12.010 Computational Methods of Scientific Programming

Lecture 11: Data Analysis

Summary

- Data analysis type problems: Often under the generic term regression.
- Concept of data and model
- ”Cost functions” How to define what is a “good” fit to the data
- Least-squares and weighted least-squares
- Relationship to Gaussian multivariate distribution.
- Explicit implementation.
- Simple: polyfit
- More general curve_fit

Concepts

- Cost functions: When we “fit” a model to data, what do we mean
- Elements of inversion:
  - Mathematical models
  - Stochastic models
  - Inversion scheme (minimize cost function/maximize information)
  - Data and Models: No distinction
- Types
  - Maximum likelihood estimator
  - Bayes estimator

Data, models and errors

- The general concept is that we have ”data,” which can be expressed as a model determined by parameters whose values we don’t know, with additive noise whose statistical characteristics we assume are known.
- Example: Drop a ball and measure its height as a function of time.
  - Data: Height measurements at known times (most often)
  - Model: initial position, maybe velocity, and acceleration due to gravity
  - Noise: Random, uncorrelated values added to each measurement.
- Inversion/regression problem: Estimate the model parameters that “best” fit the observed data and determine the uncertainties of the model parameters.

Cost functions

- What do we mean by ”best fit” to data
- Generally, a regression or inversion problem is posed as determining the “best” set of parameters to represent an observed data stream with random errors.
- The definition of best is encapsulated either in a cost function to be minimized or a likelihood or information function to be maximized.
- Generally, a cost function of the sum of squares of data residuals or weighted residuals (more later) with possibly a model function cost, often in the form of magnitude or roughness.

Creating cost function

- Here, we examine the common problem of having an observed data set and a parametric model, which we believe describes the behavior of the data.
- Some nomenclatures use ”observations” (measured data) and “observables” (the quantity observed for which there is a mathematical model). The expectation of the observations should match the mathematical model.
- The cost function is computed from the differences between the observations and the computed values of the observables with a specific set of parameter values. As the parameter values are changed, the cost function value changes.
- In addition to the mathematical models, stochastic models are often used to represent the data noise. Different stochastic models lead to different parameter estimates even when the mathematical model is unchanged.

Example cost functions

- A very common cost function is:

$$C(x) = \sum (d_i - m(x))^2$$

Where *d_i* are data, *x* are parameters and *m(x)* is the model. Regression/inversion finds x that minimalizes C(x). This an L2 norm; sometimes L1 norm used.

- A more statistical model might be

$$C(x) = \sum (d_i - m(x))^T V_d^{-1} (d_i - m(x))$$

- V_d is a data covariance matrix. Or more complicated

$$C(x) = \sum \left(d_i - m(x)\right)^T V_d^{-1} (d_i - m(x)) + \lambda \nabla m(x)^2$$

- Where $\nabla m(x)$ could be a smoothness expression and λ controls the weight between the model and data (we do not consider this case in this class)

Elements of inversion

- **Mathematical model:**
- This is the parametric model that determines what would be observed with noiseless data for a set of model parameters. Often, models have many parameters, and one decision is which of these parameters are known accurately enough that they do not need to be estimated, i.e., we use the apriori value of the parameters.
- Which model parameters are estimated depends on data sensitivity, e.g., gravity field coefficients are satisfactory for GPS but may need to be estimated for data collected on/from low Earth-orbiting satellites.

Elements of inversion

- **Stochastic model** (nominally optional)
- This is the error model for the data being collected.
- Sometimes these models are not explicitly included, which results in an assumption of uncorrelated data with all the same standard deviation.
- Often, simplified models are used because they are computationally too difficult for the full model.
- The stochastic model affects the inversion because it changes the weight between different data.
- Overall scaling of the stochastic model does not affect the parameter estimates.
- However, the scaling and stochastic model generally affects the stochastic model for the parameter estimates.

Elements of inversion

- **Inversion scheme** (min/max problem)
- This is a scheme for determining the model parameter values that minimize the cost function. (For information content or likelihood functions it is a maximization problem).
- For a large group of problems with L2 norm minimizations (sum of squares), closed-form linear algebra matrix formulas can often be used.
- For some classes of problems, an iterative search type technique is needed.

Data and models : **Data≡Model≡Data**

- In the proceeding slides, I used terms of data and models, but in practice, they are the same thing, or more appropriately, what we call data these days are really the model parameters from an earlier inversion, e.g., GPS phase and range measurements are outputs (model parameters) from a Kalman filter running in the GPS receiver.
- In this light, a cost function of the form:

$$C(x) = \sum \left(d_i - m(x)\right)^T V_d^{-1}(d_i - m(x)) + \lambda \nabla m(x)^2$$

can be considered as just two types of data being used in the cost function.

Estimator types

- Maximum likelihood estimator is a method that determines values for the parameters of a model. The parameter values are found such that they maximize the likelihood that the process described by the model produced the data that were observed.
- The likelihood function is usually the probability density function(PDF) evaluated at the data points. Parameters in the PDF (e.g., mean) are varied to maximize the PDF.
- For Gaussian PDFs this can be easily done. By taking the log$_e$ of the PDF, the $f(x;\mu,\sigma) \propto \frac{1}{\sigma\sqrt{2\pi}}\prod e^{-(x_i-\mu)^2/(2\sigma^2)}$ the product becomes a sum which can differentiated with respect to say μ (results in mean being maximum likelihood estimator.

Gaussian MLE

- For multi-variant Gaussian is
$$f(\boldsymbol{x};\boldsymbol{\mu}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{V}|}}e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{V}^{-\boldsymbol{1}}(\boldsymbol{x}-\boldsymbol{\mu})}$$
maximizing $f(\boldsymbol{x};\boldsymbol{\mu})$ with respect to$\boldsymbol{\mu}$ ($\boldsymbol{\mu}$ is the model values computed to specific parameter values happen when
$$-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{V}^{-\boldsymbol{1}}(\boldsymbol{x}-\boldsymbol{\mu})$$
is minimized. We will call this weighted least squares (WLS).
- MLE estimators find the mode of the distribution. For some distributions this values differs from the median and mean.

Estimator types

- All the problems we explore will use an L2 norm, i.e., the sum of squares of residuals or model deviations.
- In terms of model effects, the L2 norm makes the parameter estimates sensitive to data outliers (the model changes trying to make them small because of the squared effect).
- To overcome L2 norm outlier problems, some approaches use an L1 norm, but there are no direct matrix solutions to these problems. These are often referred to as robust estimators. The non-continuous derivative of the L1 norm makes normal differentiation to find the minimum difficult.

Minimize residuals squared*

- Algebraic approach (dates to Gauss).
- Mathematical model: We have n observables, y, that are modeled as a linear combination of m parameters, x, through a matrix A

  $$y\ [n\times 1] = A\ [n\times m]\ x\ [m\times 1]$$

  where the values in [] are dimensions
- Cost function is

  $$C(x) = \sum (y - Ax)^2$$

- Expanding the sum square and differentiating to find minimum yields

  $$A^T A x\ - A^T y = 0$$

  $$\hat{x} = (A^T A)^{-1} A^T y$$

- This is standard least-squares.
- Jupyter notebook: Lec12_Regression.ipynb

Weighted minimization: Gaussian*

- The concept here is to "weight" the data according to quality. With the assumption of a data covariance function

$$f(y;\boldsymbol{\mu}(\boldsymbol{x})) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{V}|}} e^{-\frac{1}{2}\left(y-\boldsymbol{\mu}(x)\right)^T \boldsymbol{V}^{-\mathbf{1}}(y-\boldsymbol{\mu}(x))}$$

- The minimization here is for $\left(y-\boldsymbol{\mu}(x)\right)^T \boldsymbol{V}^{-\mathbf{1}}(y-\boldsymbol{\mu}(x))$ where $\boldsymbol{\mu}(x)$is the model value of y. For linear model $\boldsymbol{\mu}(x) = \mathrm{Ax}$.
- Weighted least squares is

$$\hat{x} = \underset{\text{Normal Equations}}{\underleftrightarrow{(A^T V^{-1} A)}}{}^{-1} \underset{\text{B-vector}}{\underleftrightarrow{A^T V^{-1} y}}$$

Slightly non-linear problem

- The standard method for non-linear problems is to use a Taylor series expansion.

$$y = f(x_0) + \frac{\partial f}{\partial x}\Delta x = y_0 + A\Delta x$$

where A is the partials, design or Jacobian matrix

- The WLS estimate of the change to the parameter values is

$$\begin{aligned}
&\Delta y = y - f(x_0) \\
&\widehat{\Delta x} = (A^T V^{-1} A)^{-1} A^T V^{-1} \Delta y \\
&\hat{x} = x_0 + \widehat{\Delta x}
\end{aligned}$$

- System is iterated until $\widehat{\Delta x}$ is small compared to its sigma.

Python implementations

- Notebook so far has explicitly code the solutions using numpy matrix routines.
- Now examine to common Python methods: Both have MATLAB counter parts:
    - polyfit (numpy)
    - curve_fit (scipy ; curvefit MATLAB curvefitting tool box.
- Quality of fit: Sum of residuals squared divided by sigma-squared; call chi-squared.
- Look at cases in notebook and problems with non-linear.