Bayesian Gaussian Process Latent Variable Model

Titsias, Lawrence, 2010

Heiko

Gatsby MLJC

March 19, 2014

Intro

GPLVM

Inference

Prediction

Experiments

Introduction

- GPs - usually supervised learning.
- GP-LVM
  - Multiple output GP regression where only outputs are observed
  - inputs unobserved, treated as latent variables
  - Can be seen as non-linear extension to PPCA
- Variational Bayesian Inference challenging: Need to integrate out latent variable that appear non-linearly in inverse kernel matrix of GP
  - Trick: GP prior is augmented to include auxiliary inducing variables (c.f. last session)
  - Leads to closed form lower bound on marginal likelihood
  - Avoid overfitting, infer dimensionality of non-linear latent space

Gaussian Process Latent Variable Model

- N data of dimension D, latent dimension $Q \ll D$
- Observed: $Y \in \mathbb{R}^{N \times D}$, latent: $X \in \mathbb{R}^{N \times Q}$
- Generative model, independent in features. The likelihood is:

$$p(Y|X) = \prod_{d=1}^{D} p(\mathbf{y}_d|X) = \prod_{d=1}^{D} \mathcal{N}(\mathbf{y}_d|\mathbf{0}, K_{NN} + \beta^{-1} I_N),$$

  $\mathbf{y}_d$ is d-th column of Y, K_{NN} is covariance/kernel of GP.
- ARD exponentiated square kernel:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{q=1}^{Q} \alpha_q (x_q - x_q')^2\right)$$

Gaussian Process Latent Variable Model

- Prior

$$p(X) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n | \mathbf{0}, I_Q),$$

  $\mathbf{x}_n$ is n-th row of X

- Joint model

$$p(Y, X) = p(Y|X)p(X)$$

- Hyperparameters: $\theta = (\sigma_f, \alpha_1, \ldots, \alpha_Q)$
- Usually, MAP estimate of X (Lawrence, 2005). Now, marginalise out.

Variational Inference

- Goal intractable

$$p(Y) = \int p(Y|X)p(X)dX$$
$$= \prod_{d=1}^{D} \mathcal{N}(\mathbf{y}_d|\mathbf{0}, K_{NN} + \beta^{-1} I_N) \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mathbf{0}, I_Q)$$

- Approximation $q(X)$ to posterior $p(X|Y)$

$$q(X) = \prod_{n=1}^{N} \mathcal{N}(\mathbf{x}_n|\mu_n, S_n)$$

with variational parameters μ_n, S_n gives lower bound

Lower bound on $p(Y)$

$$\begin{aligned}\log p(Y) &\geq F(q) \\ &= \int q(X) \log p(Y|X) dX - \int q(X) \log \frac{q(X)}{p(X)} dX \\ &= \tilde{F}(q) - KL(q||p)\end{aligned}$$

and

$$\tilde{F}(q) = \sum_{d=1}^{D} \int q(X) \log p(\mathbf{y}_d|X) dX = \sum_{d=1}^{D} \tilde{F}_d(q)$$

$KL(q||p)$ analytical, $\tilde{F}(q)$ still contains intractable integration.

Augmented Model

- Augment via $\mathbf{f}_d \in \mathbb{R}^N$ associated with $\mathbf{y}_d$ (noise corrupted)

$$p(\mathbf{y}_d, \mathbf{f}_d|X) = p(\mathbf{y}_d|\mathbf{f}_d)p(\mathbf{f}_d|X),$$

where $p(\mathbf{y}_d|\mathbf{f}_d) = \mathcal{N}(\mathbf{y}_d|\mathbf{f}_d, \beta^{-1}I_N)$ and
$p(\mathbf{f}_d|X) = \mathcal{N}(\mathbf{f}_d|\mathbf{0}, K_{NN})$

- Likelihood $p(\mathbf{y}_d|X)$ is marginal likelihood of this GP regression model

$$\begin{aligned}\int p(\mathbf{y}_d|\mathbf{f}_d)p(\mathbf{f}_d|X) &= \int \mathcal{N}(\mathbf{y}_d|\mathbf{f}_d, \beta^{-1}I_N)\mathcal{N}(\mathbf{f}_d|\mathbf{0}, K_{NN})d\mathbf{f}_d \\ &= \mathcal{N}(\mathbf{y}_d|\mathbf{0}, K_{NN} + \beta^{-1}I_N) \\ &= p(\mathbf{y}_d|X)\end{aligned}$$

Inducing Variables

- Sparse approximation via M inducing variables $\mathbf{u}_d \in \mathbb{R}^M$ at input locations $Z \in \mathbb{R}^{M \times Q}$

$$p(\mathbf{y}_d, \mathbf{f}_d, \mathbf{u}_d | X, Z) = p(\mathbf{y}_d | \mathbf{f}_d) p(\mathbf{f}_d | \mathbf{u}_d, X, Z) p(\mathbf{u}_d | Z)$$

- Std conditioning of (jointly Gaussian)

$$p(\mathbf{f}_d, \mathbf{u}_d, | X, Z) = p(\mathbf{f}_d | \mathbf{u}_d, X, Z) p(\mathbf{u}_d | Z)$$

gives

$$p(\mathbf{f}_d | \mathbf{u}_d, X, Z) = \mathcal{N}(\mathbf{f}_d | \alpha_d, K_{NN} - K_{NM} K_{MM}^{-1} K_{MN})$$

where $\alpha_d = K_{NM} K_{MM}^{-1} \mathbf{u}_d$ and $p(\mathbf{u}_d | Z) = \mathcal{N}(\mathbf{u}_d | \mathbf{0}, K_{MM})$

Sparse Variational Inference

- Idea: $p(\mathbf{y}_d|X)$ can be equivalently computed from marginalising out $(\mathbf{f}_d, \mathbf{u}_d)$ in the augmented model.
- True for any inducing input Z. Therefore, Z are neither random variables nor hyperparameters, but variational parameters.
- See Titsias 2009.

Sparse Variational Inference

Lower bound via (again) approximating posterior $p(\mathbf{f}_d, \mathbf{u}_d|\mathbf{y}_d, X) = p(\mathbf{f}_d|\mathbf{u}_d, \mathbf{y}_d, X)p(\mathbf{u}_d|\mathbf{y}_d, X)$ with

$$q(\mathbf{f}_d, \mathbf{u}_d) = p(\mathbf{f}_d|\mathbf{u}_d, X)\phi(\mathbf{u}_d).$$

$\phi(\mathbf{u}_d)$ is variational distribution over inducing variables (independent of X). The bound is (?)

$$\log p(\mathbf{y}_d|X) \geq \int \phi(\mathbf{u}_d) \log \frac{p(\mathbf{u}_d)\mathcal{N}(\mathbf{y}_d|\alpha_d, \beta^{-1}I_N)}{\phi(\mathbf{u}_d)} d\mathbf{u}_d$$
$$- \frac{\beta}{2}\mathrm{Tr}(K_{NN} - K_{NM}K_{MM}^{-1}K_{MN})$$

Mean Field Approach

- Plug into original lower bound (using trace rotation)

$$\tilde{F}_d(q) \geq \int q(X) \left[\int \phi(\mathbf{u}_d) \log \frac{p(\mathbf{u}_d)\mathcal{N}(\mathbf{y}_d|\alpha_d, \beta^{-1} I_N)}{\phi(\mathbf{u}_d)} d\mathbf{u}_d \right.$$
$$\left. - \frac{\beta}{2}\mathrm{Tr}(K_{NN}) + \frac{\beta}{2}\mathrm{Tr}(K_{MM}^{-1} K_{MN} K_{NM}) \right] dX$$

- $\phi(\mathbf{u}_d)$ does not depend on X, can change integration order

$$\tilde{F}_d(q) \geq \int \phi(\mathbf{u}_d) \left[\langle \log \mathcal{N}(\mathbf{y}_d|\alpha_d, \beta^{-1} I_N) \rangle_{q(X)} + \log \frac{p(\mathbf{u}_d)}{\phi(\mathbf{u}_d)} \right] d\mathbf{u}_d$$
$$\left. - \frac{\beta}{2}\mathrm{Tr}\left(\langle K_{NN} \rangle_{q(X)} \right) + \frac{\beta}{2}\mathrm{Tr}\left(K_{MM}^{-1} \langle K_{MN} K_{NM} \rangle_{q(X)} \right) \right]$$

Mean Field Approach

- Optimal setting is $\phi(\mathbf{u}_d) \propto e^{\langle \log \mathcal{N}(\mathbf{y}_d | \alpha_d, \beta^{-1} I_N) \rangle_{q(X)}} p(\mathbf{u}_d)$, reverse Jensen (?) to get

$$\tilde{F}_d(q) \geq \log \left(\int e^{\langle \log \mathcal{N}(\mathbf{y}_d | \alpha_d, \beta^{-1} I_N) \rangle_{q(X)}} p(\mathbf{u}_d) d\mathbf{u}_d \right) - \frac{\beta}{2} \mathrm{Tr} \left(\langle K_{NN} \rangle_{q(X)} \right) + \frac{\beta}{2} \mathrm{Tr} \left(K_{MM}^{-1} \langle K_{MN} K_{NM} \rangle_{q(X)} \right)$$

- RHS can be computed using statistics $\Psi_0 = \mathrm{Tr} \left(\langle K_{NN} \rangle_{q(X)} \right)$, $\Psi_1 = \langle K_{NM} \rangle_{q(X)}$, $\Psi_2 \langle K_{MN} K_{NM} \rangle_{q(X)}$
- Closed form for exponentiated square kernels
- LHS is a quadratic function in $\mathbf{u}_d$ that depends on above statistics Ψ_1, Ψ_2.

Finally, optimise

- Plug into original bound

$$\log p(Y) \geq F(q) = \tilde{F}(q) - KL(q||p)$$

- Optimise wrt variational parameters $(\{\mu_n, S_n\}_{n=1}^N, Z)$ and hyperparameters (β, θ) "by applying gradient based optimisation techniques"
- I skip the details on the Ψ-statistics here. Technical, based on Gaussian convolutions.

Prediction and computation of proabilities in test data

Two applications:

1. Predict $p(\mathbf{y}_*|Y)$ of some observed test data $\mathbf{y}_* \in \mathbb{R}^D$
2. Predict missing values of partially observed test output $\mathbf{y}_* = (\mathbf{y}_*^O, \mathbf{y}_*^U) \in \mathbb{R}^D$. E.g. noise removal.

Predict $p(\mathbf{y}_*|Y)$

- Latent variables X correspond to training points Y, latent variable $\mathbf{x}_*$ corresponds to test output
- Marginalise over both

$$p(\mathbf{y}_*|Y) = \frac{p(\mathbf{y}_*, Y)}{p(Y)} = \frac{\int\int p(\mathbf{y}_*, Y|X, \mathbf{x}_*)p(X, \mathbf{x}_*)dXd\mathbf{x}_*}{\int p(Y|X)p(X)dX}$$

- Denominator: Solved above. Gives variational distribution $q(X)$. Fixed during test time.
- Numerator: Optimise wrt (μ_*, S_*) of $q(\mathbf{x}_*)$. Local minima avoided by clever initialisation based on neighbours of $\mathbf{y}_*$.
- Solution: Ratio of lower bounds

$$p(\mathbf{y}_*|Y) \approx q(\mathbf{y}_*|Y) = e^{F(q(X,\mathbf{x}_*))-F(q(X))}$$

Oil data

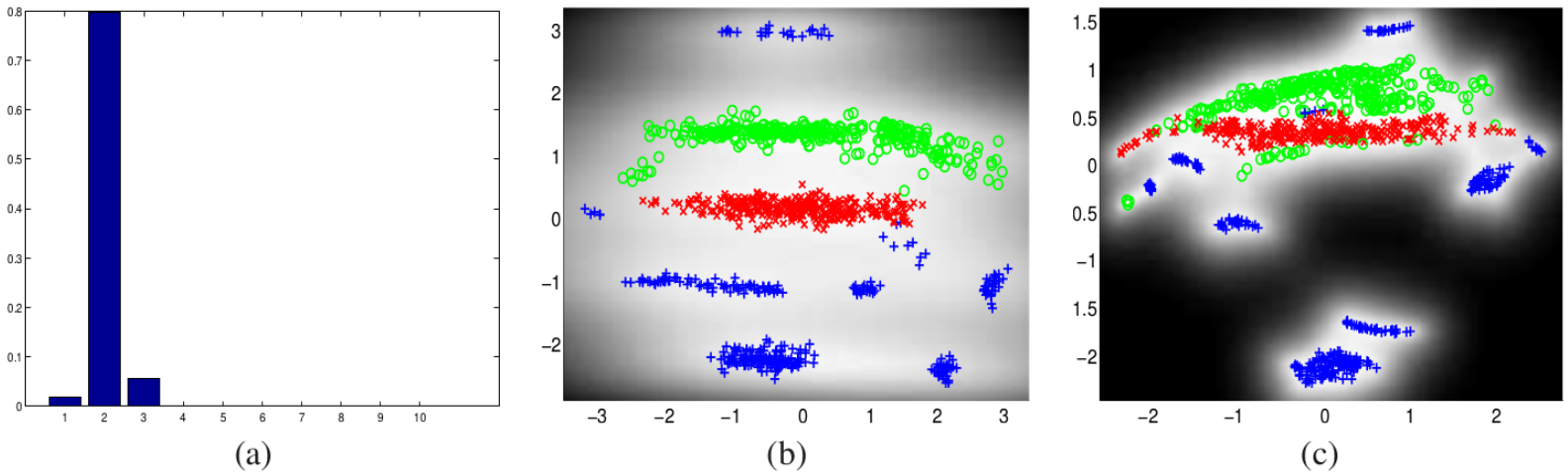

Figure 1: Panel (a) shows the inverse lengthscales found by applying the Bayesian GP-LVM with ARD SE kernel on the oil flow data. Panel (b) shows the visualization achieved by keeping the most dominant latent dimensions (2 and 3) which have the largest inverse lengthscale value. Dimension 2 is plotted on the y-axis and 3 and on the x-axis. Plot (c) shows the visualization found by standard sparse GP-LVM.