Statistics for Astronomers: Lecture 21, 2019.05.16

Prof. Sundar Srinivasan

IRyA/UNAM

Recall: Basic Monte Carlo methods

We use Monte Carlo methods in order to either sample from a distribution or compute an expectation value of a function over a distribution.

Simple MC: $\mathbb{E}[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(X_i)$, where $X_i \sim p_X(x)$.

Problem: $p_X(x)$ may be too complicated (esp. multidimensional), and/or difficult to sample from.
Solution: rejection sampling, importance sampling – sample from a proposal distribution instead of the target distribution.

Problem: "curse of high dimensionality" – the proposal needs to be as close as possible to the target; as d increases, the discrepancy increases exponentially.
Solution: Markov Chain Monte Carlo (MCMC); explore multidimensional parameter space by sampling ("travelling") along regions/zones of high probability.

Recall: Markov Chain Monte Carlo

The purpose is to generate draws from a target distribution $p_X(x)$. Algorithms are framed in such a way that the Markov process asymptotically approaches a unique stationary distribution $\pi(x)$ such that $\pi(x) = p_X(x)$.

After N steps (iterations), $\mathbb{E}[f(X)] \approx \frac{1}{N}\sum_{i=1}^{N} f(X_i)$, where X_i is the states explored at step i. As N increases, the average converges to $\mathbb{E}[f(X)]$ due to the Ergodic Theorem.

Markov Chain: the decision as to where to advance in parameter space depends only on the current location. The next "link" in the chain is decided using a jump distribution.

Monte Carlo: Pseudorandom numbers are generated in order to sample the target distribution.

Dependent sampling: Future step depends on present step.

The algorithm for each iteration:

1. Select starting point/state (parameter value θ_0).
2. Evaluate unnormalised posterior probability at this point.
3. Draw new parameter value θ_{j+1} from a proposal distribution.
4. Evaluate unnormalised posterior probability for this value.
5. Decide whether you will accept the new value.

Recall: The Metropolis-Hastings Algorithm

One of the oldest MCMC implementations.

1. Select starting point/state (parameter value θ_j).
2. Evaluate unnormalised posterior probability at this point.
3. Draw new parameter value θ_{j+1} from a proposal distribution ("jump distribution" centered on current value).
   For the Metropolis algorithm, the jump distribution must be symmetric: $p(\theta_{j+1}|\theta_j) = p(\theta_j|\theta_{j+1})$. Usually, $\theta_{j+1} \sim \mathcal{N}(\theta_j, \sigma^2)$, with σ the characteristic "step size". The results may depend on σ (small = high acceptance rate but more iterations required, and vice versa).
4. Evaluate unnormalised posterior probability for this value.
5. Decide whether you will accept the new value accept with probability $\alpha = p(\theta')/p(\theta_j)$.
   This is implemented by comparing α to a uniform random variable $u \sim U(0,1)$. If $\alpha > u$, the new value is accepted. If not, the old value is retained.
   This is because $p(\alpha \geq u) = p(u \leq \alpha) \equiv F_u(\alpha) = \alpha$ for $U(0,1)$.

Regression

Terminology

X variable(s): predictor, regressor, feature, independent variable (**).
Y variable: outcome, response, target, dependent variable.
Independent variable fallacy (see Hogg et al. 2010): pick the one with lower uncertainties as independent variable.
Feigelsen & Babu: Careful when interpreting data from a new method/instrument – this needs to incorporate the effect of improving technology when compared with older observations!

Regression function: $Y(x) = \mathbb{E}[Y|X = x]$
Regression can be nonparametric (*e.g.*, kriging, interpolation using machine learning) or parametric (*e.g.*, χ^2 fitting, MLE).
Once regression is performed, we can estimate Y for a new set of X values. The Y can be discrete ("classification") or continuous ("prediction").

Regression model: $Y = f(X) + \epsilon$; $\mathbb{E}[\epsilon] = 0$. (Even if X isn't random, Y is, because of ϵ.)
ϵ can be a combination of measurement error and intrinsic variation. Typically one is much smaller and can be ignored compared to the other.

"Linear" parametric regression: linear in parameters, not necessary in the regressor.
$Y = mX + c$, $Y = \alpha\sqrt{X^2 + 1}$, $Y = \mathrm{constant}$ are examples of linear regression models, while $Y = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2\right]$, $Y = AX/B$, $Y = \delta + \left(\frac{X-\xi}{\Delta v}\right)^2$ are not.

General procedure and notation

1. Dependent variable decision
2. Model choice(s)
3. Method of parameter estimation/choice of goodness-of-fit ("objective function")
   Method of moments, least-squares, MLE, Bayesian inference.
   – estimate best values for parameters and their variances; confidence intervals.
4. Occam's Razor model validation/selection.
5. Prediction/classification can then be performed – for a given set of x values, compute y along with variances and thus confidence intervals for these y.

Notation

In expressions involving matrix products, bold lowercase symbols refer to column vectors, and bold uppercase symbols are matrices.

Henceforth, the lowercase bold **x** refers to the row vector containing the parameters, not the values x_i of the dependent variable (which are stored in a matrix **A**).

Heteroskedasticity

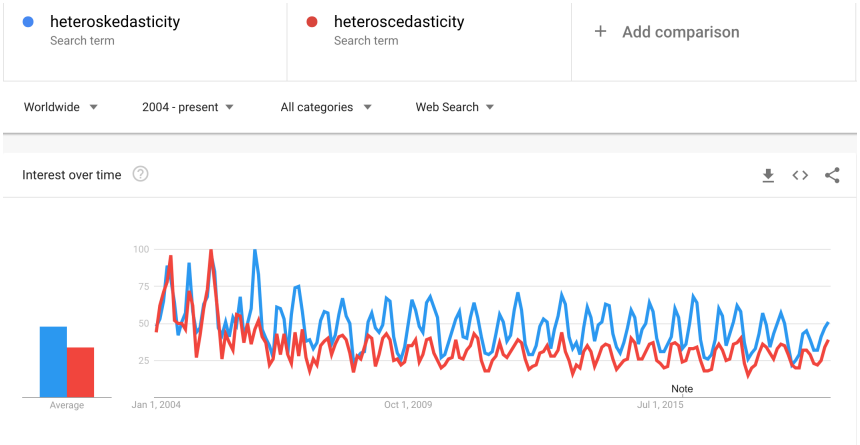
If the uncertainties ϵ_i associated with the dependent variable values y_i are drawn from a distribution with the same variance, the uncertainties are said to be homoskedastic. If they are drawn from distributions with differing variances, they are heteroskedastic.

Typical measurements in astronomy come with heteroskedastic uncertainties – they may be drawn from the same distribution, but their variances are not identical.
Example: photometric measurements of a galaxy-wide stellar population span a large range in magnitude. While we assume that the magnitudes are drawn from a normal distribution, the associated standard deviations are not the same if the magnitudes themselves are wildly different.

Heteros*edasticity

*In a brief article in Econometrica[1985b], J. Huston McCulloch advanced that "[t]he most pressing issue in econometric orthography today is whether heteros*edasticity should be spelled with a k or with a c. At the time his note went into press, the majority of published manuscripts spelled it as heteroscedasticity.*

*– "When did we begin to spell heteros*edasticity correctly?" A. R. Paloyo, Ruhr Economic Papers 0300, 2011.*



Ordinary least-squares (OLS) estimation

The objective function in least-squares estimation is the sum of the squares of the differences between the observed data and the model prediction ("residues").
As a first step, we ignore the effect of the (homoskedastic) uncertainties. In typical situations, however, we weight each residue by the inverse of the standard deviation of the associated uncertainty.

The objective function for the problem is the residual sum of squares: $RSS = \sum_{i=1}^{N} \left(y_i - y_{\mathrm{mod},i}\right)^2$.

We can rewrite this in terms of a matrix product:

$$RSS = \sum_{i=1}^{N} \left(y_i - y_{\mathrm{mod},i}\right) \cdot 1 \cdot \left(y_i - y_{\mathrm{mod},i}\right) = (\mathbf{y} - \mathbf{y}_{\mathrm{mod}})^T \cdot \mathbf{\Sigma}^{-1} \cdot (\mathbf{y} - \mathbf{y}_{\mathrm{mod}}),$$

where $\mathbf{y}$ and $\mathbf{y}_{\mathrm{mod}}$ are $N \times 1$ column vectors and $\mathbf{\Sigma}^{-1}$ is, in this case, the $N \times N$ identity matrix $\mathbb{I}_{N \times N}$.

$\mathbf{y}_{\mathrm{mod}}$ may depend on one or more parameters $\boldsymbol{\theta}$, which can be solved for by minimising RSS. This is done by setting the derivative w.r.t. each parameter to zero.

In general, Σ is the covariance matrix that incorporates information about correlated uncertainties.

OLS estimation – linear model

The model is linear in the regressor: $y_{\mathrm{mod},i} = mx_i + b$. m and b are the slope and intercept of the line, respectively. The errors ϵ are homoskedastic with variance σ.

$$RSS = \sum_{i=1}^{N} \left(y_i - y_{\mathrm{mod},i}\right)^2 = \sum_{i=1}^{N} \left(y_i - mx_i - b\right)^2.$$

Optimisation:

$$\left.\frac{\partial RSS}{\partial m}\right|_{(m,b)=(\hat{m},\hat{b})} \propto \sum_{i=1}^{N} \left(y_i - \hat{m}x_i - \hat{b}\right) \cdot x_i = 0 \Longrightarrow \hat{m} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \equiv \frac{S_{xy}}{S_{xx}}.$$

$$\left.\frac{\partial RSS}{\partial b}\right|_{(m,b)=(\hat{m},\hat{b})} \propto \sum_{i=1}^{N} \left(y_i - \hat{m}x_i - \hat{b}\right) = 0 \Longrightarrow \hat{b} = \bar{y} - \hat{m}\bar{x}.$$

RSS is distributed as $\chi^2(N-2)$ (we determined two parameters using the data). The estimate for the variance in $\mathbf{y}$ is then $S^2 = \frac{RSS}{N-2} \sim \sigma^2 \frac{\chi^2_{N-2}}{N-2}$.

OLS linear model – variances

Recall: The y_i are random variables because of the uncertainties ϵ_i, which have variance σ^2.

$$S_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \bar{x}\bar{y}$$

$$Var(S_{xy}) = \frac{1}{N^2}\sum_{i=1}^{N} x_i^2\sigma^2 - \bar{x}^2\frac{\sigma^2}{N} = \sigma^2\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \bar{x}^2\right) = \sigma^2 S_{xx}.$$

$$\Longrightarrow Var(\hat{m}) = Var\left(\frac{S_{xy}}{S_{xx}}\right) = \frac{1}{S_{xx}^2}Var(S_{xy}) = \frac{\sigma^2}{S_{xx}}$$

$$\Longrightarrow Var(\hat{b}) = Var(\bar{y} - \hat{m}\bar{x}) = Var(\bar{y}) + \bar{x}^2 Var(\hat{m}) = \frac{\sigma^2}{N} + \bar{x}^2\frac{\sigma^2}{S_{xx}} = \sigma^2\left(\frac{1}{N} + \frac{\bar{x}^2}{S_{xx}}\right)$$

For large N, the best-fit parameter values are normally distributed about their means with variances as given above. If σ is unknown, the it also has to be estimated from the data: $\hat{\sigma} = S$. The slope and intercept are then drawn from a t distribution.

Prediction: the predicted value of y for a value of x is t-distributed around $mx + b$ with standard deviation $S\sqrt{1 + \frac{1}{N}\frac{(x-\bar{x})^2}{S_{xx}}}$.

OLS in matrix notation

For multivariate problems, it's much easier to work with matrices.

In general, the regression relation becomes $\mathbf{y} = \mathbf{A}\mathbf{x}$.

Linear case: $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_N \end{bmatrix}_{N\times 1}$ $\mathbf{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_N \end{bmatrix}_{N\times 2}$ $\mathbf{x} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}_{2\times 1}$ (θ_1 = intercept, θ_2 = slope)

The general covariance matrix $\mathbf{\Sigma}$ for heteroskedasstic uncertainties is such that $\Sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$, where ρ_{ij} is the correlation coefficient between σ_i and σ_j.

For uncorrelated uncertainties, $\mathbf{\Sigma}$ is diagonal: $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \cdots & & & \\ 0 & 0 & \cdots & \sigma_N^2 \end{bmatrix}_{N\times N}$

$RSS \propto (\mathbf{y} - \mathbf{A}\mathbf{x})^T \mathbf{\Sigma}^{-1} (\mathbf{y} - \mathbf{A}\mathbf{x})$, where $\mathbf{\Sigma}^{-1} = \frac{1}{\sigma^2}\mathbb{I}$ (homoskedastic uncorrelated uncertainties).

If RSS is minimized w.r.t. $\mathbf{x}$, we get the matrix product version of the results obtained in the previous slide: $\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{y} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{y}$.

$(\mathbf{A}^T\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}$ is the covariance matrix for the parameters.

OLS – comparison to MLE

If the uncertainties ϵ are normally distributed and homoskedastic, the associated likelihood is

$$\mathscr{L}(m, b) = \prod_{i=1}^{N} \exp\left[-\frac{1}{2}\left(\frac{y_i - mx_i - b}{\sigma}\right)^2\right] \Rightarrow \ln\mathscr{L} = \text{constant} + \sum_{i=1}^{N}\left(\frac{y_i - mx_i - b}{\sigma}\right)^2.$$

The objective function RSS is related to $\ln\mathscr{L}$, so the results from optimising RSS are equivalent to the maximum likelihood estimate for this problem.

For heteroskedastic uncertainties, we replace σ with N distinct values σ_i. The matrix product version of the log-likelihood is

$\ln\mathscr{L} = \text{constant} + (\mathbf{y} - \mathbf{A}\mathbf{x})^T\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x})$.

"Inverse" OLS and OLS bisector

"Inverse" OLS – switch the dependent variable with the independent one and derive slope and intercept for this case.

OLS bisector – Derive slope and intercept such that this line bisects the OLS and Inverse OLS lines. These two methods are useful when the uncertainties along both axes are comparable.

The Hogg et al. paper warns against using either method!