General Formula

Notation

D : the data set
D_i : the i th data point in data set
$\boldsymbol{\theta}$: the model parameters

Formula

Our aim is to infer the posterior probability of the model from the data. According to Bayes formula, we have

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(D)}$$

Where the likelihood can be rewritten as

$$P(D|\boldsymbol{\theta}) = P(\{D_i\}|\boldsymbol{\theta})$$

If each data point is independent of each other (if not see Dependent data points in Bayesian Regression), we have

$$P(\{D_i\}|\boldsymbol{\theta}) = \prod_i^m P(D_i|\boldsymbol{\theta})$$

Hence we focus on one of data point, say D_i. In D_i, there is one dependent observed variable $y_{obs,i}$, and n independent observed variables $\{x_{obs,ij}\}$. Thus

$$P(D_i|\boldsymbol{\theta}) = P(y_{obs,i}, \{x_{obs,ij}\}|\boldsymbol{\theta})$$

In general case, we have errors in our data, and our model is usually established for intrinsic y and $\{x_j\}$. Hence we should link our model to the data:

$$\begin{aligned} P(y_{obs,i}, \{x_{obs,ij}\}|\boldsymbol{\theta}) = \\ \int P(y_{obs,i},\{x_{obs,ij}\}|y, \{x_j\})\, P(y, \{x_j\}|\boldsymbol{\theta})\, dy\, d\{x_j\} \end{aligned}$$

The best (but rare) case is that this integration has analytic solution, if not, we need to do it numerically.

$P(y_{obs,i}, \{x_{obs,ij}\}|y, \{x_j\})$ is related to error properties, and $P(y, \{x_j\}|\boldsymbol{\theta})$ is determined by our model.

If we have Gaussian error, this integration can be calculated by Gauss–Hermite quadrature. If not, the Monte Carlo integral (with importance sampling) is a working choice.

In next step, we decompose $P(y, \{x_j\}|\boldsymbol{\theta})$ as

$$P(y, \{x_j\}|\boldsymbol{\theta}) = P(y|\{x_j\}, \boldsymbol{\theta_1})P(\{x_j\}|\boldsymbol{\theta_2})$$

Where the first item in right describes how do $\{x_j\}$ determine y, the second item shows the distribution of $\{x_j\}$, and $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}$ are related parameters for first and second part separately.

According to the equation above, we should know **though we are not usually interested in the distribution of $\{x_j\}$, it does have its own contribution to $P(D_i|\boldsymbol{\theta})$.**

Linear Regression

If the model that relates y to $\{x_j\}$ is a linear function, we say the regression is linear regression. Furtherly, if the error for y_{obs} is gaussian like, and there is no error for $\{x_{obs}\}$, we have analytic solution for posterior, which can be found in wikipedia of Bayesian linear regression.

Astronomy example: Simple Stellar Population (SSP) fitting at given kinematics and extinction.

Note

If there is no error for $\{x_{obs}\}$, we can regard it as a Dirac Delta function. By using the properties of Dirac Delta function we have

$$\begin{aligned}P(y_{obs,i}, \{x_{obs,ij}\}|\boldsymbol{\theta}) &= \int P(y_{obs,i}|y)\ P(y, \{x_{obs,ij}\}|\boldsymbol{\theta})\ dy \\ &= \int P(y_{obs,i}|y)\ P(y|\{x_{obs,ij}\}, \boldsymbol{\theta_1})\ dy\ P(\{x_{obs,ij}\}|\boldsymbol{\theta_2})\end{aligned}$$

In thie case, the distribution of $\{x_j\}$ is just constant coefficient of likelihood (consequently posterior), so we can safely ignore it.

In most case, the analytic solution is much efficient then other method such as MCMC. So try to use it, it will save a lot of time and computation resource for you (and your organization).

Useful results

When deal with the posterior of linear regression, we may need do some integration of multi-dimensional Gaussian function (e.g. you may want to marginalize some variables).

Astronomy example: we do SSP fitting, we can marginalize the SSP weights.

Some of useful results can be found in here.

Challenge

These results are for full space intergration, the SSP weights are considered as non-negative. The analytic solution for intergration ove positive reals can be real challenge.

Refer to this question

Deal with The Error of Independent Variables

Ignore it

If the typical scale of the error of independent variables is smaller than the scale at which the $P(y, \{x_j\}|\boldsymbol{\theta})$ changes significantly.

An intuitive explanation is in this integral

$$\int P(y_{obs,i}, \{x_{obs,ij}\}|y, \{x_j\})\ P(y, \{x_j\}|\boldsymbol{\theta})\ dy\, d\{x_j\}$$

if $P(y, \{x_j\}|\boldsymbol{\theta})$ does not change approximately in the region where $P(y_{obs,i}, \{x_{obs,ij}\})$ has a non-zero value (this region usually center at $\{x_j\}$), we can regard it as a constant. Thus we have the same results as no error case:

$$\begin{aligned}
&\int P(y_{obs,i}, \{x_{obs,ij}\}|y, \{x_j\})\ P(y, \{x_j\}|\boldsymbol{\theta})\ dy\, d\{x_j\} \\
\approx &\int P(y_{obs,i}, \{x_{obs,ij}\}|y, \{x_j\})\ P(y, \{x_{obs,ij}\}|\boldsymbol{\theta})\ dy\, d\{x_j\} \\
\approx &\int P(y_{obs,i}|y)\ P(y, \{x_{obs,ij}\}|\boldsymbol{\theta})\ dy
\end{aligned}$$

The second $\approx$ use the fact that the $P(y_{obs,i}, \{x_{obs,ij}\}|y, \{x_j\})$ is a probability so it normalizes to 1 (ignore the correlation between $y_{obs,i}$ and $\{x_{obs,ij}\}$.

A more mathematical way to think about it is to refer to First mean value theorem for definite integrals

Deal with it

In this case should establish the model for $P(y_{obs,i}, \{x_{obs,ij}\}|y, \{x_j\})$ and $P(\{x_j\}|\boldsymbol{\theta_2})$.

It is usually not a problem for first item if you know the properties of error and the error level is part of your data set.

For second item, if you know the intrinsic distribution of $\{x_j\}$, it will be easy. Just use the known distribution and sample or optimize the $\boldsymbol{\theta}_2$ together with $\boldsymbol{\theta}_1$.

If you do not know the intrinsic distribution, for 1d case, you can use the Generalized Beta Distribution, which is very studied and has enough flexibility to back to a variety of common distributions.

For high demission case, normalizing flow can be a better choice.

If deal with $\boldsymbol{\theta}_2$ together with $\boldsymbol{\theta}_1$ leads to curse of dimensionality, you can constrain $\boldsymbol{\theta}_2$ use the information in $\{x_{obs,ij}\}$:

$$P(\{x_{obs,ij}\}|\boldsymbol{\theta}_2) = \int P(\{x_{obs,ij}\}|\{x_j\})\ P(\{x_j\}|\boldsymbol{\theta_2})\ d\{x_j\}$$

Sampler

You need sampler to sample your posterior, refer to Sampler.