Dimensionality reduction for survival data via the Gaussian process latent variable model (GPLVM)

James Barrett

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

4th June 2014, KCL
1. Dimensionality and overfitting

2. The GPLVM-WPHM

3. Simulation studies
Dimensionality and overfitting

- Observe data $\mathbf{Y} \in \mathbb{R}^{N \times d}$ with $N$ individuals and $d$ covariates.
- Event times $\tau_i > 0$.
- Indicator variable: $\Delta_i = 1$ if event occurred, $\Delta_i = 0$ if $i$ is censored.
- Assume one risk with independent right censoring.

Overfitting

Inference of relationships between covariates and outcomes that are due to chance, are not biologically meaningful, and fail to exist in test data.

- Overfitting can be severe when $d \gg N$.
- But can also occur when $d < N$.
- Also can be a problem when complex models are used.
Spurious Correlations 1

Divorce rate in Maine correlates with Per capita consumption of margarine (US)

<table>
<thead>
<tr>
<th>Year</th>
<th>Divorce rate in Maine (Divorces per 1000 people, US Census)</th>
<th>Per capita consumption of margarine (US (Pounds, USDA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>5</td>
<td>8.2</td>
</tr>
<tr>
<td>2001</td>
<td>4.7</td>
<td>7</td>
</tr>
<tr>
<td>2002</td>
<td>4.6</td>
<td>6.5</td>
</tr>
<tr>
<td>2003</td>
<td>4.4</td>
<td>5.3</td>
</tr>
<tr>
<td>2004</td>
<td>4.3</td>
<td>5.2</td>
</tr>
<tr>
<td>2005</td>
<td>4.1</td>
<td>4</td>
</tr>
<tr>
<td>2006</td>
<td>4.2</td>
<td>4.6</td>
</tr>
<tr>
<td>2007</td>
<td>4.2</td>
<td>4.5</td>
</tr>
<tr>
<td>2008</td>
<td>4.2</td>
<td>4.2</td>
</tr>
<tr>
<td>2009</td>
<td>4.1</td>
<td>3.7</td>
</tr>
</tbody>
</table>

Correlation: 0.992558

Source: Tyler Vigen (http://www.tylervigen.com)
Spurious Correlations 2

**Age of Miss America**
correlates with
**Murders by steam, hot vapours and hot objects**

![Graph showing correlation between Age of Miss America and Murders by steam, hot vapours and hot objects](image)

<table>
<thead>
<tr>
<th>Year</th>
<th>Age of Miss America</th>
<th>Murders by steam, hot vapours and hot objects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>2000</td>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>2001</td>
<td>24</td>
<td>7</td>
</tr>
<tr>
<td>2002</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>2003</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>2004</td>
<td>21</td>
<td>3</td>
</tr>
<tr>
<td>2005</td>
<td>24</td>
<td>8</td>
</tr>
<tr>
<td>2006</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>2007</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>2008</td>
<td>19</td>
<td>3</td>
</tr>
<tr>
<td>2009</td>
<td>22</td>
<td>2</td>
</tr>
</tbody>
</table>

Correlation: 0.870127

Source: Tyler Vigen ([http://www.tylervigen.com](http://www.tylervigen.com))
Per capita consumption of cheese (US) correlates with Number of people who died by becoming tangled in their bedsheets

<table>
<thead>
<tr>
<th>Year</th>
<th>Cheese (US)</th>
<th>Deaths (US)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>29.8</td>
<td>327</td>
</tr>
<tr>
<td>2001</td>
<td>30.1</td>
<td>456</td>
</tr>
<tr>
<td>2002</td>
<td>30.5</td>
<td>509</td>
</tr>
<tr>
<td>2003</td>
<td>30.6</td>
<td>497</td>
</tr>
<tr>
<td>2004</td>
<td>31.3</td>
<td>596</td>
</tr>
<tr>
<td>2005</td>
<td>31.7</td>
<td>573</td>
</tr>
<tr>
<td>2006</td>
<td>32.6</td>
<td>661</td>
</tr>
<tr>
<td>2007</td>
<td>33.1</td>
<td>741</td>
</tr>
<tr>
<td>2008</td>
<td>32.7</td>
<td>809</td>
</tr>
<tr>
<td>2009</td>
<td>32.8</td>
<td>717</td>
</tr>
</tbody>
</table>

Correlation: 0.947091

Source: Tyler Vigen (http://www.tylervigen.com)
Spurious Correlations 4

Number of films Nicolas Cage appeared in correlates with Number of people who drowned by falling into a swimming-pool.

**Table:**

<table>
<thead>
<tr>
<th>Year</th>
<th>Number people who drowned by falling into a swimming-pool (Deaths (US) (CDC))</th>
<th>Number of films Nicolas Cage appeared in (Films (IMDB))</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>109</td>
<td>2</td>
</tr>
<tr>
<td>2000</td>
<td>102</td>
<td>2</td>
</tr>
<tr>
<td>2001</td>
<td>102</td>
<td>2</td>
</tr>
<tr>
<td>2002</td>
<td>98</td>
<td>3</td>
</tr>
<tr>
<td>2003</td>
<td>85</td>
<td>1</td>
</tr>
<tr>
<td>2004</td>
<td>95</td>
<td>1</td>
</tr>
<tr>
<td>2005</td>
<td>96</td>
<td>2</td>
</tr>
<tr>
<td>2006</td>
<td>98</td>
<td>3</td>
</tr>
<tr>
<td>2007</td>
<td>123</td>
<td>4</td>
</tr>
<tr>
<td>2008</td>
<td>94</td>
<td>1</td>
</tr>
<tr>
<td>2009</td>
<td>102</td>
<td>4</td>
</tr>
</tbody>
</table>

Correlation: 0.666004

Source: Tyler Vigen (http://www.tylervigen.com)
A similar problem can occur with biomedical survival data:

- Automated image analysis software $d \sim O(100)$ or $d \sim O(1,000)$
- Gene expression data with $d \sim O(10,000)$.
- Single-nucleotide polymorphism (SNP) data $d \sim O(100,000)$
A similar problem can occur with biomedical survival data:

- Automated image analysis software \( d \sim \mathcal{O}(100) \) or \( d \sim \mathcal{O}(1,000) \)
- Gene expression data with \( d \sim \mathcal{O}(10,000) \).
- Single-nucleotide polymorphism (SNP) data \( d \sim \mathcal{O}(100,000) \)

Strategies for high dimensional survival data:

- Feature selection: univariate analysis, random forest...
- Regularised regression: \( L_1 \) penalised regression, elastic nets...
- Dimensionality reduction: Principal component analysis, latent variable models...

See Witten (2010) for a good review.
1 Dimensionality and overfitting

2 The GPLVM-WPHM

3 Simulation studies
The Gaussian process latent variable model (GPLVM)

The GPLVM (Lawrence, 2005) tries to represent high dimensional data in terms of a small number of latent variables.

- Non-parametric, flexible
- Non-linear
- Probabilistic (can do Bayesian model selection)
- Detect and extract low dimensional structure
- Combine multiple datasets
The Gaussian process latent variable model (GPLVM)

We extend the GPLVM in a number of ways:

- Incorporation of survival data by combining the GPLVM with a Weibull proportional hazards model (WPHM)
- Extract low dimensional structure that is relevant to survival outcomes
- Construct a Laplace approximation of marginal likelihood.
- Compare model likelihood for different values of $q$
- Combine multiple datasets
The Gaussian process latent variable model (GPLVM)

High dimensional data are written as a function of the latent variables:

\[ y_{i\mu}^s = f^s(x_i) + \text{noise}. \]

We assume a Gaussian process (GP) prior over the functions, and Gaussian noise. The data likelihood is

\[
p(\{Y_s\}|X, \{\beta_s^2\}, \theta) = \prod_{s=1}^{S} \prod_{\mu=1}^{d_s} \frac{e^{-\frac{1}{2}y_{:,\mu}^s \cdot K_s^{-1}y_{:,\mu}^s}}{(2\pi)^{\frac{N}{2}} |K_s|^{\frac{1}{2}}}. \]

This is a product of Gaussian processes that map the latent variables to each observed covariate.
High dimensional data are written as a function of the latent variables:

\[ y_{i\mu}^s = f^s(x_i) + \text{noise}. \]

We assume a Gaussian process (GP) prior over the functions, and Gaussian noise. The data likelihood is

\[
p(\{Y_s\} | X, \{\beta_s^2\}, \theta) = \prod_{s=1}^{S} \prod_{\mu=1}^{d_s} \frac{e^{-\frac{1}{2}y_{s,\mu}^s \cdot k_{s}^{-1} y_{s,\mu}^s}}{(2\pi)^{N/2} |K_s|^{1/2}}. \]

This is a product of Gaussian processes that map the latent variables to each observed covariate. A GP prior can be regarded as a prior over functions and is completely specified by it’s mean function, \( m(x_i) = \text{mean}(x_i) \) (zero in this case), and the kernel function \( k(x_i, x_j) = \text{cov}(x_i, x_j) \). The kernel functions considered in this paper are

- \( k(x_i, x_j) = \sigma x_i \cdot x_j \) linear,
- \( k(x_i, x_j) = \sigma (1 + x_i \cdot x_j)^2 \) polynomial (of second order),
- \( k(x_i, x_j) = \sigma \exp(-l(x_i - x_j)^2/2) \) squared exponential.
The Weibull Proportional Hazards Model (WPHM)

The hazard rate for individual $i$ is

$$h_i(t|x_i, \nu, \rho, b) = \lambda_0(t) e^{b \cdot x_i},$$

with

- Base hazard rate $\lambda_0(t) = (\nu/\rho)(t/\rho)^{\nu-1}$
- Scale parameter $\rho$
- Shape parameter $\nu$
- Regression coefficients $b \in \mathbb{R}^q$.

The data likelihood is

$$p(D|X, b, \rho, \nu) = \prod_{i=1}^N [\lambda_0(t_i) e^{b \cdot x_i}]^{\Delta_i} \exp(-\Lambda_0(t_i) e^{b \cdot x_i})$$

where the integrated base hazard rate as $\Lambda_0(t) = (t/\rho)^\nu$ and the survival data are $D = \{(t_1, \Delta_1), \ldots, (t_N, \delta_N)\}$. 

James Barrett (IMMB, KCL)
The combined model (GPLVM-WPHM)

Quantities we wish to infer:

- Latent variables $X$, WPHM parameters $b, \rho, \nu$.
- Hyperparameters $\{\beta^2_s\}, \theta$.
- Models (i.e. choice of $q$ and kernel function)
The combined model (GPLVM-WPHM)

Quantities we wish to infer:
- Latent variables $\mathbf{X}$, WPHM parameters $\mathbf{b}, \rho, \nu$.
- Hyperparameters $\{\beta_s^2\}, \theta$.
- Models (i.e. choice of $q$ and kernel function)

Using bayes’ theorem

$$
p(\mathbf{X}, \mathbf{b}, \rho, \nu|\{\mathbf{Y}_s\}, D, \theta, \{\beta_s\}) = \frac{\text{data likelihood}}{\text{priors}} = \frac{p(\{\mathbf{Y}_s\}, D|\mathbf{X}, \mathbf{b}, \rho, \nu, \theta, \{\beta_s\}) p(\mathbf{X})p(\mathbf{b})p(\rho)p(\nu)}{p(\{\mathbf{Y}_s\}, D|\theta, \{\beta_s\})}
$$

where

$$
p(\{\mathbf{Y}_s\}, D|\theta, \{\beta_s\}) = \int d\mathbf{X} \, d\mathbf{b} \, d\rho \, d\nu \, p(\{\mathbf{Y}_s\}, D|\mathbf{X}, \mathbf{b}, \rho, \nu, \theta, \{\beta_s\}) p(\mathbf{X})p(\mathbf{b})p(\rho)p(\nu).
$$
The combined model (GPLVM-WPHM)

Quantities we wish to infer:
- Latent variables $\mathbf{X}$, WPHM parameters $\mathbf{b}, \rho, \nu$.
- Hyperparameters $\{\beta_s^2\}, \theta$.
- Models (i.e. choice of $q$ and kernel function)

Using bayes’ theorem

\[
p(\mathbf{X}, \mathbf{b}, \rho, \nu|\{\mathbf{Y}_s\}, D, \theta, \{\beta_s\}) = \frac{p(\{\mathbf{Y}_s\}, D|\mathbf{X}, \mathbf{b}, \rho, \nu, \theta, \{\beta_s\})p(\mathbf{X})p(\mathbf{b})p(\rho)p(\nu)}{p(\{\mathbf{Y}_s\}, D|\theta, \{\beta_s\})}
\]

where

\[
p(\{\mathbf{Y}_s\}, D|\theta, \{\beta_s\}) = \int d\mathbf{X} d\mathbf{b} d\rho d\nu p(\{\mathbf{Y}_s\}, D|\mathbf{X}, \mathbf{b}, \rho, \nu, \theta, \{\beta_s\})p(\mathbf{X})p(\mathbf{b})p(\rho)p(\nu).
\]

We now make a key assumption of conditional independence between the observed covariates and the survival data given the latent variables:

\[
p(\{\mathbf{Y}_s\}, D|\mathbf{X}, \mathbf{b}, \rho, \nu, \theta, \{\beta_s\}) = p(\{\mathbf{Y}_s\}|\mathbf{X}, \theta, \{\beta_s\}) p(D|\mathbf{X}, \mathbf{b}, \rho, \nu, \theta).
\]
Inference of parameters and hyperparameters

Numerically maximise \( p(X, b, \rho, \nu|\{Y_s\}, D, \theta, \{\beta_s\}) \) with respect to \( X \) and \( b, \rho, \nu \).

Laplace approximation of the marginal likelihood:

\[
\int d\{\beta_s\} \int d\theta \int dX \int db \int d\rho \int d\nu p(X, b, \rho, \nu|\{Y_s\}, D, \theta, \{\beta_s\}) p(X) p(b) p(\rho) p(\nu) \]

approximate with a Gaussian.
Inference of parameters and hyperparameters

Numerically maximise $p(\mathbf{X}, \mathbf{b}, \rho, \nu|\{\mathbf{Y}_s\}, D, \theta, \{\beta_s\})$ with respect to $\mathbf{X}$ and $\mathbf{b}, \rho, \nu$. The posterior over hyperparameters is given by

$$p(\{\beta_s^2\}, \theta|\{\mathbf{Y}_s\}, D) = \frac{p(\{\mathbf{Y}_s\}, D|\{\beta_s^2\}, \theta)p(\{\beta_s^2\})p(\theta)}{\int d\{\beta_s^2\}d\theta' p(\{\mathbf{Y}_s\}, D|\{\beta_s^2\}, \theta')p(\{\beta_s^2\})p(\theta')}$$

Laplace approximation of the the marginal likelihood:

$$p(\{\mathbf{Y}_s\}, D|\theta, \{\beta_s\}) = \int d\mathbf{X} \, d\mathbf{b} \, d\rho \, d\nu \, p(\{\mathbf{Y}_s\}, D|\mathbf{X}, \mathbf{b}, \rho, \nu, \theta, \{\beta_s\})p(\mathbf{X})p(\mathbf{b})p(\rho)p(\nu).$$

approximate with a Gaussian
Predictions for new individuals

If we observe a new patient with $y^*$ we predict the corresponding event time $t^*$ via

$$y^* \xrightarrow{\text{GPLVM}} x^* \xrightarrow{\text{WPHM}} t^*.$$

Maximise the GP predictive distribution w.r.t. $x^*$:

$$p(y^*|x^*) = \mathcal{N}(m, \kappa^{-2} I)$$

with

$$m_\mu = k \cdot K^{-1} y_\mu$$
$$\kappa^2 = k(x^*, x^*) - k \cdot K^{-1} k + \beta^2.$$  

Numerically computing the mean of the event time density corresponding to $x^*$:

$$t^* = \langle t \rangle = \int_0^\infty ds \, s \lambda_0(s) e^{\hat{b} \cdot x^*} \exp(-\Lambda_0(s) e^{\hat{b} \cdot x^*}).$$
Dimensionality and overfitting

The GPLVM-WPHM

Simulation studies
Generation of synthetic data

- Choose specific $X$ or generate randomly
- Choose values for hyperparameters
- Generate $Y$ by drawing covariates from a GP prior.
- Generate survival times from the inverse cumulative distribution $C_i(t) = 1 - \exp(-\Lambda_0(t)e^{b \cdot x_i})$. That is, generate $z \in [0, 1]$ from a uniform density and then

$$t_i = \rho \left( -e^{-b \cdot x_i} \log(1 - z) \right)^{1/\nu}.$$
Retrieval accuracy

Figure: High dimensional data with $d = 10$ and a linear kernel. Gaussian noise with variance $\beta^2 = 0.1$ is added.

We can define some ad hoc quantitative measures of retrieval accuracy

$$\varepsilon_{\text{radial}} = 0.0051, \ \varepsilon_{\text{angular}} = 0.0086 \ \text{and} \ \varepsilon_{\text{linear}} = 0.0288.$$
Integration of datasets and survival outcomes

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$\mathcal{E}_{\text{radial}}$</th>
<th>$\mathcal{E}_{\text{angular}}$</th>
<th>$\mathcal{E}_{\text{linear}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>$-7.3%$</td>
<td>$-5.3%$</td>
<td>$-6.1%$</td>
</tr>
<tr>
<td>0.5</td>
<td>$-14.5%$</td>
<td>$-17.1%$</td>
<td>$-19.5%$</td>
</tr>
<tr>
<td>1.0</td>
<td>$-16.0%$</td>
<td>$-24.8%$</td>
<td>$-15.7%$</td>
</tr>
</tbody>
</table>

**Table:** Decrease in misalignment error using the GPLVM-WPHM compared to the GPLVM.
Integration of datasets and survival outcomes

Table: Decrease in misalignment error using the GPLVM-WPHM compared to the GPLVM.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$\mathcal{E}_{\text{radial}}$</th>
<th>$\mathcal{E}_{\text{angular}}$</th>
<th>$\mathcal{E}_{\text{linear}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>$-7.3%$</td>
<td>$-5.3%$</td>
<td>$-6.1%$</td>
</tr>
<tr>
<td>0.5</td>
<td>$-14.5%$</td>
<td>$-17.1%$</td>
<td>$-19.5%$</td>
</tr>
<tr>
<td>1.0</td>
<td>$-16.0%$</td>
<td>$-24.8%$</td>
<td>$-15.7%$</td>
</tr>
</tbody>
</table>

Table: Decrease in misalignment errors when two datasets are combined.

<table>
<thead>
<tr>
<th>$\mathbf{Y}_1$ ($d = 10, \beta_1^2 = 0.1$)</th>
<th>$\mathcal{E}_{\text{radial}}$</th>
<th>$\mathcal{E}_{\text{angular}}$</th>
<th>$\mathcal{E}_{\text{linear}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathbf{Y}_2$ ($d = 100, \beta_2^2 = 1.0$)</td>
<td>0.0071</td>
<td>0.0093</td>
<td>0.0270</td>
</tr>
<tr>
<td>$\mathbf{Y}_2 &amp; \mathbf{Y}_2$</td>
<td>0.0244</td>
<td>0.0148</td>
<td>0.0509</td>
</tr>
<tr>
<td></td>
<td>0.0046</td>
<td>0.0052</td>
<td>0.0146</td>
</tr>
</tbody>
</table>
Overfitting I

- Split dataset into training and test set.
- Compute the mean square error (MSE) between predicted and reported event times in test data.
- Compare MSE in $X$ and $Y$ spaces.

<table>
<thead>
<tr>
<th>$d$</th>
<th>$d = 10$</th>
<th>$d = 25$</th>
<th>$d = 50$</th>
<th>$d = 100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>+1.2%</td>
<td>+14.7%</td>
<td>+26.6%</td>
<td>+43.4%</td>
</tr>
</tbody>
</table>

**Table:** Percentage change in the mean squared error between $X$ and $Y$ spaces. (Linear kernel, $\beta^2 = 0.01$, $N = 200$, $q = 2$.)

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>$\beta = 0.01$</th>
<th>$\beta = 0.1$</th>
<th>$\beta = 0.5$</th>
<th>$\beta = 1.0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE</td>
<td>+1.2%</td>
<td>+2.7%</td>
<td>+38.0%</td>
<td>+5.67</td>
</tr>
</tbody>
</table>

**Table:** Percentage change in the mean squared error between $X$ and $Y$ spaces. (Linear kernel, $d = 10$, $N = 200$, $q = 2$.)
Overfitting II

![Kaplan-Meier survival curves](image)

**Figure:** Kaplan-Meier survival curves obtained in the latent variable space \( q = 1 \) (left) and observed data space with \( d = 2 \) (right). A squared exponential kernel was used.

Generating hyperparameters: \((\beta^2, \sigma, l, b, \rho, \nu) = (0.0010, 1.00, 1.00, -1.00, 10.0, 10.0)\)

Inferred hyperparameters: \((\beta^2, \sigma, l, b, \rho, \nu) = (0.0006, 1.23, 1.11, -0.68, 9.70, 10.3)\)
Figure: The minimum log marginal likelihood versus q. Both models correctly identify $q^* = 2$ with a linear kernel as the best model.
Conclusion

- Detect and extract intrinsic low dimensional structure from high dimensional data.
- Extract non-linear low dimensional structure.
- Combine multiple data sources with survival information.
- Increase predictive accuracy of predictions for new individuals.
- Diminish the effects of overfitting.

References


Acknowledgments

Thanks to my supervisor Ton Coolen, King’s College London.

This work was funded by the European Union FP7 Imagint Project, Grant 259881.