Lecture 18

- Examples of least-squares models
- Nonlinear modeling

Reading (already assigned):

Chapter 10: Linear model fitting
Chapter 11: Nonlinear model fitting
Chapter 3: The how-to of Bayesian inference
Upcoming Topics

The remaining lectures will include these topics:

• Model fitting and statistical inference (frequentist and Bayesian)
  – Model definition
  – Linear and non-linear least squares, maximum likelihood
  – Parameter errors (credible intervals, Fisher matrix)
  – Parameter space exploration

• Markov processes and stochastic resonance
• MCMC (Hastings-Metropolis, etc.)
• Cholesky decomposition
• Principal component analysis (PCA)
• Wavelets
• Localization/Matched filtering
  – 1D, 2D problems (time, frequency/wavelength, image)
• Phase retrieval and Hilbert transforms
• Radon transform
• Extreme value statistics

SA = simulated annealing, GA = Genetic Algorithm, GHS = Guided Hamiltonian Sampler
Projects

General idea:
• Develop an analysis method analytically, as much as possible, and apply it to real or simulated data.
• Explore a theoretical issue and illustrate it numerically.
Guidelines:
• Written in the form of a journal article (abstract, intro, main, conclusions)
• Full reference list
Examples of topics:
1. Compare performance of spectral estimators for a well-defined problem or specific data set
2. Diagnose bootstrap sampling from frequentist and Bayesian perspectives
3. Compare actual vs. predicted performance for a localization algorithm (e.g. radial velocity, arrival time, sky location)
4. Detection problems: consider a survey of some kind (objects within images, spectral lines, time-domain bursts) and analyze detection and false-alarm probabilities; ROC curves, etc.
5. Bayesian assessment of the `reality’ of a particular pattern in a set of points using the odds ratio to compare two hypotheses.
6. Bayesian inference to a qualitative assertion
   Examples:
   i. Climate change from human activity is occurring
   ii. The face on Mars is real
   iii. Pulsars are really quark stars, not neutron stars
Least Squares Examples

I. A polynomial model for a times series $y_i$ with errors $\epsilon_i$:

$$y_i = \sum_{j=1}^{k} X_{ij} \theta_j + \epsilon_i = \sum_{j=1}^{k} t_i^{j-1} \theta_j + \epsilon_i.$$ 

The model is linear in the parameters even though it is nonlinear in the independent variable, $t_i$. If the polynomial order is $p$, then $k = p + 1$ and

The design matrix and parameter vector are:

$$X = \begin{pmatrix}
1 & t_1 & t_1^2 & \cdots & t_1^p \\
1 & t_2 & t_2^2 & \cdots & t_2^p \\
\vdots & & & & \\
1 & t_n & t_n^2 & \cdots & t_n^p
\end{pmatrix}$$

$$\theta = \begin{pmatrix}
\theta_1 \\
\theta_2 \\
\vdots \\
\theta_{p+1}
\end{pmatrix}$$

Define $T_k = \sum_{i=1}^{n} t_i^k$ and $\overline{t^k y} = \frac{1}{n} \sum_{i=1}^{n} t_i^k y_i$.

Need $1/n$ in front of last sum.
Then the product of the design matrix with itself is the $k \times k = (p + 1) \times (p + 1)$ matrix

$$X^\dagger X = \begin{pmatrix}
T_0 & T_1 & T_2 & \cdots & T_p \\
T_1 & T_2 & T_3 & \cdots & T_{p+1} \\
T_2 & T_3 & T_4 & \cdots & T_{p+2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
T_p & T_{p+1} & T_{p+2} & \cdots & T_{2p}
\end{pmatrix}$$

and

$$X^\dagger y = X^\dagger \begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{pmatrix} = \begin{pmatrix}
\sum_{i=1}^{n} y_i \\
\sum_{i=1}^{n} t_i y_i \\
\vdots \\
\sum_{i=1}^{n} t_i^p y_i
\end{pmatrix} \equiv n \begin{pmatrix}
y \\
ty \\
\vdots \\
t^p y
\end{pmatrix}$$

The least-squares solution $\hat{\theta} = (X^\dagger X)^{-1} X^\dagger y$ requires the inverse of $X^\dagger X$ that will exist if the determinant is nonzero.
First-order polynomial: something we can easily solve.

\[ y_i = \theta_1 + \theta_2 t_i \]

\[ X^\dagger X = \begin{pmatrix} T_0 & T_1 \\ T_1 & T_2 \end{pmatrix} \]

\[ (X^\dagger X)^{-1} = \frac{1}{\det(X^\dagger X)} \times \text{(matrix of cofactors)} \]

\[ = \frac{1}{(T_0 T_2 - T_1^2)} \begin{pmatrix} T_2 & -T_1 \\ -T_1 & T_0 \end{pmatrix} \]
Then
\[
\hat{\theta} = (X^\dagger X)^{-1} X^\dagger y = \frac{n}{(T_0 T_2 - T_1^2)} \begin{pmatrix} T_2 & -T_1 \\ -T_1 & T_0 \end{pmatrix} \begin{pmatrix} \bar{y} \\ ty \end{pmatrix}
\]

\[
= \frac{n}{(T_0 T_2 - T_1^2)} \begin{pmatrix} \bar{y} T_2 - ty T_1 \\ -\bar{y} T_1 + \bar{y} T_0 \end{pmatrix}
\]

\[
\equiv \frac{n}{(n T_2 - T_1^2)} \begin{pmatrix} \bar{y} T_2 - \bar{y} T_1 \\ -\bar{y} T_1 + \bar{y} n \end{pmatrix}
\]

So the individual parameters are
\[
\hat{\theta}_1 = \frac{n (\bar{y} T_2 - \bar{y} T_1)}{(n T_2 - T_1^2)} \quad \text{and} \quad \hat{\theta}_2 = \frac{n (n t y - \bar{y} T_1)}{(n T_2 - T_1^2)}. \]
Assuming the errors $\epsilon_i$ are stationary and statistically independent with variance $\langle \epsilon_i^2 \rangle = \sigma^2$, the covariance matrix of the parameters is

$$P = \sigma^2 (X^\dagger X)^{-1} = \begin{pmatrix} \langle \theta_1^2 \rangle & \langle \theta_1 \theta_2 \rangle \\ \langle \theta_1 \theta_2 \rangle & \langle \theta_2^2 \rangle \end{pmatrix} \equiv \begin{pmatrix} \sigma_{\theta_1}^2 & \sigma_{\theta_1} \sigma_{\theta_2} \rho_{\theta_1 \theta_2} \\ \sigma_{\theta_1} \sigma_{\theta_2} \rho_{\theta_1 \theta_2} & \sigma_{\theta_2}^2 \end{pmatrix}$$

where

$$\sigma_{\theta_1} = \sigma \left( \frac{T_2}{nT_2 - T_1^2} \right)^{1/2}$$

$$\sigma_{\theta_2} = \sigma \left( \frac{n}{nT_2 - T_1^2} \right)^{1/2}$$

$$\rho_{\theta_1 \theta_2} = \frac{\langle \theta_1 \theta_2 \rangle}{\sigma_{\theta_1} \sigma_{\theta_2}} = \frac{-T_1}{\sqrt{nT_2}} \quad \text{(negatively correlated)}$$
For \( n \gg 1 \) and uniform sampling \( t_i = i, i = 1, \ldots, n \),
\[
T_1 \approx \frac{n^2}{2} \quad \text{and} \quad T_2 \approx \frac{n^3}{3}
\]
so
\[
\sigma_{\theta_1} \approx \frac{2\sigma}{\sqrt{n}} \quad \sigma_{\theta_2} \approx \frac{\sqrt{12}\sigma}{n^{3/2}}
\]
\[
\rho_{\theta_1\theta_2} \approx \frac{-\frac{n^2}{2}}{\sqrt{n} \cdot \frac{n^3}{3}} = -\frac{\sqrt{3}}{2} \approx -0.87 \quad (\text{highly anticorrelated})
\]
The anticorrelation means that any error in one parameter is compensated by the error in the other.
Better parameterization for the first-order polynomial: orthogonal polynomials.

E.g.

\[ y_i = \theta_1 + \theta_2(t_i - \bar{t}) \quad \text{where} \quad \bar{t} = \frac{1}{n} \sum_{i=1}^{n} t_i \]

Now the design matrix and the various products are

\[
\begin{pmatrix}
1 & t_1 - \bar{t} \\
1 & t_2 - \bar{t} \\
\vdots \\
1 & t_n - \bar{t}
\end{pmatrix}, \quad
\begin{pmatrix}
T_0 & T_1 \\
T_1 & T_2
\end{pmatrix} = \begin{pmatrix}
T_0 & 0 \\
0 & T_2
\end{pmatrix}, \quad
\begin{pmatrix}
T_0 & T_1 \\
T_1 & T_2
\end{pmatrix}^{-1} = \frac{1}{T_0 T_2} \begin{pmatrix}
T_2 & 0 \\
0 & T_0
\end{pmatrix}
\]

and the solution is now

\[
\theta_1 = \bar{y}, \quad \theta_2 = \frac{n(t - \bar{t})y}{T_2}
\]

The errors on \( \theta_1,2 \) are the same but the parameters are now uncorrelated, \( \rho_{\theta_1 \theta_2} = 0 \).
II. Sinusoids

Consider the linear model $y = X\theta + \epsilon$ where $X$ comprises complex exponentials and the parameter vector $\theta$ comprises Fourier amplitudes:

$$X_{nm} = e^{2\pi inm/N}, \quad n = 0, \ldots, N - 1, \quad m = 0, \ldots, k - 1$$

$$\equiv W_{Nm}^n$$

where $W_N \equiv e^{2\pi i/N}$ is the $N^{th}$ root of 1 on the unit circle. For $k \neq N$ we have

$$X = \begin{pmatrix}
1 & 1 & \cdots & 1 \\
1 & W_N & \cdots & W_N^k \\
1 & W_N^2 & \cdots & W_N^{2k} \\
\vdots & \vdots & \ddots & \vdots \\
1 & W_N^{N-1} & \cdots & W_N^{(N-1)k}
\end{pmatrix} \quad \rightarrow \quad k = N \begin{pmatrix}
1 & 1 & \cdots & 1 \\
1 & W_N & \cdots & W_N^{N-1} \\
1 & W_N^2 & \cdots & W_N^{2(N-1)} \\
\vdots & \vdots & \ddots & \vdots \\
1 & W_N^{N-1} & \cdots & W_N^{(N-1)^2}
\end{pmatrix}$$
The product matrix is

\[ X^\dagger X = \begin{pmatrix}
  s_0 & s_1 & \cdots & s_{N-1} \\
  s_1^* & s_0 & \cdots & s_{N-2} \\
  s_2^* & s_1^* & s_0 & s_{N-3} \\
  \vdots & \vdots & \vdots & \vdots \\
  s_{N-1}^* & s_{N-2}^* & \cdots & s_0
\end{pmatrix} \]

where

\[ s_p \equiv \sum_{j=0}^{N-1} W_{N}^{pj}. \]

The off-diagonal terms all sum to zero because the sums are over integer multiples of the periods of the complex sinusoids. Therefore

\[ X^\dagger X = NI \quad \text{and} \quad (X^\dagger X)^{-1} = N^{-1}I. \]
We also have

\[ x^\dagger y = \begin{pmatrix} 
\sum_{j=0}^{N-1} y_j \\
\sum_{j=0}^{N-1} W_N^{-j} y_j \\
\sum_{j=0}^{N-1} W_N^{-2j} y_j \\
\vdots \\
\sum_{j=0}^{N-1} W_N^{-(N-1)j} y_j 
\end{pmatrix} \]

The least-squares coefficients are then

\[ \hat{\theta} = (X^\dagger X)^{-1} X^\dagger y = N^{-1} X^\dagger y \]

which is just the DFT of \( y \) expressed in vector form.

The parameter error vector is

\[ P = \sigma^2 (X^\dagger X)^{-1} = N^{-1} \sigma^2 I. \]
Weighted mean as an example of optimal weights.

Consider measurements $x_i = \langle x \rangle + \epsilon_i$ where the measurement errors are white noise with variance $\sigma_i^2$ (i.e. not necessarily stationary over the index $i$). Then $\langle x_i^2 \rangle = \langle x \rangle^2 + \sigma_i^2$.

Define the weighted sum

$$\bar{x} = \frac{\sum_i w_i x_i}{\sum_i w_i}.$$

As with any estimator we are interested in its first and second moments. We have

$$\langle \bar{x} \rangle = \langle x \rangle, \quad \sigma_{\bar{x}}^2 = \langle \bar{x}^2 \rangle - \langle \bar{x} \rangle^2$$
So we have

\[
\langle x^2 \rangle = \frac{\sum_i \sum_j w_i w_j \langle x_i x_j \rangle}{\left( \sum_i w_i \right)^2}
= \frac{\sum_i w_i^2 \langle x_i^2 \rangle + \sum_{i \neq j} w_i w_j \langle x_i \rangle \langle x_j \rangle}{\left( \sum_i w_i \right)^2}
= \frac{\sum_i w_i^2 (\langle x^2 \rangle + \sigma_i^2) + \sum_{i \neq j} w_i w_j \langle x \rangle^2}{\left( \sum_i w_i \right)^2}
= \langle x \rangle^2 + \frac{\sum_i w_i^2 \sigma_i^2}{\left( \sum_i w_i \right)^2}
\]

So

\[
\sigma_x^2 = \frac{\sum_i w_i^2 \sigma_i^2}{\left( \sum_i w_i \right)^2}
\]  
(1)
Now let’s minimize the error in the mean estimate with respect to the weights:

\[
\partial_{w_j} \sigma^2_x = 0 = \frac{\left( \sum_i w_i \right)^2 \partial_{w_j} \sum_i w_i^2 \sigma_i^2 - \sum_i w_i^2 \sigma_i^2 \partial_{w_j} \left( \sum_i w_i \right)^2}{\left( \sum_i w_i \right)^4}
\]

The numerator gives

\[
\left( \sum_i w_i \right)^2 \left[ 2 w_j \sigma_j^2 - \sum_i w_i^2 \sigma_i^2 \right] \sum_i w_i^2 = 0
\]

which gives

\[
w_j = \frac{\sum_i w_i^2 \sigma_i^2}{\sigma_i^2 \sum_i w_i}.
\]

A consistent solution is therefore

\[
w_j = \sigma_i^{-2},
\]

which is the standard result that the weighted mean involves weights proportional to the inverse variance of each data point.
Example of a bad model: Consider

\[ y_i = \theta_1 + \theta_2 x_i + a(\sin \theta_3 x_i) \approx \theta_1 + \theta_3 x_i + a x_i (\cos \theta_3 x_i) \delta \theta_3 \]

where the parameters of the linearized function are \( \theta_1, \theta_2, \theta'_3 = a \delta \theta_3 \). The design matrix and product matrix are

\[
\begin{align*}
x & = \begin{pmatrix}
1 & x_1 & x_1 \cos \theta_3 x_1 \\
1 & x_2 & x_2 \cos \theta_3 x_2 \\
& \vdots & \\
1 & x_N & x_N \cos \theta_3 x_N
\end{pmatrix} \\
X^\dagger X & = \begin{pmatrix}
N & \sum_i x_i & \sum_i x_i \cos \theta_3 x_i \\
\sum_i x_i & \sum_i x_i^2 & \sum_i x_i^2 \cos \theta_3 x_i \\
\sum_i x_i \cos \theta_3 x_i & \sum_i x_i^2 \cos \theta_3 x_i & \sum_i x_i^2 (\cos \theta_3 x_i)^2
\end{pmatrix}
\end{align*}
\]

For the case where \( \theta_3 x_i \ll 1 \) for all \( x_i \), the elements involving \( \cos \theta_3 x_i \approx 1 - (\theta_3 x_i)^2 / 2 \) so for very small \( \theta_3 x_i \) the cosine factors will be very close to unity. The elements in the matrix are then degenerate with neighboring elements because the sine term in the model is degenerate with the linear term. In this case the design matrix is ill-conditioned and the determinant of \( X^\dagger X \to 0 \). For cases like this the fitting function should be redefined or singular value decomposition may be used.
Continue with Lecture 17 material