Introduction to Gaussian Processes

Neil D. Lawrence

MLSS, Cadiz
12th May 2016
Outline

Gaussian Processes

GP Non-Gaussian

GP Limitations

Kalman Filter

Dimensionality Reduction
Outline

Gaussian Processes

GP Non-Gaussian

GP Limitations

Kalman Filter

Dimensionality Reduction
Rasmussen and Williams (2006)
\[ y = mx + c \]
$y = mx + c$
\[ y = mx + c \]
\[ y = mx + c \]
$y = mx + c$
\[ y = mx + c \]
\( y = mx + c \)
$y = mx + c$
$y = mx + c$

point 1: $x = 1, y = 3$

$$3 = m + c$$

point 2: $x = 3, y = 1$

$$1 = 3m + c$$

point 3: $x = 2, y = 2.5$

$$2.5 = 2m + c$$
SUR LES PROBABILITÉS.

riens. L'opinion contraire est une illusion de l'esprit qui, perdant de vue les raisons fugitives du choix de la volonté dans les choses indifférentes, se persuade qu'elle s'est déterminée d'elle-même et sans motifs.

Nous devons donc envisager l'état présent de l'univers, comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui, pour un instant donné, connaîtrait toutes les forces dont la nature est animée, et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'univers et ceux du plus léger atome : rien ne serait incertain pour elle, et l'avenir comme le passé, serait présent à ses yeux. L'esprit humain offre, dans la perfection qu'il a su donner à l'Astronomie, une faible esquisse de cette intelligence. Ses découvertes en Mécanique et en Géométrie, jointes à celle de la pesanteur universelle, l'ont mis à portée de comprendre dans les mêmes expressions analytiques, les états passés et futurs du système du monde. En appliquant la même méthode à quelques autres objets de ses connaissances, il est parvenu à ramener à des lois générales, les phénomènes observés, et à prévoir ceux que des circonstances données doivent faire éclore. Tous ces efforts dans la recherche de la vérité, tendent à le rapprocher sans cesse de l'intelligence que nous venons de concevoir, mais dont il restera toujours infiniment éloigné. Cette tendance propre à l'espèce humaine, est ce qui la rend supérieure aux animaux ; et ses progrès en ce genre, distinguent les nations et les siècles, et font leur véritable gloire.

Rappelons-nous qu'autrefois, et à une époque qui
other, we say that its choice is an effect without a cause. It is then, says Leibnitz, the blind chance of the Epicureans. The contrary opinion is an illusion of the mind, which, losing sight of the evasive reasons of the choice of the will in indifferent things, believes that choice is determined of itself and without motives.

We ought then to regard the present state of the universe as the effect of its anterior state and as the cause of the one which is to follow. Given for one instant an intelligence which could comprehend all the forces by which nature is animated and the respective situation of the beings who compose it—an intelligence sufficiently vast to submit these data to analysis—it would embrace in the same formula the movements of the greatest bodies of the universe and those of the lightest atom; for it, nothing would be uncertain and the future, as the past, would be present to its eyes. The human mind offers, in the perfection which it has been able to give to astronomy, a feeble idea of this intelligence. Its discoveries in mechanics and geometry, added to that of universal gravity, have enabled it to comprehend in the same analytical expressions the past and future states of the system of the world. Applying the same method to some other objects of its knowledge, it has succeeded in referring to general laws observed phenomena and in foreseeing those which given circumstances ought to produce. All these efforts in the search for truth tend to lead it back continually to the vast intelligence which we have just mentioned, but from which it will always remain infinitely removed. This tendency, peculiar to the human race, is that which renders it superior to animals; and their progress
The day will come when, by study pursued through several ages, the things now concealed will appear with evidence; and posterity will be astonished that truths so clear had escaped us.' Clairaut then undertook to submit to analysis the perturbations which the comet had experienced by the action of the two great planets, Jupiter and Saturn; after immense calculations he fixed its next passage at the perihelion toward the beginning of April, 1759, which was actually verified by observation. The regularity which astronomy shows us in the movements of the comets doubtless exists also in all phenomena.

The curve described by a simple molecule of air or vapor is regulated in a manner just as certain as the planetary orbits; the only difference between them is that which comes from our ignorance.

Probability is relative, in part to this ignorance, in part to our knowledge. We know that of three or a greater number of events a single one ought to occur; but nothing induces us to believe that one of them will occur rather than the others. In this state of indecision it is impossible for us to announce their occurrence with certainty. It is, however, probable that one of these events, chosen at will, will not occur because we see several cases equally possible which exclude its occurrence, while only a single one favors it.

The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of
\[ y = mx + c + \epsilon \]

**point 1:** \( x = 1, \ y = 3 \)
\[ 3 = m + c + \epsilon_1 \]

**point 2:** \( x = 3, \ y = 1 \)
\[ 1 = 3m + c + \epsilon_2 \]

**point 3:** \( x = 2, \ y = 2.5 \)
\[ 2.5 = 2m + c + \epsilon_3 \]
What about two unknowns and *one* observation?

\[ y_1 = mx_1 + c \]
Can compute $m$ given $c$.

$$m = \frac{y_1 - c}{x}$$
Can compute $m$ given $c$.

$c = 1.75 \implies m = 1.25$
Can compute $m$ given $c$.

$c = -0.777 \implies m = 3.78$
Can compute $m$ given $c$.

$c = -4.01 \implies m = 7.01$
Can compute $m$ given $c$.

$$c = -0.718 \implies m = 3.72$$
Can compute $m$ given $c$.

$c = 2.45 \implies m = 0.545$
Can compute $m$ given $c$.

\[ c = -0.657 \implies m = 3.66 \]
Can compute $m$ given $c$. 

$c = -3.13 \implies m = 6.13$
Underdetermined System

Can compute $m$ given $c$.

\[ c = -1.47 \implies m = 4.47 \]
Can compute $m$ given $c$. Assume

$$c \sim \mathcal{N}(0, 4),$$

we find a distribution of solutions.
Gaussian Process

\[ y_i(x_i) = f(x_i) + \epsilon_i \]

- Place a prior over the process as well as the noise.
- Leads to models that are not i.i.d.
- Contrast with classical model’s objective function:

\[
\sum_{i=1}^{n} \left( 1 - y_i(w^\top x_i - b) \right)_+ + \lambda w^\top w
\]
I’m keen on the idea of a conceptual separation model and algorithm.

Model is how you encode the regularities of the universe.

Algorithm is how you combine that model with data.

data + model → prediction

Of course often we are restricted in modeling choice due to lack of algorithms.
Gaussian Processes: Extremely Short Overview
Gaussian Processes: Extremely Short Overview
Gaussian Processes: Extremely Short Overview
Gaussian Processes: Extremely Short Overview
Multi-variate Gaussians

- We will consider a Gaussian with a particular structure of covariance matrix.
- Generate a single sample from this 25 dimensional Gaussian distribution, $f = [f_1, f_2 \ldots f_{25}]$.
- We will plot these points against their index.
Gaussian Distribution Sample

(a) A 25 dimensional correlated random variable (values plotted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.
Figure: A sample from a 25 dimensional Gaussian distribution.
Gaussian Distribution Sample

(a) A 25 dimensional correlated random variable (values plotted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.
Gaussian Distribution Sample

(a) A 25 dimensional correlated random variable (values plotted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.
Gaussian Distribution Sample

(a) A 25 dimensional correlated random variable (values plotted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.
Gaussian Distribution Sample

(a) A 25 dimensional correlated random variable (values plotted against index)

(b) colormap showing correlations between dimensions.

Figure: A sample from a 25 dimensional Gaussian distribution.
Figure: A sample from a 25 dimensional Gaussian distribution.
A 25 dimensional correlated random variable (values plotted against index)

(b) correlation between $f_1$ and $f_2$.

Figure: A sample from a 25 dimensional Gaussian distribution.
 Prediction of $f_2$ from $f_1$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
Prediction of $f_2$ from $f_1$

The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.

We observe that $f_1 = -0.313$. 
Prediction of $f_2$ from $f_1$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_2 | f_1 = -0.313)$.
The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_2)$.

We observe that $f_1 = -0.313$.

Conditional density: $p(f_2 | f_1 = -0.313)$. 
Prediction of $f_2$ from $f_1$ requires conditional density.

Conditional density is also Gaussian.

$$p(f_2|f_1) = \mathcal{N}\left(f_2|\frac{k_{1,2}}{k_{1,1}}f_1, k_{2,2} - \frac{k_{1,2}^2}{k_{1,1}}\right)$$

where covariance of joint density is given by

$$K = \begin{bmatrix} k_{1,1} & k_{1,2} \\ k_{2,1} & k_{2,2} \end{bmatrix}$$
Prediction of $f_5$ from $f_1$

The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
The single contour of the Gaussian density represents the joint distribution, \( p(f_1, f_5) \).

We observe that \( f_1 = -0.313 \).
The single contour of the Gaussian density represents the joint distribution, \( p(f_1, f_5) \).

We observe that \( f_1 = -0.313 \).

Conditional density: \( p(f_5|f_1 = -0.313) \).
Prediction of $f_5$ from $f_1$

- The single contour of the Gaussian density represents the joint distribution, $p(f_1, f_5)$.
- We observe that $f_1 = -0.313$.
- Conditional density: $p(f_5|f_1 = -0.313)$. 

$$
\begin{bmatrix}
1 & 0.57375 \\
0.57375 & 1
\end{bmatrix}
$$
Prediction with Correlated Gaussians

- Prediction of $f_*$ from $f$ requires multivariate conditional density.
- Multivariate conditional density is also Gaussian.

$$p(f_*|f) = \mathcal{N}(f_*|K_{*,f}K_{f,f}^{-1}f, K_{*,*} - K_{*,f}K_{f,f}^{-1}K_{f,*})$$

- Here covariance of joint density is given by

$$K = \begin{bmatrix} K_{f,f} & K_{*,f} \\ K_{f,*} & K_{*,*} \end{bmatrix}$$
Prediction with Correlated Gaussians

- Prediction of \( f_* \) from \( f \) requires multivariate conditional density.
- Multivariate conditional density is also Gaussian.

\[
p(f_*|f) = \mathcal{N}(f_*|\mu, \Sigma)
\]

\[
\mu = K_{*,f}K_{f,f}^{-1}f
\]

\[
\Sigma = K_{*,*} - K_{*,f}K_{f,f}^{-1}K_{f,*}
\]

- Here covariance of joint density is given by

\[
K = \begin{bmatrix}
K_{f,f} & K_{*,f} \\
K_{f,*} & K_{*,*}
\end{bmatrix}
\]
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

\[ k(x, x') = \alpha \exp \left( -\frac{||x - x'||^2}{2\ell^2} \right) \]

- Covariance matrix is built using the inputs to the function \( x \).
- For the example above it was based on Euclidean distance.
- The covariance function is also known as a kernel.
Covariance Functions

Where did this covariance matrix come from?

**Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)**

\[ k(x, x') = \alpha \exp \left( -\frac{\|x - x'\|^2}{2\ell^2} \right) \]

- Covariance matrix is built using the *inputs* to the function \( x \).
- For the example above it was based on Euclidean distance.
- The covariance function is also known as a kernel.
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right) \]

\[ x_1 = -3.0, \quad x_1 = -3.0 \]

\[ k_{1,1} = 1.00 \times \exp\left(-\frac{(-3.0-(-3.0))^2}{2 \times 2.00^2}\right) \]

\[ x_1 = -3.0, \quad x_2 = 1.20, \quad \text{and} \quad x_3 = 1.40 \quad \text{with} \quad \ell = 2.00 \quad \text{and} \quad \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[ k_{1,1} = 1.00 \times \exp\left(-\frac{(-3.0-(-3.0))^2}{2\times2.00^2}\right) \]

\[ x_1 = -3.0, x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 2.00 \text{ and } \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp \left( -\frac{||x_i-x_j||^2}{2\ell^2} \right) \]

1.00

\[
\begin{bmatrix}
1.00 \\
\end{bmatrix}
\]

\[ x_2 = 1.20, \ x_1 = -3.0 \]

\[ k_{2,1} = 1.00 \times \exp \left( -\frac{(1.20-(-3.0))^2}{2 \times 2.00^2} \right) \]

\[
\begin{bmatrix}
\end{bmatrix}
\]

\[ x_1 = -3.0, \ x_2 = 1.20, \ \text{and} \ x_3 = 1.40 \ \text{with} \ \ell = 2.00 \ \text{and} \ \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp \left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)
\]

\[
x_2 = 1.20, \ x_1 = -3.0
\]

\[
k_{2,1} = 1.00 \times \exp \left(-\frac{(1.20 - -3.0)^2}{2 \times 2.00^2}\right)
\]

\[
x_1 = -3.0, \ x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 2.00 \text{ and } \alpha = 1.00.
\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.00 & 0.110 \\
0.110 & 1.00
\end{bmatrix}
\]

\[ k_{2,1} = 1.00 \times \exp\left(-\frac{(1.20 - -3.0)^2}{2 \times 2.00^2}\right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 2.00 \text{ and } \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right)
\]

\[
x_2 = 1.20, \quad x_2 = 1.20
\]

\[
k_{2,2} = 1.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2 \times 2.00^2}\right)
\]

\[
x_1 = -3.0, \quad x_2 = 1.20, \quad \text{and} \quad x_3 = 1.40 \quad \text{with} \quad \ell = 2.00 \quad \text{and} \quad \alpha = 1.00.
\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left( -\frac{||x_i - x_j||^2}{2\ell^2} \right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \text{ and } x_3 = 1.40 \] with \( \ell = 2.00 \) and \( \alpha = 1.00 \).
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right)
\]

\[
\begin{bmatrix}
1.00 & 0.110 \\
0.110 & 1.00
\end{bmatrix}
\]

\[x_3 = 1.40, x_1 = -3.0\]

\[k_{3,1} = 1.00 \times \exp\left(-\frac{(1.40-(-3.0))^2}{2 \times 2.00^2}\right)\]

\[x_1 = -3.0, x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 2.00 \text{ and } \alpha = 1.00.\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[ k_{3,1} = 1.00 \times \exp\left(-\frac{(1.40-3.0)^2}{2 \times 2.00^2}\right) \]

\[ \begin{bmatrix} 1.00 & 0.110 \\ 0.110 & 1.00 \end{bmatrix} \]

\[ 0.0889 \]

\[ x_1 = -3.0, \ x_2 = 1.20, \ \text{and} \ x_3 = 1.40 \ \text{with} \ \ell = 2.00 \ \text{and} \ \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right) \]

\[ \begin{bmatrix}
  1.00 & 0.110 & 0.0889 \\
  0.110 & 1.00 \\
  0.0889 \\
\end{bmatrix} \]

\[ x_3 = 1.40, \quad x_1 = -3.0 \]

\[ k_{3,1} = 1.00 \times \exp\left( -\frac{(1.40 - -3.0)^2}{2 \times 2.00^2} \right) \]

\[ x_1 = -3.0, \quad x_2 = 1.20, \quad \text{and} \quad x_3 = 1.40 \quad \text{with} \quad \ell = 2.00 \quad \text{and} \quad \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.00 & 0.110 & 0.0889 \\
0.110 & 1.00 \\
0.0889 & 1.00
\end{bmatrix}
\]

\[ x_3 = 1.40, \ x_2 = 1.20 \]

\[ k_{3,2} = 1.00 \times \exp\left(-\frac{(1.40-1.20)^2}{2\times2.00^2}\right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 2.00 \text{ and } \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp \left( -\frac{||x_i - x_j||^2}{2\ell^2} \right) \]

\[ x_3 = 1.40, \ x_2 = 1.20 \]

\[ k_{3,2} = 1.00 \times \exp \left( -\frac{(1.40 - 1.20)^2}{2 \times 2.00^2} \right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \ \text{and} \ x_3 = 1.40 \ \text{with} \ \ell = 2.00 \ \text{and} \ \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.00 & 0.110 & 0.0889 \\
0.110 & 1.00 & 0.995 \\
0.0889 & 0.995 & \end{bmatrix}
\]

\[ x_3 = 1.40, \ x_2 = 1.20 \]

\[ k_{3,2} = 1.00 \times \exp\left(-\frac{(1.40-1.20)^2}{2\times2.00^2}\right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \ \text{and} \ x_3 = 1.40 \ \text{with} \ \ell = 2.00 \ \text{and} \ \alpha = 1.00. \]
**Covariance Functions**

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.00 & 0.110 & 0.0889 \\
0.110 & 1.00 & 0.995 \\
0.0889 & 0.995 & 1.00
\end{bmatrix}
\]

\[ x_3 = 1.40, x_3 = 1.40 \]

\[ k_{3,3} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times2.00^2}\right) \]

\[ k_{3,3} = 0.0889 \times 0.995 \]

\[ x_1 = -3.0, x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 2.00 \text{ and } \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.00 & 0.110 & 0.0889 \\
0.110 & 1.00 & 0.995 \\
0.0889 & 0.995 & 1.00
\end{bmatrix}
\]

\[ x_3 = 1.40, \quad x_3 = 1.40 \]

\[ k_{3,3} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times2.00^2}\right) \]

\[ x_1 = -3.0, \quad x_2 = 1.20, \quad \text{and} \quad x_3 = 1.40 \quad \text{with} \quad \ell = 2.00 \quad \text{and} \quad \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{|x_i-x_j|^2}{2\ell^2}\right) \]

\[ x_3 = 1.40, \quad x_3 = 1.40 \]

\[ k_{3,3} = 1.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2 \times 2.00^2}\right) \]

\[ x_1 = -3.0, \quad x_2 = 1.20, \quad \text{and} \quad x_3 = 1.40 \quad \text{with} \quad \ell = 2.00 \quad \text{and} \quad \alpha = 1.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[ x_1 = -3, x_1 = -3 \]

\[ k_{1,1} = 1.0 \times \exp\left(-\frac{(-3+3)^2}{2\times2.0^2}\right) \]

\[ x_1 = -3, x_2 = 1.2, x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left( -\frac{||x_i - x_j||^2}{2\ell^2} \right) \]

\[ x_1 = -3, \quad x_1 = -3 \]

\[ k_{1,1} = 1.0 \times \exp\left( -\frac{(-3-(-3))^2}{2 \times 2.0^2} \right) \]

\[ x_1 = -3, \quad x_2 = 1.2, \quad x_3 = 1.4, \quad \text{and} \quad x_4 = 2.0 \quad \text{with} \quad \ell = 2.0 \quad \text{and} \quad \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[ x_2 = 1.2, \ x_1 = -3 \]

\[ k_{2,1} = 1.0 \times \exp\left(-\frac{(1.2-(-3))^2}{2 \times 2.0^2}\right) \]

\[ x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[ x_2 = 1.2, \ x_1 = -3 \]

\[ k_{2,1} = 1.0 \times \exp\left(-\frac{(1.2-(-3))^2}{2\times2.0^2}\right) \]

\[ x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp \left( -\frac{||x_i - x_j||^2}{2\ell^2} \right) \]

\[
\begin{bmatrix}
1.0 & 0.11 \\
0.11 & 1.0
\end{bmatrix}
\]

\[
k_{2,1} = 1.0 \times \exp \left( -\frac{(1.2 - -3)^2}{2 \times 2.0^2} \right)
\]

\[
x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \ \text{and} \ x_4 = 2.0 \ \text{with} \ \ell = 2.0 \ \text{and} \ \alpha = 1.0.
\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp \left( -\frac{||x_i-x_j||^2}{2\ell^2} \right) \]

\[
\begin{bmatrix}
1.0 & 0.11 \\
0.11 & \cdot \cdot \cdot
\end{bmatrix}
\]

\[ x_2 = 1.2, \quad x_2 = 1.2 \]

\[ k_{2,2} = 1.0 \times \exp \left( -\frac{(1.2-1.2)^2}{2\times2.0^2} \right) \]

\[ x_1 = -3, \quad x_2 = 1.2, \quad x_3 = 1.4, \quad \text{and} \quad x_4 = 2.0 \quad \text{with} \quad \ell = 2.0 \quad \text{and} \quad \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.0 & 0.11 \\
0.11 & 1.0
\end{bmatrix}
\]

\[ x_2 = 1.2, \ x_2 = 1.2 \]

\[ k_{2,2} = 1.0 \times \exp\left(-\frac{(1.2-1.2)^2}{2 \times 2.0^2}\right) \]

\[ x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left( -\frac{||x_i - x_j||^2}{2\ell^2} \right) \]

\[ \begin{bmatrix} 1.0 & 0.11 \\ 0.11 & 1.0 \end{bmatrix} \]

\[ k_{3,1} = 1.0 \times \exp\left( -\frac{(1.4 - -3)^2}{2 \times 2.0^2} \right) \]

\[ x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \ \text{and} \ x_4 = 2.0 \ \text{with} \ \ell = 2.0 \ \text{and} \ \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.0 & 0.11 \\
0.11 & 1.0 \\
0.089 & \\
\end{bmatrix}
\]

\[ x_3 = 1.4, x_1 = -3 \]

\[ k_{3,1} = 1.0 \times \exp\left(-\frac{(1.4-(-3))^2}{2\times2.0^2}\right) \]

\[ x_1 = -3, x_2 = 1.2, x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions
Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 \\
0.11 & 1.0 & \\
0.089 & & \\
\end{bmatrix}
\]

\[
x_3 = 1.4, x_1 = -3
\]

\[ k_{3,1} = 1.0 \times \exp\left(-\frac{(1.4-(-3))^2}{2 \times 2.0^2}\right) \]

\[
x_1 = -3, x_2 = 1.2, x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0.\]
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)
\]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 \\
0.11 & 1.0 & \\
0.089 & & \\
\end{bmatrix}
\]

\[
x_3 = 1.4, x_2 = 1.2
\]

\[
k_{3,2} = 1.0 \times \exp\left(-\frac{(1.4-1.2)^2}{2\times2.0^2}\right)
\]

\[
x_1 = -3, x_2 = 1.2, x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0.
\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
x_3 = 1.4, \ x_2 = 1.2
\]

\[ k_{3,2} = 1.0 \times \exp\left(-\frac{(1.4-1.2)^2}{2 \times 2.0^2}\right) \]

\[
x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0.
\]
Covariance Functions
Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right)
\]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 \\
0.11 & 1.0 & 1.0 \\
0.089 & 1.0
\end{bmatrix}
\]

\[
x_3 = 1.4, \ x_2 = 1.2
\]

\[
k_{3,2} = 1.0 \times \exp\left(-\frac{(1.4-1.2)^2}{2\times2.0^2}\right)
\]

\[
x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0.
\]
Covariance Functions
Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 \\
0.11 & 1.0 & 1.0 \\
0.089 & 1.0 & 1.0
\end{bmatrix}
\]

\[ x_3 = 1.4, \ x_3 = 1.4 \]

\[ k_{3,3} = 1.0 \times \exp\left(-\frac{(1.4-1.4)^2}{2\times2.0^2}\right) \]

\[ x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 \\
0.11 & 1.0 & 1.0 \\
0.089 & 1.0 & 1.0
\end{bmatrix}
\]

\[ x_3 = 1.4, \quad x_3 = 1.4 \]

\[ k_{3,3} = 1.0 \times \exp\left(-\frac{(1.4-1.4)^2}{2\times2.0^2}\right) \]

\[ x_1 = -3, \quad x_2 = 1.2, \quad x_3 = 1.4, \quad \text{and} \quad x_4 = 2.0 \quad \text{with} \quad \ell = 2.0 \quad \text{and} \quad \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{pmatrix}
1.0 & 0.11 & 0.089 \\
0.11 & 1.0 & 1.0 \\
0.089 & 1.0 & 1.0 \\
\end{pmatrix}
\]

\[ k_{4,1} = 1.0 \times \exp\left(-\frac{(2.0-3)^2}{2 \times 2.0^2}\right) \]

\[ x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \ and \ x_4 = 2.0 \ with \ \ell = 2.0 \ and \ \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 \\
0.11 & 1.0 & 1.0 \\
0.089 & 1.0 & 1.0 \\
0.044 & \end{bmatrix}
\]

\[ x_4 = 2.0, \ x_1 = -3 \]

\[ k_{4,1} = 1.0 \times \exp\left(-\frac{(2.0 - -3)^2}{2 \times 2.0^2}\right) \]

\[ x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right)
\]

\[
\begin{pmatrix}
1.0 & 0.11 & 0.089 & 0.044 \\
0.11 & 1.0 & 1.0 & \\
0.089 & 1.0 & 1.0 & \\
0.044 & \\
\end{pmatrix}
\]

\[
x_4 = 2.0, \ x_1 = -3
\]

\[
k_{4,1} = 1.0 \times \exp\left(-\frac{(2.0-(-3))^2}{2\times2.0^2}\right)
\]

\[
x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0.
\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 & 0.044 \\
0.11 & 1.0 & 1.0 & \\
0.089 & 1.0 & 1.0 & \\
0.044 & \\
\end{bmatrix}
\]

\[ x_4 = 2.0, \, x_2 = 1.2 \]

\[ k_{4,2} = 1.0 \times \exp\left(-\frac{(2.0-1.2)^2}{2\times2.0^2}\right) \]

\[ x_1 = -3, \, x_2 = 1.2, \, x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 & 0.044 \\
0.11 & 1.0 & 1.0 & \\
0.089 & 1.0 & 1.0 & \\
0.044 & 0.92 & \\
\end{bmatrix}
\]

\[ k_{4,2} = 1.0 \times \exp\left(-\frac{(2.0-1.2)^2}{2\times2.0^2}\right) \]

\[ x_4 = 2.0, x_2 = 1.2 \]

\[ x_1 = -3, x_2 = 1.2, x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right)
\]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 & 0.044 \\
0.11 & 1.0 & 1.0 & 0.92 \\
0.089 & 1.0 & 1.0 \\
0.044 & 0.92
\end{bmatrix}
\]

\[
k_{4,2} = 1.0 \times \exp\left(-\frac{(2.0-1.2)^2}{2 \times 2.0^2}\right)
\]

\[
x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0.
\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp \left( -\frac{||x_i-x_j||^2}{2\ell^2} \right) \]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 & 0.044 \\
0.11 & 1.0 & 1.0 & 0.92 \\
0.089 & 1.0 & 1.0 & \\
0.044 & 0.92 & \\
\end{bmatrix}
\]

\[ x_4 = 2.0, \ x_3 = 1.4 \]

\[ k_{4,3} = 1.0 \times \exp \left( -\frac{(2.0-1.4)^2}{2 \times 2.0^2} \right) \]

\[ x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right)
\]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 & 0.044 \\
0.11 & 1.0 & 1.0 & 0.92 \\
0.089 & 1.0 & 1.0 \\
0.044 & 0.92 & 0.96 \\
\end{bmatrix}
\]

\[x_1 = -3, \ x_2 = 1.2, \ x_3 = 1.4, \ \text{and} \ x_4 = 2.0 \ \text{with} \ \ell = 2.0 \ \text{and} \ \alpha = 1.0.\]
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right)
\]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 & 0.044 \\
0.11 & 1.0 & 1.0 & 0.92 \\
0.089 & 1.0 & 1.0 & 0.96 \\
0.044 & 0.92 & 0.96 \\
\end{bmatrix}
\]

\[
x_4 = 2.0, x_3 = 1.4\]

\[
k_{4,3} = 1.0 \times \exp\left(-\frac{(2.0-1.4)^2}{2\times2.0^2}\right)
\]

\[
x_1 = -3, x_2 = 1.2, x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0.
\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
k_{4,4} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2 \times 2.0^2}\right)
\]

\[
x_1 = -3, x_2 = 1.2, x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0.
\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp \left( -\frac{||x_i - x_j||^2}{2\ell^2} \right) \]

\[
\begin{bmatrix}
1.0 & 0.11 & 0.089 & 0.044 \\
0.11 & 1.0 & 1.0 & 0.92 \\
0.089 & 1.0 & 1.0 & 0.96 \\
0.044 & 0.92 & 0.96 & 1.0
\end{bmatrix}
\]

\[ x_4 = 2.0, x_4 = 2.0 \]

\[ k_{4,4} = 1.0 \times \exp \left( -\frac{(2.0-2.0)^2}{2 \times 2.0^2} \right) \]

\[ x_1 = -3, x_2 = 1.2, x_3 = 1.4, \text{ and } x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[ x_4 = 2.0, \quad x_4 = 2.0 \]

\[ k_{4,4} = 1.0 \times \exp\left(-\frac{(2.0-2.0)^2}{2 \times 2.0^2}\right) \]

\[ x_1 = -3, \quad x_2 = 1.2, \quad x_3 = 1.4, \quad \text{and} \quad x_4 = 2.0 \text{ with } \ell = 2.0 \text{ and } \alpha = 1.0. \]
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp\left( -\frac{||x_i - x_j||^2}{2\ell^2} \right)
\]

\[x_1 = -3.0, x_1 = -3.0\]

\[
k_{1,1} = 4.00 \times \exp\left( -\frac{(-3.0 - -3.0)^2}{2\times5.00^2} \right)
\]

\[x_1 = -3.0, x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 5.00 \text{ and } \alpha = 4.00.\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[ x_1 = -3.0, x_1 = -3.0 \]

\[ k_{1,1} = 4.00 \times \exp\left(-\frac{(-3.0-(-3.0))^2}{2 \times 5.00^2}\right) \]

\[ x_1 = -3.0, x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 5.00 \text{ and } \alpha = 4.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp \left( -\frac{||x_i - x_j||^2}{2\ell^2} \right) \]

\[ x_2 = 1.20, \ x_1 = -3.0 \]

\[ k_{2,1} = 4.00 \times \exp \left( -\frac{(1.20 - -3.0)^2}{2 \times 5.00^2} \right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \ \text{and} \ x_3 = 1.40 \ \text{with} \ \ell = 5.00 \ \text{and} \ \alpha = 4.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[ x_2 = 1.20, \ x_1 = -3.0 \]

\[ k_{2,1} = 4.00 \times \exp\left(-\frac{(1.20-(-3.0))^2}{2\times5.00^2}\right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \ \text{and} \ x_3 = 1.40 \ \text{with} \ \ell = 5.00 \ \text{and} \ \alpha = 4.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
4.00 & 2.81 \\
2.81 & 2.81
\end{bmatrix}
\]

\[ k_{2,1} = 4.00 \times \exp\left(-\frac{(1.20-(-3.0))^2}{2\times5.00^2}\right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \ \text{and} \ x_3 = 1.40 \ \text{with} \ \ell = 5.00 \ \text{and} \ \alpha = 4.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
4.00 & 2.81 \\
2.81 & \quad \quad \\
2.81 & \\
\end{bmatrix}
\]

\[ x_2 = 1.20, \quad x_2 = 1.20 \]

\[ k_{2,2} = 4.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2\times5.00^2}\right) \]

\[
\begin{bmatrix}
\quad \\
\quad \\
\quad \\
\end{bmatrix}
\]

\[ x_1 = -3.0, \quad x_2 = 1.20, \quad \text{and} \quad x_3 = 1.40 \quad \text{with} \quad \ell = 5.00 \quad \text{and} \quad \alpha = 4.00. \]
Covariance Functions
Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
4.00 & 2.81 \\
2.81 & 4.00
\end{bmatrix}
\]

\[ x_2 = 1.20, x_2 = 1.20 \]

\[ k_{2,2} = 4.00 \times \exp\left(-\frac{(1.20-1.20)^2}{2 \times 5.00^2}\right) \]

\[ x_1 = -3.0, x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 5.00 \text{ and } \alpha = 4.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
4.00 & 2.81 \\
2.81 & 4.00 \\
\end{bmatrix}
\]

\[ x_3 = 1.40, \ x_1 = -3.0 \]

\[ k_{3,1} = 4.00 \times \exp\left(-\frac{(1.40 - -3.0)^2}{2 \times 5.00^2}\right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 5.00 \text{ and } \alpha = 4.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp \left( -\frac{||x_i-x_j||^2}{2\ell^2} \right) \]

\[
\begin{bmatrix}
4.00 & 2.81 \\
2.81 & 4.00 \\
\end{bmatrix}
\]

\[ k_{3,1} = 4.00 \times \exp \left( -\frac{(1.40-3.0)^2}{2 \times 5.00^2} \right) \]

\[
\begin{bmatrix}
4.00 & 2.81 \\
2.81 & 4.00 \\
\end{bmatrix}
\]

\[ x_1 = -3.0, \ x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 5.00 \text{ and } \alpha = 4.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
4.00 & 2.81 & 2.72 \\
2.81 & 4.00 & \\
2.72 & & \\
\end{bmatrix}
\]

\[
x_3 = 1.40, x_1 = -3.0
\]

\[
k_{3,1} = 4.00 \times \exp\left(-\frac{(1.40-(-3.0))^2}{2\times5.00^2}\right) = 2.72
\]

\[
x_1 = -3.0, x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 5.00 \text{ and } \alpha = 4.00.
\]
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp \left(-\frac{||x_i-x_j||^2}{2\ell^2}\right)
\]

\[
\begin{bmatrix}
4.00 & 2.81 & 2.72 \\
2.81 & 4.00 \\
2.72 & & \\
\end{bmatrix}
\]

\[
k_{3,2} = 4.00 \times \exp \left(-\frac{(1.40-1.20)^2}{2\times5.00^2}\right)
\]

\[
\begin{bmatrix}
2.72 \\
\end{bmatrix}
\]

\[
x_1 = -3.0, \; x_2 = 1.20, \; \text{and} \; x_3 = 1.40 \; \text{with} \; \ell = 5.00 \; \text{and} \; \alpha = 4.00.
\]
Covariance Functions

Where did this covariance matrix come from?

\[
k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right)
\]

\[
\begin{bmatrix}
4.00 & 2.81 & 2.72 \\
2.81 & 4.00 & \\
2.72 & 4.00 \\
\end{bmatrix}
\]

\[x_3 = 1.40, x_2 = 1.20\]

\[k_{3,2} = 4.00 \times \exp\left(-\frac{(1.40-1.20)^2}{2\times5.00^2}\right)\]

\[x_1 = -3.0, x_2 = 1.20\text{, and } x_3 = 1.40 \text{ with } \ell = 5.00 \text{ and } \alpha = 4.00.\]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
4.00 & 2.81 & 2.72 \\
2.81 & 4.00 & 4.00 \\
2.72 & 4.00 & \\
\end{bmatrix}
\]

\[ x_3 = 1.40, \ x_2 = 1.20 \]

\[ k_{3,2} = 4.00 \times \exp\left(-\frac{(1.40-1.20)^2}{2\times5.00^2}\right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \ \text{and} \ x_3 = 1.40 \ \text{with} \ \ell = 5.00 \ \text{and} \ \alpha = 4.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{\|x_i - x_j\|^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
4.00 & 2.81 & 2.72 \\
2.81 & 4.00 & 4.00 \\
2.72 & 4.00 & \\
\end{bmatrix}
\]

\[ x_3 = 1.40, x_3 = 1.40 \]

\[ k_{3,3} = 4.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 5.00^2}\right) \]

\[ x_1 = -3.0, x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 5.00 \text{ and } \alpha = 4.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i-x_j||^2}{2\ell^2}\right) \]

\[
\begin{bmatrix}
4.00 & 2.81 & 2.72 \\
2.81 & 4.00 & 4.00 \\
2.72 & 4.00 & 4.00
\end{bmatrix}
\]

\[ k_{3,3} = 4.00 \times \exp\left(-\frac{(1.40-1.40)^2}{2\times5.00^2}\right) \]

\[ x_1 = -3.0, x_2 = 1.20, \text{ and } x_3 = 1.40 \text{ with } \ell = 5.00 \text{ and } \alpha = 4.00. \]
Covariance Functions

Where did this covariance matrix come from?

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{||x_i - x_j||^2}{2\ell^2}\right) \]

\[ x_3 = 1.40, \ x_3 = 1.40 \]

\[ k_{3,3} = 4.00 \times \exp\left(-\frac{(1.40 - 1.40)^2}{2 \times 5.00^2}\right) \]

\[ x_1 = -3.0, \ x_2 = 1.20, \ \text{and} \ x_3 = 1.40 \ \text{with} \ \ell = 5.00 \ \text{and} \ \alpha = 4.00. \]
Gaussian Process Interpolation

Figure: Real example: BACCO (see e.g. (Oakley and O’Hagan, 2002)). Interpolation through outputs from slow computer simulations (e.g. atmospheric carbon levels).
Figure: Real example: BACCO (see e.g. (Oakley and O’Hagan, 2002)). Interpolation through outputs from slow computer simulations (e.g. atmospheric carbon levels).
Figure: Real example: BACCO (see e.g. (Oakley and O’Hagan, 2002)). Interpolation through outputs from slow computer simulations (e.g. atmospheric carbon levels).
Gaussian Process Interpolation

![Figure: Real example: BACCO (see e.g. (Oakley and O’Hagan, 2002)). Interpolation through outputs from slow computer simulations (e.g. atmospheric carbon levels).]
Gaussian Process Interpolation

**Figure:** Real example: BACCO (see e.g. (Oakley and O’Hagan, 2002)). Interpolation through outputs from slow computer simulations (e.g. atmospheric carbon levels).
Gaussian Process Interpolation

Figure: Real example: BACCO (see e.g. (Oakley and O’Hagan, 2002)). Interpolation through outputs from slow computer simulations (e.g. atmospheric carbon levels).
Gaussian Process Interpolation

Figure: Real example: BACCO (see e.g. (Oakley and O’Hagan, 2002)). Interpolation through outputs from slow computer simulations (e.g. atmospheric carbon levels).
Gaussian Process Interpolation

Figure: Real example: BACCO (see e.g. (Oakley and O’Hagan, 2002)). Interpolation through outputs from slow computer simulations (e.g. atmospheric carbon levels).
Gaussian Noise

- Gaussian noise model,

\[ p(y_i|f_i) = \mathcal{N}(y_i|f_i, \sigma^2) \]

where \( \sigma^2 \) is the variance of the noise.

- Equivalent to a covariance function of the form

\[ k(x_i, x_j) = \delta_{i,j} \sigma^2 \]

where \( \delta_{i,j} \) is the Kronecker delta function.

- Additive nature of Gaussians means we can simply add this term to existing covariance matrices.
Gaussian Process Regression

Figure: Examples include WiFi localization, C14 calibration curve.
Gaussian Process Regression

Figure: Examples include WiFi localization, C14 callibration curve.
Gaussian Process Regression

Figure: Examples include WiFi localization, C14 callibration curve.
Gaussian Process Regression

Figure: Examples include WiFi localization, C14 calibration curve.
Figure: Examples include WiFi localization, C14 callibration curve.
Gaussian Process Regression

Figure: Examples include WiFi localization, C14 calibration curve.
Gaussian Process Regression

Figure: Examples include WiFi localization, C14 calibration curve.
Gaussian Process Regression

Figure: Examples include WiFi localization, C14 calibration curve.
Gaussian Process Regression

Figure: Examples include WiFi localization, C14 calibration curve.
Gaussian Process Fit to Olympic Marathon Data

![Graph showing marathon times from 1892 to 2012 with a trend line and shaded area indicating uncertainty.](image)
Outline

Gaussian Processes

GP Non-Gaussian

GP Limitations

Kalman Filter

Dimensionality Reduction
General Noise Models

Graph of a GP

- Relates input variables, $X$, to vector, $y$, through $f$ given kernel parameters $\theta$.
- Plate notation indicates independence of $y_i|f_i$.
- In general $p(y_i|f_i)$ is non-Gaussian.
- We approximate with Gaussian $p(y_i|f_i) \approx \mathcal{N}(m_i|f_i, \beta_i^{-1})$.

Figure: The Gaussian process depicted graphically.
Figure: Inclusion of a data point with Gaussian noise.
Gaussian Noise

$\text{Figure: Inclusion of a data point with Gaussian noise.}$
**Gaussian Noise**

**Figure:** Inclusion of a data point with Gaussian noise.
Local Moment Matching

- Easiest to consider a single previously unseen data point, $y^*, x^*$.
- Before seeing data point, prediction of $f^*$ is a GP, $q(f^* | y, X)$.
- Update prediction using Bayes’ Rule,

$$p (f^*|y, y^*, X, x^*) = \frac{p (y^*|f^*) p (f^*|y, X, x^*)}{p (y, y^*|X, x^*)}.$$

This posterior is not a Gaussian process if $p (y^*|f^*)$ is non-Gaussian.
**Classification Noise Model**

**Probit Noise Model**

![Probit model plot](image)

**Figure:** The probit model (classification). The plot shows $p(y_i|f_i)$ for different values of $y_i$. For $y_i = 1$ we have

$$p(y_i|f_i) = \phi (f_i) = \int_{-\infty}^{f_i} \mathcal{N}(z|0,1)\,dz.$$
Match Moments

- Idea behind EP — approximate with a Gaussian process at this stage by matching moments.
- This is equivalent to minimizing the following KL divergence where \( q(f_*|y, y_*, X, x_*) \) is constrained to be a GP.

\[
q(f_*|y, y_*, X, x_*) = \arg\min_{q(f_*|y, y_*, X, x_*)} \text{KL}(p(f_*|y, y_*, X, x_*) || q(f_*|y, y_*, X, x_*))
\]

- This is equivalent to setting

\[
\langle f_* \rangle_{q(f_*|y, y_*, X, x_*)} = \langle f_* \rangle_{p(f_*|y, y_*, X, x_*)}
\]

\[
\langle f_*^2 \rangle_{q(f_*|y, y_*, X, x_*)} = \langle f_*^2 \rangle_{p(f_*|y, y_*, X, x_*)}
\]
This is achieved by replacing $p(y_*|f_*)$ with a Gaussian distribution

$$p(f_*|y, y_*, X, x_*) = \frac{p(y_*|f_*) p(f_*|y, X, x_*)}{p(y, y_*|X, x_*)}$$

does not become

$$q(f_*|y, y_*, X, x_*) = \frac{\mathcal{N}(m_*, f_*, \beta^{-1}_m) p(f_*|y, X, x_*)}{p(y, y_*|X, x_*)}.$$
Classification

Figure: An EP style update with a classification noise model.
Classification

Figure: An EP style update with a classification noise model.

$p(f_*|X, x_*, y)$
$p(y_* = 1|f_*)$
Figure: An EP style update with a classification noise model.

\[ p(f_*|X, x_*, y) \]
\[ p(y_* = 1|f_*) \]
\[ p(f_*|X, x_*, y, y_*) \]
Figure: An EP style update with a classification noise model.
**Ordinal Noise Model**

**Ordered Categories**

**Figure:** The ordered categorical noise model (ordinal regression). The plot shows $p(y_i|f_i)$ for different values of $y_i$. Here we have assumed three categories.
Laplace Approximation

- Equivalent Gaussian is found by making a local 2nd order Taylor approximation at the mode.
- Laplace was the first to suggest this\(^1\), so it’s known as the *Laplace approximation*. 
Learning Covariance Parameters
Can we determine covariance parameters from the data?

\[
N(y|0, K) = \frac{1}{(2\pi)^{n/2} |K|^{1/2}} \exp\left(-\frac{y^\top K^{-1} y}{2}\right)
\]

The parameters are *inside* the covariance function (matrix).

\[
k_{i,j} = k(x_i, x_j; \theta)
\]
Can we determine covariance parameters from the data?

\[ N(y|0, K) = \frac{1}{(2\pi)^{n/2}|K|^{1/2}} \exp \left( -\frac{y^\top K^{-1}y}{2} \right) \]

The parameters are *inside* the covariance function (matrix).

\[ k_{i,j} = k(x_i, x_j; \theta) \]
Learning Covariance Parameters
Can we determine covariance parameters from the data?

\[
\log \mathcal{N}(y|0, K) = -\frac{1}{2} \log |K| - \frac{y^\top K^{-1} y}{2} - \frac{n}{2} \log 2\pi
\]

The parameters are \textit{inside} the covariance function (matrix).

\[
k_{i,j} = k(x_i, x_j; \theta)
\]
Learning Covariance Parameters

Can we determine covariance parameters from the data?

$$E(\theta) = \frac{1}{2} \log |K| + \frac{y^T K^{-1} y}{2}$$

The parameters are *inside* the covariance function (matrix).

$$k_{i,j} = k(x_i, x_j; \theta)$$
Eigendecomposition of Covariance

A useful decomposition for understanding the objective function.

\[ K = R \Lambda^2 R^\top \]

Diagonal of \( \Lambda \) represents distance along axes.
\( R \) gives a rotation of these axes.
Capacity control: $\log|\mathbf{K}|$

\[ \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \]
Capacity control: $\log |\mathbf{K}|$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$
Capacity control: $\log |\mathbf{K}|$

\[ \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \]
Capacity control: $\log |K|$

\[
\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}
\]

$|\Lambda| = \lambda_1 \lambda_2$
Capacity control: $\log|\mathbf{K}|$

\[ \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \]

\[ |\Lambda| = \lambda_1 \lambda_2 \]
Capacity control: $\log |\mathbf{K}|$

\[ \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \]

\[ |\Lambda| = \lambda_1 \lambda_2 \]
Capacity control: $\log |\mathbf{K}|$

$$\Psi = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

$$|\Psi| = \lambda_1 \lambda_2$$
Capacity control: $\log |\mathbf{K}|$

$$
\Lambda = \begin{bmatrix}
\lambda_1 & 0 & 0 \\
0 & \lambda_2 & 0 \\
0 & 0 & \lambda_3
\end{bmatrix}
$$

$$
|\Lambda| = \lambda_1 \lambda_2 \lambda_3
$$
Capacity control: $\log |K|$ 

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$ 

$$|\Lambda| = \lambda_1 \lambda_2$$
Capacity control: $\log |K|$ 

\[ R\Lambda = \begin{bmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{bmatrix} \]

\[ |R\Lambda| = \lambda_1 \lambda_2 \]
Data Fit: $\frac{y^\top K^{-1} y}{2}$
Data Fit: $\frac{y^T K^{-1} y}{2}$
Data Fit: \( \frac{y^\top K^{-1} y}{2} \)
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^T K^{-1} y}{2} \]
Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1}y}{2} \]
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1} y}{2} \]
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1} y}{2} \]
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1}y}{2} \]
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

$$E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1}y}{2}$$
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1} y}{2} \]
Learning Covariance Parameters
Can we determine length scales and noise levels from the data?

\[
E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1}y}{2}
\]
Learning Covariance Parameters

Can we determine length scales and noise levels from the data?

\[ E(\theta) = \frac{1}{2} \log |K| + \frac{y^\top K^{-1}y}{2} \]
Gene Expression Example

- Given expression levels in the form of a time series from Della Gatta et al. (2008).
- Want to detect if a gene is expressed or not, fit a GP to each gene (Kalaitzis and Lawrence, 2011).
A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression

Alfredo A Kalaitzis* and Neil D Lawrence*

Abstract

Background: The analysis of gene expression from time series underpins many biological studies. Two basic forms of analysis recur for data of this type: removing inactive (quiet) genes from the study and determining which genes are differentially expressed. Often these analysis stages are applied disregarding the fact that the data is drawn from a time series. In this paper we propose a simple model for accounting for the underlying temporal nature of the data based on a Gaussian process.

Results: We review Gaussian process (GP) regression for estimating the continuous trajectories underlying in gene expression time-series. We present a simple approach which can be used to filter quiet genes, or for the case of time series in the form of expression ratios, quantify differential expression. We assess via ROC curves the rankings produced by our regression framework and compare them to a recently proposed hierarchical Bayesian model for the analysis of gene expression time-series (BATS). We compare on both simulated and experimental data showing that the proposed approach considerably outperforms the current state of the art.
Contour plot of Gaussian process likelihood.
Optima: length scale of 1.2221 and $\log_{10}$ SNR of 1.9654 log likelihood is -0.22317.
Optima: length scale of 1.5162 and log\(_{10}\) SNR of 0.21306 log likelihood is -0.23604.
Optima: length scale of 2.9886 and log_{10} SNR of -4.506 log likelihood is -2.1056.
Radial basis functions commonly have the form

\[ \phi_k(x_i) = \exp\left(-\frac{|x_i - \mu_k|^2}{2\ell^2}\right). \]

- Basis function maps data into a "feature space" in which a linear sum is a non linear function.

Figure: A set of radial basis functions with width \( \ell = 2 \) and location parameters \( \mu = [-4 \ 0 \ 4]^T \).
Basis Function Representations

- Represent a function by a linear sum over a basis,

\[ f(x_i; \mathbf{w}) = \sum_{k=1}^{m} w_k \phi_k(x_i;), \quad (1) \]

- Here: \( m \) basis functions and \( \phi_k(\cdot) \) is \( k \)th basis function and

\[ \mathbf{w} = [w_1, \ldots, w_m]^\top. \]

- For standard linear model: \( \phi_k(x_i;) = x_{i,k}. \)
Random Functions

Functions derived using:

\[ f(x) = \sum_{k=1}^{m} w_k \phi_k(x), \]

where elements of \( \mathbf{w} \) are independently sampled from a Gaussian density,

\[ w_k \sim \mathcal{N}(0, \alpha). \]

**Figure:** Functions sampled using the basis set from figure 9. Each line is a separate sample, generated by a weighted sum of the basis set. The weights, \( \mathbf{w} \) are sampled from a Gaussian density with variance \( \alpha = 1 \).
1. Sum of Gaussian variables is also Gaussian.

\[ y_i \sim \mathcal{N}(\mu_i, \sigma^2_i) \]
1. Sum of Gaussian variables is also Gaussian.

\[ y_i \sim N(\mu_i, \sigma_i^2) \]

\[ \sum_{i=1}^{n} y_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \]
1. Sum of Gaussian variables is also Gaussian.

\[ y_i \sim N\left(\mu_i, \sigma_i^2\right) \]

\[ \sum_{i=1}^{n} y_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \]

2. Scaling a Gaussian leads to a Gaussian.
Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.
   \[ y_i \sim \mathcal{N} \left( \mu_i, \sigma_i^2 \right) \]
   \[
   \sum_{i=1}^{n} y_i \sim \mathcal{N} \left( \sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right)
   \]

2. Scaling a Gaussian leads to a Gaussian.
   \[ y \sim \mathcal{N} \left( \mu, \sigma^2 \right) \]
Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

\[ y_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \]

\[ \sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \]

2. Scaling a Gaussian leads to a Gaussian.

\[ y \sim \mathcal{N}(\mu, \sigma^2) \]

\[ wy \sim \mathcal{N}(w\mu, w^2\sigma^2) \]
Multivariate Consequence

- If

\[ x \sim \mathcal{N} (\mu, \Sigma) \]
Multivariate Consequence

- If

\[ x \sim \mathcal{N}(\mu, \Sigma) \]

- And

\[ y = Wx \]
Multivariate Consequence

- If
  \[ x \sim \mathcal{N}(\mu, \Sigma) \]

- And
  \[ y = Wx \]

- Then
  \[ y \sim \mathcal{N}(W\mu, W\Sigma W^\top) \]
Covariance Functions

RBF Basis Functions

\[ k(x, x') = \alpha \phi(x)^\top \phi(x') \]

\[ \phi_k(x) = \exp \left( -\frac{\|x - \mu_k\|_2^2}{\ell^2} \right) \]

\[ \mu = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \]
Covariance Functions

RBF Basis Functions

\[ k(x, x') = \alpha \phi(x)^\top \phi(x') \]

\[ \phi_k(x) = \exp \left( -\frac{\|x - \mu_k\|^2}{\ell^2} \right) \]

\[ \mu = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \]
1. Sum of Gaussian variables is also Gaussian.

\[ y_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \]
Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

\[ y_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \]

\[ \sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \]
1. Sum of Gaussian variables is also Gaussian.

\[ y_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \]

\[ \sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \]

2. Scaling a Gaussian leads to a Gaussian.
Recall Univariate Gaussian Properties

1. Sum of Gaussian variables is also Gaussian.

\[ y_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \]

\[ \sum_{i=1}^{n} y_i \sim \mathcal{N} \left( \sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2 \right) \]

2. Scaling a Gaussian leads to a Gaussian.

\[ y \sim \mathcal{N}(\mu, \sigma^2) \]
1. Sum of Gaussian variables is also Gaussian.

\[ y_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \]

\[ \sum_{i=1}^{n} y_i \sim \mathcal{N}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \]

2. Scaling a Gaussian leads to a Gaussian.

\[ y \sim \mathcal{N}(\mu, \sigma^2) \]

\[ wy \sim \mathcal{N}(w\mu, w^2\sigma^2) \]
Multivariate Consequence

If

\[ x \sim \mathcal{N}(\mu, \Sigma) \]

And

\[ y = Wx \]

Then

\[ y \sim \mathcal{N}(W\mu, W\Sigma W^\top) \]
Multivariate Consequence

- If

\[ x \sim \mathcal{N}(\mu, \Sigma) \]

- And

\[ y = Wx \]
If \( x \sim \mathcal{N}(\mu, \Sigma) \)

And \( y = Wx \)

Then \( y \sim \mathcal{N}(W\mu, W\Sigma W^\top) \)
Selecting Number and Location of Basis

- Need to choose
  1. location of centers
  2. number of basis functions

Restrict analysis to 1-D input, $x$.

- Consider uniform spacing over a region:

$$ k(x_i, x_j) = \alpha \phi_k(x_i)^\top \phi_k(x_j) $$
Selecting Number and Location of Basis

- Need to choose
  1. location of centers
  2. number of basis functions

Restrict analysis to 1-D input, \( x \).

- Consider uniform spacing over a region:

\[
k(x_i, x_j) = \alpha \sum_{k=1}^{m} \phi_k(x_i)\phi_k(x_j)
\]
Selecting Number and Location of Basis

- Need to choose
  1. location of centers
  2. number of basis functions

Restrict analysis to 1-D input, $x$.

- Consider uniform spacing over a region:

$$k(x_i, x_j) = \alpha \sum_{k=1}^{m} \exp\left(-\frac{(x_i - \mu_k)^2}{2\ell^2}\right) \exp\left(-\frac{(x_j - \mu_k)^2}{2\ell^2}\right)$$
Selecting Number and Location of Basis

- Need to choose
  1. location of centers
  2. number of basis functions

Restrict analysis to 1-D input, $x$.

- Consider uniform spacing over a region:

$$ k(x_i, x_j) = \alpha \sum_{k=1}^{m} \exp \left( -\frac{(x_i - \mu_k)^2}{2\ell^2} - \frac{(x_j - \mu_k)^2}{2\ell^2} \right) $$
Selecting Number and Location of Basis

- Need to choose
  1. location of centers
  2. number of basis functions

Restrict analysis to 1-D input, $x$.

- Consider uniform spacing over a region:

$$k(x_i, x_j) = \alpha \sum_{k=1}^{m} \exp\left(-\frac{x_i^2 + x_j^2 - 2\mu_k(x_i + x_j) + 2\mu_k^2}{2\ell^2}\right),$$
Set each center location to

\[ \mu_k = a + \Delta \mu \cdot (k - 1). \]
Uniform Basis Functions

- Set each center location to

\[ \mu_k = a + \Delta \mu \cdot (k - 1). \]

- Specify the basis functions in terms of their indices,

\[
k(x_i, x_j) = \alpha' \Delta \mu \sum_{k=1}^{m} \exp \left( - \frac{x_i^2 + x_j^2}{2\ell^2} \right) \left( 2 \left( a + \Delta \mu \cdot (k - 1) \right) \left( x_i + x_j \right) + 2 \left( a + \Delta \mu \cdot (k - 1) \right)^2 \right) \]
Uniform Basis Functions

- Set each center location to

\[ \mu_k = a + \Delta \mu \cdot (k - 1). \]

- Specify the basis functions in terms of their indices,

\[ k(x_i, x_j) = \alpha' \Delta \mu \sum_{k=1}^{m} \exp \left( - \frac{x_i^2 + x_j^2}{2\ell^2} \right) \]

\[ - 2 (a + \Delta \mu \cdot (k - 1)) \left( x_i + x_j \right) + 2 \left( a + \Delta \mu \cdot (k - 1) \right)^2 \]

\[ \frac{2\ell^2}{2\ell^2} \].

- Here we’ve scaled variance of process by \( \Delta \mu \).
Take

$$
\mu_1 = a \quad \text{and} \quad \mu_m = b \quad \text{so} \quad b = a + \Delta \mu \cdot (m - 1)
$$
Infinite Basis Functions

- Take

\[ \mu_1 = a \text{ and } \mu_m = b \text{ so } b = a + \Delta \mu \cdot (m - 1) \]

- This implies

\[ b - a = \Delta \mu(m - 1) \]
Infinite Basis Functions

- Take
  \[ \mu_1 = a \text{ and } \mu_m = b \text{ so } b = a + \Delta \mu \cdot (m - 1) \]

- This implies
  \[ b - a = \Delta \mu (m - 1) \]
  and therefore
  \[ m = \frac{b - a}{\Delta \mu} + 1 \]
Infinite Basis Functions

- Take
  \[ \mu_1 = a \text{ and } \mu_m = b \text{ so } b = a + \Delta \mu \cdot (m - 1) \]

- This implies
  \[ b - a = \Delta \mu (m - 1) \]
  and therefore
  \[ m = \frac{b - a}{\Delta \mu} + 1 \]

- Take limit as \( \Delta \mu \to 0 \) so \( m \to \infty \)
Infinite Basis Functions

- Take \( \mu_1 = a \) and \( \mu_m = b \) so \( b = a + \Delta \mu \cdot (m - 1) \)

- This implies \( b - a = \Delta \mu (m - 1) \) and therefore
  \[
  m = \frac{b - a}{\Delta \mu} + 1
  \]

- Take limit as \( \Delta \mu \to 0 \) so \( m \to \infty \)

\[
k(x_i, x_j) = \alpha' \int_a^b \exp \left( -\frac{x_i^2 + x_j^2}{2\ell^2} + \frac{2(\mu - \frac{1}{2}(x_i + x_j))^2 - \frac{1}{2}(x_i + x_j)^2}{2\ell^2} \right) d\mu,
\]

where we have used \( a + k \cdot \Delta \mu \to \mu \).
Performing the integration leads to

\[ k(x_i, x_j) = \alpha' \sqrt{\pi \ell^2} \exp \left( -\frac{(x_i - x_j)^2}{4\ell^2} \right) \]

\[ \times \frac{1}{2} \left[ \text{erf} \left( \frac{b - \frac{1}{2} (x_i + x_j)}{\ell} \right) - \text{erf} \left( \frac{a - \frac{1}{2} (x_i + x_j)}{\ell} \right) \right], \]
Performing the integration leads to

\[ k(x_i, x_j) = \alpha' \sqrt{\pi \ell^2} \exp \left( -\frac{(x_i - x_j)^2}{4\ell^2} \right) \times \frac{1}{2} \left[ \text{erf} \left( \frac{b - \frac{1}{2} (x_i + x_j)}{\ell} \right) - \text{erf} \left( \frac{a - \frac{1}{2} (x_i + x_j)}{\ell} \right) \right], \]

Now take limit as \( a \to -\infty \) and \( b \to \infty \)
Performing the integration leads to

\[ k(x_i, x_j) = \alpha' \sqrt{\pi \ell^2} \exp\left(-\frac{(x_i - x_j)^2}{4\ell^2}\right) \]

\times \frac{1}{2} \left[ \text{erf}\left(\frac{b - \frac{1}{2}(x_i + x_j)}{\ell}\right) - \text{erf}\left(\frac{a - \frac{1}{2}(x_i + x_j)}{\ell}\right) \right],

Now take limit as \(a \to -\infty\) and \(b \to \infty\)

\[ k(x_i, x_j) = \alpha \exp\left(-\frac{(x_i - x_j)^2}{4\ell^2}\right). \]

where \(\alpha = \alpha' \sqrt{\pi \ell^2}\).
An RBF model with infinite basis functions is a Gaussian process.
An RBF model with infinite basis functions is a Gaussian process.

The covariance function is given by the exponentiated quadratic covariance function.

\[ k(x_i, x_j) = \alpha \exp \left( -\frac{(x_i - x_j)^2}{4\ell^2} \right). \]
An RBF model with infinite basis functions is a Gaussian process.

The covariance function is the exponentiated quadratic.

**Note:** The functional form for the covariance function and basis functions are similar.

- this is a special case,
- in general they are very different

Similar results can obtained for multi-dimensional input models Williams (1998); Neal (1996).
Covariance Functions

RBF Basis Functions

\[ k(x, x') = \alpha \phi(x)^\top \phi(x') \]

\[ \phi_k(x) = \exp \left( -\frac{\|x - \mu_k\|_2^2}{\ell^2} \right) \]

\[ \mu = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \]
Covariance Functions

RBF Basis Functions

\[ k(x, x') = \alpha \phi(x)^\top \phi(x') \]

\[ \phi_k(x) = \exp \left( -\frac{\|x - \mu_k\|_2^2}{\ell^2} \right) \]

\[ \mu = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \]
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

\[ k(x, x') = \alpha \exp\left(-\frac{||x - x'||^2}{2\ell^2}\right) \]

- Covariance matrix is built using the inputs to the function \( x \).
- For the example above it was based on Euclidean distance.
- The covariance function is also known as a kernel.
Covariance Functions
Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

\[ k(x, x') = \alpha \exp \left( -\frac{\|x - x'\|^2}{2\ell^2} \right) \]

- Covariance matrix is built using the inputs to the function \( x \).
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.
Covariance Functions

MLP Covariance Function

\[ k(x, x') = \alpha \sin \left( \frac{wx^\top x' + b}{\sqrt{wx^\top x + b + 1} \sqrt{wx'^\top x' + b + 1}} \right) \]

- Based on infinite neural network model.

\[ w = 40 \]

\[ b = 4 \]
Covariance Functions

MLP Covariance Function

\[ k(x, x') = \alpha \text{asin} \left( \frac{wx^\top x' + b}{\sqrt{wx^\top x + b + 1} \sqrt{wx'^\top x' + b + 1}} \right) \]

- Based on infinite neural network model.
  
  \[ w = 40 \]
  
  \[ b = 4 \]
Constructing Covariance Functions

- Sum of two covariances is also a covariance function.

\[ k(x, x') = k_1(x, x') + k_2(x, x') \]
Product of two covariances is also a covariance function.

\[ k(x, x') = k_1(x, x')k_2(x, x') \]
Multiply by Deterministic Function

- If $f(x)$ is a Gaussian process.
- $g(x)$ is a deterministic function.
- $h(x) = f(x)g(x)$
- Then

$$k_h(x, x') = g(x)k_f(x, x')g(x')$$

where $k_h$ is covariance for $h(\cdot)$ and $k_f$ is covariance for $f(\cdot)$. 
Covariance Functions

MLP Covariance Function

\[ k(x, x') = \alpha \sin \left( \frac{wx^\top x' + b}{\sqrt{wx^\top x + b + 1} \sqrt{wx'^\top x' + b + 1}} \right) \]

- Based on infinite neural network model.
  \[ w = 40 \]
  \[ b = 4 \]
Covariance Functions

MLP Covariance Function

\[
k(x, x') = \alpha \sin \left( \frac{wx^\top x' + b}{\sqrt{wx^\top x + b + 1} \sqrt{wx'^\top x' + b + 1}} \right)
\]

- Based on infinite neural network model.

\[
wx = 40
\]

\[
b = 4
\]
Covariance Functions

Linear Covariance Function

\[ k(x, x') = \alpha x^\top x' \]

- Bayesian linear regression.

\[ \alpha = 1 \]
Covariance Functions

Linear Covariance Function

\[ k(x, x') = \alpha x^\top x' \]

- Bayesian linear regression.
  \[ \alpha = 1 \]
Outline

Gaussian Processes

GP Non-Gaussian

GP Limitations

Kalman Filter

Dimensionality Reduction
Limitations of Gaussian Processes

- Inference is $O(n^3)$ due to matrix inverse (in practice use Cholesky).
- Gaussian processes don’t deal well with discontinuities (financial crises, phosphorylation, collisions, edges in images).
- Widely used exponentiated quadratic covariance (RBF) can be too smooth in practice (but there are many alternatives!!).
Outline

Gaussian Processes

GP Non-Gaussian

GP Limitations

Kalman Filter

Dimensionality Reduction
Simple Markov Chain

- Assume 1-d latent state, a vector over time, $\mathbf{x} = [x_1 \ldots x_T]$.

- Markov property,

  $$
  x_i = x_{i-1} + \epsilon_i,
  $$

  $$
  \epsilon_i \sim \mathcal{N}(0, \alpha)
  $$

  $$
  \implies x_i \sim \mathcal{N}(x_{i-1}, \alpha)
  $$

- Initial state,

  $$
  x_0 \sim \mathcal{N}(0, \alpha_0)
  $$

- If $x_0 \sim \mathcal{N}(0, \alpha)$ we have a Markov chain for the latent states.

- Markov chain it is specified by an initial distribution (Gaussian) and a transition distribution (Gaussian).
Gauss Markov Chain

\[ x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0, 1) \]

\[ x_0 = 0.000, \quad \epsilon_1 = -2.24 \]

\[ x_1 = 0.000 - 2.24 = -2.24 \]
Gauss Markov Chain

\[ x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0, 1) \]

\[ x_1 = -2.24, \quad \epsilon_2 = 0.457 \]

\[ x_2 = -2.24 + 0.457 = -1.78 \]
Gauss Markov Chain

\[ x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1) \]

\[ x_2 = -1.78, \quad \epsilon_3 = 0.178 \]

\[ x_3 = -1.78 + 0.178 = -1.6 \]
Gauss Markov Chain

\[ x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0, 1) \]

\[ x_3 = -1.6, \quad \epsilon_4 = -0.292 \]

\[ x_4 = -1.6 - 0.292 = -1.89 \]
Gauss Markov Chain

\[ x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0,1) \]

\[ x_4 = -1.89, \quad \epsilon_5 = -0.501 \]

\[ x_5 = -1.89 - 0.501 = -2.39 \]
Gauss Markov Chain

\[ x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0, 1) \]

\[ x_5 = -2.39, \quad \epsilon_6 = 1.32 \]

\[ x_6 = -2.39 + 1.32 = -1.08 \]
Gauss Markov Chain

\[ x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0, 1) \]

\[ x_6 = -1.08, \quad \epsilon_7 = 0.989 \]

\[ x_7 = -1.08 + 0.989 = -0.0881 \]
Gauss Markov Chain

\[ x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0, 1) \]

\[ x_7 = -0.0881, \quad \epsilon_8 = -0.842 \]

\[ x_8 = -0.0881 - 0.842 = -0.93 \]
Gauss Markov Chain

\[ x_0 = 0, \quad \epsilon_i \sim \mathcal{N}(0, 1) \]

\[ x_8 = -0.93, \quad \epsilon_9 = -0.41 \]

\[ x_9 = -0.93 - 0.410 = -1.34 \]
If
\[ z \sim \mathcal{N}(\mu, C) \]
and
\[ x = Wz + b \]
then
\[ x \sim \mathcal{N}(W\mu + b, WCW^T) \]
Simplified: If

\[ z \sim \mathcal{N}(0, \sigma^2 I) \]

and

\[ x = Wz \]

then

\[ x \sim \mathcal{N}(0, \sigma^2 WW^\top) \]
Matrix Representation of Latent Variables

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3 \\
  x_4 \\
  x_5 \\
\end{bmatrix} =
\begin{bmatrix}
  1 & 0 & 0 & 0 & 0 \\
  1 & 1 & 0 & 0 & 0 \\
  1 & 1 & 1 & 0 & 0 \\
  1 & 1 & 1 & 1 & 0 \\
  1 & 1 & 1 & 1 & 1 \\
\end{bmatrix}
\times
\begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \epsilon_4 \\
  \epsilon_5 \\
\end{bmatrix}
\]

\[
x_1 = \epsilon_1
\]
Matrix Representation of Latent Variables

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3 \\
  x_4 \\
  x_5 \\
\end{bmatrix}
= \begin{bmatrix}
  1 & 0 & 0 & 0 & 0 \\
  1 & 1 & 0 & 0 & 0 \\
  1 & 1 & 1 & 0 & 0 \\
  1 & 1 & 1 & 1 & 0 \\
  1 & 1 & 1 & 1 & 1 \\
\end{bmatrix}
\times
\begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \epsilon_4 \\
  \epsilon_5 \\
\end{bmatrix}
\]

\[x_2 = \epsilon_1 + \epsilon_2\]
Matrix Representation of Latent Variables

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  \textcolor{red}{x_3} \\
  x_4 \\
  x_5 
\end{bmatrix}
= 
\begin{bmatrix}
  1 & 0 & 0 & 0 & 0 \\
  1 & 1 & 0 & 0 & 0 \\
  1 & 1 & 1 & 0 & 0 \\
  1 & 1 & 1 & 1 & 0 \\
  1 & 1 & 1 & 1 & 1 
\end{bmatrix}
\times
\begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \textcolor{red}{\epsilon_3} \\
  \epsilon_4 \\
  \epsilon_5 
\end{bmatrix}
\]

\[x_3 = \epsilon_1 + \epsilon_2 + \epsilon_3\]
Matrix Representation of Latent Variables

\[
\begin{bmatrix}
  x_1 \\
  x_2 \\
  x_3 \\
  x_4 \\
  x_5 \\
\end{bmatrix}
= \begin{bmatrix}
  1 & 0 & 0 & 0 & 0 \\
  1 & 1 & 0 & 0 & 0 \\
  1 & 1 & 1 & 0 & 0 \\
  \boxed{1 & 1 & 1 & 1 & 0} \\
  1 & 1 & 1 & 1 & 1 \\
\end{bmatrix}
\times
\begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \epsilon_4 \\
  \epsilon_5 \\
\end{bmatrix}
\]

\[x_4 = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4\]
Matrix Representation of Latent Variables

\[
x_1 \\
x_2 \\
x_3 \\
x_4 \\
x_5
\begin{bmatrix}
  1 & 0 & 0 & 0 & 0 \\
  1 & 1 & 0 & 0 & 0 \\
  1 & 1 & 1 & 0 & 0 \\
  1 & 1 & 1 & 1 & 0 \\
  1 & 1 & 1 & 1 & 1 \\
\end{bmatrix} \times \begin{bmatrix}
  \epsilon_1 \\
  \epsilon_2 \\
  \epsilon_3 \\
  \epsilon_4 \\
  \epsilon_5
\end{bmatrix}
\]

\[x_5 = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4 + \epsilon_5\]
$x = L_1 \times \epsilon$
Multivariate Process

- Since $\mathbf{x}$ is linearly related to $\epsilon$ we know $\mathbf{x}$ is a Gaussian process.
- Trick: we only need to compute the mean and covariance of $\mathbf{x}$ to determine that Gaussian.
Latent Process Mean

\[ x = L_1 \epsilon \]
Latent Process Mean

\[ \langle x \rangle = \langle L_1 \epsilon \rangle \]
\langle x \rangle = L_1 \langle \epsilon \rangle
Latent Process Mean

\[ \langle x \rangle = L_1 \langle \epsilon \rangle \]

\[ \epsilon \sim \mathcal{N}(0, \alpha I) \]
\langle x \rangle = L_{10}
Latent Process Mean

\[ \langle x \rangle = 0 \]
Latent Process Covariance

\[ xx^T = L_1 \epsilon \epsilon^T L_1^T \]

\[ x^T = \epsilon^T L^T \]
Latent Process Covariance

\[ \langle xx^\top \rangle = \langle L_1 \epsilon \epsilon^\top L_1^\top \rangle \]
\[ \langle xx^\top \rangle = L_1 \langle \epsilon \epsilon^\top \rangle L_1^\top \]
Latent Process Covariance

\[
\langle xx^\top \rangle = L_1 \langle \epsilon \epsilon^\top \rangle L_1^\top
\]

\[
\epsilon \sim \mathcal{N}(0, \alpha I)
\]
Latent Process Covariance

\[ \langle xx^\top \rangle = \alpha L_1 L_1^\top \]
Latent Process

\[ x = L_1 \epsilon \]
Latent Process

\[ x = L_1 \epsilon \]

\[ \epsilon \sim \mathcal{N} (0, \alpha I) \]
Latent Process

\[ x = L_1 \epsilon \]

\[ \epsilon \sim \mathcal{N}(0, \alpha I) \]

\[ \implies \]
Latent Process

\[ x = L_1 \epsilon \]

\[ \epsilon \sim \mathcal{N}(0, \alpha I) \]

\[ \implies \]

\[ x \sim \mathcal{N}(0, \alpha L_1 L_1^T) \]
Covariance for Latent Process II

- Make the variance dependent on time interval.
- Assume variance grows *linearly* with time.
- Justification: sum of two Gaussian distributed random variables is distributed as Gaussian with sum of variances.
- If variable’s movement is additive over time (as described) variance scales linearly with time.
Given
\[ \epsilon \sim \mathcal{N}(0, \alpha I) \implies \epsilon \sim \mathcal{N}(0, \alpha L_1 L_1^T). \]

Then
\[ \epsilon \sim \mathcal{N}(0, \Delta t \alpha I) \implies \epsilon \sim \mathcal{N}(0, \Delta t \alpha L_1 L_1^T). \]

where \( \Delta t \) is the time interval between observations.
Covariance for Latent Process II

\[ \epsilon \sim \mathcal{N}(0, \alpha \Delta t I), \quad x \sim \mathcal{N}\left(0, \alpha \Delta t L_1 L_1^T\right) \]
Covariance for Latent Process II

\[ \epsilon \sim \mathcal{N}(0, \alpha \Delta t I), \quad x \sim \mathcal{N}(0, \alpha \Delta t L_1 L_1^T) \]

\[ K = \alpha \Delta t L_1 L_1^T \]
Covariance for Latent Process II

\[ \epsilon \sim \mathcal{N}(0, \alpha \Delta t \mathbf{I}), \quad \mathbf{x} \sim \mathcal{N}(0, \alpha \Delta t \mathbf{L}_1 \mathbf{L}_1^T) \]

\[ \mathbf{K} = \alpha \Delta t \mathbf{L}_1 \mathbf{L}_1^T \]

\[ k_{i,j} = \alpha \Delta t \mathbf{l}_{i,k} \mathbf{l}_{j,k} \]

where \( \mathbf{l}_{:,:k} \) is a vector from the \( k \)th row of \( \mathbf{L}_1 \): the first \( k \) elements are one, the next \( T - k \) are zero.
\[ \epsilon \sim \mathcal{N}(0, \alpha \Delta t \mathbf{I}), \quad \mathbf{x} \sim \mathcal{N}(0, \alpha \Delta t \mathbf{L}_1 \mathbf{L}_1^T) \]

\[ \mathbf{K} = \alpha \Delta t \mathbf{L}_1 \mathbf{L}_1^T \]

\[ k_{i,j} = \alpha \Delta t \mathbf{l}_{.,i} \mathbf{l}_{.,j} \]

where \( \mathbf{l}_{.,k} \) is a vector from the \( k \)th row of \( \mathbf{L}_1 \): the first \( k \) elements are one, the next \( T - k \) are zero.

\[ k_{i,j} = \alpha \Delta t \min(i, j) \]

define \( \Delta t_i = t_i \) so

\[ k_{i,j} = \alpha \min(t_i, t_j) = k(t_i, t_j) \]
Covariance Functions

Where did this covariance matrix come from?

Markov Process

$$k(t, t') = \alpha \min(t, t')$$

- Covariance matrix is built using the inputs to the function $t$. 
Covariance Functions

Where did this covariance matrix come from?

Markov Process

$$k(t, t') = \alpha \min(t, t')$$

- Covariance matrix is built using the inputs to the function $t$. 

![Graph showing covariance functions]
Markov Process

Visualization of inverse covariance (precision).

- Precision matrix is sparse: only neighbours in matrix are non-zero.
- This reflects *conditional* independencies in data.
- In this case *Markov* structure.
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

\[ k(x, x') = \alpha \exp \left( -\frac{||x - x'||^2}{2\ell^2} \right) \]

- Covariance matrix is built using the inputs to the function \(x\).
- For the example above it was based on Euclidean distance.
- The covariance function is also known as a kernel.
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

\[
k(x, x') = \alpha \exp \left( -\frac{\|x - x'\|^2}{2\ell^2} \right)
\]

▶ Covariance matrix is built using the inputs to the function \(x\).
▶ For the example above it was based on Euclidean distance.
▶ The covariance function is also know as a kernel.
Covariance Functions

Where did this covariance matrix come from?

Exponentiated Quadratic

Visualization of inverse covariance (precision).

- Precision matrix is not sparse.
- Each point is dependent on all the others.
- In this case non-Markovian.
Covariance Functions

Where did this covariance matrix come from?

Markov Process

Visualization of inverse covariance (precision).

- Precision matrix is sparse: only neighbours in matrix are non-zero.
- This reflects *conditional* independencies in data.
- In this case *Markov* structure.
We have state vector \( \mathbf{X} = [x_1 \ldots x_q] \in \mathbb{R}^{T \times q} \) and if each state evolves independently we have

\[
p(\mathbf{X}) = \prod_{i=1}^{q} p(x_{:,i})
\]

\[
p(x_{:,i}) = \mathcal{N}(x_{:,i}|0, \mathbf{K}).
\]

We want to obtain outputs through:

\[
y_{i,:} = Wx_{i,:}
\]
- Represent with a ‘stacked’ system:

\[ p(x) = \mathcal{N}(x|0, I \otimes K) \]

where the stacking is placing each column of \( X \) one on top of another as

\[
\begin{bmatrix}
  x_{:,1} \\
  x_{:,2} \\
  \vdots \\
  x_{:,q}
\end{bmatrix}
\]
Kronecker Product

\[
\begin{bmatrix}
  a & b \\
  c & d \\
\end{bmatrix} \otimes K =
\begin{bmatrix}
  aK & bK \\
  cK & dK \\
\end{bmatrix}
\]
Kronecker Product

\[
\begin{bmatrix}
\begin{array}{cc}
\text{black} & \text{grey} \\
\text{white} & \text{white}
\end{array}
\end{bmatrix} \otimes 
\begin{bmatrix}
\begin{array}{cc}
\text{red} & \text{green} \\
\text{blue} & \text{blue}
\end{array}
\end{bmatrix} =
\begin{bmatrix}
\begin{array}{cccc}
\text{black} & \text{black} & \text{black} & \text{black} \\
\text{black} & \text{black} & \text{black} & \text{black} \\
\text{black} & \text{black} & \text{black} & \text{black} \\
\text{black} & \text{black} & \text{black} & \text{black}
\end{array}
\end{bmatrix}
\]
Represent with a ‘stacked’ system:

\[ p(x) = \mathcal{N}(x|0, I \otimes K) \]

where the stacking is placing each column of \( X \) one on top of another as

\[
    x = \begin{bmatrix}
        x_{:,1} \\
        x_{:,2} \\
        \vdots \\
        x_{:,q}
    \end{bmatrix}
\]
Column Stacking

\[\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \]
For this stacking the marginal distribution over time is given by the block diagonals.
For this stacking the marginal distribution over *time* is given by the block diagonals.
For this stacking the marginal distribution over *time* is given by the block diagonals.
For this stacking the marginal distribution over time is given by the block diagonals.
For this stacking the marginal distribution over *time* is given by the block diagonals.
Can also stack each row of $X$ to form column vector:

$$
x = \begin{bmatrix}
    x_{1,:} \\
    x_{2,:} \\
    \vdots \\
    x_{T,:}
\end{bmatrix}
$$

$$
p(x) = \mathcal{N}(x|0, K \otimes I)
$$
Row Stacking

\[
\begin{array}{c}
\begin{array}{ccc}
\otimes & = \\
\end{array}
\end{array}
\]
For this stacking the marginal distribution over the latent *dimensions* is given by the block diagonals.
For this stacking the marginal distribution over the latent dimensions is given by the block diagonals.
For this stacking the marginal distribution over the latent dimensions is given by the block diagonals.
For this stacking the marginal distribution over the latent dimensions is given by the block diagonals.
For this stacking the marginal distribution over the latent dimensions is given by the block diagonals.
The observations are related to the latent points by a linear mapping matrix,

$$y_{i,:} = Wx_{i,:} + \epsilon_{i,:}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2 I)$$
Mapping from Latent Process to Observed

\[
\begin{bmatrix}
W & 0 & 0 \\
0 & W & 0 \\
0 & 0 & W
\end{bmatrix}
\times
\begin{bmatrix}
x_{1,:} \\
x_{2,:} \\
x_{3,:}
\end{bmatrix}
= 
\begin{bmatrix}
Wx_{1,:} \\
Wx_{2,:) \\
Wx_{3,:}
\end{bmatrix}
\]
Output Covariance

This leads to a covariance of the form

\[(I \otimes W)(K \otimes I)(I \otimes W^T) + I\sigma^2\]

Using \((A \otimes B)(C \otimes D) = AC \otimes BD\) This leads to

\[K \otimes WW^T + I\sigma^2\]

or

\[y \sim \mathcal{N}\left(0, WW^T \otimes K + I\sigma^2\right)\]
Kernels for Vector Valued Outputs: A Review

By Mauricio A. Álvarez,
Lorenzo Rosasco and Neil D. Lawrence
This Kronecker structure leads to several published models.

\[(K(x, x'))_{j,j'} = k(x, x')k_T(j, j'),\]

where \(k\) has \(x\) and \(k_T\) has \(i\) as inputs.

Can think of multiple output covariance functions as covariances with augmented input.

Alongside \(x\) we also input the \(j\) associated with the output of interest.
Taking $\mathbf{B} = \mathbf{W}\mathbf{W}^\top$ we have a matrix expression across outputs.

$$K(x, x') = k(x, x')\mathbf{B},$$

where $\mathbf{B}$ is a $p \times p$ symmetric and positive semi-definite matrix.

$\mathbf{B}$ is called the *coregionalization* matrix.

We call this class of covariance functions *separable* due to their product structure.
In the same spirit a more general class of kernels is given by

\[ K(x, x') = \sum_{j=1}^{q} k_j(x, x')B_j. \]

This can also be written as

\[ K(X, X) = \sum_{j=1}^{q} B_j \otimes k_j(X, X), \]

This is like several Kalman filter-type models added together, but each one with a different set of latent functions.

We call this class of kernels sum of separable kernels (SoS kernels).
Geostatistics

- Use of GPs in Geostatistics is called kriging.
- These multi-output GPs pioneered in geostatistics: prediction over vector-valued output data is known as cokriging.
- The model in geostatistics is known as the *linear model of coregionalization* (LMC, Journel and Huijbregts (1978); Goovaerts (1997)).
- Most machine learning multitask models can be placed in the context of the LMC model.
In the linear model of coregionalization (LMC) outputs are expressed as linear combinations of independent random functions.

In the LMC, each component $f_j$ is expressed as a linear sum

$$f_j(x) = \sum_{j=1}^{q} w_{j,j}u_j(x).$$

where the latent functions are independent and have covariance functions $k_j(x, x')$.

The processes $\{f_j(x)\}_{j=1}^{q}$ are independent for $q \neq j'$. 

Weighted sum of Latent Functions
The Kalman filter is an example of the LMC where $u_i(x) \rightarrow x_i(t)$.

I.e. we’ve moved from time input to a more general input space.

In matrix notation:

1. Kalman filter
   \[ F = WX \]

2. LMC
   \[ F = WU \]

where the rows of these matrices $F$, $X$, $U$ each contain $q$ samples from their corresponding functions at a different time (Kalman filter) or spatial location (LMC).
If one covariance used for latent functions (like in Kalman filter).

This is called the intrinsic coregionalization model (ICM, Goovaerts (1997)).

The kernel matrix corresponding to a dataset $X$ takes the form

$$K(X, X) = B \otimes k(X, X).$$
If outputs are noise-free, maximum likelihood is equivalent to independent fits of $B$ and $k(x, x')$ (Helterbrand and Cressie, 1994).

In geostatistics this is known as autokrigeability (Wackernagel, 2003).

In multitask learning its the cancellation of intertask transfer (Bonilla et al., 2008).
Intrinsic Coregionalization Model

\[ K(X, X) = \mathbf{w}\mathbf{w}^T \otimes k(X, X). \]

\[ \mathbf{w} = \begin{bmatrix} 1 \\ 5 \end{bmatrix} \]

\[ \mathbf{B} = \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix} \]
Intrinsic Coregionalization Model

\[ K(X, X) = \mathbf{w}\mathbf{w}^T \otimes k(X, X). \]

\[ \mathbf{w} = \begin{bmatrix} 1 \\ 5 \end{bmatrix} \]

\[ \mathbf{B} = \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix} \]
Intrinsic Coregionalization Model

\[ K(X, X) = w w^T \otimes k(X, X). \]

\[ w = \begin{bmatrix} 1 \\ 5 \end{bmatrix} \]

\[ B = \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix} \]
Intrinsic Coregionalization Model

$$K(X, X) = w w^T \otimes k(X, X).$$

$$w = \begin{bmatrix} 1 \\ 5 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix}$$
Intrinsic Coregionalization Model

\[ K(X, X) = w w^\top \otimes k(X, X). \]

\[ w = \begin{bmatrix} 1 \\ 5 \end{bmatrix} \]

\[ B = \begin{bmatrix} 1 & 5 \\ 5 & 25 \end{bmatrix} \]
Intrinsic Coregionalization Model

\[ K(X, X) = B \otimes k(X, X). \]

\[ B = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{bmatrix} \]
Intrinsic Coregionalization Model

\[ K(X, X) = B \otimes k(X, X). \]

\[ B = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{bmatrix} \]
Intrinsic Coregionalization Model

\[ K(X, X) = B \otimes k(X, X). \]

\[ B = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{bmatrix} \]
Intrinsic Coregionalization Model

\[
K(X, X) = B \otimes k(X, X).
\]

\[
B = \begin{bmatrix}
1 & 0.5 \\
0.5 & 1.5
\end{bmatrix}
\]
Intrinsic Coregionalization Model

\[ K(X, X) = B \otimes k(X, X). \]

\[ B = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.5 \end{bmatrix} \]
LMC Samples

\[ K(X, X) = B_1 \otimes k_1(X, X) + B_2 \otimes k_2(X, X) \]

\[ B_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.2 \end{bmatrix} \]

\[ \ell_1 = 1 \]

\[ B_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3 \end{bmatrix} \]

\[ \ell_2 = 0.2 \]
\[ \mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{B}_1 \otimes k_1(\mathbf{X}, \mathbf{X}) + \mathbf{B}_2 \otimes k_2(\mathbf{X}, \mathbf{X}) \]

\[
\mathbf{B}_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.2 \end{bmatrix}
\]

\[ \ell_1 = 1 \]

\[
\mathbf{B}_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3 \end{bmatrix}
\]

\[ \ell_2 = 0.2 \]
\[ K(X, X) = B_1 \otimes k_1(X, X) + B_2 \otimes k_2(X, X) \]

\[ B_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.2 \end{bmatrix} \]
\[ \ell_1 = 1 \]

\[ B_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3 \end{bmatrix} \]
\[ \ell_2 = 0.2 \]
LMC Samples

\[ K(X, X) = B_1 \otimes k_1(X, X) + B_2 \otimes k_2(X, X) \]

\[ B_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.2 \end{bmatrix} \]

\[ \ell_1 = 1 \]

\[ B_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3 \end{bmatrix} \]

\[ \ell_2 = 0.2 \]
\[ K(X, X) = B_1 \otimes k_1(X, X) + B_2 \otimes k_2(X, X) \]

\[
B_1 = \begin{bmatrix} 1.4 & 0.5 \\ 0.5 & 1.2 \end{bmatrix}
\]

\[ \ell_1 = 1 \]

\[
B_2 = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1.3 \end{bmatrix}
\]

\[ \ell_2 = 0.2 \]
Used in machine learning for GPs for multivariate regression and in statistics for computer emulation of expensive multivariate computer codes.

Imposes the correlation of the outputs explicitly through the set of coregionalization matrices.

Setting $B = I_p$ assumes outputs are conditionally independent given the parameters $\theta$. (Minka and Picard, 1997; Lawrence and Platt, 2004; Yu et al., 2005).

More recent approaches for multiple output modeling are different versions of the linear model of coregionalization.
Coregionalization matrices are rank 1. Rewrite equation (???) as

\[ K(X, X) = \sum_{j=1}^{q} w_{:,j} w_{:,j}^\top \otimes k_j(X, X). \]

Like the Kalman filter, but each latent function has a different covariance.

Authors suggest using an exponentiated quadratic characteristic length-scale for each input dimension.
Semiparametric Latent Factor Model Samples

\[ K(X, X) = w_{:,1} w_{:,1}^\top \otimes k_1(X, X) + w_{:,2} w_{:,2}^\top \otimes k_2(X, X) \]

\[ w_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \]

\[ w_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} \]
$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathbf{w}_{:,1} \mathbf{w}_{:,1}^\top \otimes \mathbf{k}_1(\mathbf{X}, \mathbf{X}) + \mathbf{w}_{:,2} \mathbf{w}_{:,2}^\top \otimes \mathbf{k}_2(\mathbf{X}, \mathbf{X})$

$\mathbf{w}_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$

$\mathbf{w}_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$
Semiparametric Latent Factor Model Samples

$$K(X, X) = w_{:,1} w_{:,1}^\top \otimes k_1(X, X) + w_{:,2} w_{:,2}^\top \otimes k_2(X, X)$$

$$w_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}$$

$$w_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}$$
Semiparametric Latent Factor Model Samples

\[
K(X, X) = w_{:,1} w_{:,1}^\top \otimes k_1(X, X) + w_{:,2} w_{:,2}^\top \otimes k_2(X, X)
\]

\[
w_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix}
\]

\[
w_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}
\]
Semiparametric Latent Factor Model Samples

\[ K(X, X) = \mathbf{w}_{:,1} \mathbf{w}_{:,1}^\top \otimes k_1(X, X) + \mathbf{w}_{:,2} \mathbf{w}_{:,2}^\top \otimes k_2(X, X) \]

\[ \mathbf{w}_1 = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \]

\[ \mathbf{w}_2 = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix} \]
Bonilla et al. (2008) suggest ICM for multitask learning.
Use a PPCA form for $B$: similar to our Kalman filter example.
Refer to the autokrigeability effect as the cancellation of inter-task transfer.
Also discuss the similarities between the multi-task GP and the ICM, and its relationship to the SLFM and the LMC.
Multitask Classification

- Mostly restricted to the case where the outputs are conditionally independent given the hyperparameters $\phi$ (Minka and Picard, 1997; Williams and Barber, 1998; Lawrence and Platt, 2004; Seeger and Jordan, 2004; Yu et al., 2005; Rasmussen and Williams, 2006).

- Intrinsic coregionalization model has been used in the multiclass scenario. Skolidis and Sanguinetti (2011) use the intrinsic coregionalization model for classification, by introducing a probit noise model as the likelihood.

- Posterior distribution is no longer analytically tractable: approximate inference is required.
Computer Emulation

- A statistical model used as a surrogate for a computationally expensive computer model.
- Higdon et al. (2008) use the linear model of coregionalization to model images representing the evolution of the implosion of steel cylinders.
- In Conti and O’Hagan (2009) use the ICM to model a vegetation model: called the Sheffield Dynamic Global Vegetation Model (Woodward et al., 1998).
Bochner’s Theorem

Given a positive finite Borel measure $\mu$ on the real line $\mathbb{R}$, the Fourier transform $Q$ of $\mu$ is the continuous function

$$Q(t) = \int_{\mathbb{R}} e^{-itx} \, d\mu(x).$$

$Q$ is continuous since for a fixed $x$, the function $e^{-itx}$ is continuous and periodic. The function $Q$ is a positive definite function, i.e. the kernel $k(x, x') = Q(x' - x)$ is positive definite.

Bochner’s theorem says the converse is true, i.e. every positive definite function $Q$ is the Fourier transform of a positive finite Borel measure. A proof can be sketched as follows.
Covariance Functions

Where did this covariance matrix come from?

**Ornstein-Uhlenbeck (stationary Gauss-Markov) covariance function**

\[ k(x, x') = \alpha \exp\left(-\frac{|x - x'|}{2\ell^2}\right) \]

- In one dimension arises from a stochastic differential equation. Brownian motion in a parabolic tube.
- In higher dimension a Fourier filter of the form \( \frac{1}{\pi(1+x^2)} \).
Covariance Functions

Where did this covariance matrix come from?

Ornstein-Uhlenbeck (stationary Gauss-Markov) covariance function

\[ k(x, x') = \alpha \exp\left(-\frac{|x - x'|}{2\ell^2}\right) \]

- In one dimension arises from a stochastic differential equation. Brownian motion in a parabolic tube.
- In higher dimension a Fourier filter of the form \[ \frac{1}{\pi(1+x^2)}. \]
Matern 3/2 Covariance Function

\[ k(x, x') = \alpha \left(1 + \sqrt{3}r\right) \exp\left(-\sqrt{3}r\right) \quad \text{where} \quad r = \frac{||x - x'||_2}{\ell} \]

- Matern 3/2 is a once differentiable covariance.
- Matern family constructed with Student-t filters in Fourier space.
Covariance Functions

Where did this covariance matrix come from?

**Matern 3/2 Covariance Function**

\[
k(x, x') = \alpha \left( 1 + \sqrt{3}r \right) \exp \left( -\sqrt{3}r \right) \quad \text{where} \quad r = \frac{\|x - x'\|_2}{\ell}
\]

- Matern 3/2 is a once differentiable covariance.
- Matern family constructed with Student-\(t\) filters in Fourier space.
Matern 5/2 Covariance Function

\[ k(x, x') = \alpha \left( 1 + \sqrt{5}r + \frac{5}{3}r^2 \right) \exp \left( - \sqrt{5}r \right) \quad \text{where} \quad r = \frac{\|x - x'\|_2}{\ell} \]

- Matern 5/2 is a twice differentiable covariance.
- Matern family constructed with Student-\(t\) filters in Fourier space.
Matern 5/2 Covariance Function

\[ k(x, x') = \alpha \left( 1 + \sqrt{5}r + \frac{5}{3}r^2 \right) \exp \left( -\sqrt{5}r \right) \]

where \( r = \frac{\|x - x'\|_2}{\ell} \)

- Matern 5/2 is a twice differentiable covariance.
- Matern family constructed with Student-\( t \) filters in Fourier space.
Covariance Functions
Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

\[ k(x, x') = \alpha \exp \left( -\frac{\|x - x'\|^2}{2\ell^2} \right) \]

- Covariance matrix is built using the inputs to the function \( x \).
- For the example above it was based on Euclidean distance.
- The covariance function is also known as a kernel.
Covariance Functions
Where did this covariance matrix come from?

Exponentiated Quadratic Kernel Function (RBF, Squared Exponential, Gaussian)

\[ k(x, x') = \alpha \exp \left( -\frac{||x - x'||^2}{2\ell^2} \right) \]

- Covariance matrix is built using the inputs to the function \( x \).
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.
Outline

Gaussian Processes

GP Non-Gaussian

GP Limitations

Kalman Filter

Dimensionality Reduction
Motivation for Non-Linear Dimensionality Reduction

USPS Data Set Handwritten Digit

- 3648 Dimensions
  - 64 rows by 57 columns

![Handwritten Digit]
Motivation for Non-Linear Dimensionality Reduction

USPS Data Set Handwritten Digit

- 3648 Dimensions
  - 64 rows by 57 columns
  - Space contains more than just this digit.
Motivation for Non-Linear Dimensionality Reduction

USPS Data Set Handwritten Digit

- 3648 Dimensions
  - 64 rows by 57 columns
  - Space contains more than just this digit.
  - Even if we sample every nanosecond from now until the end of the universe, you won’t see the original six!
Motivation for Non-Linear Dimensionality Reduction

USPS Data Set Handwritten Digit

- 3648 Dimensions
  - 64 rows by 57 columns
  - Space contains more than just this digit.
  - Even if we sample every nanosecond from now until the end of the universe, you won’t see the original six!
Simple Model of Digit

Rotate a 'Prototype'
Simple Model of Digit

Rotate a 'Prototype'
Simple Model of Digit

Rotate a 'Prototype'
Simple Model of Digit

Rotate a 'Prototype'
Simple Model of Digit

Rotate a 'Prototype'
Simple Model of Digit

Rotate a 'Prototype'
Simple Model of Digit

Rotate a 'Prototype'
Simple Model of Digit

Rotate a 'Prototype'
Simple Model of Digit

Rotate a 'Prototype'
MATLAB Demo

demDigitsManifold([1 2], 'all')
MATLAB Demo

demDigitsManifold([1 2], 'all')
demDigitsManifold([1 2], 'sixnine')
Low Dimensional Manifolds

Pure Rotation is too Simple

- In practice the data may undergo several distortions.
  - *e.g.* digits undergo ‘thinning’, translation and rotation.
- For data with ‘structure’:
  - we expect fewer distortions than dimensions;
  - we therefore expect the data to live on a lower dimensional manifold.
- Conclusion: deal with high dimensional data by looking for lower dimensional non-linear embedding.
Existing Methods

Spectral Approaches

- Classical Multidimensional Scaling (MDS) (Mardia et al., 1979).
  - Uses eigenvectors of similarity matrix.
    - Isomap (Tenenbaum et al., 2000) is MDS with a particular proximity measure.
- Kernel PCA (Schölkopf et al., 1998)
  - Provides a representation and a mapping — dimensional expansion.
  - Mapping is implied through the use of a kernel function as a similarity matrix.
- Locally Linear Embedding (Roweis and Saul, 2000).
  - Looks to preserve locally linear relationships in a low dimensional space.
Existing Methods II

Iterative Methods

- Multidimensional Scaling (MDS)
  - Iterative optimisation of a stress function (Kruskal, 1964).
  - Sammon Mappings (Sammon, 1969).
    - Strictly speaking not a mapping — similar to iterative MDS.
- NeuroScale (Lowe and Tipping, 1997)
  - Augmentation of iterative MDS methods with a mapping.
Existing Methods III

Probabilistic Approaches

- Probabilistic PCA *(Tipping and Bishop, 1999; Roweis, 1998)*
  - A linear method.
Probabilistic Approaches

- **Probabilistic PCA** (Tipping and Bishop, 1999; Roweis, 1998)
  - A linear method.

- **Density Networks** (MacKay, 1995)
  - Use importance sampling and a multi-layer perceptron.
Probabilistic Approaches

- **Probabilistic PCA** *(Tipping and Bishop, 1999; Roweis, 1998)*
  - A linear method.
- **Density Networks** *(MacKay, 1995)*
  - Use importance sampling and a multi-layer perceptron.
- **Generative Topographic Mapping (GTM)** *(Bishop et al., 1998)*
  - Uses a grid based sample and an RBF network.
Existing Methods III

Probabilistic Approaches

- **Probabilistic PCA** (Tipping and Bishop, 1999; Roweis, 1998)
  - A linear method.

- **Density Networks** (MacKay, 1995)
  - Use importance sampling and a multi-layer perceptron.

- **Generative Topographic Mapping (GTM)** (Bishop et al., 1998)
  - Uses a grid based sample and an RBF network.

Difficulty for Probabilistic Approaches

- Propagate a probability distribution through a non-linear mapping.
The New Model

A Probabilistic Non-linear PCA

- PCA has a probabilistic interpretation (Tipping and Bishop, 1999; Roweis, 1998).
- It is difficult to ‘non-linearise’.

Dual Probabilistic PCA

- We present a new probabilistic interpretation of PCA (Lawrence, 2005).
- This interpretation can be made non-linear.
- The result is non-linear probabilistic PCA.
**Notation**

- $q$ — dimension of latent/embedded space
- $p$ — dimension of data space
- $n$ — number of data points

Centred data, $Y = [y_{1:}, \ldots, y_{n:}]^\top = [y_{:,1}, \ldots, y_{:,p}] \in \mathbb{R}^{n \times p}$

Latent variables, $X = [x_{1:}, \ldots, x_{n:}]^\top = [x_{:,1}, \ldots, x_{:,q}] \in \mathbb{R}^{n \times q}$

Mapping matrix, $W \in \mathbb{R}^{p \times q}$

- $a_{i:}$ is a vector from the $i$th row of a given matrix $A$
- $a_{:,j}$ is a vector from the $j$th row of a given matrix $A$
X and Y are design matrices

- Covariance given by $n^{-1}Y^TY$.
- Inner product matrix given by $YY^T$. 
Linear Latent Variable Model

- Represent data, $Y$, with a lower dimensional set of latent variables $X$.
- Assume a linear relationship of the form

$$y_{i:} = Wx_{i:} + \epsilon_{i:},$$

where

$$\epsilon_{i:} \sim \mathcal{N}(0, \sigma^2 I).$$
Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.

\[
p(Y|X, W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|Wx_{i,:}, \sigma^2 I)
\]
Linear Latent Variable Model

Probabilistic PCA

- Define linear-Gaussian relationship between latent variables and data.

- **Standard** Latent variable approach:

\[ p(Y|X, W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|Wx_{i,:}, \sigma^2 I) \]
Linear Latent Variable Model

Probabilistic PCA

- Define linear-Gaussian relationship between latent variables and data.
- **Standard** Latent variable approach:
  - Define Gaussian prior over latent space, X.

\[
p(Y|X, W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|Wx_{i,:}, \sigma^2 I)
\]

\[
p(X) = \prod_{i=1}^{n} \mathcal{N}(x_{i,:}|0, I)
\]
Linear Latent Variable Model

Probabilistic PCA

- Define linear-Gaussian relationship between latent variables and data.

- **Standard** Latent variable approach:
  - Define Gaussian prior over latent space, $X$.
  - Integrate out latent variables.

\[
p(Y|X, W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|Wx_{i,:}, \sigma^2 I)
\]

\[
p(X) = \prod_{i=1}^{n} \mathcal{N}(x_{i,:}|0, I)
\]

\[
p(Y|W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|0, WW^T + \sigma^2 I)
\]
Computation of the Marginal Likelihood

\[ y_{i,:} = Wx_{i,:} + \epsilon_{i,:}, \quad x_{i,:} \sim \mathcal{N}(0, I), \quad \epsilon_{i,:} \sim \mathcal{N}(0, \sigma^2 I) \]
Computation of the Marginal Likelihood

\[ y_{i,:} = W x_{i,:} + \epsilon_{i,:}, \quad x_{i,:} \sim \mathcal{N}(0, I), \quad \epsilon_{i,:} \sim \mathcal{N}(0, \sigma^2 I) \]

\[ W x_{i,:} \sim \mathcal{N}(0, WW^\top), \]
Computation of the Marginal Likelihood

\[
y_{i,:} = Wx_{i,:} + \epsilon_{i,:}, \quad x_{i,:] \sim \mathcal{N}(0,I), \quad \epsilon_{i,:} \sim \mathcal{N}(0,\sigma^2I)
\]

\[
Wx_{i,:} \sim \mathcal{N}(0,WW^\top),
\]

\[
Wx_{i,:} + \epsilon_{i,:} \sim \mathcal{N}(0,WW^\top + \sigma^2I)
\]
Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

\[ p(Y|W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|0, WW^T + \sigma^2 I) \]
Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

\[ p(Y|W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|0, C), \quad C = WW^T + \sigma^2 I \]
Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

\[
p(Y|W) = \prod_{i=1}^{n} \mathcal{N}(y_i; |0, C), \quad C = WW^\top + \sigma^2 I
\]

\[
\log p(Y|W) = -\frac{n}{2} \log |C| - \frac{1}{2} \text{tr}(C^{-1}Y^\top Y) + \text{const.}
\]
Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

\[
p(Y|W) = \prod_{i=1}^{n} \mathcal{N}(y_{i:}|0, C), \quad C = WW^\top + \sigma^2 I
\]

\[
\log p(Y|W) = -\frac{n}{2} \log |C| - \frac{1}{2} \text{tr} \left( C^{-1}Y^\top Y \right) + \text{const.}
\]

If \( U_q \) are first \( q \) principal eigenvectors of \( n^{-1}Y^\top Y \) and the corresponding eigenvalues are \( \Lambda_q \),
Probabilistic PCA Max. Likelihood Soln (Tipping and Bishop, 1999)

\[
p(Y|W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|0, C), \quad C = WW^\top + \sigma^2 I
\]

\[
\log p(Y|W) = -\frac{n}{2} \log |C| - \frac{1}{2} \text{tr}\left(C^{-1}Y^\top Y\right) + \text{const.}
\]

If \(U_q\) are first \(q\) principal eigenvectors of \(n^{-1}Y^\top Y\) and the corresponding eigenvalues are \(\Lambda_q\),

\[
W = U_q L R^\top, \quad L = \left(\Lambda_q - \sigma^2 I\right)^{\frac{1}{2}}
\]

where \(R\) is an arbitrary rotation matrix.
Dual Probabilistic PCA

- Define linear-Gaussian relationship between latent variables and data.

\[ p(Y|X, W) = \prod_{i=1}^{n} N(y_{i,:}|Wx_{i,:}, \sigma^2 I) \]
Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.

- **Novel** Latent variable approach:

\[
p(Y|X, W) = \prod_{i=1}^{n} N(y_{i,:}|Wx_{i,:}, \sigma^2 I)
\]
Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.

- **Novel** Latent variable approach:
  - Define Gaussian prior over parameters, $W$.

\[
p(Y|X, W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|w_{x_{i,:}}, \sigma^2 I)
\]

\[
p(W) = \prod_{i=1}^{p} \mathcal{N}(w_{i,:}|0, I)
\]
Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.

- **Novel** Latent variable approach:
  - Define Gaussian prior over *parameters*, $W$.
  - Integrate out *parameters*.

\[
p(Y|X, W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|Wx_{i,:}, \sigma^2 I)
\]

\[
p(W) = \prod_{i=1}^{p} \mathcal{N}(w_{i,:}|0, I)
\]

\[
p(Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{j,:}|0, XX^\top + \sigma^2 I)
\]
Computation of the Marginal Likelihood

\[ y_{:,j} = Xw_{:,j} + \epsilon_{:,j}, \quad w_{:,j} \sim \mathcal{N}(0, I), \quad \epsilon_{i,:) \sim \mathcal{N}(0, \sigma^2 I) \]
Computation of the Marginal Likelihood

\[ y_{:,j} = Xw_{:,j} + \epsilon_{:,j}, \quad w_{:,j} \sim \mathcal{N}(0, I), \quad \epsilon_{i,:} \sim \mathcal{N}(0, \sigma^2 I) \]

\[ Xw_{:,j} \sim \mathcal{N}(0, XX^\top), \]
Computation of the Marginal Likelihood

\[ y_{:,j} = Xw_{:,j} + \epsilon_{:,j}, \quad w_{:,j} \sim \mathcal{N}(0, I), \quad \epsilon_{i:,} \sim \mathcal{N}(0, \sigma^2 I) \]

\[ Xw_{:,j} \sim \mathcal{N}(0, XX^\top) , \]

\[ Xw_{:,j} + \epsilon_{:,j} \sim \mathcal{N}(0, XX^\top + \sigma^2 I) \]
Dual Probabilistic PCA Max. Likelihood Soln (Lawrence, 2004, 2005)

\[
p(Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{j}|0, XX^\top + \sigma^2 I)
\]
Dual PPCA Max. Likelihood Soln (Lawrence, 2004, 2005)

\[ p(Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{:,j}|0, K), \quad K = XX^T + \sigma^2 I \]
**Linear Latent Variable Model IV**

**PPCA Max. Likelihood Soln** (Tipping and Bishop, 1999)

\[
p(Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{:,j}|0, K), \quad K = XX^\top + \sigma^2 I
\]

\[
\log p(Y|X) = -\frac{p}{2} \log |K| - \frac{1}{2} \text{tr}(K^{-1}YY^\top) + \text{const.}
\]
PPCA Max. Likelihood Soln

\[ p (Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{:,j}|0, K), \quad K = XX^\top + \sigma^2 I \]

\[ \log p (Y|X) = -\frac{p}{2} \log |K| - \frac{1}{2} \text{tr} \left( K^{-1} YY^\top \right) + \text{const.} \]

If \( U'_q \) are first \( q \) principal eigenvectors of \( p^{-1} YY^\top \) and the corresponding eigenvalues are \( \Lambda_q \),
PPCA Max. Likelihood Soln

\[
p(Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{:,j}|0, K), \quad K = XX^\top + \sigma^2 I
\]

\[
\log p(Y|X) = -\frac{p}{2} \log |K| - \frac{1}{2} \text{tr} \left( K^{-1} YY^\top \right) + \text{const.}
\]

If \( U' \) are first \( q \) principal eigenvectors of \( p^{-1} YY^\top \) and the corresponding eigenvalues are \( \Lambda_q \),

\[
X = U'_q LR^\top, \quad L = \left( \Lambda_q - \sigma^2 I \right)^{\frac{1}{2}}
\]

where \( R \) is an arbitrary rotation matrix.
Dual PPCA Max. Likelihood Soln (Lawrence, 2004, 2005)

\[
p(Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{:,j}|0, K) , \quad K = XX^\top + \sigma^2 I
\]

\[
\log p(Y|X) = -\frac{p}{2} \log |K| - \frac{1}{2} \text{tr}(K^{-1}YY^\top) + \text{const.}
\]

If \(U'_q\) are first \(q\) principal eigenvectors of \(p^{-1}YY^\top\) and the corresponding eigenvalues are \(\Lambda_q\),

\[
X = U'_q LR^\top , \quad L = \left(\Lambda_q - \sigma^2 I\right)^{\frac{1}{2}}
\]

where \(R\) is an arbitrary rotation matrix.
PPCA Max. Likelihood Soln  (Tipping and Bishop, 1999)

\[
p (Y|W) = \prod_{i=1}^{n} \mathcal{N} (y_i; 0, C), \quad C = WW^T + \sigma^2 I
\]

\[
\log p (Y|W) = -\frac{n}{2} \log |C| - \frac{1}{2} \text{tr} \left( C^{-1} Y^T Y \right) + \text{const.}
\]

If \( U_q \) are first \( q \) principal eigenvectors of \( n^{-1} Y^T Y \) and the corresponding eigenvalues are \( \Lambda_q \),

\[
W = U_q LR^T, \quad L = \left( \Lambda_q - \sigma^2 I \right)^{\frac{1}{2}}
\]

where \( R \) is an arbitrary rotation matrix.
The Eigenvalue Problems are equivalent

- Solution for Probabilistic PCA (solves for the mapping)
  \[ Y^T Y U_q = U_q \Lambda_q \quad W = U_q L R^T \]

- Solution for Dual Probabilistic PCA (solves for the latent positions)
  \[ Y Y^T U'_q = U'_q \Lambda_q \quad X = U'_q L R^T \]

- Equivalence is from
  \[ U_q = Y^T U'_q \Lambda_q^{-\frac{1}{2}} \]
Non-Linear Latent Variable Model

Dual Probabilistic PCA

- Define *linear-Gaussian relationship* between latent variables and data.
- **Novel** Latent variable approach:
  - Define Gaussian prior over parameters, \( W \).
  - Integrate out parameters.

\[
p(Y|X, W) = \prod_{i=1}^{n} \mathcal{N}(y_{i,:}|Wx_{i,:}, \sigma^2 I)
\]

\[
p(W) = \prod_{i=1}^{p} \mathcal{N}(w_{i,:}|0, I)
\]

\[
p(Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{j,:}|0, XX^\top + \sigma^2 I)
\]
Non-Linear Latent Variable Model

Dual Probabilistic PCA

- Inspection of the marginal likelihood shows ...

\[
p (Y|X) = \prod_{j=1}^{p} \mathcal{N} (y_{:,j}|0, XX^T + \sigma^2 I)
\]
Non-Linear Latent Variable Model

Dual Probabilistic PCA

- Inspection of the marginal likelihood shows ...
  - The covariance matrix is a covariance function.

\[
p(Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_j|0, K)
\]

\[
K = XX^T + \sigma^2 I
\]
Non-Linear Latent Variable Model

Dual Probabilistic PCA

- Inspection of the marginal likelihood shows ...
  - The covariance matrix is a covariance function.
  - We recognise it as the ‘linear kernel’.

\[
p (Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{:,j}|0, K)
\]

\[
K = XX^T + \sigma^2 I
\]

This is a product of Gaussian processes with linear kernels.
Non-Linear Latent Variable Model

**Dual Probabilistic PCA**

- Inspection of the marginal likelihood shows ...
  - The covariance matrix is a covariance function.
  - We recognise it as the ‘linear kernel’.
  - We call this the Gaussian Process Latent Variable model (GP-LVM).

Replace linear kernel with non-linear kernel for non-linear model.
Exponentiated Quadratic (EQ) Covariance

- The EQ covariance has the form $k_{i,j} = k\left(x_{i,:}, x_{j,:}\right)$, where

$$
k\left(x_{i,:}, x_{j,:}\right) = \alpha \exp\left(-\frac{\|x_{i,:) - x_{j,:}\|^2}{2\ell^2}\right).
$$

- No longer possible to optimise wrt $X$ via an eigenvalue problem.
- Instead find gradients with respect to $X, \alpha, \ell$ and $\sigma^2$ and optimise using conjugate gradients.
Applications

Style Based Inverse Kinematics

▷ Facilitating animation through modeling human motion
  (Grochow et al., 2004)

Tracking

▷ Tracking using human motion models (Urtasun et al., 2005, 2006)

Assisted Animation

▷ Generalizing drawings for animation (Baxter and Anjyo, 2006)

Shape Models

▷ Inferring shape (e.g. pose from silhouette). (Ek et al., 2008b,a; Priacuriu and Reid, 2011a,b)
Example: Latent Doodle Space

(Baxter and Anjyo, 2006)
Generalization with much less Data than Dimensions

- Powerful uncertainly handling of GPs leads to surprising properties.
- Non-linear models can be used where there are fewer data points than dimensions without overfitting.
We introduce a prior that is based on the Fisher criteria

\[ p(X) \propto \exp \left\{ -\frac{1}{\sigma_d^2} \text{tr} \left( S_w^{-1} S_b \right) \right\}, \]

with \( S_b \) the between class matrix and \( S_w \) the within class matrix.
Prior for Supervised Learning

(Urtasun and Darrell, 2007)

We introduce a prior that is based on the Fisher criteria

\[
p(X) \propto \exp \left\{ - \frac{1}{\sigma^2} \text{tr} \left( S_w^{-1} S_b \right) \right\},
\]

with \( S_b \) the between class matrix and \( S_w \) the within class matrix

\[
S_w = \sum_{i=1}^{L} \frac{n_i}{n} (M_i - M_0)(M_i - M_0)^\top
\]

where \( X^{(i)} = [x_1^{(i)}, \ldots, x_{n_i}^{(i)}] \) are the \( n_i \) training points of class \( i \), \( M_i \) is the mean of the elements of class \( i \), and \( M_0 \) is the mean of all the training points of all classes.
We introduce a prior that is based on the Fisher criteria

\[ p(X) \propto \exp \left\{ -\frac{1}{\sigma^2} \text{tr} \left( S_w^{-1} S_b \right) \right\}, \]

with \( S_b \) the between class matrix and \( S_w \) the within class matrix

\[ S_w = \sum_{i=1}^{L} \frac{n_i}{n} (M_i - M_0)(M_i - M_0)^\top \]

\[ S_b = \sum_{i=1}^{L} \frac{n_i}{n} \left[ \frac{1}{n_i} \sum_{k=1}^{n_i} (x_{ik} - M_i)(x_{ik} - M_i)^\top \right] \]

where \( X^{(i)} = [x_{1}^{(i)}, \ldots, x_{n_i}^{(i)}] \) are the \( n_i \) training points of class \( i \), \( M_i \) is the mean of the elements of class \( i \), and \( M_0 \) is the mean of all training points of all classes.
We introduce a prior that is based on the Fisher criteria

\[ p(X) \propto \exp \left\{ -\frac{1}{\sigma_d^2} \text{tr} \left( S_w^{-1} S_b \right) \right\}, \]

with \( S_b \) the between class matrix and \( S_w \) the within class matrix.
First system to surpass human performance on cropped Learning Faces in Wild Data.
http://tinyurl.com/nkt9a38

Lots of feature engineering, followed by a Discriminative GP-LVM.

Figure 4: The ROC curve on LFW. Our method achieves the best performance, beating human-level performance.

Figure 5: The two rows present examples of matched and mismatched pairs respectively from LFW that were incorrectly classified by the GaussianFace model.

Conclusion and Future Work

This paper presents a principled Multi-Task Learning approach based on Gaussian Processes and Gaussian Process Latent Variable Models, for face verification. The GaussianFace model achieves high performance on real-world face verification tasks, surpassing human-level performance on some specific scenarios. Future work includes developing more efficient and scalable algorithms for large-scale face verification tasks, as well as exploring the use of additional data from multiple source domains to improve performance.
Graph diffusion prior for enforcing connectivity between motions.

\[
\log p(X) = w_c \sum_{i,j} \log K_{ij}^d
\]

with the graph diffusion kernel \( K_{ij}^d \) obtained from

\[
K_{ij}^d = \exp(\beta H) \quad \text{with} \quad H = -T^{-1/2}LT^{-1/2}
\]

the graph Laplacian, and \( T \) is a diagonal matrix with

\[
T_{ii} = \sum_j w(x_i, x_j),
\]

\[
L_{ij} = \begin{cases} 
\sum_k w(x_i, x_k) & \text{if } i = j \\
-w(x_i, x_j) & \text{otherwise.}
\end{cases}
\]

and \( w(x_i, x_j) = ||x_i - x_j||^{-p} \) measures similarity.
Character Control: Results
Other Topics

- Local distance preservation ▶ Details
- Dynamical models ▶ Details
- Hierarchical models ▶ Details
- Bayesian GP-LVM ▶ Details
Local Distance Preservation \textsuperscript{(Lawrence and Quiñonero Candela, 2006)}

- Most dimensional reduction techniques preserve local distances.
- The GP-LVM does not.
- GP-LVM maps smoothly from latent to data space.
  - Points close in latent space are close in data space.
  - This does not imply points close in data space are close in latent space.
- Kernel PCA maps smoothly from data to latent space.
  - Points close in data space are close in latent space.
  - This does not imply points close in latent space are close in data space.
Forward Mapping (**demBackMapping** in oxford toolbox)

- Mapping from 1-D latent space to 2-D data space.

\[
y_1 = x^2 - 0.5, \quad y_2 = -x^2 + 0.5
\]
Forward Mapping (demBackMapping in oxford toolbox)

- Mapping from 1-D latent space to 2-D data space.

\[ y_1 = x^2 - 0.5, \quad y_2 = -x^2 + 0.5 \]
Forward Mapping (**demBackMapping** in oxford toolbox)

- Mapping from 1-D latent space to 2-D data space.

\[ y_1 = x^2 - 0.5, \quad y_2 = -x^2 + 0.5 \]
Backward Mapping (demBackMapping in oxford toolbox)

- Mapping from 2-D data space to 1-D latent.

\[ x = 0.5\left(y_1^2 + y_2^2 + 1\right) \]
Backward Mapping (demBackMapping in oxford toolbox)

- Mapping from 2-D data space to 1-D latent.

\[ x = 0.5(y_1^2 + y_2^2 + 1) \]
Back Constraints II

Backward Mapping (demBackMapping in oxford toolbox)

- Mapping from 2-D data space to 1-D latent.

\[ x = 0.5(y_1^2 + y_2^2 + 1) \]
Multi-Dimensional Scaling with a Mapping

- Lowe and Tipping (1997) made latent positions a function of the data.
  \[ x_{i,j} = f_j (y_{i,:}; v) \]

- Function was either multi-layer perceptron or a radial basis function network.

- Their motivation was different from ours:
  - They wanted to add the advantages of a true mapping to multi-dimensional scaling.
Back Constraints

- We can use the same idea to force the GP-LVM to respect local distances. (Lawrence and Quiñonero Candela, 2006)
  - By constraining each $x_i$ to be a ‘smooth’ mapping from $y_i$, local distances can be respected.

- This works because in the GP-LVM we maximise wrt latent variables, we don’t integrate out.

- Can use any ‘smooth’ function:
  1. Neural network.
  2. RBF Network.
Computing Gradients

- GP-LVM normally proceeds by optimising

\[ L(X) = \log p(Y|X) \]

with respect to \( X \) using \( \frac{dL}{dX} \).

- The back constraints are of the form

\[ x_{i,j} = f_j(y_{i,:}; v) \]

where \( v \) are parameters.

- We can compute \( \frac{dL}{dv} \) via chain rule and optimise parameters of mapping.
Motion Capture Results

demStick1 and demStick3

Figure: The latent space for the motion capture data with (right) and without (left) back constraints.
Motion Capture Results

demStick1 and demStick3

Figure: The latent space for the motion capture data with (right) and without (left) back constraints.
Projection into data space from four points in the latent space. The inclination of the runner changes becoming more upright.
MAP Solutions for Dynamics Models

- Data often has a temporal ordering.
- Markov-based dynamics are often used.
- For the GP-LVM
  - Marginalising such dynamics is intractable.
  - But: MAP solutions are trivial to implement.
- Many choices: Kalman filter, Markov chains etc..
Gaussian Process Dynamics

GP-LVM with Dynamics

- Autoregressive Gaussian process mapping in latent space between time points.
GP-LVM with Dynamics

- Autoregressive Gaussian process mapping in latent space between time points.
GP-LVM with Dynamics

- Autoregressive Gaussian process mapping in latent space between time points.
Motion Capture Results

demStick1 and demStick2

**Figure:** The latent space for the motion capture data without dynamics (*left*), with auto-regressive dynamics (*right*) based on an exponentiated quadratic kernel.
Motion Capture Results

demStick1 and demStick2

Figure: The latent space for the motion capture data without dynamics (left), with auto-regressive dynamics (right) based on an exponentiated quadratic kernel.
Regressive Dynamics

Inner Groove Distortion

▶ Autoregressive unimodal dynamics, $p(x_t|x_{t-1})$.
▶ Forces spiral visualisation.
▶ Poorer model due to inner groove distortion.
Direct use of Time Variable

- Instead of auto-regressive dynamics, consider regressive dynamics.
- Take $t$ as an input, use a prior $p(X|t)$.
- Use a Gaussian process prior for $p(X|t)$.
- Also allows us to consider variable sample rate data.
Motion Capture Results

demStick1, demStick2 and demStick5

Figure: The latent space for the motion capture data without
dynamics (left), with auto-regressive dynamics (middle) and with
recessive dynamics (right) based on an exponentiated quadratic
kernel.
Motion Capture Results

demStick1, demStick2 and demStick5

Figure: The latent space for the motion capture data without dynamics (left), with auto-regressive dynamics (middle) and with regressive dynamics (right) based on an exponentiated quadratic kernel.
Hierarchical GP-LVM

Stacking Gaussian Processes

- Regressive dynamics provides a simple hierarchy.
  - The input space of the GP is governed by another GP.
- By stacking GPs we can consider more complex hierarchies.
- Ideally we should marginalise latent spaces
  - In practice we seek MAP solutions.
Two Correlated Subjects

(Lawrence and Moore, 2007)

Figure: Hierarchical model of a 'high five'.
Within Subject Hierarchy

Decomposition of Body

Figure: Decomposition of a subject.
Figure: Hierarchical model of a walk and a run.

(Lawrence and Moore, 2007)
Selecting Data Dimensionality

- GP-LVM Provides probabilistic non-linear dimensionality reduction.
- How to select the dimensionality?
- Need to estimate marginal likelihood.
- In standard GP-LVM it increases with increasing $q$. 
Integrate Mapping Function and Latent Variables

Bayesian GP-LVM

- Start with a standard GP-LVM.

\[
p(Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{:,j}|0, K)
\]
Bayesian GP-LVM

- Start with a standard GP-LVM.
- Apply standard latent variable approach:
  - Define Gaussian prior over latent space, $X$.

$$p(Y|X) = \prod_{j=1}^{p} \mathcal{N}(y_{*,j}|0, K)$$
Integrate Mapping Function and Latent Variables

Bayesian GP-LVM

- Start with a standard GP-LVM.
- Apply standard latent variable approach:
  - Define Gaussian prior over latent space, $X$.
  - Integrate out latent variables.

![Diagram]

\[
p(Y|X) = \prod_{j=1}^{p} N(y_{:,j}|0, K)
\]

\[
p(X) = \prod_{j=1}^{q} N(x_{:,j}|0, \alpha_i^{-2}I)
\]
Bayesian GP-LVM

- Start with a standard GP-LVM.
- Apply standard latent variable approach:
  - Define Gaussian prior over latent space, $X$.
  - Integrate out latent variables.
  - Unfortunately integration is intractable.

\[
p(Y|X) = \prod_{j=1}^{p} N(y_{.,j}|0, K)
\]
\[
p(X) = \prod_{j=1}^{q} N(x_{.,j}|0, \alpha_i^{-2}I)
\]
\[
p(Y|\alpha) = ??
\]
Standard Variational Approach Fails

- Standard variational bound has the form:

\[
\mathcal{L} = \langle \log p(y|X) \rangle_{q(X)} + \text{KL} (q(X) \parallel p(X))
\]
Standard Variational Approach Fails

- Standard variational bound has the form:

\[ \mathcal{L} = \langle \log p(y|X) \rangle_{q(X)} + \text{KL} (q(X) \parallel p(X)) \]

- Requires expectation of \( \log p(y|X) \) under \( q(X) \).

\[
\log p(y|X) = -\frac{1}{2} y^\top \left( K_{f,f} + \sigma^2 I \right)^{-1} y - \frac{1}{2} \log |K_{f,f} + \sigma^2 I| - \frac{n}{2} \log 2\pi
\]

Extremely difficult to compute because \( K_{f,f} \) is dependent on \( X \) and appears in the inverse.
Standard Variational Approach Fails

- Standard variational bound has the form:

\[
\mathcal{L} = \langle \log p(y|X) \rangle_{q(X)} + \text{KL} (q(X) \parallel p(X))
\]

- Requires expectation of \( \log p(y|X) \) under \( q(X) \).

\[
\log p(y|X) = -\frac{1}{2} y^\top (K_{f,f} + \sigma^2 I)^{-1} y - \frac{1}{2} \log |K_{f,f} + \sigma^2 I| - \frac{n}{2} \log 2\pi
\]

- Extremely difficult to compute because \( K_{f,f} \) is dependent on \( X \) and appears in the inverse.
Consider collapsed variational bound,

\[
p(y) \geq \prod_{i=1}^{n} c_i \int \mathcal{N}(y|\langle f \rangle, \sigma^2 I) p(u)du
\]
Consider collapsed variational bound,

\[ p(y|X) \geq \prod_{i=1}^{n} c_i \int \mathcal{N}(y|\langle f \rangle_{p(f|u,X)}, \sigma^2 I) p(u) du \]
Consider collapsed variational bound,
\[ \int p(y | X) p(X) dX \geq \int \prod_{i=1}^{n} c_i \mathcal{N}(y | \langle f \rangle_{p(f|u,X)}, \sigma^2 I) p(X) dX p(u) du \]
Consider collapsed variational bound,

\[
\int p(y|X)p(X)dX \geq \int \prod_{i=1}^{n} c_i \mathcal{N}(y|\langle f \rangle_{p(f|u,X)}, \sigma^2 I) p(X)dX p(u)du
\]

Apply variational lower bound to the inner integral.
Consider collapsed variational bound,

\[
\int p(y|X)p(X)\,dX \geq \int \prod_{i=1}^{n} c_i \mathcal{N}(y|\langle f \rangle_{p(f|u,X)}, \sigma^2 I) \, p(X)\,dX p(u)\,du
\]

Apply variational lower bound to the inner integral.

\[
\int \prod_{i=1}^{n} c_i \mathcal{N}(y|\langle f \rangle_{p(f|u,X)}, \sigma^2 I) \, p(X)\,dX \\
\geq \left\langle \sum_{i=1}^{n} \log c_i \right\rangle_{q(X)} \\
+ \left\langle \log \mathcal{N}(y|\langle f \rangle_{p(f|u,X)}, \sigma^2 I) \right\rangle_{q(X)} \\
+ \text{KL} (q(X) \parallel p(X))
\]
Consider collapsed variational bound,

\[ \int p(y|X)p(X)dX \geq \int \prod_{i=1}^{n} c_i \mathcal{N} \left( y|\langle f \rangle_{p(f|u,X)}, \sigma^2 \right) p(X)dXp(u)du \]

Apply variational lower bound to the inner integral.

\[ \int \prod_{i=1}^{n} c_i \mathcal{N} \left( y|\langle f \rangle_{p(f|u,X)}, \sigma^2 \right) p(X)dX \geq \left\langle \sum_{i=1}^{n} \log c_i \right\rangle_{q(X)} + \left\langle \log \mathcal{N} \left( y|\langle f \rangle_{p(f|u,X)}, \sigma^2 \right) \right\rangle_{q(X)} + KL \left( q(X) \| p(X) \right) \]

Which is analytically tractable for Gaussian \( q(X) \) and some covariance functions.
Required Expectations

- Need expectations under $q(X)$ of:

$$\log c_i = \frac{1}{2\sigma^2} \left[ k_{i,i} - k_{i,u}^T K_{u,u}^{-1} k_{i,u} \right]$$

and

$$\log \mathcal{N}\left(y|\langle f \rangle_{p(f|u,Y)}, \sigma^2 I\right) = -\frac{1}{2} \log 2\pi \sigma^2 - \frac{1}{2\sigma^2} \left(y_i - K_{f,u} K_{u,u}^{-1} u\right)^2$$

- This requires the expectations

$$\langle K_{f,u} \rangle_{q(X)}$$

and

$$\langle K_{f,u} K_{u,u}^{-1} K_{u,f} \rangle_{q(X)}$$

which can be computed analytically for some covariance functions.
Priors for Latent Space

Variational marginalization of $X$ allows us to learn parameters of $p(X)$.

Standard GP-LVM where $X$ learnt by MAP, this is not possible (see e.g. Wang et al., 2008).

First example: learn the dimensionality of latent space.
Graphical Representations of GP-LVM
Graphical Representations of GP-LVM
Graphical Representations of GP-LVM
Graphical Representations of GP-LVM

The diagram illustrates the relationship between the data space and the latent space in a GP-LVM model. The variables $x_1, x_2, x_3, x_4, x_5, x_6$ are observed in the data space, which maps to the latent space through the node $y$. The variance $\sigma^2$ is associated with the latent space, and $w$ represents the weight or influence in the data space.
Graphical Representations of GP-LVM

\[ w \sim \mathcal{N}(0, \alpha I) \quad x \sim \mathcal{N}(0, I) \]

\[ y \sim \mathcal{N}(x^T w, \sigma^2) \]
Graphical Representations of GP-LVM

\[
\begin{align*}
    \mathbf{w} &\sim \mathcal{N}(0, \mathbf{I}) \quad \mathbf{x} \sim \mathcal{N}(0, \alpha \mathbf{I}) \\
    \mathbf{y} &\sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)
\end{align*}
\]
Graphical Representations of GP-LVM

\[
\begin{align*}
\mathbf{w} & \sim \mathcal{N}(0, \mathbf{I}) \quad x_i \sim \mathcal{N}(0, \alpha_i) \\
y & \sim \mathcal{N}(\mathbf{x}^\top \mathbf{w}, \sigma^2)
\end{align*}
\]
Graphical Representations of GP-LVM

\[ w_i \sim \mathcal{N}(0, \alpha_i) \quad x \sim \mathcal{N}(0, I) \]

\[ y \sim \mathcal{N}(x^T w, \sigma^2) \]
Non-linear $f(x)$

- In linear case equivalence because $f(x) = w^\top x$

\[ p(w_i) \sim \mathcal{N}(0, \alpha_i) \]

- In non linear case, need to scale columns of $X$ in prior for $f(x)$.

- This implies scaling columns of $X$ in covariance function

\[ k(x_{i,:}, x_{j,:}) = \exp \left( -\frac{1}{2} (x_{i,:} - x_{j,:})^\top A (x_{i,:} - x_{j,:}) \right) \]

$A$ is diagonal with elements $\alpha_i^2$. Now keep prior spherical

\[ p(X) = \prod_{j=1}^{q} \mathcal{N}(x_{j,:}|0, I) \]

- Covariance functions of this type are known as ARD (see e.g. Neal, 1996; MacKay, 2003; Rasmussen and Williams, 2006).
Automatic dimensionality detection

- Achieved by employing an *Automatic Relevance Determination (ARD)* covariance function for the prior on the GP mapping

\[ f \sim GP(0, k_f) \text{ with } k_f(x_i, x_j) = \sigma^2 \exp \left( -\frac{1}{2} \sum_{q=1}^{Q} w_q (x_{i,q} - x_{j,q})^2 \right) \]

- Example
Gaussian Process Dynamical Systems

(Damianou et al., 2011)
Gaussian Process over Latent Space

- Assume a GP prior for $p(X)$.
- Input to the process is time, $p(X|t)$. 
Interpolation of HD Video
Modeling Multiple ‘Views’

- Single space to model correlations between two different data sources, e.g., images & text, image & pose.
- Shared latent spaces: (Shon et al., 2006; Navaratnam et al., 2007; Ek et al., 2008b)

![Diagram](image)

- Effective when the ‘views’ are correlated.
- But not all information is shared between both ‘views’.
- PCA applied to concatenated data vs CCA applied to data.
In real scenarios, the ‘views’ are neither fully independent, nor fully correlated.

Shared models

- either allow information relevant to a single view to be mixed in the shared signal,
- or are unable to model such private information.

Solution: Model shared and private information (Virtanen et al., 2011; Ek et al., 2008a; Leen and Fyfe, 2006; Klami and Kaski, 2007, 2008; Tucker, 1958)

Probabilistic CCA is case when dimensionality of $Z$ matches $Y^{(i)}$ (cf Inter Battery Factor Analysis (Tucker, 1958)).
Manifold Relevance Determination

Damianou et al. (2012)
Separate ARD parameters for mappings to $Y^{(1)}$ and $Y^{(2)}$. 

Shared GP-LVM
Example: Yale faces

- Dataset Y: 3 persons under all illumination conditions
- Dataset Z: As above for 3 different persons
- Align datapoints $x_n$ and $z_n$ only based on the lighting direction

Deep Gaussian processes
Results

• Latent space $X$ initialised with 14 dimensions

• Weights define a segmentation of $X$

• Video / demo...

Deep Gaussian processes
Potential applications..?


References IV


C. Lu and X. Tang. Surpassing human-level face verification performance on LFW with GaussianFace. Technical report,


References VI


M. Seeger and M. I. Jordan. Sparse Gaussian Process Classification With Multiple Classes. Technical Report 661, Department of Statistics, University of California at Berkeley,


