T-61.5040

Learning Models and Methods, (5 cr) L

Spring 2007

Lecturer: **Petteri Pajunen**

Course Assistant: **Ville Viitaniemi**

Lecture 7: 28.2.2007

Bayesian Modeling: Multivariate and Hierarchical Models

- *no teaching next week due to midterm exam week*
- next lecture is on March 14th and next exercises on March 16th

Multivariate Models

- many uncertain quantities in a vector-valued θ
- in principle, Bayesian inference proceeds as before
- specify the prior $p(\theta)$ and the likelihood $p(D|\theta)$, then compute the posterior $p(\theta|D)$

- however, some practical difficulties arise
- assume $\theta = (\theta_1, \ldots, \theta_n)'$ contains some components that are not interesting
- these are called *nuisance parameters*

- as before, we can obtain the *full posterior* $p(\theta|y)$
- at least we can write it as a function of θ using Bayes' theorem $p(\theta|y) \propto p(y|\theta)p(\theta)$
- the posterior is a multivariate function of θ
- unless conjugate priors are used, it is not generally a simple standard distribution

- in practice, various computations on the posterior are required
- for example, we might be interested in θ_1 only
- marginalize out variables $\theta_2, \ldots, \theta_n$
- done by integrating the full posterior over these variables

- the full posterior may be, and often is, impossible to integrate in closed-form
- this integration must be done in one way or another
- for example, the probability $p(0 \leq \theta_1 \leq 1|D)$ cannot be computed from the full posterior without integration

- example: Normally distributed data, mean θ and variance σ^2 unknown
- standard estimates for θ consider σ^2 as a constant
- corresponds to a known σ^2 from Bayes point of view
- unrealistic to assume that σ^2 is known

- suppose only the mean θ is interesting
- this makes σ^2 a *nuisance parameter*
- must be in the model, since it can affect results
- we don't want it in the posterior

- data: $y_1, \ldots, y_n : y_i \sim N(\theta, \sigma^2)$
- Jeffreys' prior: $p(\theta, \sigma^2) \propto \sigma^{-2}$
- the marginalized posterior $p(\theta | y_1, \ldots, y_n)$ can be solved analytically
- the result is not a Normal distribution, unlike for constant σ^2

- the full posterior is easy to write as

$$p(\theta, \sigma^2 | y) \propto \sigma^{-2} \prod_i N(y_i | \theta, \sigma^2)$$

- marginalizing over σ^2 yields a *Student-t distribution*

$$p(\theta | y) = t_{n-1}(\mu_y, s^2/n)$$

where

$$\mu_y = \frac{1}{n} \sum_i y_i \text{ (sample mean)}$$

$$s^2 = \frac{1}{n-1} \sum_i (y_i - \mu_y)^2 \text{ (sample variance)}$$

- Student-t distribution has "heavy tails", meaning that $p(\theta|y) \to 0$ slowly when $|\theta - \mu_y|$ increases
- very large (or small) values of θ are more probable than for a Normal distribution
- special case of one observation is calculated in the exercises
- general result is in Gelman's book, pp. 66-69
- demo: `student.R`

- example: assume that $y_1, \dots, y_n$ are Normally distributed: $y_i \sim N(\theta, \sigma_i^2)$
- note that *each observation y_i has its own variance σ_i^2*
- number of parameters is $n + 1$ and the number of observations is n
- it seems that there is too much freedom in the model

- posterior is obtained in closed form using a Gamma distribution for $p(\sigma_i^2)$ (demo exercise)
- the demo below compares this to estimating θ using a Normal model with constant σ^2
- Bayes result is less sensitive to an outlier
- demo: `robust.R`

Hierarchical Models

- a multivariate model with a special structure
- often prior information is easiest to represent using a hierarchical model
- the structure of the problem may also suggest a hierarchical model

- some benefits of HM's:
  - posterior can be written in a form suitable for simulation (discussed in a later lecture)
  - models can be constructed using simple distributions as building blocks

- example: coin toss
- toss $n = 20$ times and observe $y = 12$ heads
- probability of heads is θ
- you suspect the coin may be biased: denote a biased coin by $b = 1$ and unbiased by $b = 0$
- use a prior $p(b = 0) = p(b = 1) = 0.5$

- assume that biasedness defines θ as

$$p(\theta = 0.5|b = 0) = 1$$
$$p(\theta = 0.7|b = 1) = 0.5$$
$$p(\theta = 0.6|b = 1) = 0.5$$

- now the model is ready: the unknowns are b, θ and we know $y = 12$

- find the probability $p(b = 1|y)$
- the marginal posterior is $p(b|y) = \int p(b, \theta|y)d\theta$
- compute the full posterior:

$$\begin{aligned} p(b, \theta|y) &\propto p(y|b, \theta)p(b, \theta) \\ &= p(y|\theta)p(\theta|b)p(b) \\ &= \mathrm{Bin}(y|n, \theta)p(\theta|b)p(b) \end{aligned}$$

- the equation

$$p(\theta, b|y) \propto p(y|\theta)p(\theta|b)p(b)$$

  is generally true for a hierarchical model
- i.e. data y depends on θ, θ depends on an *unknown* b, and finally there is a prior for b
- the parameter b is called a *hyperparameter*, since it only affects another parameter directly but not data
- this is what makes the model hierarchical

- since $p(\theta|b)$ is nonzero only for $\theta \in \{0.5, 0.6, 0.7\}$, compute the posterior at these values:

$$p(b = 1, \theta = 0.6|y) \propto \text{Bin}(y|n, \theta = 0.6)p(\theta = 0.6|b = 1) \approx 0.045$$
$$p(b = 1, \theta = 0.7|y) \propto \text{Bin}(y|n, \theta = 0.7)p(\theta = 0.7|b = 1) \approx 0.029$$
$$p(b = 0, \theta = 0.5|y) \propto \text{Bin}(y|n, \theta = 0.5)p(\theta = 0.5|b = 0) \approx 0.060$$

- *full posterior* says that values $b = 0, \theta = 0.5$ are most probable, implying that the coin is unbiased
- the marginal posterior is obtained from the full posterior values:

$$p(b = 1|y) \approx 0.074$$
$$p(b = 0|y) \approx 0.060$$

- this is maximized by $b = 1$, which says that the coin is probably biased

- few important things to notice:
  - the data y *does not depend on* b, when θ is known
  - this means that $p(y|\theta, b) = p(y|\theta)$
  - the MAP value of the full posterior conflicts with the MAP of the marginal posterior
  - it is possible to compute $p(\theta) = \int p(\theta|b)p(b)db$, but in general this is more complicated than using a HM

- the full posterior of a HM is often impossible to marginalize by integration
- HM's can be constructed so that they can be approximated by simulation (discussed later)
- one such method requires that we can draw random values from the conditional posteriors $p(\theta|b, y)$ and $p(b|\theta, y)$

- these can be computed as

$$p(b|\theta, y) = \frac{p(y|\theta, b)p(b|\theta)}{p(y|\theta)} = \frac{p(y|\theta)p(b|\theta)}{p(y|\theta)} \propto p(\theta|b)p(b)$$

$$p(\theta|b, y) = \frac{p(y|\theta, b)p(\theta|b)}{p(y|b)} \propto p(y|\theta, b)p(\theta|b)$$

- $p(b|\theta, y)$ can be solved if $p(b)$ is conjugate to $p(\theta|b)$
- $p(\theta|b, y)$ is the non-HM posterior (just consider b as known): conjugate prior gives also this in closed form

Data with a Hierarchical Structure

- example: average length of produced parts
- two sets of measurements (days 1 and 2)
- assume that length varies more between days than within a day
- interesting quantity is the mean length (assume constant variance)

- two easy solutions:
  - consider *all data* as identically distributed
  - consider the data from each day separately
- first case: mean length is the same on both days
- second case: mean length on day 1 has nothing to do with mean length on day 2
- both assumptions unrealistic

- the average length is assumed to change between days
- but we also assume it does not change very much
- this suggests a hierarchical model
- data distribution: day one as $N(\theta_1, \sigma^2)$ and day two as $N(\theta_2, \sigma^2)$

- lets use a prior $p(\theta_i) = N(\theta_i|\mu, \tau^2)$
- the hyperparameters μ, τ are *unknown*
- note that fixed values μ, τ^2 prevent any dependency between the data on different days
- the hierarchical model gives different results than either of the easy solutions

- lets specify the HM
- day 1 data $x_1, \ldots, x_m$ are from $N(\theta_1, \sigma^2)$ and day 2 data $y_1, \ldots, y_n$ are from $N(\theta_2, \sigma^2)$
- prior: $p(\theta_i) = N(\theta_i|\mu, \tau^2)$
- prior for hyperparameters: $p(\mu, \tau^2) \propto \tau^{-1}$

- with known σ^2, the model can be solved (Gelman's book, section 5.4)
- with unknown σ^2 and a prior $p(\sigma^2) \propto \sigma^{-2}$, a closed-form solution is impossible
- but we can compute the conditional posteriors for simulation (exercise problem)

- necessary conditional posteriors are

$$p(\theta_i|\mu, \sigma, \tau, D)$$
$$p(\mu|\theta_1, \theta_2, \sigma, \tau, D)$$
$$p(\sigma^2|\theta_1, \theta_2, \mu, \tau, D)$$
$$p(\tau^2|\theta_1, \theta_2, \mu, \sigma, D)$$

- with some thought some simplifications can be

made:

$$p(\mu|\theta_1, \theta_2, \sigma, \tau, D) = p(\mu|\theta_1, \theta_2, \tau)$$
$$p(\sigma^2|\theta_1, \theta_2, \mu, \tau, D) = p(\sigma^2|\theta_1, \theta_2, D)$$
$$p(\tau^2|\theta_1, \theta_2, \mu, \sigma, D) = p(\tau^2|\theta_1, \theta_2, \mu)$$

- it is relatively straightforward to find these distributions (exercise problem)
- one can use repeatedly the Bayes' Theorem and the product rule and eliminate unneeded parameters

Model Uncertainty

- *model selection* means selecting a model out of several candidates using data
- this is heuristic if done using only the information contained in the model(s) and data
- proper way is to use *one model* $p(\theta, D)$ as discussed before
- so there is nothing to select in theory

- what if one has more than one possible model?
- example: linear vs. nonlinear model
- if data is known to be linear, then the correct model is obviously linear
- but one might be uncertain of the linearity
- linear model underfits nonlinear data, so perhaps it is better to use a nonlinear model?

- major problem with "flat" nonlinear models: they overfit linear data
- *if the model corresponds to information about the problem*, then this overfitting is not a problem
- a nonlinear model does not generally put a high probability on the possibility that data is linear

- we can construct a hierarchical model which includes both a linear and a nonlinear model
- a hierarchical nonlinear regression model:

$$
\begin{aligned}
p(Y|\theta, \sigma^2) &= N(Y|X\theta, \sigma^2 K) \\
p(\theta|\sigma^2, \tau^2, W, \theta_0) &= N(\theta|\theta_0, \sigma^2\tau^2 W) \\
p(\theta_0) &= N(\theta_0|\mu, B)
\end{aligned}
$$

- hyperparameters τ^2, σ^2 have inverse Gamma prior distributions, and the covariance matrix W has an inverse Wishart prior distribution

- the model is constructed using conjugate priors, so that conditional posteriors are obtained in closed form (see Gelman et al. for more details on the conjugate priors)
- the nonlinearity of the model comes from the covariance matrix K
- the elements of K are inner products of feature vectors (as in SVM)
- these methods will be discussed in more detail in a later lecture

- the hyperparameters increase the uncertainty in the model
- for example, in a flat model the prior mean of $p(\theta)$ would be fixed e.g. to zero
- the hierarchical prior does not explicitly fix $p(\theta)$ as a zero-mean Normal distribution
- generally a hierarchical prior amounts to a more noninformative prior

- the above nonlinear model includes a linear model as a special case
- if the covariance matrix K is selected to be the identity matrix, then a hierarchical linear regression model is obtained
- these models may be *combined* simply by adding a binary indicator, which selects one of the models
- the indicator is a hyperparameter, and thus requires a hyperprior

- a predictive distribution can be computed as before: all parameters are integrated out, also the indicator variable
- this means that predictions will be made using both models, but the weighting is affected by data
- demo: gpmixture.R

- summary of HM's:
  - learning problem may have a hierarchical structure
  - allow models with no closed-form posterior to be constructed with closed-form conditional posteriors
  - suitable for approximation by certain simulation methods
  - combining models, adding uncertainty lead to HM's