The Gaussian process latent variable model with Cox regression

James Barrett

Institute for Mathematical and Molecular Biomedicine (IMMB), King's College London

Layout

1. Introduction
2. GPLVM
3. Results
4. Extension to Cox proportional hazards model

Integrating Multiple Data Sources via the GPLVM

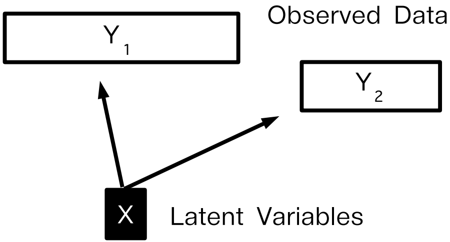
The Gaussian process latent variable model (Lawrence, 2005) is a flexible non-parametric probabilistic dimensionality reduction method.

We want to:

- Represent each dataset in terms of latent variables.
- Extract information common to each data source.
- Retain information unique to each source.
- Account for dimension mismatch between multiple datasets.

Also:

- Detect any intrinsic low dimensional structure.



Layout

Model Definition

- Observe S datasets $\mathbf{Y}_1 \in \mathbb{R}^{N \times d_1}, \ldots, \mathbf{Y}_S \in \mathbb{R}^{N \times d_S}$.
- It is assumed each column of $\mathbf{Y}_s$ is normalised to zero mean and unit variance.
- Represent these data in terms of q latent variables $\mathbf{x}$ where $q < \min_s(d_s)$.

For individual i and covariate μ in source s we write

$$y^s_{i\mu} = \sum_{m=1}^{M} w^s_{\mu m} \phi^s_m(\mathbf{x}_i) + \xi^s_{i\mu}$$

Where

- $\phi^s_m : \mathbb{R}^q \to \mathbb{R}^M$ are non-linear mappings that may depend on hyperparameters $\boldsymbol{\phi}_s$
- $w^s_{\mu m}$ are mapping coefficients
- $\xi^s_{i\mu}$ are noise variables.

Data Likelihood

Assume Gaussian priors for $p(\mathbf{W}_s)$ and $p(\boldsymbol{\xi}_s|\beta_s)$ with zero mean and covariances given by

$$\left\langle w^{s}_{\mu m} w^{s'}_{\nu n} \right\rangle = \delta_{ss'}\delta_{\mu\nu}\delta_{mn} \quad \text{and} \quad \left\langle \xi^{s}_{i\mu}\xi^{s'}_{j\nu} \right\rangle = \beta_s^{-1}\delta_{ss'}\delta_{ij}\delta_{\mu\nu}.$$

For notational simplicity we define $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_S\}$, $\boldsymbol{\Phi} = \{\phi_1, \ldots, \phi_S\}$, $\mathbf{W} = \{\mathbf{W}_1, \ldots, \mathbf{W}_S\}$, $\boldsymbol{\xi} = \{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_S\}$ and $\mathbf{Y} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_S\}$. The data likelihood factorises over samples

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{W}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\Phi}) = \prod_{i=1}^{N} p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{W}, \boldsymbol{\xi}_i, \boldsymbol{\beta}, \boldsymbol{\Phi})$$

Marginalising $\mathbf{W}$ and $\boldsymbol{\xi}$ we get a Gaussian distribution for $\mathbf{Y}$ with mean $\langle y_{i\mu} \rangle = 0$ and covariance

$$\begin{aligned}\left\langle y^{s}_{i\mu} y^{s'}_{j\nu} \right\rangle &= \delta_{ss'}\delta_{\mu\nu}\left(\sum_{m} \phi^{s}_{m}(\mathbf{x}_i)\phi^{s}_{m}(\mathbf{x}_j) + \beta_s^{-1}\delta_{ij}\right) \\ &= \delta_{ss'}\delta_{\mu\nu} K_s(\mathbf{x}_i, \mathbf{x}_j)\end{aligned}$$

The data likelihood can then be written as

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Phi}) = \prod_{s=1}^{S}\prod_{\mu=1}^{d_s} \frac{e^{-\frac{1}{2}\mathbf{y}^{s}_{:,\mu}\mathbf{K}_s^{-1}\mathbf{y}^{s}_{:,\mu}}}{(2\pi)^{\frac{N}{2}}|\mathbf{K}_s|^{\frac{1}{2}}}$$

Bayesian Inference

We specify three levels of uncertainty:

- Microscopic parameters: $\{\mathbf{X}\}$
- Hyperparameters: $\{\boldsymbol{\beta}, \mathbf{\Phi}\}$
- Models: $H = \{q, \phi_m\}$

Posterior distributions are

$$p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\beta}, \mathbf{\Phi}, H) = \frac{p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}, \mathbf{\Phi}, H) p(\mathbf{X}|H)}{\int \mathrm{d}\mathbf{X}' p(\mathbf{Y}|\mathbf{X}', \boldsymbol{\beta}, \mathbf{\Phi}, H) p(\mathbf{X}'|H)}$$

$$p(\boldsymbol{\beta}, \mathbf{\Phi}|\mathbf{Y}, H) = \frac{p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{\Phi}, H) p(\boldsymbol{\beta}, \mathbf{\Phi}|H)}{\int \mathrm{d}\boldsymbol{\beta}' \mathrm{d}\mathbf{\Phi}'\, p(\mathbf{Y}|\boldsymbol{\beta}', \mathbf{\Phi}', H) p(\boldsymbol{\beta}', \mathbf{\Phi}'|H)}$$

$$P(H|\mathbf{Y}) = \frac{p(\mathbf{Y}|H) p(H)}{\sum_{H'} p(\mathbf{Y}|H') p(H')},$$

where

$$p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{\Phi}, H) = \int \mathrm{d}\mathbf{X} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\beta}, \mathbf{\Phi}, H) p(\mathbf{X}|H)$$

$$p(\mathbf{Y}|H) = \int \mathrm{d}\boldsymbol{\beta} \mathrm{d}\mathbf{\Phi}\, p(\mathbf{Y}|\boldsymbol{\beta}, \mathbf{\Phi}, H) p(\boldsymbol{\beta}, \mathbf{\Phi}|H).$$

Inferring latent variables

To find the optimal latent variable representation, $\mathbf{X}^\star$ we will numerically minimise the negative log likelihood of $p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Phi}, H)$

$$\mathcal{L}_X(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\Phi}) = \sum_s \left[\frac{d_s}{2N} \mathrm{tr}(\mathbf{K}_s^{-1}\mathbf{S}_s) + \frac{d_s}{2N} \log |\mathbf{K}_s| + \frac{d_s}{2} \log 2\pi \right]$$

where $\mathbf{S}_s = \frac{1}{d_s}\mathbf{Y}_s\mathbf{Y}_s^T$. Should we rescale the contribution from each source by d_{tot}/d_s where $d_{tot} = \sum_s d_s$? Expand to second order

$$\mathcal{L}_X(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\Phi}) \approx \mathcal{L}_X(\mathbf{X}^\star; \boldsymbol{\beta}, \boldsymbol{\Phi}) + \frac{1}{2} \sum_{i,j}^{N} \sum_{\mu,\nu}^{q} (x_{i\mu}^\star - x_{i\mu})(x_{j\nu}^\star - x_{j\nu}) A_{i\mu,j\nu}$$

where

$$A_{i\mu,j\nu} = \frac{\partial^2}{\partial x_{i\mu} \partial x_{j\nu}} \mathcal{L}_X(\mathbf{X}; \boldsymbol{\beta}, \boldsymbol{\Phi}\}) \Big|_{\mathbf{X}=\mathbf{X}^\star}$$

$$\begin{aligned} p(\mathbf{Y}|\boldsymbol{\beta}, \boldsymbol{\Phi}, H) &= \int \mathrm{d}\mathbf{X} e^{-N\mathcal{L}_X(\mathbf{X};\boldsymbol{\beta},\boldsymbol{\Phi})} \\ &= p(\mathbf{Y}|\mathbf{X}^\star, \boldsymbol{\beta}, \boldsymbol{\Phi}, H) \int \mathrm{d}\mathbf{X} e^{-\frac{1}{2}\sum_{ij}\sum_{\mu\nu}(x_{i\mu}^\star - x_{i\mu})(x_{j\nu}^\star - x_{j\nu})A_{i\mu,j\nu}} \\ &= p(\mathbf{Y}|\mathbf{X}^\star, \boldsymbol{\beta}, \boldsymbol{\Phi}, H)(2\pi)^{Nq/2} |\mathbf{A}(\mathbf{X}^\star, \boldsymbol{\beta}, \boldsymbol{\Phi})|^{-1/2} \end{aligned}$$

Invariance under Unitary Transformations

The kernel functions considered here are all invariant under arbitrary unitary transformations. Let $\mathbf{U}$ be a unitary matrix, such that $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{U}\mathbf{U}^{\mathsf{T}} = \mathbf{I}$ and let $\tilde{\mathbf{x}} = \mathbf{U}\mathbf{x}$. Then

$$\tilde{\mathbf{x}}_i \cdot \tilde{\mathbf{x}}_j = \mathbf{x}_i \mathbf{U}^{\mathsf{T}} \mathbf{U} \mathbf{x}_j = \mathbf{x}_i \cdot \mathbf{x}_j$$

and

$$(\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j)^2 = (\mathbf{x}_i - \mathbf{x}_j)\mathbf{U}^{\mathsf{T}}\mathbf{U}(\mathbf{x}_i - \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^2.$$

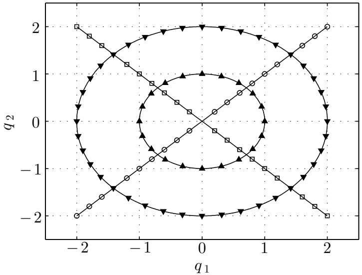
This invariance under unitary transformations induces symmetries in the posterior search space of $\mathbf{X} \in \mathbb{R}^{N \times q}$. Fix with

$$\mathbf{X} = \begin{pmatrix} x_{11} & 0 & 0 & 0 \\ x_{21} & x_{22} & 0 & 0 \\ x_{31} & x_{32} & x_{33} & 0 \\ x_{41} & x_{42} & x_{43} & x_{44} \\ \vdots & & & \vdots \end{pmatrix}.$$

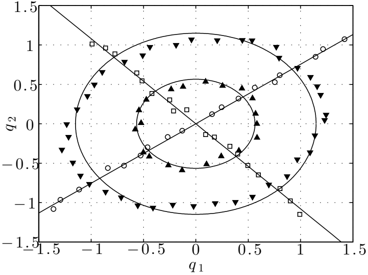
We 'pin down' the latent variables and optimise over the $Nq - (q^2 - q)/2$ non zero entries.

Layout

Synthetic Data



(a) 'True' latent variables



(b) Retrieved latent variables

We can define three ad hoc error measures

$$\mathcal{E}_{radial} = \frac{1}{|C|} \sum_{i \in C} \frac{|\mathbf{x}_i| - \tilde{r}}{\tilde{r}} \qquad \mathcal{E}_{angular} = \frac{1}{|C|} \sum_{i \in C} \frac{\Delta\theta_i - \tilde{\theta}}{\tilde{\theta}} \qquad \mathcal{E}_{linear} = \frac{SS_{err}}{SS_{tot}}$$

where $SS_{err} = \sum (x_{i2} - \alpha x_{i1})^2$ and $SS_{tot} = \sum (x_{i2} - \bar{x}_{i2})^2$.

Dependence on β and d

β	$\mathcal{E}_{radial}$	$\mathcal{E}_{angular}$	$\mathcal{E}_{linear}$
0.1	0.0060	0.0046	0.0079
0.5	0.0766	0.0813	0.2577
1.0	0.0998	0.1701	0.3263

(a) Dependence on β

d	$\mathcal{E}_{radial}$	$\mathcal{E}_{angular}$	$\mathcal{E}_{linear}$
10	0.0944	0.0454	0.5491
100	0.0061	0.0051	0.0108
1000	0.0004	0.0008	0.0016

(b) Dependence on d

Table: (a) The magnitude of the errors increases as more noise is added (for fixed d) to the synthetic data. (b) For fixed noise levels the greater d is the better the extraction of the 'true' low dimensional structure from a dataset.

Dimensionality detection

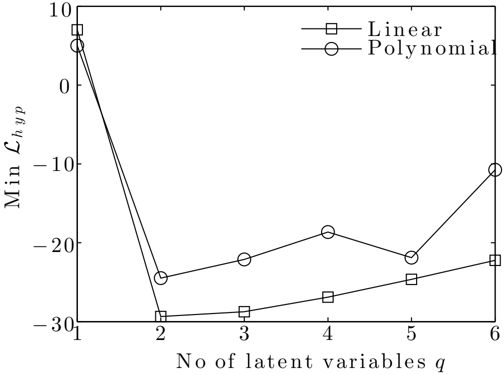
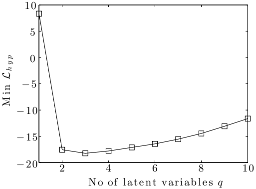
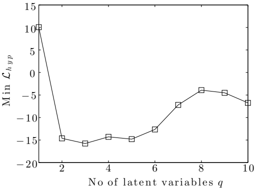


Figure: Plot of the minimal values of $\mathcal{L}_{hyp}$ obtained for different values of q and two different kernels, the linear kernel and the polynomial kernel. Both the kernel types detect that $q = 2$ is the optimal dimension. Furthermore, the model can distinguish that the linear kernel offers the best explanation of the data in this case.

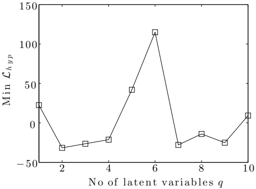
Integration of two sources



(a) Linear



(b) Polynomial



(c) Combined

Classification experiment

We generated $N = 100$ samples with $q = 2$.

- 50 samples from a Gaussian with unit variance and mean $(1, 1)$ with class $+1$.
- 50 samples from a unit variance Gaussians with means $(-\frac{1}{2}, -\frac{1}{2})$ and $(\frac{1}{2}, -\frac{1}{2})$ with class -1.
- Projected into $d = 100$ space with a linear mapping.

	$\mathbf{Y}$ $(d = 100)$	$\mathbf{X}^\star$ $(q = 2)$
Training Success	86.3%	86.6%
Validation Success	74.0%	83.0%

We repeated this 300 times. The mean improvement between the validation success on $\mathbf{Y}$ and the success on $\mathbf{X}^\star$ was found to be 8.7% with a standard deviation of 4.8%.

Layout

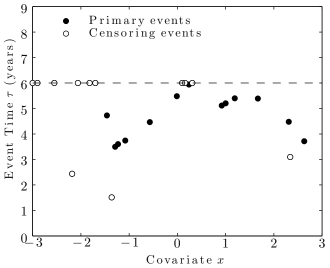
1. Introduction
2. GPLVM
3. Results
4. Extension to Cox proportional hazards model

What is Survival Analysis?

Suppose we have a group of N cancer patients. For each individual i we measure:

- The time $\tau_i \geq 0$ until an event of interest occurs, for example the time to metastasis.
- A vector of covariates (also called features or input variables) $\mathbf{x}_i \in \mathbb{R}^d$
- We will assume one risk and use an indicator variable

$$\Delta_i = \begin{cases} 0 & \text{if } i \text{ is censored} \\ 1 & \text{if the primary risk occurs} \end{cases}$$



Aim

To extract any statistical relationship between $\mathbf{x}$ and τ for each risk.

Challenges:

- How can we incorporate information from censored individuals?
- How can we deal with non-negative outputs?

Linking Survival Data to Latent Variables

For each patient we observe covariates **y**, time to event t, and event type Δ.

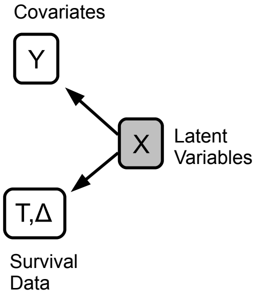
$$p(\mathbf{X}|\mathbf{Y},\mathbf{t}) \propto p(\mathbf{Y},\mathbf{t}|\mathbf{X})p(\mathbf{X}) = \underbrace{p(\mathbf{Y}|\mathbf{X})}_{\text{GPLVM}}\underbrace{p(\mathbf{t}|\mathbf{X})}_{\text{Cox}}p(\mathbf{X})$$

where as above

$$p(\mathbf{Y}|\mathbf{X}) = \prod_{\mu=1}^{d} \frac{e^{-\frac{1}{2}\mathbf{y}_{:,\mu}^{\mathsf{T}}\mathbf{K}^{-1}\mathbf{y}_{:,\mu}}}{(2\pi)^{\frac{N}{2}}|\mathbf{K}|^{\frac{1}{2}}}$$

and for the Cox model

$$p(\mathbf{t}|\mathbf{X}) = \prod_{i=1}^{N} \lambda_0(t)e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}e^{-e^{\boldsymbol{\beta}\cdot\mathbf{x}_i}\Lambda_0(t)}$$



Predictions

If we observe a new patient with $\mathbf{y}^\star$ we predict the corresponding event time $t^\star$ via

$$\mathbf{y}^\star \xrightarrow{GPLVM} \mathbf{x}^\star \xrightarrow{Cox} t^\star$$

We can also use Cox to generate survival curves: