Learning with Probabilities

Neil D. Lawrence

School of Computer Science, University of Manchester, U.K.
(from August 1st: Sheffield Institute for Translational Neuroscience and the University of Sheffield)

Machine Learning and CogSci Summer School, Pula, Sardinia

7th May 2010

Outline

Error Functions to Probabilities

- Last time we introduced different learning scenarios using error functions.
- In this lecture we will reinterpret those error functions through probability.
- The error function can be seen as a logarithm of a probability density function.
- Before we take that perspective we will first review probability.

Outline

Probability Review I

- We are interested in trials which result in two random variables, X and Y, each of which has an 'outcome' denoted by x or y.
- We summarise the notation and terminology for these distributions in the following table.

Terminology	Notation	Description
Joint Probability	$P(X = x, Y = y)$	'The probability that $X = x$ *and* $Y = y$'
Marginal Probability	$P(X = x)$	'The probability that $X = x$ *regardless of* Y'
Conditional Probability	$P(X = x \mid Y = y)$	'The probability that $X = x$ *given that* $Y = y$'

Table: The different basic probability distributions.

A Pictorial Definition of Probability

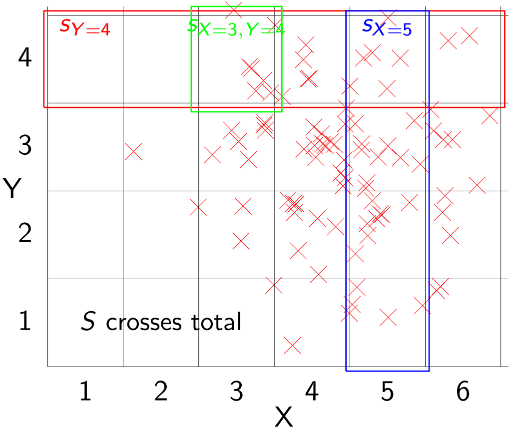


Figure: Representation of joint and conditional probabilities.

Different Distributions

Terminology	Definition
Joint Probability	$\lim_{S\to\infty} \frac{s_{X=3,Y=4}}{S} = P(X=3, Y=4)$
Marginal Probability	$\lim_{S\to\infty} \frac{s_{X=5}}{S} = P(X=5)$
Conditional Probability	$\lim_{S\to\infty} \frac{s_{X=3,Y=4}}{s_{Y=4}} = P(X=3 \vert Y=4)$

Table: Definition of probability distributions from Table 1 in terms of the system depicted in Figure 1.

Notational Details

- Typically we should write out $P\left(X=x, Y=y\right)$.
- In practice, we often use $P\left(x, y\right)$.
- This looks very much like we might write a multivariate function, *e.g.* $f\left(x, y\right)=\frac{x}{y}$.
  - For a multivariate function though, $f\left(x, y\right) \neq f\left(y, x\right)$.
  - However $P\left(x, y\right)=P\left(y, x\right)$ because $P\left(X=x, Y=y\right)=P\left(Y=y, X=x\right)$.
- We now quickly review the 'rules of probability'.

Normalization

All distributions are normalized. This is clear from the fact that $\sum_x s_x = S$, which gives

$$\sum_x P(x) = \frac{\sum_x s_x}{S} = \frac{S}{S} = 1.$$

A similar result can be derived for the marginal and conditional distributions.

The Sum Rule

- The marginal probability $P(y)$ is $\frac{s_y}{S}$ (ignoring the limit).
- The joint distribution $P(x, y)$ is $\frac{s_{x,y}}{S}$.
- $s_y = \sum_x s_{x,y}$ so
$$\frac{s_y}{S} = \sum_x \frac{s_{x,y}}{S},$$
in other words
$$P(y) = \sum_x P(x, y).$$
This is known as the sum rule of probability.

The Product Rule

- $P(x|y)$ is
$$\frac{s_{x,y}}{s_y}.$$
- $P(x,y)$ is
$$\frac{s_{x,y}}{S} = \frac{s_{x,y}}{s_y}\frac{s_y}{S}$$
or in other words
$$P(x,y) = P(x|y)\,P(y)\,.$$
This is known as the product rule of probability.

Bayes' Rule

- From the product rule,

$$P(x, y) = P(y, x) = P(y|x)\, P(x),$$

so

$$P(x|y)\, P(y) = P(y|x)\, P(x)$$

which leads to Bayes' rule,

$$P(x|y) = \frac{P(y|x)\, P(x)}{P(y)}.$$

Expectations

- We use a probabilistic model to summarizes our beliefs about states.
- We compute expected values by evaluating function under the distribution.

$$\langle f(x) \rangle_{P(x)} = \sum_x P(x) f(x).$$

You will also see expectations written in the form $E\{f(x)\}$.
- The mean is $\langle x \rangle_{P(x)}$, the variance is $\langle x^2 \rangle - \langle x \rangle^2$.

Distribution Representation

- We can represent probabilities as tables

x	0	1	2
$P(x)$	0.2	0.5	0.3

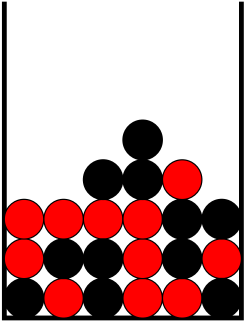
- But sometimes we prefer to represent them as functions.

Binomial Distribution

- Jakob Bernoulli: black and red balls in an urn. Proportion of red is π.
- Sample with replacement. Binomial gives the distribution of number of reds, y, from S extractions

$$p(y|\pi, S) = \frac{S!}{y!(S-y)!}\pi^{y}(1-\pi)^{(S-y)}$$

- Mean is given by $S\pi$ and variance $S\pi(1-\pi)$.



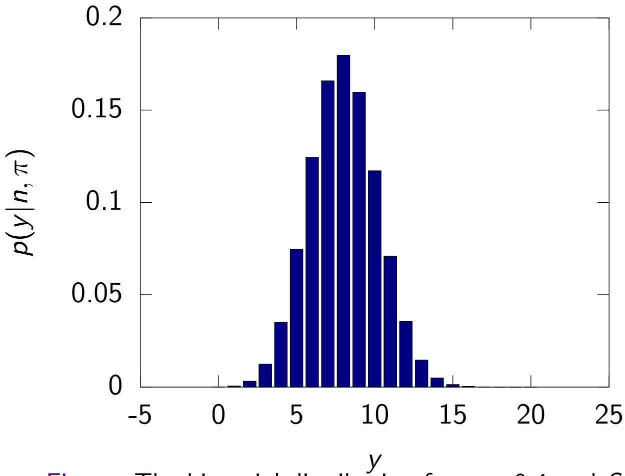


Figure: The binomial distribution for $\pi = 0.4$ and $S = 20$.

Continuous Variables

- So far discrete values of x or y.
- For continuous models we use the *probability density function* (PDF).
- Discrete case: defined probability distributions over a discrete number of states.
- How do we represent continuous as probability?
- Student heights:
  - Develop a representation which could answer *any* question we chose to ask about a student's height.
- PDF is a positive function, integral over the region of interest is one[1].

[1] In what follows we shall use the word distribution to refer to both discrete probabilities and continuous probability density functions.

Manipulating PDFs

- Same rules for PDFs as distributions *e.g.*

$$p(y|x) = \frac{p(x|y)\,p(y)}{p(x)}$$

  where $p(x, y) = p(x|y)\,p(y)$ and for continuous variables $p(x) = \int p(x, y)\,\mathrm{d}y$.

- Expectations under a PDF

$$\langle f(x) \rangle_{p(x)} = \int f(x)\,p(x)\,\mathrm{d}x$$

  where the integral is over the region for which our PDF for x is defined.

The Gaussian Density

- Perhaps the most common probability density.

$$p(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$
$$= \mathcal{N}\left(y|\mu,\sigma^2\right)$$

- Also available in multivariate form.
- First proposed maybe by de Moivre but also used by Laplace.

The Gaussian Density

- Perhaps the most common probability density.

$$
\begin{aligned}
p(\mathbf{y}|\boldsymbol{\mu}, \mathbf{C}) &= \frac{1}{(2\pi)^{\frac{p}{2}} |\mathbf{C}|^{\frac{1}{2}}} \exp\left(-\frac{(\mathbf{y}-\boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}-\boldsymbol{\mu})}{2}\right) \\
&= \mathcal{N}\left(\mathbf{y}|\boldsymbol{\mu}, \mathbf{C}\right)
\end{aligned}
$$

- Also available in multivariate form.
- First proposed maybe by de Moivre but also used by Laplace.

Gaussian PDF I

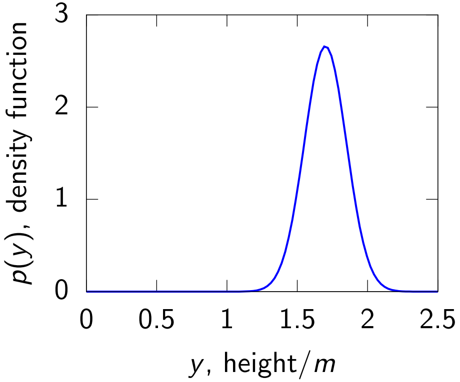


Figure: The Gaussian PDF with $\mu = 1.7$ and variance $\sigma^2 = 0.0225$. It might represent the heights of a population of students.

Outline

Sample Based Approximations I

- Sample based approximation

$$\langle f(y) \rangle_{P(y)} \approx \frac{1}{S} \sum_{i=1}^{S} f(y_i).$$

- Special cases of this include the *sample mean*, often denoted by $\bar{y}$, and computed as

$$\bar{y} = \frac{1}{S} \sum_{i=1}^{S} y_i,$$

Sample Mean vs True Mean

- This is an approximation to the true distribution mean

$$\langle y \rangle \approx \bar{y}.$$

- The same approximations can used for continuous PDFs, so we have

$$\begin{aligned} \langle f(x) \rangle_{p(x)} &= \int f(x)\, p(x)\, \mathrm{d}x \\ &\approx \frac{1}{S} \sum_{i=1}^{S} f(x_i), \end{aligned}$$

where x_i are independently obtained samples from the distribution $p(x)$.

- Approximation gets better for increasing S and worse if the samples from $P(y)$ are *not* independent.

Outline

Regression Revisited

- We introduced an error function of the form

$$E(\mathbf{w}) = \sum_{i=1}^{n} \left(\mathbf{w}^\top \phi_i - y_i\right)^2$$

- Quadratic error functions can be seen as Gaussian noise models.
- Imagine we are seeing data given by,

$$y(\mathbf{x}_i) = \mathbf{w}^\top \phi_i + \epsilon$$

where ϵ is Gaussian noise with standard deviation σ,

$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right).$$

Noise Corrupted Mapping

- This implies that

$$y_i \sim \mathcal{N}\left(\mathbf{w}^\top \boldsymbol{\phi}_i, \sigma^2\right)$$

- Which we also write

$$p(y_i|\mathbf{w}, \sigma) = \mathcal{N}\left(y_i|\mathbf{w}^\top \boldsymbol{\phi}_i, \sigma^2\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \phi_i, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|\mathbf{w}, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{w}^\top \phi_i)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \boldsymbol{\phi}_i, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|\mathbf{w}, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \boldsymbol{\phi}_i, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|\mathbf{w}, \sigma) \propto \prod_{i=1}^{n} \exp\left(-\frac{(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|\mathbf{w},\sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top\boldsymbol{\phi}_i,\sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|\mathbf{w},\sigma) \propto \prod_{i=1}^{n} \exp\left(-\frac{(y_i - \mathbf{w}^\top\boldsymbol{\phi}_i)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \phi_i, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|\mathbf{w}, \sigma) \propto \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \mathbf{w}^\top \phi_i)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \phi_i, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|\mathbf{w}, \sigma) \propto \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \mathbf{w}^\top \phi_i)^2}{2\sigma^2}\right)$$

Gaussian Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \boldsymbol{\phi}_i, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$p(\mathbf{y}|\mathbf{w}, \sigma) \propto \exp\left(-\sum_{i=1}^{n} \frac{(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i)^2}{2\sigma^2}\right)$$

Gaussian Log Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \boldsymbol{\phi}_i, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$\log p(\mathbf{y}|\mathbf{w}, \sigma) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i)^2 + \text{const}$$

Gaussian Log Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \boldsymbol{\phi}_i, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$-\log p(\mathbf{y}|\mathbf{w}, \sigma) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mathbf{w}^\top \boldsymbol{\phi}_i\right)^2 + \text{const}$$

Gaussian Log Likelihood

- If the noise is sampled independently for each data point from the same density we have

$$p(\mathbf{y}|\mathbf{w},\sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \phi_i, \sigma^2\right)$$

- This is an i.i.d. assumption about the noise.
- Writing the functional form we have

$$-\log p(\mathbf{y}|\mathbf{w},\sigma) = \frac{1}{2\sigma^2} E(\mathbf{w}) + \text{const}$$

Probabilistic Interpretation of the Error Function

- Probabilistic Interpretation for Error Function is Negative Log Likelihood.
- *Minimizing* error function is equivalent to *maximizing* log likelihood.
- Maximizing *log likelihood* is equivalent to maximizing the *likelihood* because log is monotonic.
- Probabilistic interpretation: Minimizing error function is equivalent to maximum likelihood with respect to parameters.

Consistency of Maximum Likelihood

- If data was really generated according to probability we specified.
- Correct parameters will be recovered in limit as $n \rightarrow \infty$.
- This can be proven through sample based approximations (law of large numbers) of "KL divergences".
- Mainstay of classical statistics.

Outline

Bayesian Approach

- Likelihood for the regression example has the form

$$p(\mathbf{y}|\mathbf{w}, \sigma^2) = \prod_{i=1}^{n} \mathcal{N}\left(y_i|\mathbf{w}^\top \phi_i, \sigma^2\right).$$

- Suggestion was to maximize this likelihood with respect to $\mathbf{w}$.
- This can be done with gradient based optimization of the log likelihood.
- Alternative approach: integration across $\mathbf{w}$.
- Consider expected value of likelihood under a range of potential $\mathbf{w}$s.
- This is known as the *Bayesian* approach.

Note on the Term Bayesian

- We will use Bayes' rule to invert probabilities in the Bayesian approach.
  - Bayesian is not named after Bayes' rule (v. common confusion).
  - The term Bayesian refers to the treatment of the parameters as stochastic variables.
  - For early statisticians this was very controversial (Fisher et al).

Binomial Distribution Revisited

- Binomial for one trial[a] (y_i is now either 0 or 1) given by

$$p(y_i|\pi) = \pi^{y_i}(1-\pi)^{(1-y_i)}$$

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- This treatment of a parameter, π, as a random variable that was/is considered controversial.

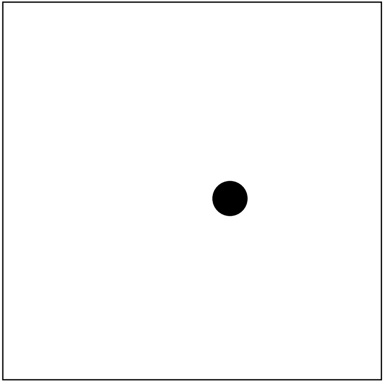
[a] Known as a Bernoulli distribution.

Binomial Distribution Revisited

- Binomial for one trial[a] (y_i is now either 0 or 1) given by

$$p(y_i|\pi) = \pi^{y_i}(1-\pi)^{(1-y_i)}$$

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



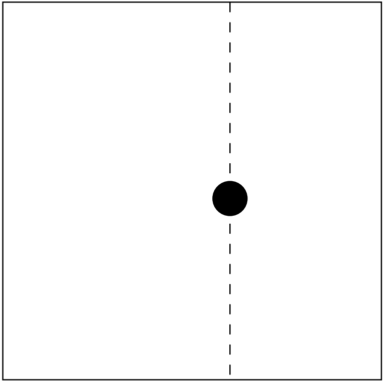
[a] Known as a Bernoulli distribution.

Binomial Distribution Revisited

- Binomial for one trial[a] (y_i is now either 0 or 1) given by

$$p(y_i|\pi) = \pi^{y_i}(1-\pi)^{(1-y_i)}$$

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



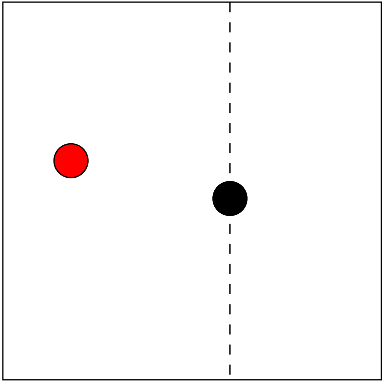
[a] Known as a Bernoulli distribution.

Binomial Distribution Revisited

- Binomial for one trial[a] (y_i is now either 0 or 1) given by

$$p(y_i|\pi) = \pi^{y_i}(1-\pi)^{(1-y_i)}$$

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



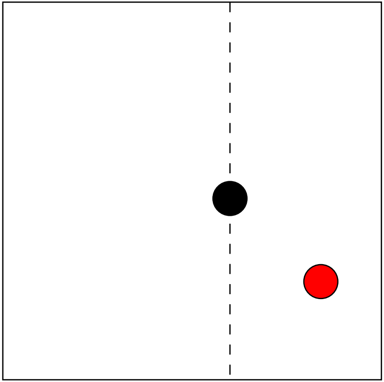
[a]Known as a Bernoulli distribution.

Binomial Distribution Revisited

- Binomial for one trial[a] (y_i is now either 0 or 1) given by

$$p(y_i|\pi) = \pi^{y_i}(1-\pi)^{(1-y_i)}$$

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



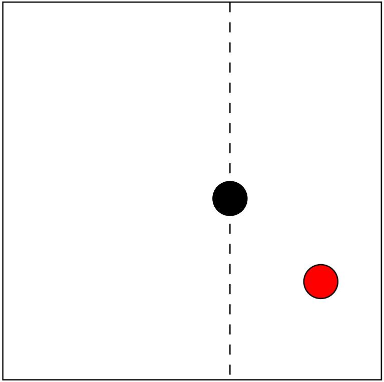
[a] Known as a Bernoulli distribution.

Binomial Distribution Revisited

- Binomial for one trial[a] (y_i is now either 0 or 1) given by

$$p(y_i|\pi) = \pi^{y_i}(1-\pi)^{(1-y_i)}$$

- Thomas Bayes considered a ball landing uniformly across a table.
- And another ball landing on the left or right (Bayes, 1763, page 385).
- This treatment of a parameter, π, as a random variable that was/is considered controversial.



[a] Known as a Bernoulli distribution.

Simple Bayesian Inference

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

- Four components:
  1. Prior distribution: represents belief about parameter values before seeing data.
  2. Likelihood: gives relation between parameters and data.
  3. Posterior distribution: represents updated belief about parameters after data is observed.
  4. Marginal likelihood: represents assessment of the quality of the model. Can be compared with other models (likelihood/prior combinations). *cf* Josh's talk. Ratios of marginal likelihoods are known as Bayes factors.

Example System: Robot Location

- Represent state (location) of the robot as $\mathbf{x}$.
- The robot makes readings using its sensors. These are stored in $\mathbf{y}$.
- Our initial belief about robot position is given by $p(\mathbf{x})$ this is the prior.
- Our expectation of sensor readings given robot location is the likelihood $p(\mathbf{y}|\mathbf{x})$.
- We combine initial picture of location, with sensor readings to get updated picture of location this is the posterior: $p(\mathbf{x}|\mathbf{y})$.

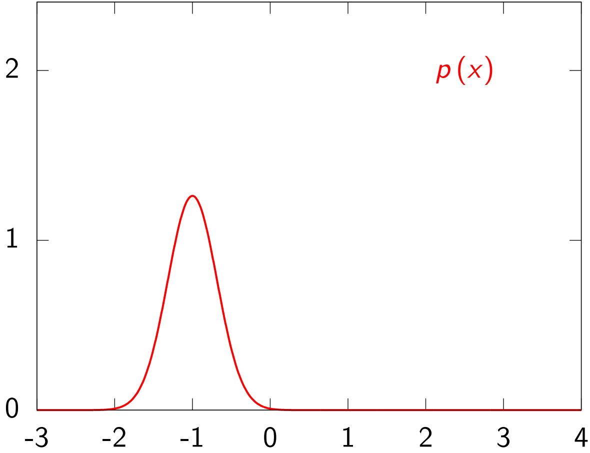


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Gaussian Noise

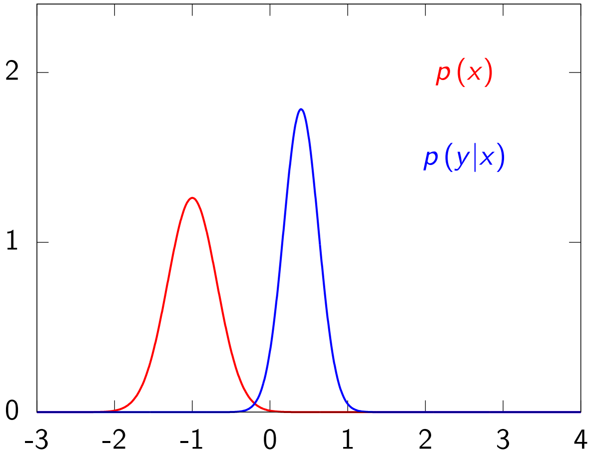


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Gaussian Noise

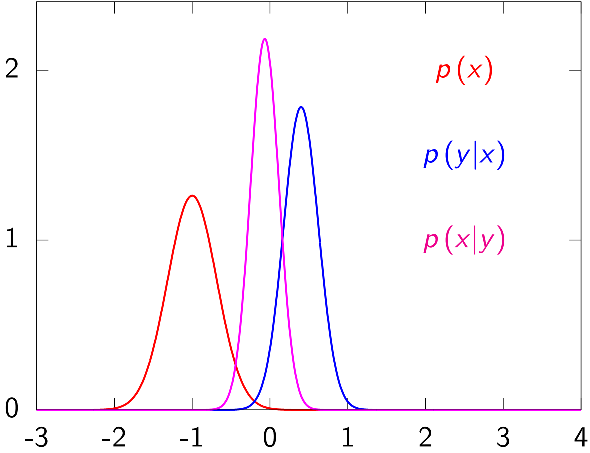


Figure: A Gaussian prior combines with a Gaussian likelihood for a Gaussian posterior.

Expectation Propagation

- Gaussian prior combines with Gaussian likelihood for Gaussian posterior.
- This Gaussian prior combines with Gaussian likelihood for Gaussian posterior.
- If likelihood is non-Gaussian one approach is to approximate the posterior distribution with a Gaussian.

Probit Likelihood

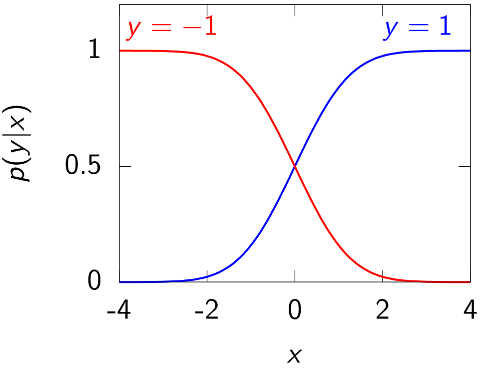


Figure: The probit likelihood. The plot shows $p\left(y|x\right)$ for different values of y. For $y = 1$ we have $p\left(y|x\right) = \phi\left(x\right) = \int_{-\infty}^{x} \mathcal{N}\left(z|0,1\right) \mathrm{d}z$.

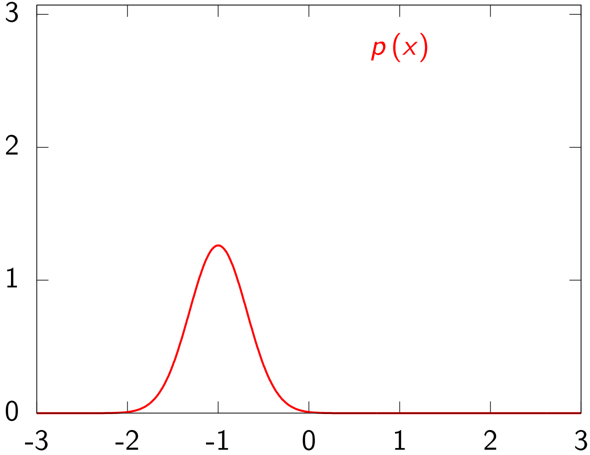


Figure: Combining a Gaussian prior with a probit likelihood.

Classification

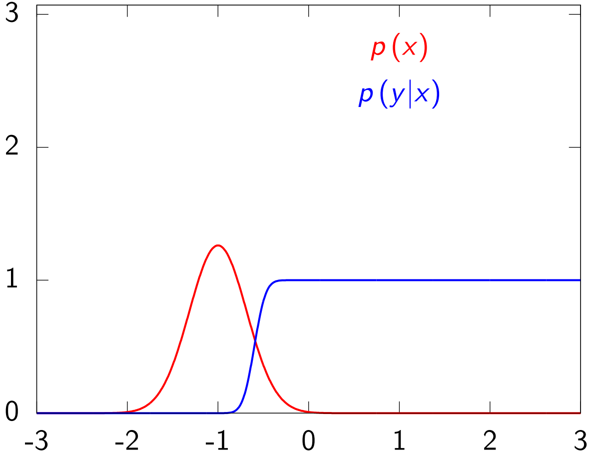


Figure: Combining a Gaussian prior with a probit likelihood.

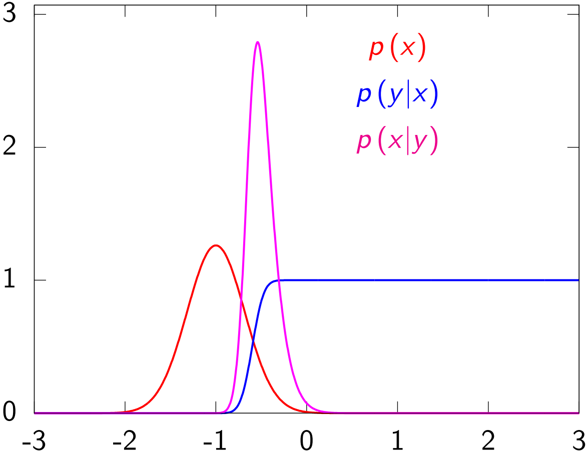


Figure: Combining a Gaussian prior with a probit likelihood.

Classification

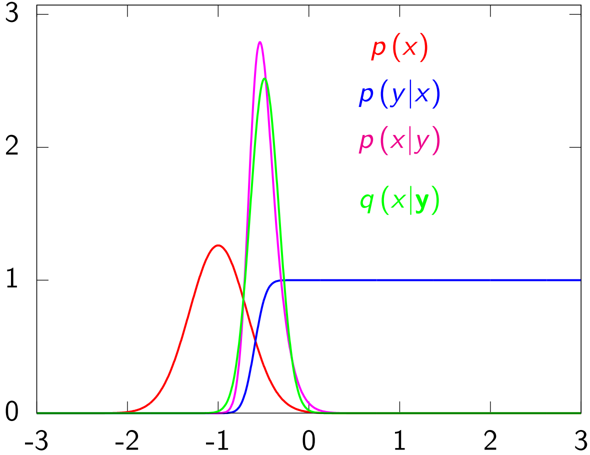


Figure: Combining a Gaussian prior with a probit likelihood.

Ordinal Noise Model

Ordered Categories

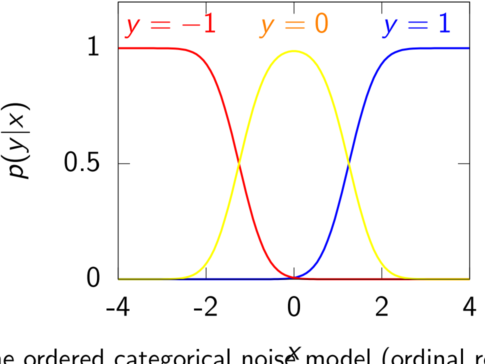


Figure: The ordered categorical noise model (ordinal regression). The plot shows $p(y|x)$ for different values of y. Here we have assumed three categories.

Ordinal Regression

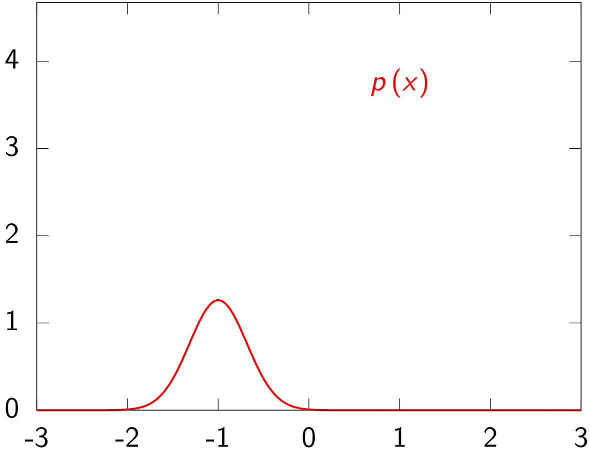


Figure: Bayesian inference with an ordinal categorical likelihood.

Ordinal Regression

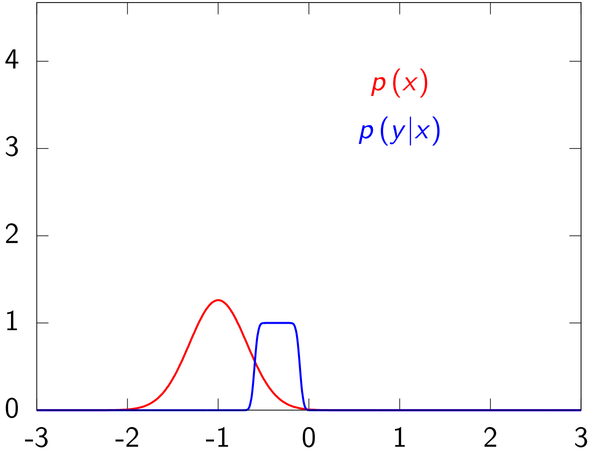


Figure: Bayesian inference with an ordinal categorical likelihood.

Ordinal Regression

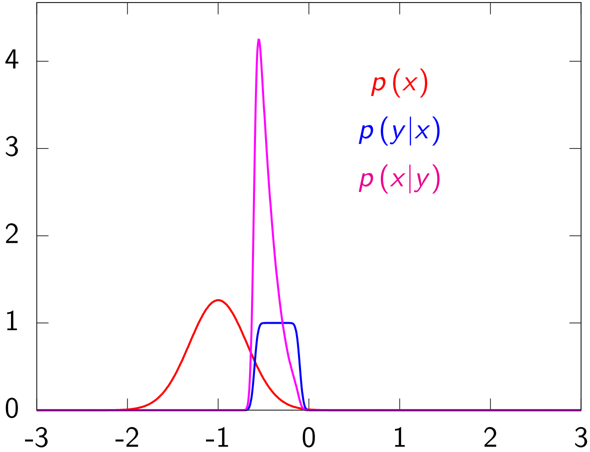


Figure: Bayesian inference with an ordinal categorical likelihood.

Ordinal Regression

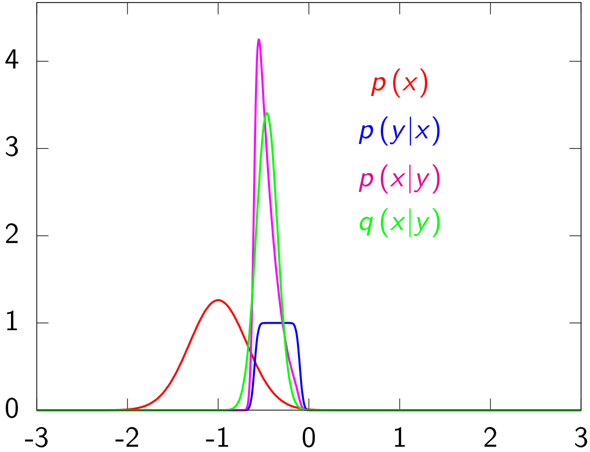


Figure: Bayesian inference with an ordinal categorical likelihood.

Outline

Bayesian Linear Regression

- Combine our regression likelihood

$$y_i \sim \mathcal{N}\left(\mathbf{w}^\top \boldsymbol{\phi}_i, \sigma^2\right)$$

- With a prior density over the parameters.

$$\mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \alpha \mathbf{I}\right)$$

- Marginal likelihood given by

$$\mathbf{y} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{K}\right)$$

where elements of $\mathbf{K}$ are given by

$$k_{i,j} = \alpha \boldsymbol{\phi}_i^\top \boldsymbol{\phi}_j + \delta_{i,j}\sigma^2$$

Marginal Likelihood

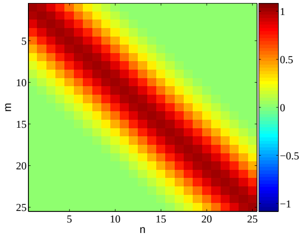
- First part of Gaussian marginal likelihood dependent on inner products
$$k_{i,j} = \alpha \boldsymbol{\phi}_i^\top \boldsymbol{\phi}_j$$
- Mercer's theorem allows us to replace this with a covariance function/kernel
$$k_{i,j} = k(\mathbf{x}_{i,:}, \mathbf{x}_{j,:})$$
- This allows us to make nonparametric models: models with infinite basis functions.

$$k(\mathbf{x}_{i,:}, \mathbf{x}_{j,:}) = \sum_{k=1}^{\infty} \phi_k(\mathbf{x}_i)\phi_k(\mathbf{x}_j)$$

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

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

- Covariance matrix is built using the *inputs* to the function t.
- For the example above it was based on Euclidean distance.
- The covariance function is also know as a kernel.



Covariance Samples

`demCovFuncSample`

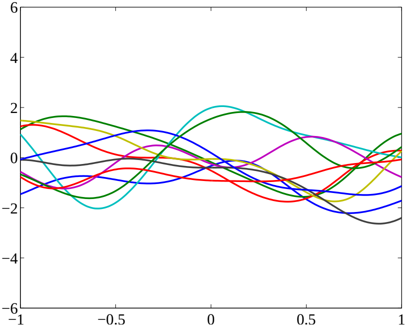


Figure: Exponentiated quadratic kernel with $\ell = 10^{-\frac{1}{2}}$, $\alpha = 1$

Covariance Samples

`demCovFuncSample`

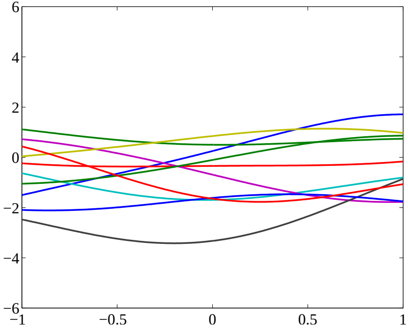


Figure: Exponentiated quadratic kernel with $\ell = 1$, $\alpha = 1$

Covariance Samples

`demCovFuncSample`

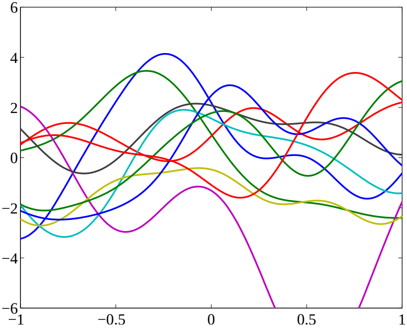


Figure: Exponentiated quadratic kernel with $\ell = 0.3$, $\alpha = 4$

Covariance Samples

`demCovFuncSample`

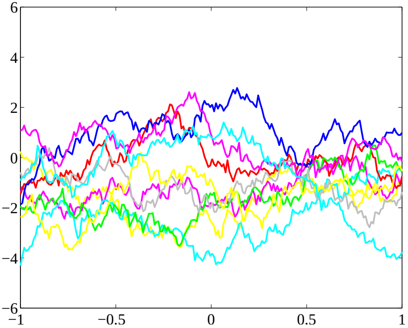


Figure: Ornstein-Uhlenbeck (stationary Gauss-Markov) covariance function $\ell = 1$, $\alpha = 4$

Gaussian Process Regression

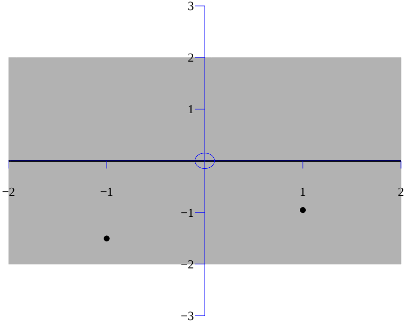


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

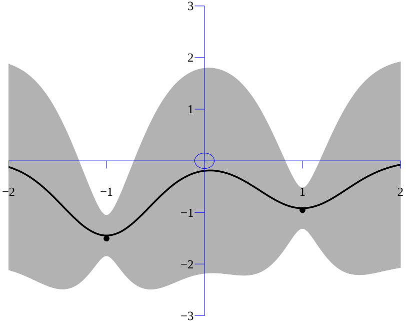


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

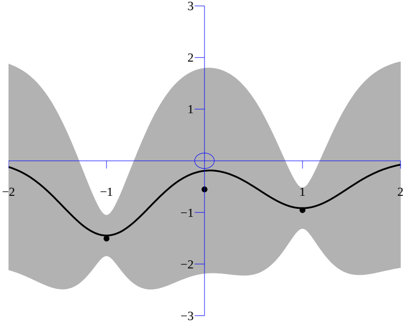


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

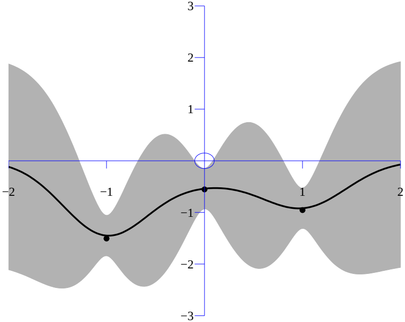


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

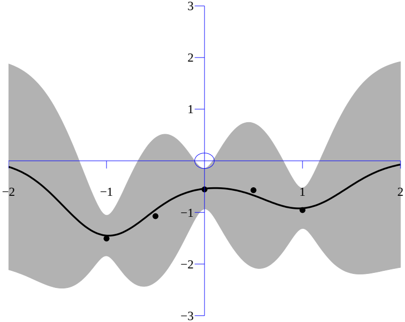


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

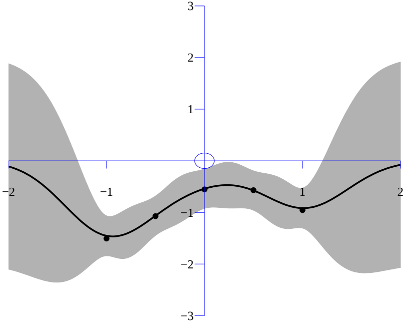


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

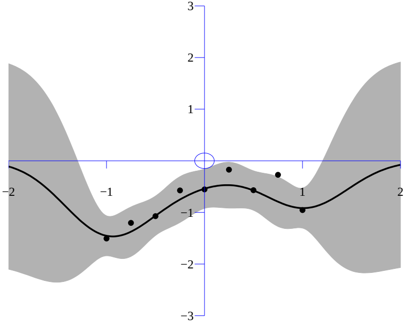


Figure: Examples include WiFi localization, C14 callibration curve.

Gaussian Process Regression

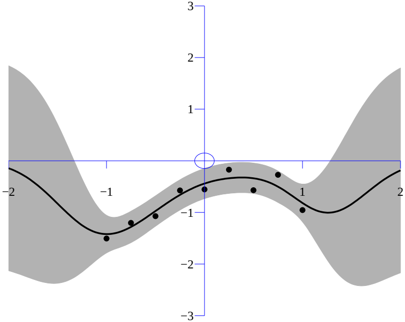
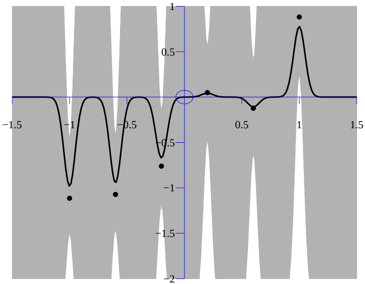
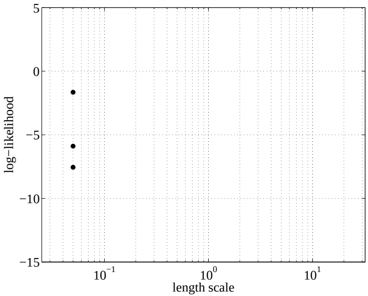


Figure: Examples include WiFi localization, C14 callibration curve.

Learning Kernel Parameters

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

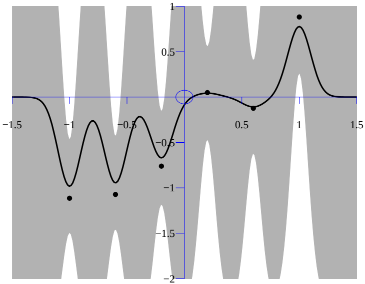


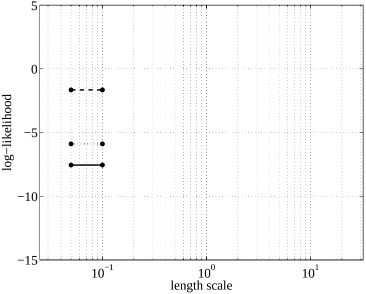


$$\log \mathcal{N}(\mathbf{y}|\mathbf{0},\mathbf{K}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Kernel Parameters

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

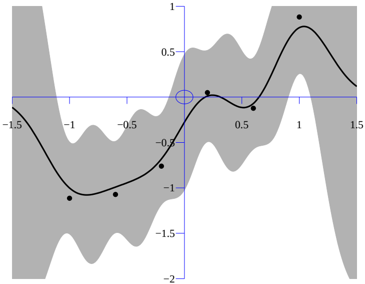


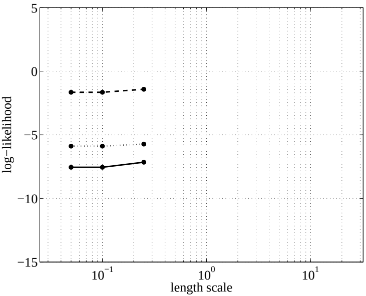


$$\log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Kernel Parameters

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

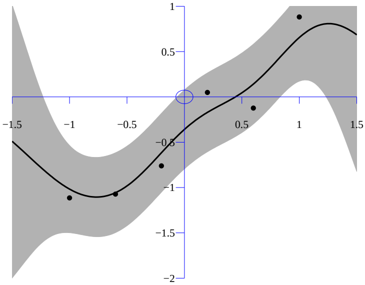


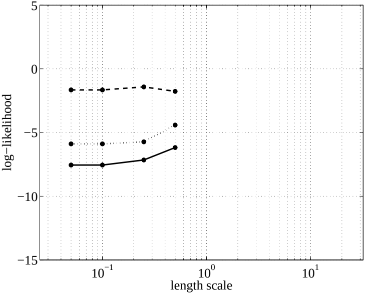


$$\log \mathcal{N}\left(\mathbf{y}|\mathbf{0},\mathbf{K}\right) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Kernel Parameters

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

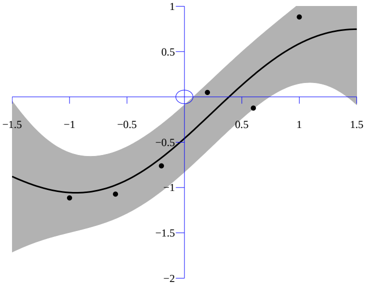


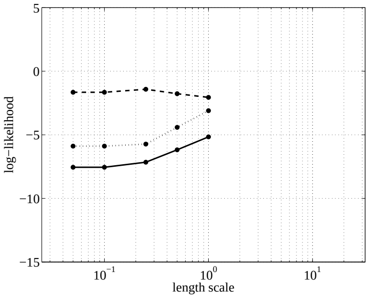


$$\log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}\right) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Kernel Parameters

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

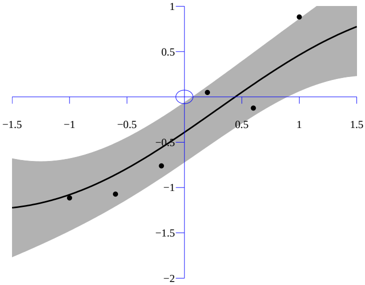


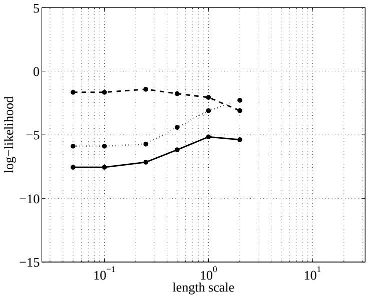


$$\log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Kernel Parameters

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

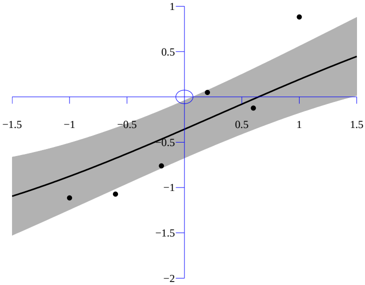


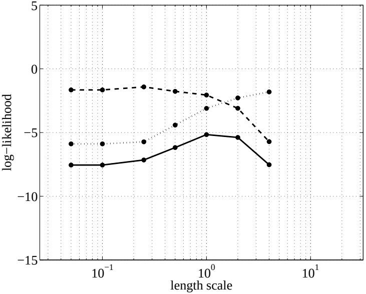


$$\log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}\right) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^\top\mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Kernel Parameters

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

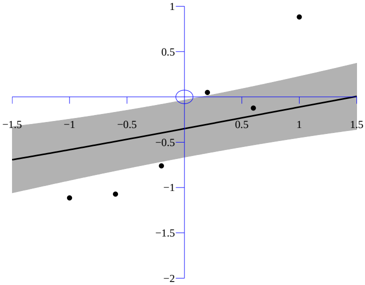


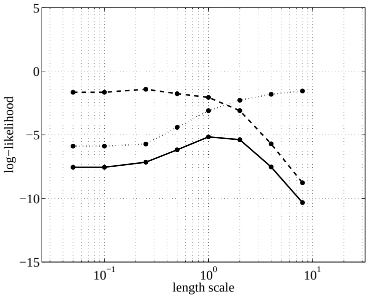


$$\log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^\top \mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Kernel Parameters

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

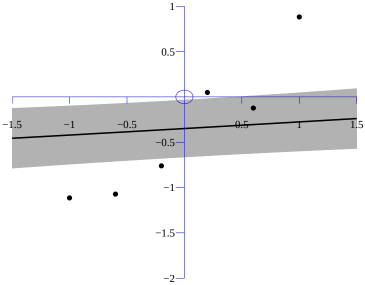


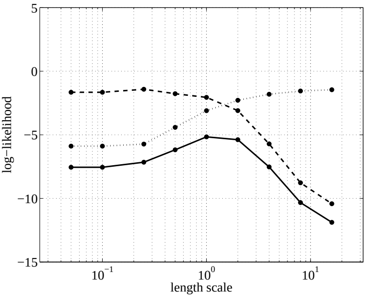


$$\log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K}\right) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^\top\mathbf{K}^{-1}\mathbf{y}}{2}$$

Learning Kernel Parameters

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





$$\log \mathcal{N}\left(\mathbf{y}|\mathbf{0},\mathbf{K}\right) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{\mathbf{y}^\top\mathbf{K}^{-1}\mathbf{y}}{2}$$

Outline

Outline

Mixture of Gaussians I

- Probabilistic clustering methods.
- Bayesian equivalent of K-means.
- Mixture of Gaussians.
- Assume data is sampled from a Gaussian density:

$$p(\mathbf{y}_i|\mathbf{s}_i) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{y}_i|\boldsymbol{\mu}_k, \mathbf{C}_k\right)^{s_{i,k}}$$

- Where $\mathbf{s}_i$ is a binary vector encoding component with 1-of-n encoding.
- Multinomial prior over $\mathbf{s}_i$

$$p(\mathbf{s}_i) = \prod_{k=1}^{K} \pi_k^{s_{i,k}}$$

EM Algorithm

- Marginal likelihood

$$\log p(\mathbf{y}_i) = \log \sum_{\mathbf{s}_i} p(\mathbf{y}_i, \mathbf{s}_i)$$

- Jensen's inequality gives a bound.
- Bound becomes equality if $q(\mathbf{s}_i) = p(\mathbf{s}_i|\mathbf{y}_i)$

$$p(\mathbf{y}_i) = \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{p(\mathbf{s}_i|\mathbf{y}_i)}$$

EM Algorithm

- Marginal likelihood

$$\log p(\mathbf{y}_i) = \log \sum_{\mathbf{s}_i} p(\mathbf{y}_i, \mathbf{s}_i)$$

$$\log p(\mathbf{y}_i) = \log \sum_{\mathbf{s}_i} q(\mathbf{s}_i) \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{q(\mathbf{s}_i)}$$

- Jensen's inequality gives a bound.
- Bound becomes equality if $q(\mathbf{s}_i) = p(\mathbf{s}_i|\mathbf{y}_i)$

$$p(\mathbf{y}_i) = \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{p(\mathbf{s}_i|\mathbf{y}_i)}$$

EM Algorithm

- Marginal likelihood

$$\log p(\mathbf{y}_i) = \log \sum_{\mathbf{s}_i} q(\mathbf{s}_i) \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{q(\mathbf{s}_i)}$$

$$\log p(\mathbf{y}_i) \geq \sum_{\mathbf{s}_i} q(\mathbf{s}_i) \log \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{q(\mathbf{s}_i)}$$

- Jensen's inequality gives a bound.
- Bound becomes equality if $q(\mathbf{s}_i) = p(\mathbf{s}_i|\mathbf{y}_i)$

$$p(\mathbf{y}_i) = \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{p(\mathbf{s}_i|\mathbf{y}_i)}$$

EM Algorithm

- Marginal likelihood

$$\log p(\mathbf{y}_i) \geq \sum_{\mathbf{s}_i} q(\mathbf{s}_i) \log \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{q(\mathbf{s}_i)}$$

$$\log p(\mathbf{y}_i) = \sum_{\mathbf{s}_i} p(\mathbf{s}_i|\mathbf{y}_i) \log \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{p(\mathbf{s}_i|\mathbf{y}_i)}$$

- Jensen's inequality gives a bound.
- Bound becomes equality if $q(\mathbf{s}_i) = p(\mathbf{s}_i|\mathbf{y}_i)$

$$p(\mathbf{y}_i) = \frac{p(\mathbf{y}_i, \mathbf{s}_i)}{p(\mathbf{s}_i|\mathbf{y}_i)}$$

EM Algorithm

- Iterate between
  1. **E Step** Set $q(\mathbf{s}_i) = p(\mathbf{s}_i|\mathbf{y}_i)$
  2. **M Step** Maximize $\sum_{\mathbf{s}_i} q(\mathbf{s}_i) \log p(\mathbf{y}_i, \mathbf{s}_i)$ with respect to parameters.

EM for Mixtures of Gaussians

- Iterate between
  1. **E Step** Set $q(\mathbf{s}_i) = \prod_{k=1}^{K} r_{i,k}^{s_{i,k}}$ where

$$r_{i,k} = \frac{\pi_k \mathcal{N}\left(\mathbf{y}_i | \boldsymbol{\mu}_k, \mathbf{C}_k\right)}{\sum_k \pi_k \mathcal{N}\left(\mathbf{y}_i | \boldsymbol{\mu}_k, \mathbf{C}_k\right)}$$

  2. **M Step** Maximize $\left\langle \log p(\mathbf{y}_i, \mathbf{s}_i) \right\rangle_{q(\mathbf{s}_i)}$ by setting

$$\pi_k = \frac{1}{n} \sum_{i=1}^{n} r_{i,k}, \quad \boldsymbol{\mu}_k = \frac{1}{\bar{n}_k} \sum_{i=1}^{n} r_{i,k} \mathbf{y}_i$$

$$\mathbf{C}_k = \frac{1}{\bar{n}_k} \sum_{i=1}^{n} r_{i,k} (\mathbf{y}_i - \boldsymbol{\mu}_k)(\mathbf{y}_i - \boldsymbol{\mu}_k)^\top$$

$$\bar{n}_k = \sum_{i=1}^{n} r_{i,k}$$

Netlab Demo

`demgmm1.m`

Variational Inference

- EM algorithm relies on computation of setting $q(\mathbf{s}_i)$ to $p(\mathbf{s}_i|y_i)$.
- In variational inference we use approximate posteriors for the $q(\cdot)$ distributions.
- This makes the algorithms tractable but non exact.

Outline

Latent Variable Models

Quoting from Hotelling, 1933, page 417:

> *Consider p variables attaching to each individual of a population. These statistical variables y_1, y_2, ... , y_p might for example be scores made by school children in tests of speed and skill in solving arithmetical problems or in reading; or they might be various physical properties of telephone poles, or the rates of exchange among various currencies. The y's will ordinarily be correlated. It is natural to ask whether some more fundamental set of independent variables exists, perhaps fewer in number than the y's, which determine the values the y's will take. If x_1, x_2, ... are such variables, we shall then have a set of relations of the form*
>
> $$y_i = f(x_1, x_2, ...) \quad (i = 1, 2, ..., p) \quad (1)$$
>
> *Quantities such as the x's have been called mental factors in recent psychological literature. However in view of the prospect of application of these ideas outside of psychology, and the conflicting usage attaching to the word "factor" in mathematics, it will be better simply to call the x's components of the complex depicted by the tests.*

Latent Variable Model

Relationship between the latent space and the data space

$$\mathbf{y}_{i,:} = \mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu} + \boldsymbol{\epsilon}_{i,:}$$

where $\mathbf{W} \in \Re^{p,q}$ is a mapping matrix and

$$\boldsymbol{\epsilon}_{i,:} \sim \mathcal{N}\left(0, \sigma^2\mathbf{I}\right).$$

Linear Dimensionality Reduction

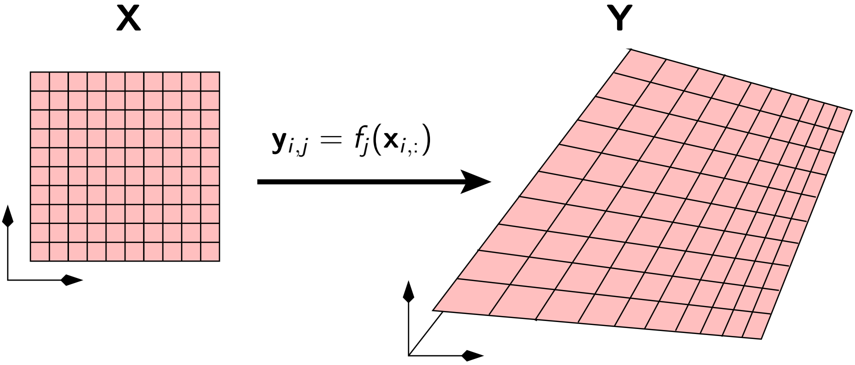


Figure: Mapping a two dimensional plane to a higher dimensional space in a linear way. Data are generated by corrupting points on the plane with noise.

Latent Variable Model

- Same likelihood as for linear regression (but multiple output now)

$$y_{i,j} \sim \mathcal{N}\left(\mathbf{w}_{j,:}^{\top}\mathbf{x}_{i,:} + \mu_j, \sigma^2\right).$$

- With independence assumptions that gives

$$p\left(\mathbf{Y}|\mathbf{X}, \mathbf{W}\right) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{y}_{i,:}|\mathbf{W}\mathbf{x}_{i,:} + \boldsymbol{\mu}, \sigma^2\mathbf{I}\right).$$

Prior in Latent Space

- The latent components (or factors are unknown).
- Use a prior distribution over them and marginalize them out.

$$x_{i,j} \sim \mathcal{N}\left(0, 1\right).$$

So the joint density for the components can be written

$$p(\mathbf{X}) = \prod_{i=1}^{n} \mathcal{N}\left(\mathbf{x}_{i,:}|\mathbf{0}, \mathbf{I}\right).$$

Marginalization of Latent Variables

- Marginal likelihood is given by

$$p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \int p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) p(\mathbf{X}) \mathrm{d}\mathbf{X}.$$

performing this integration leads to

$$\mathbf{y}_{i,:} \sim \mathcal{N}\left(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}\right).$$

define $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}$.

Maximum Likelihood

- Log likelihood is given by

$$\log p(\mathbf{Y}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = -\frac{np}{2}\log 2\pi - \frac{n}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_{i,:} - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}_{i,:} - \boldsymbol{\mu}).$$

- Error function is therefore

$$E(\mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \frac{n}{2}\log|\mathbf{C}| + \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_{i,:} - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{y}_{i,:} - \boldsymbol{\mu}).$$

- Minimize this error function.

Optimum for Mean I

- Error as function of $\boldsymbol{\mu}$

$$E(\boldsymbol{\mu}) = -\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_{i,:} - \boldsymbol{\mu})^{\top}\mathbf{C}^{-1}(\mathbf{y}_{i,:} - \boldsymbol{\mu})$$

- Compute the gradient

$$\frac{\mathrm{d}E(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}} = \mathbf{C}^{-1}\left(\sum_{i=1}^{n}\mathbf{y}_{i,:} - n\boldsymbol{\mu}\right).$$

- Find a minimum by looking for where gradients are zero,

$$\mathbf{0} = \mathbf{C}^{-1}\left(\sum_{i=1}^{n}\mathbf{y}_{i,:} - n\boldsymbol{\mu}\right)$$

Optimum for Mean II

- Implying

$$\mathbf{C}^{-1}\boldsymbol{\mu} = \mathbf{C}^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_{i,:}$$

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_{i,:}.$$

Optimizing Parameters I

- This solution allows us to set $\hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{1}\boldsymbol{\mu}^\top$.
- Substitute to give us a new "likelihood" over the centered data,

$$p\left(\hat{\mathbf{Y}}|\mathbf{W}\right) = \prod_{j=1}^{p} \mathcal{N}\left(\hat{\mathbf{y}}_{i,:}|\mathbf{0}, \mathbf{C}\right),$$

  where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}$.
- Tipping and Bishop (1999) showed that the global maximum likelihood for $\mathbf{W}$ and σ^2 can be found by an eigenvalue problem.
- Gradient of error function is

$$\frac{\mathrm{d}E(\mathbf{W}, \sigma^2)}{\mathrm{d}\mathbf{C}} = \frac{n}{2}\mathbf{C}^{-1} - \frac{1}{2}\mathbf{C}^{-1}\hat{\mathbf{Y}}^\top\hat{\mathbf{Y}}\mathbf{C}^{-1}. \tag{1}$$

Optimizing Parameters II

- Solution is given by

$$\underbrace{\mathbf{U}_q}_{\text{first } q \text{ eigenvectors}} \boldsymbol{\Lambda} = \underbrace{\frac{1}{n}\hat{\mathbf{Y}}^\top\hat{\mathbf{Y}}}_{\text{sample covariance matrix}} \mathbf{U}_q,$$

Oil Data

Homogeneous

Stratified

Annular

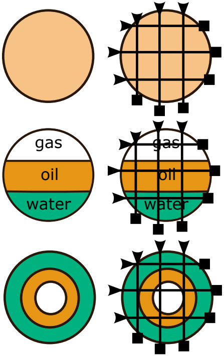


Figure: The “oil data”. The data set is artificially generated by modeling the manner in which a gamma ray’s intensity falls when it passes through a different density materials.

Probabilistic Models Allow for Missing Data

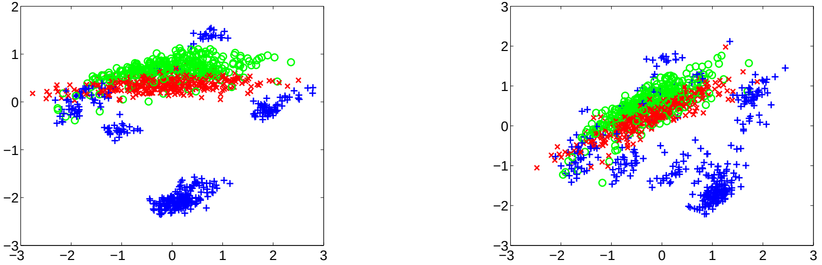


Figure: Projection of the oil data set on to $q = 2$ latent dimensions using the probabilistic PCA model. Different plots show various proportions of missing values. All values are missing at random from the design matrix $\mathbf{Y}$. *Right*: 10% missing.

Probabilistic Models Allow for Missing Data

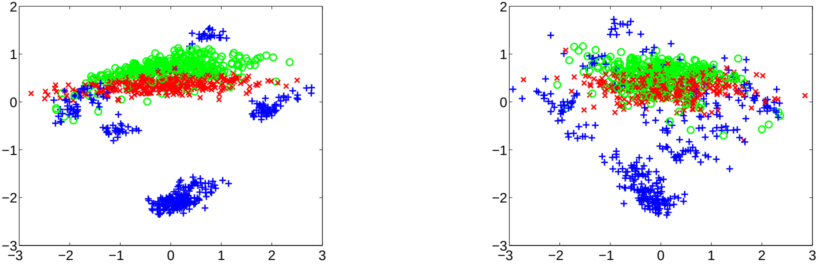


Figure: Projection of the oil data set on to $q = 2$ latent dimensions using the probabilistic PCA model. Different plots show various proportions of missing values. All values are missing at random from the design matrix $\mathbf{Y}$. *Right*: 20% missing.

Probabilistic Models Allow for Missing Data

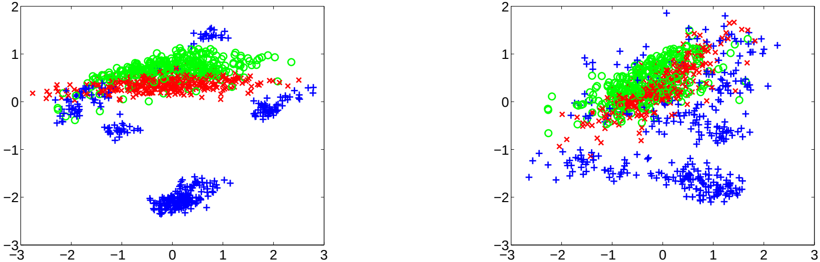
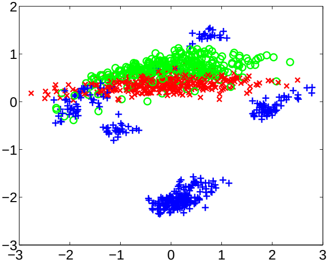


Figure: Projection of the oil data set on to $q = 2$ latent dimensions using the probabilistic PCA model. Different plots show various proportions of missing values. All values are missing at random from the design matrix $\mathbf{Y}$. *Right*: 30% missing.

Probabilistic Models Allow for Missing Data



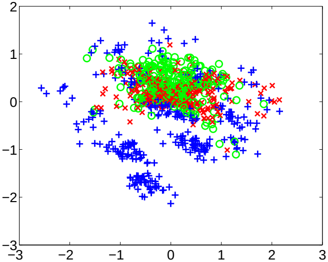


Figure: Projection of the oil data set on to $q = 2$ latent dimensions using the probabilistic PCA model. Different plots show various proportions of missing values. All values are missing at random from the design matrix $\mathbf{Y}$. *Right*: 50% missing.

Factor Analysis

- Factor Analysis is a very similar model.
- In factor analysis the likelihood allows for different variances at each output

$$p(y_{i,j}|\mathbf{w}_{j,:}, \mathbf{x}_{i,:}, \sigma_j^2) = \mathcal{N}\left(y_{i,j}|\mathbf{w}_{j,:}^\top \mathbf{x}_{i,:}, \sigma_j^2\right)$$

- This leads to a marginal covariance matrix of the form

$$\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \mathbf{D}$$

  where diagonal elements of $\mathbf{D}$ are given by σ_j^2.
- Cannot now be solved through an eigenvalue problem.

Conclusions

- Probabilistic interpretation of learning has error functions as negative log likelihood.
- Bayesian approach treats parameters as random variables.
- Learning proceeds through combination of prior and likelihood.
- Latent variable models and mixture of Gaussians are not Bayesian but use Bayes' rule.
- All these models sit in the wider family of probabilistic models.

References I

T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*, 53:370–418, 1763. [DOI].

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, B*, 6(3):611–622, 1999. [PDF]. [DOI].