ST308 - Lent term
Bayesian Inference

Kostas Kalogeropoulos

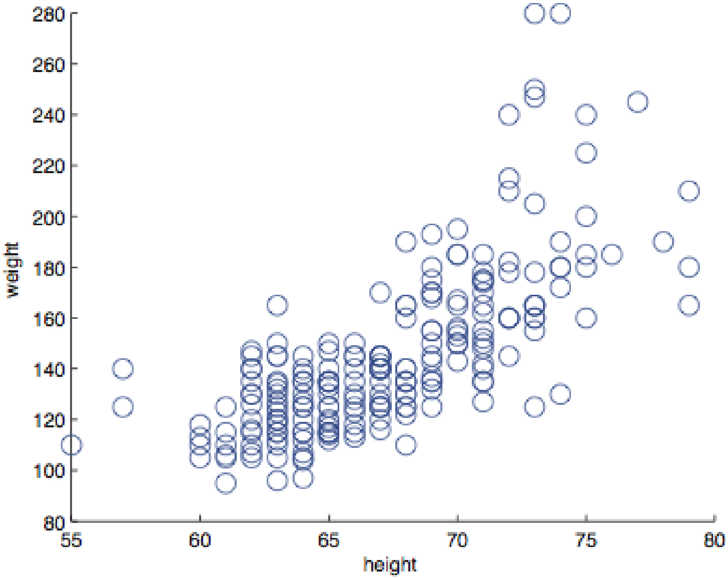
Mixture Models and the EM algorithm

Outline

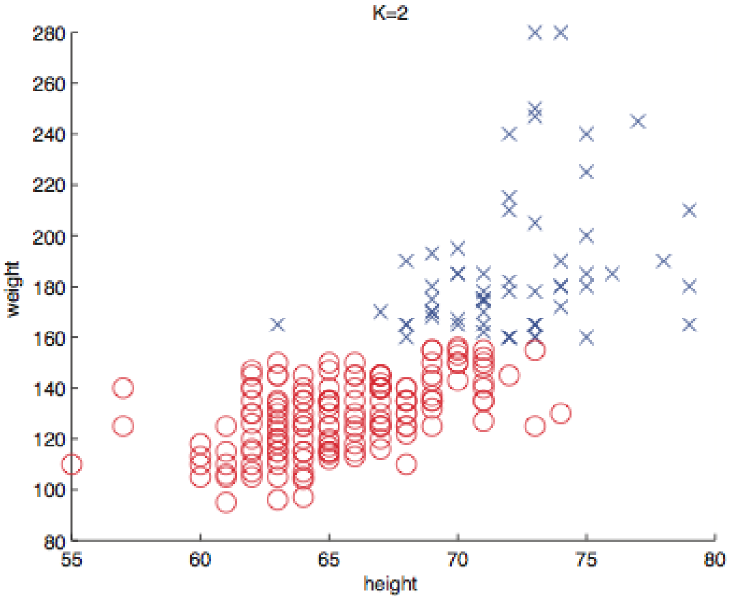
Topics: Data augmentation, Clustering, EM algorithm, Gaussian mixtures, K-means, Overfitted mixtures.

Outline

Motivating Example 1: Heights and weights



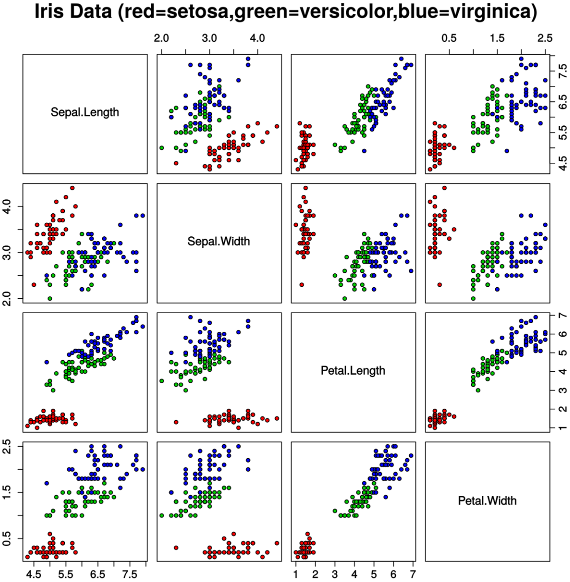
Example 1: Heights and weights



Example 2: Distinguishing Iris flower species



Example 2: Distinguishing Iris flower species

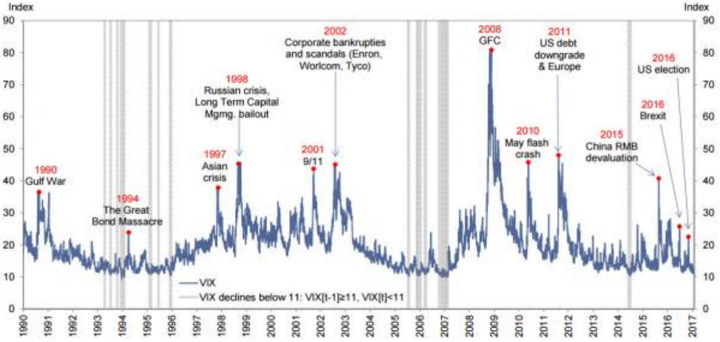


Example 3: VIX index

Volatility Index (VIX) provided by Chicago Board of Exchange (CBOE). Derived from the S&P 500 index options. Represents market's expectation of its future 30-day volatility. A measure of market risk.

Exhibit 3: VIX levels 1990-present

Shaded events represent VIX declining below 11, i.e. VIX[t-1]≥11, VIX[t]<11. Daily data from 1/2/1990– 1/27/2017.



Source: Chicago Board Options Exchange (CBOE). Goldman Sachs Global Investment Research.

Example 3: Bayesian non-parametric models

Recall the VIX index and the model we used to capture some of its stylised facts

$$Y_t = Y_{t-1} + \kappa(\mu - Y_{t-1})\delta + \sigma\epsilon_t,$$

where Y_t is VIX at time t, δ is the time interval, and ϵ_t are independent error terms.

The distribution of each ϵ_t may assumed to be a mixture of Normal distributions.

Such model is very flexible; in this case corresponds to a model with jumps.

Outline

Data augmentation

Often we want to draw inference on parameters θ based on data x from a likelihood $f(x|\theta)$ that is either intractable or expensive to compute.

Introduce an unobserved latent variable z to extend the model defining $f(z, x|\theta)$

We can then work directly with $f(z, x|\theta)$ (variational Bayes, MCMC) or approximate the integral $f(x|\theta) = \int f(z, x|\theta)dz$ in some way (simulated likelihood, EM).

Many famous examples, e.g. Ising model, factor analysis, random effects, hidden Markov models and mixtures.

Cluster/mixture analysis

- The population consists of K clusters/groups, each with distribution $f(x_i|\theta_k)$, $k = 1, \ldots, K$. Each individual $i = 1, \ldots, n$, may belong to one of them.

- Assume that for each i, there is an unobserved/latent cluster indicator z_{ik}, for $k = 1, \ldots, K$ that takes the value 1 if the individual i is in cluster k and 0 otherwise.

- Each $Z_i := (z_{i1}, \ldots, z_{iK})$ follows the Multinoulli distribution

$$\pi(Z_i|\pi_k) = \prod_{k=1}^{K} \pi_k^{z_{ik}}, \quad \text{where} \quad \sum_k \pi_k = 1.$$

- Note that if $z_{ik} = 1$, then for all $j \neq k$, $z_{ij} = 0$. Hence

$$\pi(z_{ik} = 1) = \pi_k$$

Likelihood and augmented likelihood

The augmented likelihood also includes z_i for each x_i.

$$f(Z_i, x_i|\theta) = \pi(Z_i|\pi_k) f(x_i|Z_i, \theta_k) = \prod_{k=1}^{K} \pi_k^{z_{ik}} f(x_i|\theta_k)^{z_{ik}}.$$

Note that $f(x_i|z_{ik} = 1, \theta) = f(x_i|\theta_k)$, so $f(z_{ik} = 1, x_i|\theta) = \pi_k f(x_i|\theta_k)$.

To get the likelihood $f(x_i|\theta)$ we sum out Z_i. Take z_{ik}'s and sum over k:

$$f(x_i|\theta) = \sum_{Z_i} f(Z_i, x_i|\theta) = \sum_{k=1}^{K} f(z_{ik} = 1, x_i|\theta) = \sum_{k=1}^{K} \pi_k f(x_i|\theta_k)$$

Overall we have $f(x|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$ and $f(z, x|\theta) = \prod_{i=1}^{n} f(x_i, z_i|\theta)$.

Aim: classify individuals through z_{ik}'s and estimate $\theta = (\pi_k, \theta_k)_{k=1}^{K}$.

Example: Gaussian Mixture Models

In Gaussian mixture models, we have $x_i | z_{ik} = 1 \sim N(\mu_k, \Sigma_k)$

Hence

$$f(x_i|\theta) = \sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \Sigma_k)$$

so the parameters to be estimated are $\theta = (\pi_k, \mu_k, \Sigma_k)_{k=1}^{K}$.

Due to the large number of parameters, especially for large K, restrictions are often placed on Σ_k, e.g. diagonal or tied.

Outline

Main idea

Complete Data: If we knew the cluster each person is, Z_i, then MLE is straightforward: split the data into clusters do MLE in each cluster separately.

But we don't, so we need a modified approach. The algorithm used most frequently is the EM.

A rough sketch is the one below

1. Start with a θ.
2. E step: Use Bayes theorem to find the responsibilities $\gamma_{ik} = \pi(z_i k = 1|x, \theta)$ to get the expected log likelihood.
3. M step: Maximise the expected log-likelihood and update θ.
4. Continue until convergence.

log-likelihood and augmented log-likelihod

First write down the augmented log-likelihood. Remember that

$$f(Z_i, x_i|\theta) = \prod_{k=1}^{K} \pi_k^{z_{ik}} f(x_i|\theta_k)^{z_{ik}},$$

so considering all individuals and taking log gives

$$\log f(Z, x|\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\log \pi_k + \log f(x_i|\theta_k)\right)$$

By contrast the log-likelihood is

$$\log f(x|\theta) = \sum_{i=1}^{n} \log \left[\sum_{k=1}^{K} \pi_k f(x_i|\theta_k) \right]$$

Notes

1. Can view the augmentation as way to bring log in the sum.
2. Easy to maximise the augmented log-likelihood given the z_{ik}'s.

E step

In the EM algorithm we update θ^{old} to θ^{new}. In the E step we define the expected log likelihood

$$Q(\theta, \theta^{old}) = \mathbb{E}_{z|x,\theta^{old}} \left[\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\log \pi_k + \log f(x_i|\theta_k) \right) \right]$$

Using Bayes theorem for categorical variables

$$\mathbb{E}_{z|x,\theta^{old}} [z_{ik}] = \pi(z_{ik} = 1|x_i, \theta^{old}) = \frac{\pi_k^{old} f(x_i|\theta_k^{old})}{\sum_{j=1}^{K} \pi_j^{old} f(x_i|\theta_j^{old})} = \gamma(z_{ik})$$

The $\gamma(z_{ik})$'s above are known as the responsibilities.

Hence we can write

$$Q(\theta, \theta^{old}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \gamma(z_{ik}) \left(\log \pi_k + \log f(x_i|\theta_k) \right)$$

M step

The M step: consists of simply maximising $Q(\theta, \theta^{old})$ wrt to θ. Note that the $\gamma(z_{ik})$ are known numbers based on x and θ^{old} so it is usually an easy task.

To maximising $Q(\theta, \theta^{old})$ wrt to π_k's we can use Lagrange multipliers to satisfy the restriction that they sum to one. So we set

$$L = Q(\theta, \theta^{old}) + \lambda \left(\sum_k \pi_k - 1 \right),$$

$$\begin{aligned}
\frac{\partial L}{\partial \pi_k} = 0 \quad &\leftrightarrow \quad \pi_k = \frac{\sum_i \gamma(z_{ik})}{-\lambda} \\
\frac{\partial L}{\partial \lambda} = 0 \quad &\leftrightarrow \quad \sum_k \pi_k = 1 \leftrightarrow \frac{\sum_k \sum_i \gamma(z_{ik})}{-\lambda} = 1
\end{aligned}$$

M step (cont'd)

Note that for all i, $\sum_k \gamma(z_{ik}) = 1$.

Hence

$$\sum_k \sum_i \gamma(z_{ik}) = \sum_i \sum_k \gamma(z_{ik}) = \sum_i 1 = n$$

Therefore $\lambda = -n$.

So we get that $Q(\theta, \theta^{old})$ is maximised at

$$\pi_k^{new} = \frac{\sum_i \gamma(z_{ik})}{n} = \frac{n_k}{n}$$

Example: Gaussian Mixure models

The remaining parameters depend on which type of $f(x_i|\theta_k)$ we have.

For Gaussian mixture models standard MLE methods provide

$$\begin{aligned}
\mu_k^{new} &= \frac{\sum_i \gamma(z_{ik}) x_i}{\sum_i \gamma(z_{ik})} = \frac{\sum_i \gamma(z_{ik}) x_i}{n_k} \\
\Sigma_k^{new} &= n\frac{1}{n_k} \sum_i \gamma(z_{ik})(x_i - \mu_k^{new})(x_i - \mu_k^{new})^T
\end{aligned}$$

Hence the EM algorithm initiates θ and iteratively updates from θ^{old} to θ^{new} until the log likelihood or the parameters converge.

Similar results exist for other distributions such as Bernoulli, Exponential etc.

Connection with K-means

- Mixture models classify individuals to clusters based on the responsibilities $\gamma(\zeta_{ik})$'s, i.e. the posterior probabilities of z, rather than with certainty, aka soft allocation.
- This is reflected on the estimates of θ_k where each x_i is weighted based on how likely an individual is in cluster k.
- In Gaussian mixture models if we set $\Sigma_k = \sigma^2 I_d$ and let $\sigma^2 \to 0$ we get the same solution as with the K-means approach for μ_k. Note that in this case we have hard allocation.
- If we have general $\Sigma_k' s$ the approach coincide with the elliptical k-means.

Selecting the number of clusters

- In both mixture models and K-means it is not easy to select the number of classes.
- The default criterion in the mixture models is the BIC.
- Nevertheless the approach is very sensitive to starting values as the objective is multimodal and is very likely to get trapped in local maxima.
- It is recommended to initialise parameters based on intuition, try out multiple starting points or initialise with the results of another method.

Fully Bayesian approach

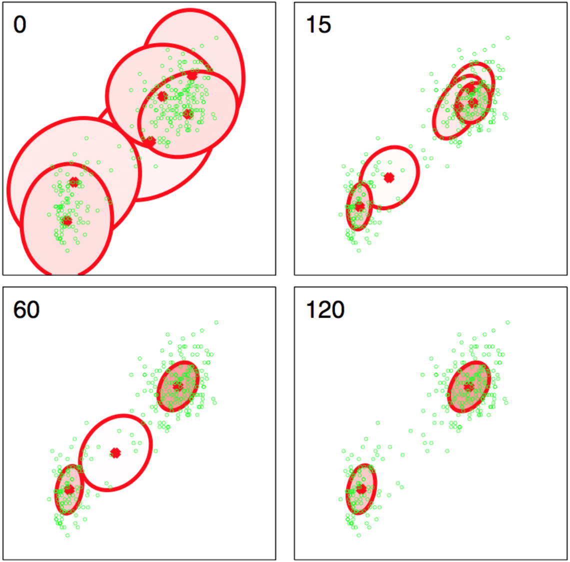
The approach so far was Bayesian with respect to z but not θ. For a fully Bayesian approach priors on θ should be specified.

In Gaussian mixture models example the conjugate priors can be used

$$\begin{aligned} \mu_k &\sim N(\mu_0, \ \Sigma_0) \\ \Sigma_k &\sim \text{IWishart}(W_0, \nu_0) \\ \pi &\sim \text{Dirichlet}(\alpha_0) \end{aligned}$$

The posterior is not available in closed form. But we can consider MCMC algorithms.

Bayesian approach - nor overfit



Today's lecture - Reading

Options reading:

Gelman et al, Chapters 22 and 23