ECE 275A: Parameter Estimation I
Parametric Statistical Modeling

Florian Meyer

Electrical and Computer Engineering Department
University of California San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Why Parametric Statistical Models?

- They **include deterministic models** as a special case.
- They **succinctly** capture & encode properties of the perceived world.
  - Allow for data compression
  - Enable efficient explanation of past measurements
  - Enable efficient prediction of future measurements
- **Statistical** models acknowledge that **uncertainty, error, and chance exist** in our understanding of the world.
- They provide "quality of fit" measures.
  - Model-mismatch measures
  - Parameter estimate quality measures

Contrast with Nonparametric Models

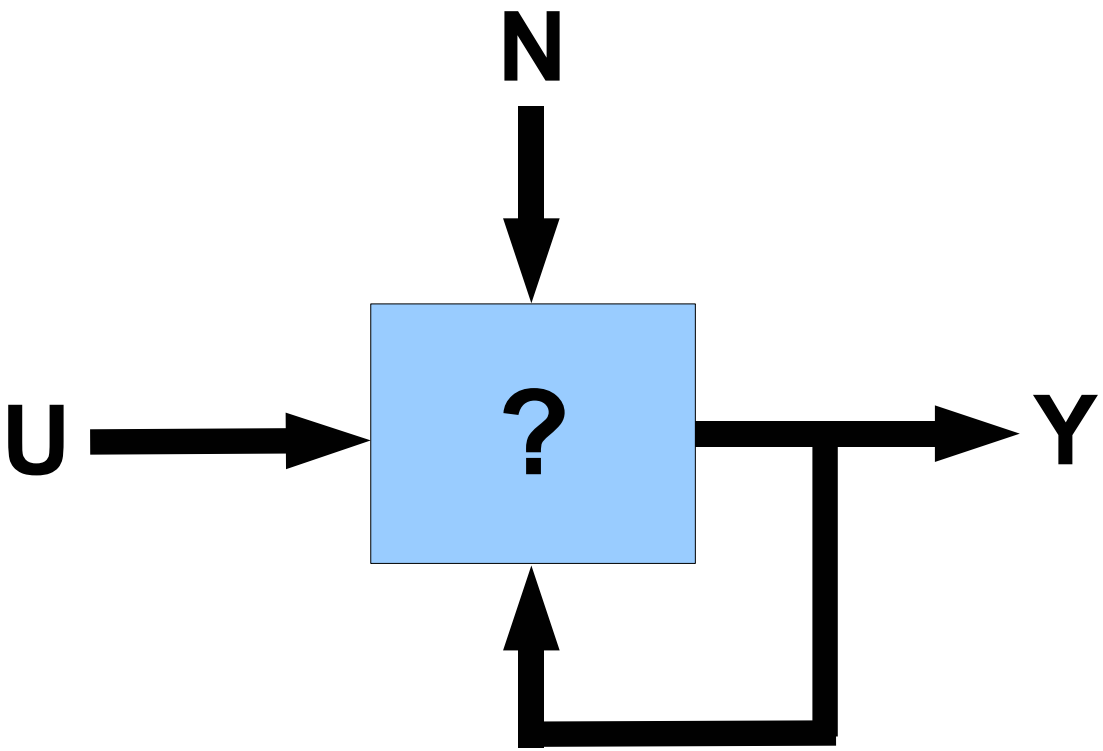
With Nonparametric Models:

- No *a prior* structural or modeling information is utilized
  - Difficult to model dynamic (nonstationary) processes.
  - Difficult to gain *insight* into physical and other processes.
- Often, all data must be kept regardless of dimensionality or amount.
  - Data processing is expensive, particularly if data is collected in an on-going, on-line manner.
- Probability density function (pdf) approximations are constructed via "binning" of data to directly form empirical density functions
  - As data is collected in an on-line manner, density-related estimates must often be recomputed via "batch processing"

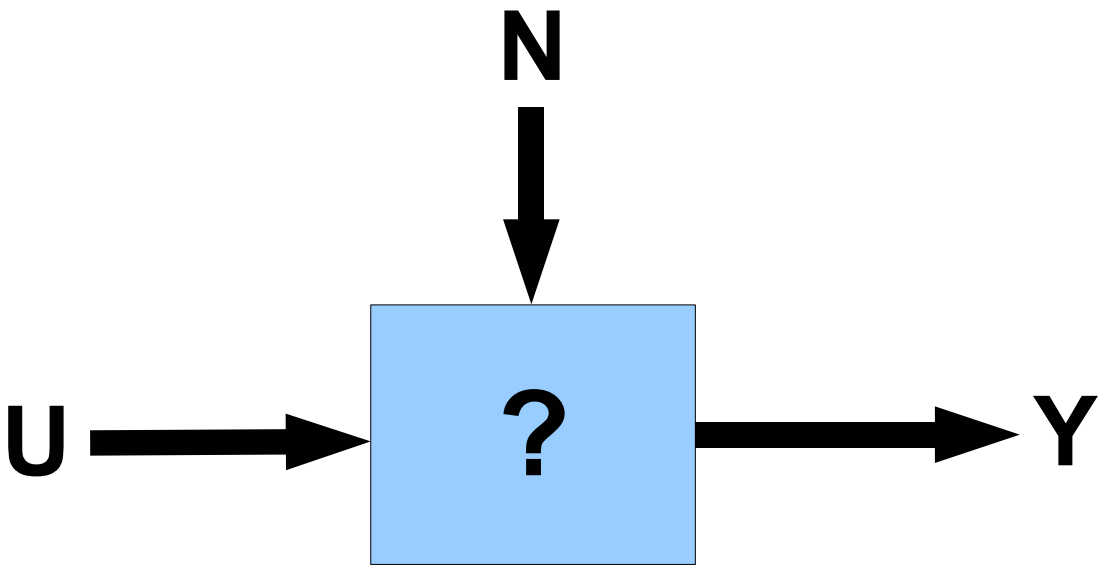
Parametric Models = Generative Models

- Parametric Probabilistic Models are posited to explain observed phenomenon.
  - For this reason they are often referred to as *Generative Models*; models that are presumed to be have generated observed data.
  - They are also known as *forward models*, as one imagines processing inputs, noise, past observations, and *parameters* in a "forward direction" to produce observed data.
  - The problem of estimating unknown model parameters given observed inputs and past observations is known as the *inverse problem.*
- In its fullness then, the problem of parameter estimation involves:
  1. Proposing and constructing a candidate generative model to explain some phenomenon of interest.
  2. Collecting data corresponding to inputs and outputs of the model.
  3. Solving the statistical inverse problem of estimating the unknown parameters of the model
  4. Validating the model. If the statistics of the outputs of the model do not match the statistics of our observed data, and/or the estimated model yields poor predictive capabilities, we must refine and improve our posited model.
- **In this course we are primarily concerned with issues 1 and 3.**

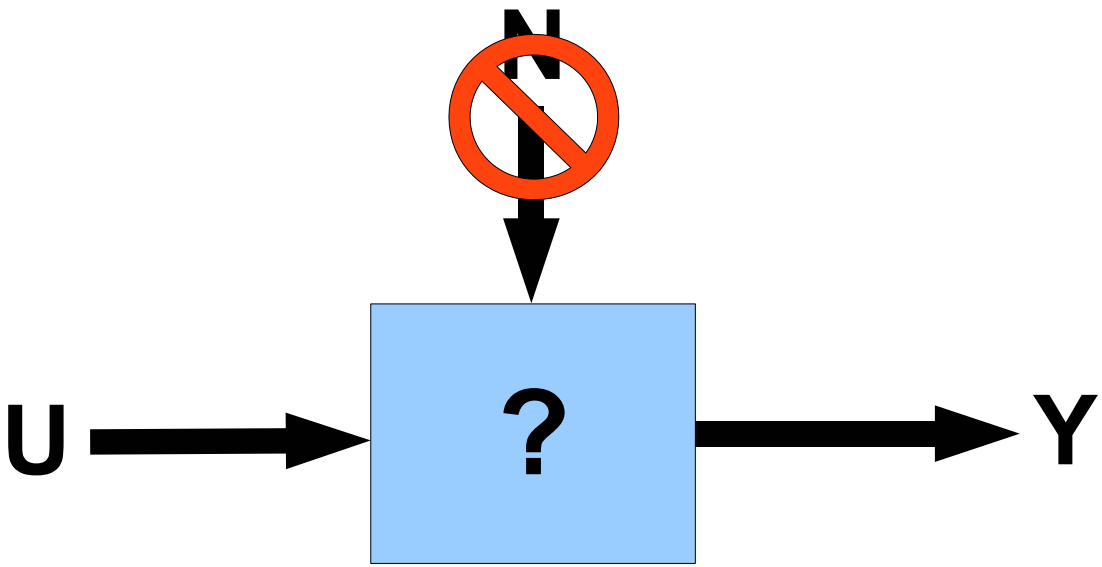
Generative Model of World or System or ...



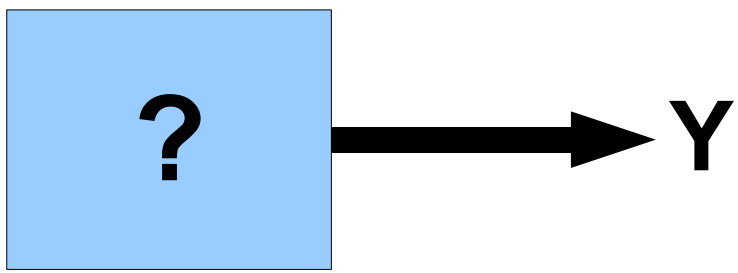
Generative Model of World or System or ...



Generative Model of World or System or ...



Generative Model of World or System or ...



Example: "Moving Average" (MA) Linear Gaussian Model

For unknown parameters x_i, $i = 1, \cdots, M$, consider the *moving average (MA)* (of $f(u[t])$) "linear" (in the *parameters*!) time-series model

$$y[t] = x_1 f(u[t]) + \cdots + x_M f(u[t-M+1]) + n[t] \quad \text{with} \quad n[t] \sim \mathrm{N}(0, \sigma^2)$$

The sequence of inputs $u[t]$ is assumed known, as is the general function $f(\cdot)$. The noise $n[t]$ is considered to be iid with σ^2 known.

Some examples are $f(x) = x$, $f(x) = \cos(x)$, $f(x) = \exp(x)$, etc.

Set $x \triangleq (x_1, \cdots, x_M)^T$ and $a[t] = (f(u[t]), \cdots, f(u[t-M+1]))^T$ then

$$y[t] = a^T[t]\boldsymbol{x} + n[t]$$

with $y[t] \sim \mathrm{N}(a^T[t]\boldsymbol{x}, \sigma^2)$,

$$p_{y[t];\boldsymbol{x}}(y[t]) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y[t] - a^T[t]\boldsymbol{x})^2\right\}$$

showing that the MA model for $y[t]$ is entirely equivalent to a probabilistic model parameterized by x.

"Moving Average" (MA) Linear Gaussian Model – Cont.

Now consider collecting a "batch" of $N > M$ samples of $y[t]$ $(t = 1, \cdots, N)$ and set $\boldsymbol{y} \triangleq (y[1] \cdots y[N])^T$.

The MA model is entirely equivalent to the vector-matrix "batch data" parametric probabilistic model

$$\boldsymbol{y} \sim \mathrm{N}(A\boldsymbol{x}, C), \quad C = \mathrm{diag}(\sigma^2, \cdots, \sigma^2) = \sigma^2 I$$

$$p_{\boldsymbol{x}}(\boldsymbol{y}) \triangleq p_{\boldsymbol{y};\boldsymbol{x}}(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^N \det C}} \exp\left\{-\frac{1}{2}\|\boldsymbol{y} - A\boldsymbol{x}\|^2_{C^{-1}}\right\}$$

where the *data matrix* A is a known $N \times M$ matrix whose N rows are comprised of the M-dimensional row vectors $a^T[t]$ and

$$\|\boldsymbol{y} - A\boldsymbol{x}\|^2_{C^{-1}} \triangleq (\boldsymbol{y} - A\,\boldsymbol{x})^T C^{-1}(\boldsymbol{y} - A\boldsymbol{x})$$

with C a known (diagonal, in this example) $N \times N$ matrix.

The Linear Gaussian Model

Linear Gaussian Model

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{v}\,, \quad v \sim \mathrm{N}(0, C), \quad C \text{ is positive definite}$$

Equivalent to

Parametric Probability Model

$$\boldsymbol{y} \sim \mathrm{N}(A\boldsymbol{x}, C), \quad p_{\boldsymbol{x}}(\boldsymbol{y}) = \frac{1}{\sqrt{(2\pi)^m \det C}} \exp\left\{-\frac{1}{2}\|\boldsymbol{y} - A\boldsymbol{x}\|^2_{C^{-1}}\right\}$$

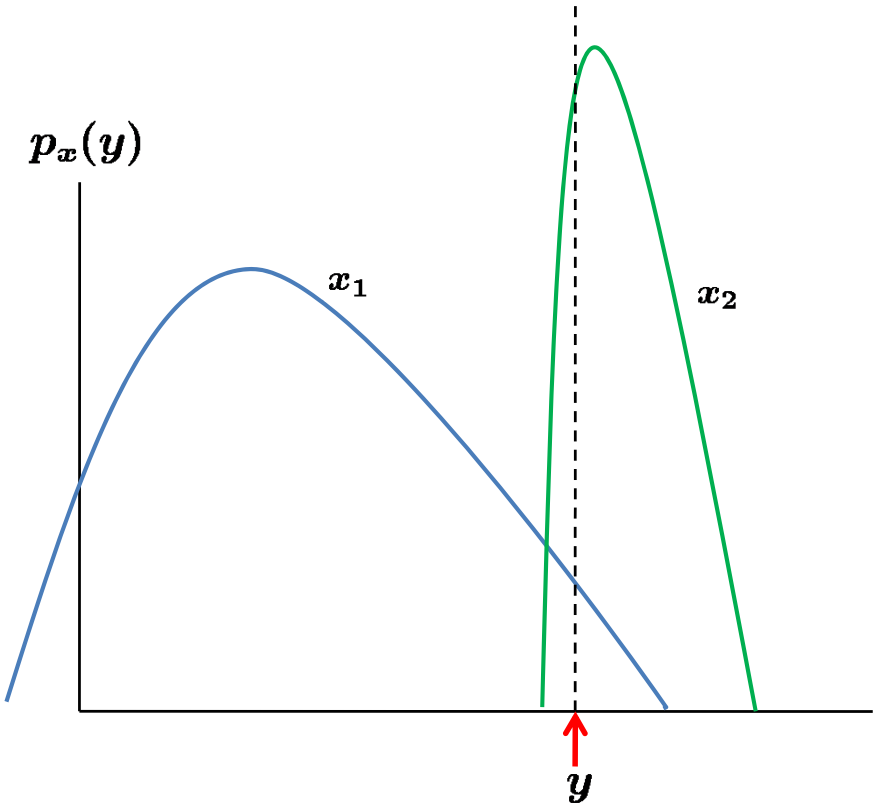
where

$$\|\boldsymbol{y} - A\boldsymbol{x}\|^2_{C^{-1}} \triangleq (\boldsymbol{y} - A\boldsymbol{x})^T C^{-1} (\boldsymbol{y} - A\boldsymbol{x})$$

The Likelihood Function

Likelihood of x given y : $\ell_y(x) \triangleq p_x(y)$

In the figure below, note that it is rational to prefer probability model $p_{x_2}(\cdot)$ over model $p_{x_1}(\cdot)$ given the observed value of y:



$$\ell_y(x_2) = p_{x_2}(y) > p_{x_1}(y) = \ell_y(x_1)$$

Model Fitting by Likelihood Maximization

The function

$$\ell_{\boldsymbol{y}}(\boldsymbol{x}) \triangleq p_{\boldsymbol{x}}(\boldsymbol{y})$$

is **the likelihood of $\boldsymbol{x}$** (i.e. of the model $P_{\boldsymbol{x}}(\cdot)$) given the measured data $\boldsymbol{y}$.

The principle of maximum likelihood estimation says to find that model (parameter $\boldsymbol{x}$) for which the likelihood function takes its maximum value, given the measured data $\boldsymbol{y}$.

Maximizing the likelihood function is equivalent to minimizing the negative logarithm of the likelihood function (the "negative log-likelihood"). For the important *linear Gaussian model with known covariance matrix* example considered above, this corresponds to finding the parameter vector $\boldsymbol{x}$ that minimizes the weighted least squares loss function

$$\|\boldsymbol{y} - A\boldsymbol{x}\|^2_{C^{-1}} \triangleq (\boldsymbol{y} - A\boldsymbol{x})^T C^{-1} (\boldsymbol{y} - A\boldsymbol{x})$$

Note that when $C = I$ this reduces to a (unweighted) least squares problem. In either case the problem is one of solving a *linear inverse problem*

$$\boldsymbol{y} \approx A\boldsymbol{x}$$

in an appropriate least squares sense.

Model Fitting by Likelihood Maximization

Maximum Likelihood Estimate of $\boldsymbol{x}$ given $\boldsymbol{y}$

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \ell_{\boldsymbol{y}}(\boldsymbol{x}) = \arg\max_{\boldsymbol{x}} p_{\boldsymbol{x}}(\boldsymbol{y})$$

For the Linear Gaussian model this is equivalent to

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - A\boldsymbol{x}\|_{C^{-1}}^2$$

Which corresponds to solving a **Linear Inverse Problem**

$$\boldsymbol{y} \approx A\boldsymbol{x}$$

in an appropriate **Minimum Norm** sense, where y and x are **Vectors**, A is a (matrix representation of) a **Linear Operator**, and $\|\cdot\|_{C^{-1}}$ is a (weighted) **Norm**.