$

$$p(\mathbf{y}_t|\theta) = \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathsf{K}_{\mathsf{fu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}} + \mathsf{D} + \sigma_{\mathsf{y}}^2\mathsf{I})$$

$$= \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathsf{K}_{\mathsf{fu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}} + \mathsf{D} + \sigma_{\mathsf{y}}^2\mathsf{I})$$

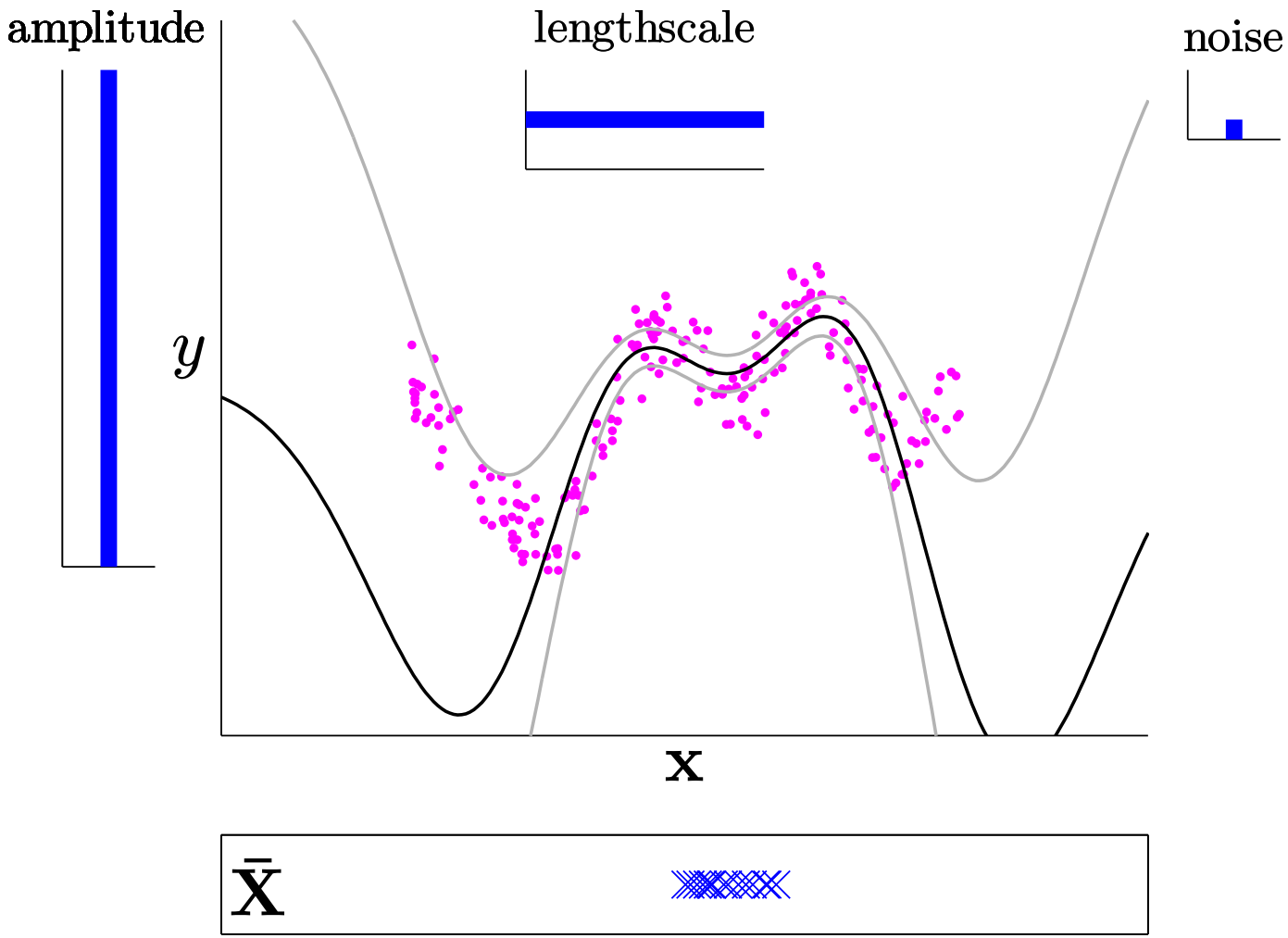


original variances along diagonal: stops variances collapsing

construct new generative model (with pseudo-data)
cheaper to perform exact learning and inference
calibrated to original

indirect posterior approximation

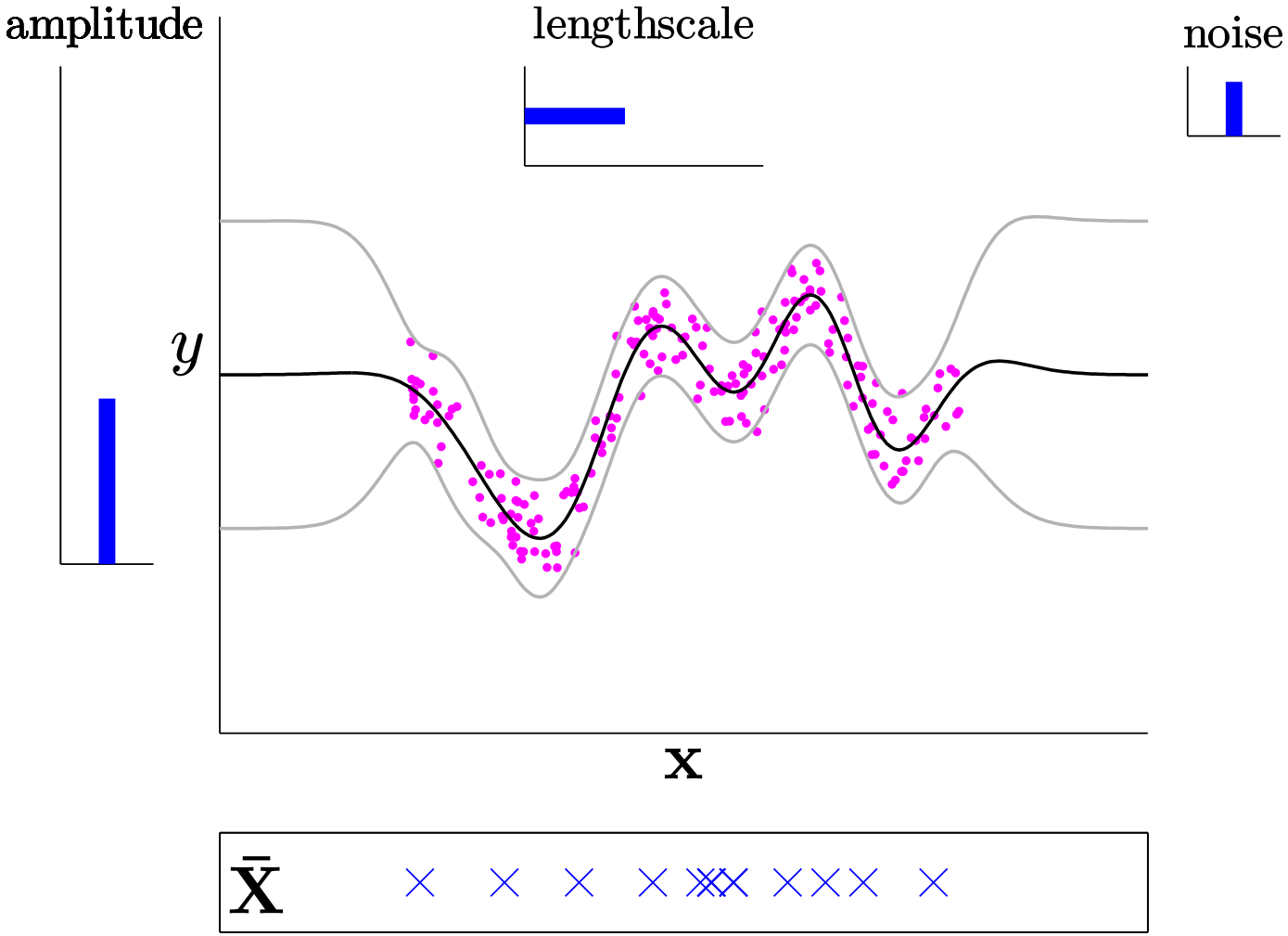
FITC: Demo (Snelson)



Initialize adversarially:

amplitude and lengthscale too big
noise too small
pseudo-inputs bunched up

FITC: Demo (Snelson)



Pseudo-inputs and hyperparameters optimized

Fully independent training conditional (FITC) approximation

- parametric (although cleverly so)
- if I see more data, should I add extra pseudo-data?
  - unnatural from a generative modelling perspective
  - natural from a prediction perspective (posterior gets more complex)

  $\Longrightarrow$ **lost elegant separation of model, inference and approximation**
- example of **prior approximation**

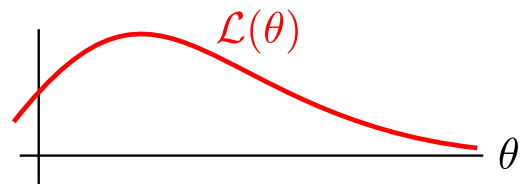
Extensions:

- inter-domain GP (pseudo-data in a different space)
- partially independent training conditional and tree-structured approximations

Variational free-energy method (VFE)

lower bound the likelihood

$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathrm{d}f \; p(\mathbf{y}, f|\theta)$$

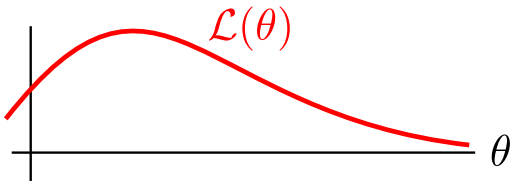


Variational free-energy method (VFE)

lower bound the likelihood

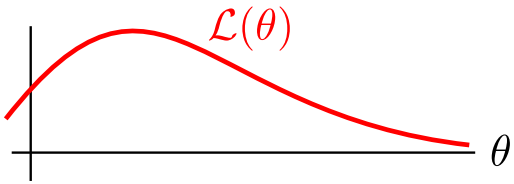
$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathsf{d}f \; p(\mathbf{y}, f|\theta)$$

$$= \log \int \mathsf{d}f \; p(\mathbf{y}, f|\theta) \frac{q(f)}{q(f)}$$



Variational free-energy method (VFE)

lower bound the likelihood

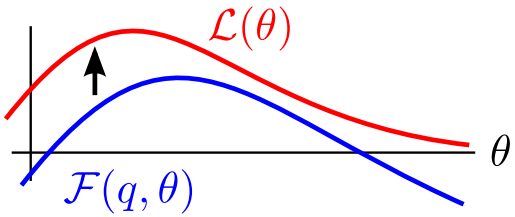


$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathsf{d}f \; p(\mathbf{y}, f|\theta)$$

$$= \log \int \mathsf{d}f \; p(\mathbf{y}, f|\theta) \frac{q(f)}{q(f)} \geq \int \mathsf{d}f \; q(f) \log \frac{p(\mathbf{y}, f|\theta)}{q(f)}$$

Variational free-energy method (VFE)

lower bound the likelihood

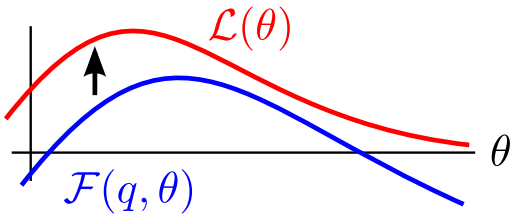


$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathsf{d}f \; p(\mathbf{y}, f|\theta)$$

$$= \log \int \mathsf{d}f \; p(\mathbf{y}, f|\theta) \frac{q(f)}{q(f)} \geq \int \mathsf{d}f \; q(f) \log \frac{p(\mathbf{y}, f|\theta)}{q(f)} = \mathcal{F}(\theta)$$

Variational free-energy method (VFE)

lower bound the likelihood



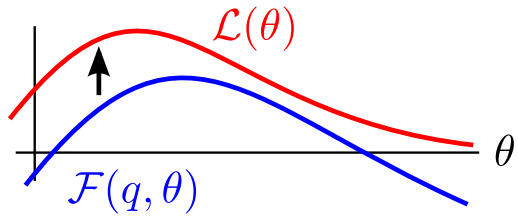
$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathrm{d}f \; p(\mathbf{y}, f|\theta)$$

$$= \log \int \mathrm{d}f \; p(\mathbf{y}, f|\theta) \frac{q(f)}{q(f)} \geq \int \mathrm{d}f \; q(f) \log \frac{p(\mathbf{y}, f|\theta)}{q(f)} = \mathcal{F}(\theta)$$

$$\mathcal{F}(\theta) = \int \mathrm{d}f \; q(f) \log \frac{p(f|\mathbf{y}, \theta) p(\mathbf{y}|\theta)}{q(f)}$$

Variational free-energy method (VFE)

lower bound the likelihood



$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathsf{d}f \; p(\mathbf{y}, f|\theta)$$

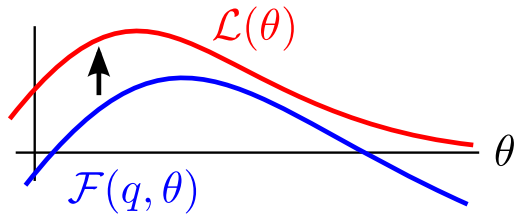
$$= \log \int \mathsf{d}f \; p(\mathbf{y}, f|\theta) \frac{q(f)}{q(f)} \geq \int \mathsf{d}f \; q(f) \log \frac{p(\mathbf{y}, f|\theta)}{q(f)} = \mathcal{F}(\theta)$$

$$\mathcal{F}(\theta) = \int \mathsf{d}f \; q(f) \log \frac{p(f|\mathbf{y}, \theta) p(\mathbf{y}|\theta)}{q(f)} = \log p(\mathbf{y}|\theta) - \mathbf{KL}(q(f)||p(f|\mathbf{y}))$$

KL between stochastic processes

Variational free-energy method (VFE)

lower bound the likelihood



$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathrm{d}f\ p(\mathbf{y}, f|\theta)$$

$$= \log \int \mathrm{d}f\ p(\mathbf{y}, f|\theta)\frac{q(f)}{q(f)} \geq \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}, f|\theta)}{q(f)} = \mathcal{F}(\theta)$$

$$\mathcal{F}(\theta) = \int \mathrm{d}f\ q(f) \log \frac{p(f|\mathbf{y}, \theta)p(\mathbf{y}|\theta)}{q(f)} = \log p(\mathbf{y}|\theta) - \mathbf{KL}(q(f)||p(f|\mathbf{y}))$$

KL between stochastic processes

assume approximate posterior factorisation with special form

$$q(f) = q(\mathbf{u}, f_{\neq \mathbf{u}}) = q(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$$

exact: $q(f_{\neq \mathbf{u}}|\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{y}, \mathbf{u})$

Variational free-energy method (VFE)

$$\mathcal{F}(\theta) \;=\; \log p(\mathbf{y}|\theta) - \mathbf{KL}(q(f)||p(f|\mathbf{y}))$$

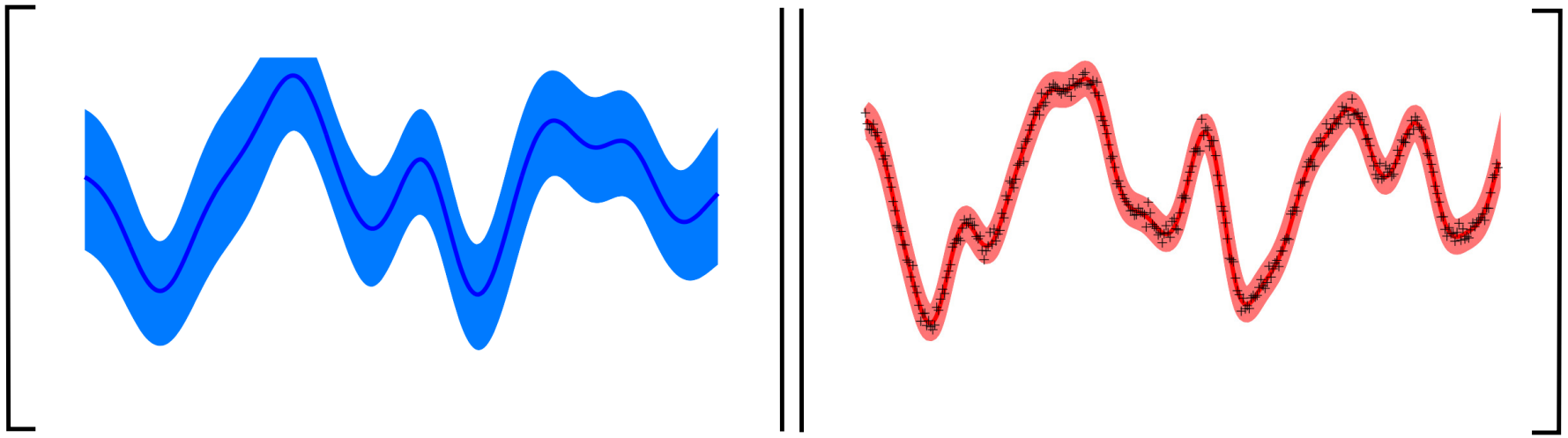
approximate posterior

$$q(f) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$$

true posterior

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

KL



Variational free-energy method (VFE)

$$\mathcal{F}(\theta) \;=\; \log p(\mathbf{y}|\theta) - \mathbf{KL}(q(f)||p(f|\mathbf{y}))$$

same form as prediction from GP-regression

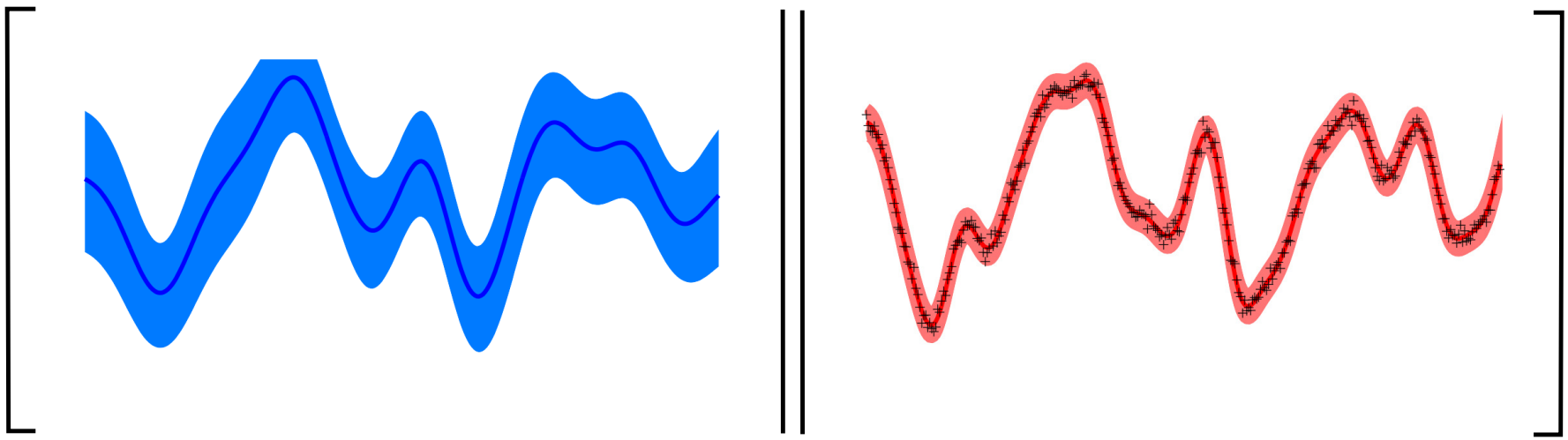
approximate posterior

$$q(f) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$$

true posterior

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

KL



Variational free-energy method (VFE)

$$\mathcal{F}(\theta) = \log p(\mathbf{y}|\theta) - \mathbf{KL}(q(f)||p(f|\mathbf{y}))$$

same form as prediction from GP-regression

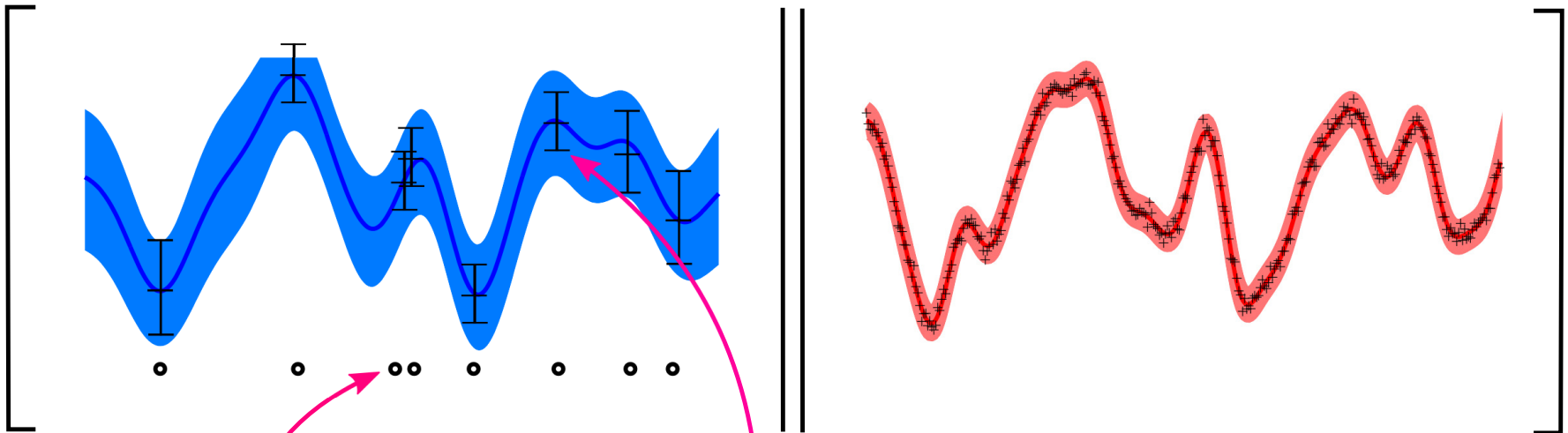
approximate posterior

$$q(f) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$$

true posterior

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

KL



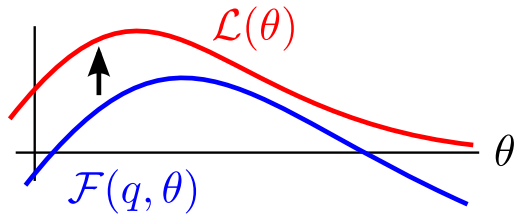
inputs locations of 'pseudo' data

output locations and covariance 'pseudo' data

optimise variational free-energy wrt to these variational parameters

Variational free-energy method (VFE)

lower bound the likelihood



$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathrm{d}f\ p(\mathbf{y}, f|\theta)$$

$$= \log \int \mathrm{d}f\ p(\mathbf{y}, f|\theta)\frac{q(f)}{q(f)} \geq \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}, f|\theta)}{q(f)} = \mathcal{F}(\theta)$$

$$\mathcal{F}(\theta) = \int \mathrm{d}f\ q(f) \log \frac{p(f|\mathbf{y}, \theta)p(\mathbf{y}|\theta)}{q(f)} = \log p(\mathbf{y}|\theta) - \mathbf{KL}(q(f)||p(f|\mathbf{y}))$$

KL between stochastic processes

assume approximate posterior factorisation with special form

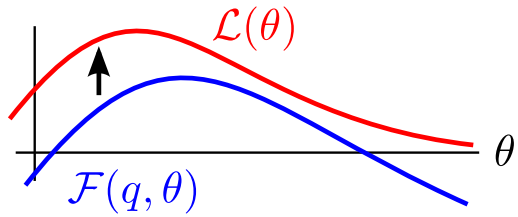
$$q(f) = q(\mathbf{u}, f_{\neq \mathbf{u}}) = q(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$$

← predictive from GP regression

exact: $q(f_{\neq \mathbf{u}}|\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{y}, \mathbf{u})$

Variational free-energy method (VFE)

lower bound the likelihood



$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathrm{d}f\ p(\mathbf{y}, f|\theta)$$

$$= \log \int \mathrm{d}f\ p(\mathbf{y}, f|\theta)\frac{q(f)}{q(f)} \geq \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}, f|\theta)}{q(f)} = \mathcal{F}(\theta)$$

$$\mathcal{F}(\theta) = \int \mathrm{d}f\ q(f) \log \frac{p(f|\mathbf{y}, \theta)p(\mathbf{y}|\theta)}{q(f)} = \log p(\mathbf{y}|\theta) - \mathbf{KL}(q(f)||p(f|\mathbf{y}))$$

KL between stochastic processes

assume approximate posterior factorisation with special form

$$q(f) = q(\mathbf{u}, f_{\neq \mathbf{u}}) = q(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$$

← predictive from GP regression

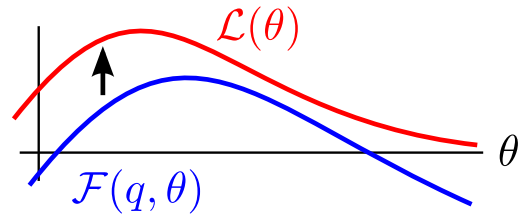
exact: $q(f_{\neq \mathbf{u}}|\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{y}, \mathbf{u})$

plug into Free-energy:

$$\mathcal{F}(\theta) = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}, f|\theta)}{p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})}$$

Variational free-energy method (VFE)

lower bound the likelihood



$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathrm{d}f\; p(\mathbf{y}, f|\theta)$$

$$= \log \int \mathrm{d}f\; p(\mathbf{y}, f|\theta)\frac{q(f)}{q(f)} \geq \int \mathrm{d}f\; q(f) \log \frac{p(\mathbf{y}, f|\theta)}{q(f)} = \mathcal{F}(\theta)$$

$$\mathcal{F}(\theta) = \int \mathrm{d}f\; q(f) \log \frac{p(f|\mathbf{y},\theta)p(\mathbf{y}|\theta)}{q(f)} = \log p(\mathbf{y}|\theta) - \mathbf{KL}(q(f)||p(f|\mathbf{y}))$$

KL between stochastic processes

assume approximate posterior factorisation with special form

$$q(f) = q(\mathbf{u}, f_{\neq \mathbf{u}}) = q(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$$

← predictive from GP regression

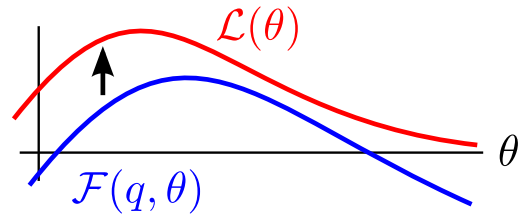
exact: $q(f_{\neq \mathbf{u}}|\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{y}, \mathbf{u})$

plug into Free-energy:

$$\mathcal{F}(\theta) = \int \mathrm{d}f\; q(f) \log \frac{p(\mathbf{y}, f|\theta)}{p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})} = \int \mathrm{d}f\; q(f) \log \frac{p(\mathbf{y}|\mathbf{f},\theta)p(f_{\neq \mathbf{u}}|\mathbf{u})p(\mathbf{u})}{p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})}$$

Variational free-energy method (VFE)

lower bound the likelihood



$$\mathcal{L}(\theta) = \log p(\mathbf{y}|\theta) = \log \int \mathrm{d}f\; p(\mathbf{y}, f|\theta)$$

$$= \log \int \mathrm{d}f\; p(\mathbf{y}, f|\theta)\frac{q(f)}{q(f)} \geq \int \mathrm{d}f\; q(f) \log \frac{p(\mathbf{y}, f|\theta)}{q(f)} = \mathcal{F}(\theta)$$

$$\mathcal{F}(\theta) = \int \mathrm{d}f\; q(f) \log \frac{p(f|\mathbf{y},\theta)p(\mathbf{y}|\theta)}{q(f)} = \log p(\mathbf{y}|\theta) - \mathbf{KL}(q(f)||p(f|\mathbf{y}))$$

KL between stochastic processes

assume approximate posterior factorisation with special form

$$q(f) = q(\mathbf{u}, f_{\neq \mathbf{u}}) = q(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$$

← predictive from GP regression

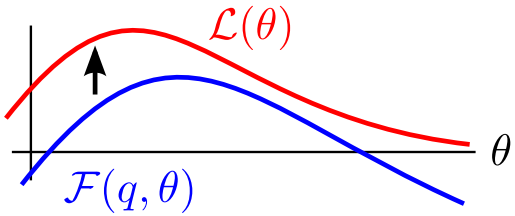
exact: $q(f_{\neq \mathbf{u}}|\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{y}, \mathbf{u})$

plug into Free-energy:

$$\mathcal{F}(\theta) = \int \mathrm{d}f\; q(f) \log \frac{p(\mathbf{y}, f|\theta)}{p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})} = \int \mathrm{d}f\; q(f) \log \frac{p(\mathbf{y}|\mathbf{f},\theta)\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}p(\mathbf{u})}{\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}q(\mathbf{u})}$$

Variational free-energy method (VFE)

lower bound the likelihood

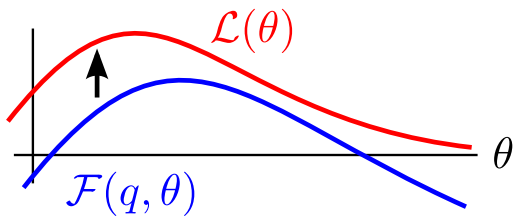


$$\mathcal{F}(\theta) = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}, f|\theta)}{p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})} = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}|\mathbf{f},\theta)\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}p(\mathbf{u})}{\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}q(\mathbf{u})}$$

where $q(f) = q(\mathbf{u}, f_{\neq \mathbf{u}}) = q(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$

Variational free-energy method (VFE)

lower bound the likelihood



$$\mathcal{F}(\theta) = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}, f|\theta)}{p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})} = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}|\mathbf{f},\theta)\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}p(\mathbf{u})}{\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}q(\mathbf{u})}$$

where $q(f) = q(\mathbf{u}, f_{\neq \mathbf{u}}) = q(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$

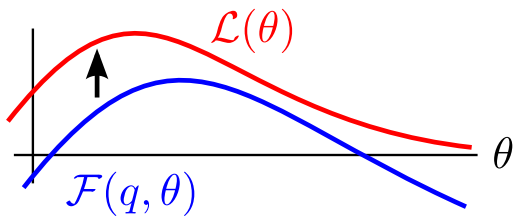
$$\mathcal{F}(\theta) = \langle \log p(\mathbf{y}|\mathbf{f},\theta) \rangle_{q(f)} - \mathbf{KL}(q(\mathbf{u})||p(\mathbf{u}))$$

average of quadratic form

KL between two multivariate Gaussians

Variational free-energy method (VFE)

lower bound the likelihood



$$\mathcal{F}(\theta) = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}, f|\theta)}{p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})} = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}|\mathbf{f},\theta)\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}p(\mathbf{u})}{\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}q(\mathbf{u})}$$

where $q(f) = q(\mathbf{u}, f_{\neq \mathbf{u}}) = q(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$

$$\mathcal{F}(\theta) = \langle \log p(\mathbf{y}|\mathbf{f},\theta) \rangle_{q(f)} - \mathbf{KL}(q(\mathbf{u})||p(\mathbf{u}))$$

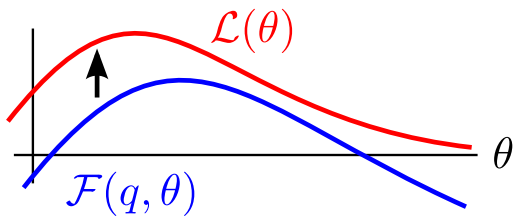
average of quadratic form

KL between two multivariate Gaussians

make bound as tight as possible: $q^*(\mathbf{u}) = \arg\max_{q(\mathbf{u})} \mathcal{F}(q,\theta)$

Variational free-energy method (VFE)

lower bound the likelihood



$$\mathcal{F}(\theta) = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}, f|\theta)}{p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})} = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}|\mathbf{f},\theta)\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}p(\mathbf{u})}{\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}q(\mathbf{u})}$$

where $q(f) = q(\mathbf{u}, f_{\neq \mathbf{u}}) = q(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$

$$\mathcal{F}(\theta) = \langle \log p(\mathbf{y}|\mathbf{f},\theta) \rangle_{q(f)} - \mathbf{KL}(q(\mathbf{u})||p(\mathbf{u}))$$

average of quadratic form

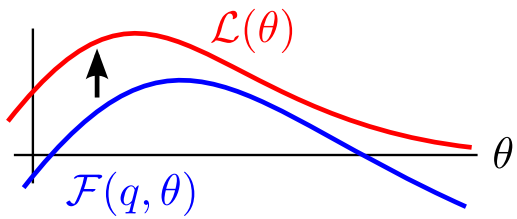
KL between two multivariate Gaussians

make bound as tight as possible: $q^*(\mathbf{u}) = \arg\max_{q(\mathbf{u})} \mathcal{F}(q,\theta)$

$q^*(\mathbf{u}) \propto p(\mathbf{u})\mathcal{N}(\mathbf{y}; \mathsf{K}_{\mathsf{fu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathbf{u}, \sigma_{\mathsf{y}}^2\mathsf{I})$ (DTC)

Variational free-energy method (VFE)

lower bound the likelihood



$$\mathcal{F}(\theta) = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}, f|\theta)}{p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})} = \int \mathrm{d}f\ q(f) \log \frac{p(\mathbf{y}|\mathbf{f},\theta)\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}p(\mathbf{u})}{\cancel{p(f_{\neq \mathbf{u}}|\mathbf{u})}q(\mathbf{u})}$$

where $q(f) = q(\mathbf{u}, f_{\neq \mathbf{u}}) = q(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u}) = p(f_{\neq \mathbf{u}}|\mathbf{u})q(\mathbf{u})$

$$\mathcal{F}(\theta) = \langle \log p(\mathbf{y}|\mathbf{f},\theta) \rangle_{q(f)} - \mathbf{KL}(q(\mathbf{u})||p(\mathbf{u}))$$

average of quadratic form

KL between two multivariate Gaussians

make bound as tight as possible: $q^*(\mathbf{u}) = \arg\max_{q(\mathbf{u})} \mathcal{F}(q,\theta)$

$q^*(\mathbf{u}) \propto p(\mathbf{u})\mathcal{N}(\mathbf{y}; \mathsf{K}_{\mathsf{fu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathbf{u}, \sigma_{\mathsf{y}}^2\mathsf{I})$ (DTC)

$$\mathcal{F}(q^*,\theta) = \log \mathcal{N}(\mathbf{y}; \mathbf{0}, \mathsf{K}_{\mathsf{fu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}}, \sigma_{\mathsf{y}}^2\mathsf{I}) - \frac{1}{2\sigma_{\mathsf{y}}^2}\mathsf{trace}(\mathsf{K}_{\mathsf{ff}} - \mathsf{K}_{\mathsf{fu}}\mathsf{K}_{\mathsf{uu}}^{-1}\mathsf{K}_{\mathsf{uf}})$$

DTC like

uncertainty based correction

Summary of VFE method

- optimisation of pseudo point inputs: **VFE has better guarantees than FITC**
- variational methods known to **underfit** (and have other **biases**)
- **no augmentation required: target is posterior over functions, which includes inducing variables**
  - pseudo-input locations are pure variational parameters (do not parameterise the generative model like they do in FITC)
  - coherent way of adding pseudo-data: more complex posteriors require more computational resources (more pseudo-points)
- Rule of thumb:
  VFE returns better mean estimates
  FITC returns better error-bar estimates
- **how should we select M = number of pseudo-points?**

State of the art

GP for big data https://www.youtube.com/watch?v=wF2jtws2nFI

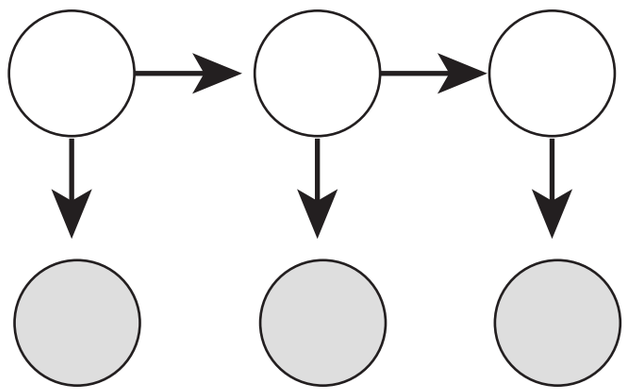
Deep GP (N. Lawrence)

Representing uncertainty in deep nets (Y.Gal, Z.Ghahramani)

GPs for time series: Rich Turner (U. Cambridge)

Efficient GPs for structure discovery: Andrew G Wilson (CMU)

Latents state models using GPs



$$\mathsf{x}_t = g(\mathsf{x}_{t-1}) + \sigma_{\mathsf{x}}\eta_t$$

$$\mathsf{y}_t = f(\mathsf{x}_t) + \sigma_{\mathbf{y}}\epsilon_t$$

$$f(\mathsf{x}), g(\mathsf{x}) \sim \mathcal{GP}(0, \mathsf{K})$$

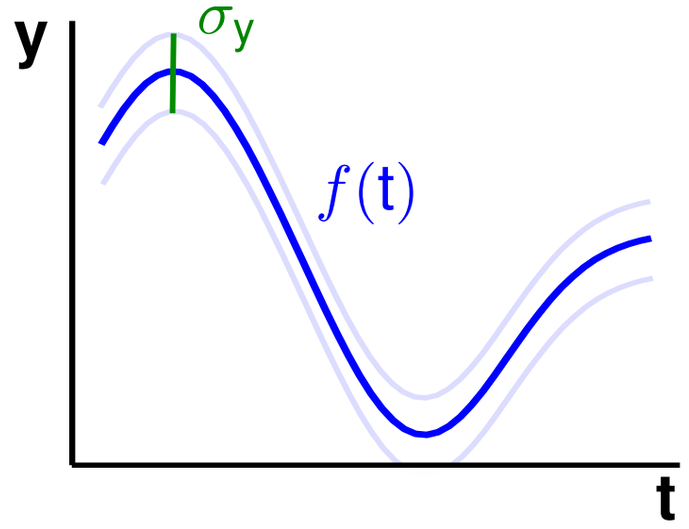
GPFA, GPLVM and co (Memming's lecture)

GP for time series: speech example

Generative model

$$\mathsf{y}(\mathsf{t}) = f(\mathsf{t}) + \epsilon\sigma_{\mathsf{y}} \quad \text{(independent Gaussian noise)}$$

$$p(\epsilon) = \mathcal{N}(0, 1)$$



place GP prior over the non-linear function

$$p(f(\mathsf{t})|\theta) = \mathcal{GP}(0, \mathsf{K}(\mathsf{t}, \mathsf{t}'))$$

$$\mathsf{K}(\mathsf{t}, \mathsf{t}') = \sigma^2 \cos(\omega(\mathsf{t} - \mathsf{t}')) \exp\left(-\frac{1}{2l^2}(\mathsf{t} - \mathsf{t}')^2\right)$$

typical time-series covariance
sinusoids with SE envelopes
power in Gaussian subband

audio time-series data