Models vs. Data

10.34 Fall 2015

by W.H. Green

Definitions

- We're comparing experimental Data points y_i mesured with knob settings "x" to model predictions $f_i(x,\theta)$ where θ are the parameters in the model we cannot control

Start by assuming large number N of repeats of each experiment…

- By central-limit-theorem of statistics, for case with a single observable, a true model, accurate parameters and knob values, and many repeats:

  $p(\langle y \rangle | \underline{x}, \underline{\theta}) = (2\pi)^{-1/2}\, \sigma^{-1} \exp(-\chi^2/2)$

- Where $\chi^2 = (\{\langle y \rangle - f(\underline{x}, \underline{\theta})\}/\sigma)^2$

  And estimated variance of the mean $\sigma = (\langle y^2 \rangle - \langle y \rangle^2)/N^2$

- For multiple quantities measured on same experiment need to consider Covariance of data: $Cov_{ij} = \{\Sigma\, (y_{i,n} - \langle y_i \rangle)(y_{j,n} - \langle y_j \rangle)\} / N$
- So we estimate Covariance of the Means (for large N): $C_{ij} \sim Cov_{ij} / N$

For many experiments dependent on same parameters θ

Each measurement repeated many times with same knob settings $\underline{x}_m$, then new knob settings; all the knob settings $\underline{x}_{m(k)}$ are stored in matrix **X.** The mean measurements are stored in a K-vector $<\underline{y}>$. If multiple observables $\{y_i, y_j, ...\}$ measured in each experiment K>M.

$$p(<\underline{y}>|\mathbf{X},\underline{\theta})=(2\pi)^{-K/2}|C|^{-1/2}\exp(-\tfrac{1}{2}\chi^2)$$

where

$$\chi^2 = \Sigma\Sigma\,(<y_k>-f_k(\underline{x}_{m(k)},\underline{\theta}))D_{kz}(<y_z>-f_z(\underline{x}_{m(z)},\underline{\theta}))$$

and D = inv(C)

- Often the covariance is ignored, then $D_{kz} = \delta_{kz}\sigma_k^{-2}$

Probability of observation depends on χ^2; if very improbable we flag a *discrepancy* between model & data

- User must decide tolerance on "improbable".
- For example: If you decide <5% chance is improbable, and you performed 12 experiments (each repeated many times to get a good average $<y_k>$ and estimate of σ_k) and adjusted 2 model parameters, then you can use Matlab function chi2inv:

  chi2max = chi2inv(0.95,12-2)

  in this case case chi2max =18.3

if measured $\chi^2 > 18.3$ you would say there is a discrepancy between the model and the data

Origin of chi2inv

The probability that a measured data set (with many repeats) would yield a χ^2<Q is given by:

$$\text{Prob}(\chi^2<Q) = \iint d^K\langle \underline{y}\rangle \; p(\langle \underline{y}\rangle | \mathbf{X}, \underline{\theta}) \; H(Q-\chi^2)$$

where H is the Heaviside function.

This K-dimensional multiple integral can be simplified by change of variables to the single integral shown in the Matlab chi2inv documentation.

If M parameters have been adjusted to fit the data it is customary to use K-M degrees of freedom when computing chi2inv (this assumes each parameter adjustment really improved the fit). If no adjustment to fit the data (a pure prediction), M=0.

If you select a desired Probability, that choice fixes the value of Q (aka chi2max).

Once we have decided the maximum χ^2 we will tolerate Q, than we have defined a "region of indifference" in parameter space

- As far as we can tell from our experiment, any θ which gives a "good enough" fit is OK, we cannot discriminate.
- To see the range of acceptable parameter values, plot the hypersurface $\chi^2(\underline{\theta})$=Q. Any $\underline{\theta}$ inside the surface is acceptable.
- For a model which depends nonlinearly on the parameters, the shape of the region can be quite convoluted….

Bayesian view

$p(\mathbf{X},\underline{\theta}|\langle\underline{y}\rangle) = p_{prior}(\underline{\theta})p_{prior}(\mathbf{X})p(\langle\underline{y}\rangle|\mathbf{X},\underline{\theta})$

where “prior” means “we have other prior information about these values, not just what we can infer from this data set”.

Usually journal readers are not interested in our imprecise knowledge of our knob settings, so we integrate this uncertainty out to get our new improved “posterior” $p(\underline{\theta})$ that we will report:

$p(\underline{\theta}) = p_{prior}(\underline{\theta}) \iint d^{W}\mathbf{X}\; p_{prior}(\mathbf{X})p(\langle\underline{y}\rangle|\mathbf{X},\underline{\theta})$

Contours of this new $p(\underline{\theta})$ can also have a very convoluted shape…

Simplifying from confidence regions to separate confidence intervals

- Often people like to report parameter values one at a time, e.g. if one θ is a heat capacity:

  $C_p(533\ K) = 89.3 \pm 0.2$ J/mol-K

- Usually people report the best-fit value as the nominal value and then need to give an estimate of the confidence interval. One way to compute the upper limit of the interval:

  $\theta_{v,max} = \max_{\theta} \theta_v$

  s.t. $\chi^2(\underline{\theta}) < Q$

Correlation of parameters

- Often two parameter values are highly correlated, e.g. you can get a good fit if θ_1 and θ_2 have some relationship e.g. $\theta_1+\theta_2$ = const or θ_1/θ_2 = const, but very poor fits for other values of (θ_1,θ_2). This information is lost if you just report the values and error bars separately.
- Sometimes you can change parameters to the appropriate well-determined combination.
- How to report the correlation of determined parameter values?

Correlation of parameters, page 2

- Usually what is done is to compute the Hessian of $\chi^2(\underline{\theta})$ evaluated at $\underline{\theta}_{bestfit}$. Diagonalize this matrix; its eigenvectors are the principal components of a hyper-ellipsoid that (exactly for a linear model, approximately otherwise) describes the region of indifference.
- If an eigenvector has large components from more than one parameter, that means the parameters are correlated. The "covariance of the parameters" is given by $C_{jk} = \Sigma\, V_{ji} V_{ki} / \lambda_i$
- This can be computed by SVD, often this is more numerically stable, see Numerical Recipes.

A note about Hessian of χ^2

- The rigorous formula for the second derivative of χ^2 includes two terms.
- Almost everyone neglects the second term, which is sensitive to noise in the data, and just uses:

  $H_{lz} \sim \Sigma\, J_{kl}\, J_{kz}\, \sigma_k^{-2}$ where $J_{kl} = \partial f_k / \partial \theta_l$
- Note now the Hessian doesn't really depend on the experimental data (at all for a linear model), you can compute it before the experiment begins…see page 413 in Beers' text. His "X" is the Jacobian of the model w.r.t. to the θ.

Are the Model & Data Consistent?

Often, the measured χ^2 is greater than the Q you would compute from chi2inv. What do you need to check before you say you have disproved the model?

1) Need to be sure you have found the very best possible values of all the parameters. Can have many local minima. Global optimization?

2) Need to be sure you have done enough repeats. If N is small probability is non-Gaussian with "fat tail".

3) χ^2 is extremely sensitive to estimate of σ (or D). Double check if you really believe these values.

4) Uncertainties in **X** and any parameters $\underline{\theta}$ you did not adjust might affect χ^2. Perhaps you can include these uncertainties in σ.

5) Often models are idealizations that do not really match experimental boundary conditions, mixing, etc. Can be tricky to try to rig up a model that really matches your experimental apparatus.

6) Be sure that you modeled 'instrumental function' or calibration of your signals carefully.

Experimental Design

- For linear models, you can compute the covariance of the model parameters BEFORE you do any experiments. Often you want to design the experiment so you are only sensitive to one or two parameters.
  - Do it! So many people do experiments and then afterwards realize they cannot possibly determine their parameter of interest from the data.
  - Sometimes you can fix the problem by using different knob settings **X**. You can play with this in your model before you do the experiments.
- Even for nonlinear models you can do this ahead of time, using the Jacobian evaluated at your prior nominal value of $\underline{\theta}$.

10.34 Numerical Methods Applied to Chemical Engineering

Fall 2015