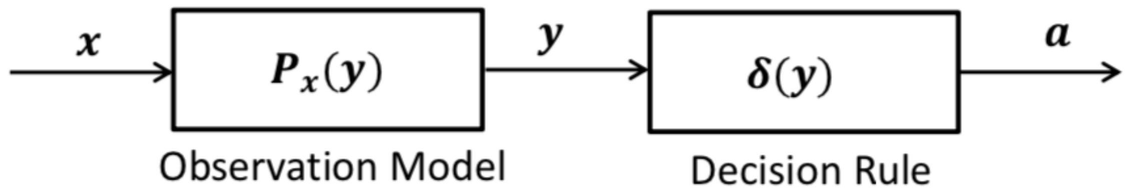
Drawing Conclusions from Data

Brian J. Harding

Space Sciences Lab, UC Berkeley



Making decisions using data

Discrete Decisions

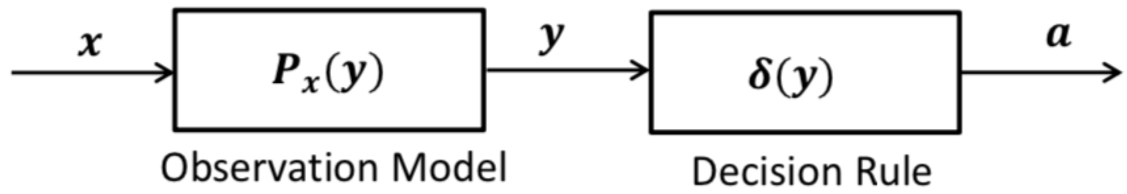
- Is there an MSTID in my data?
- Is there a relationship between atmospheric tides and electric fields?
- Did an earthquake cause an ionospheric effect?

Continuous Decisions

- What is the plasma velocity?
- What percentage of ionospheric variability can be attributed to the neutral atmosphere?
- What is a meteor's mass, based on its plasma trail?

Interpreting Decisions

- Accuracy vs Precision
- Correlation vs Causation
- Monte Carlo simulation



Making decisions using data

Discrete Decisions

- Is there an MSTID in my data?
- Is there a relationship between atmospheric tides and electric fields?
- Did an earthquake cause an ionospheric effect?

Continuous Decisions

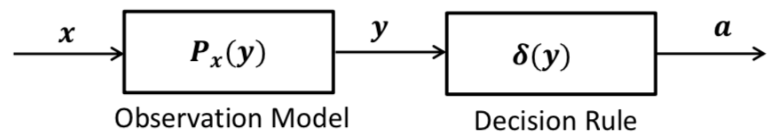
- What is the plasma velocity?
- What percentage of ionospheric variability can be attributed to the neutral atmosphere?
- What is a meteor's mass, based on its plasma trail?

Interpreting Decisions

- Accuracy vs Precision
- Correlation vs Causation
- Monte Carlo simulation

Binary Hypothesis testing

- Example: detect incoming missile using measured radar return

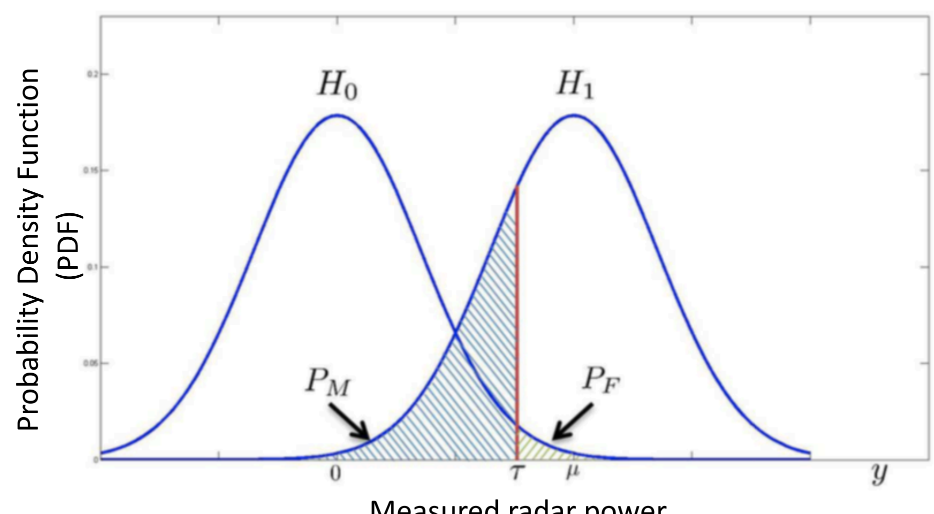


- Know:
  - *Probability Density Function (PDF) of H0*: no missile (i.e., just noise)
  - *PDF of H1*: missile incoming
- Neyman-Pearson lemma: Use Likelihood Ratio Test!
  - Even for large data sets

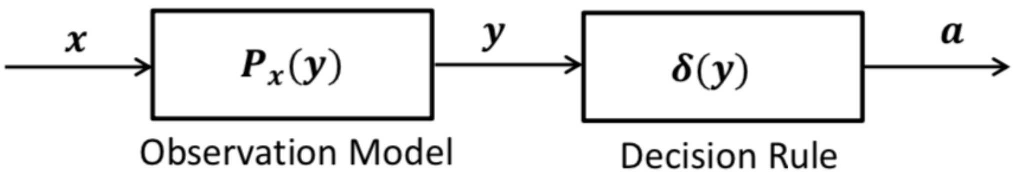
$$\text{Measurement: } y$$

$$\text{if } \frac{p_1(y)}{p_0(y)} > \tau, \text{ decide } H_1$$

$$\text{else decide } H_0$$



Binary Hypothesis testing

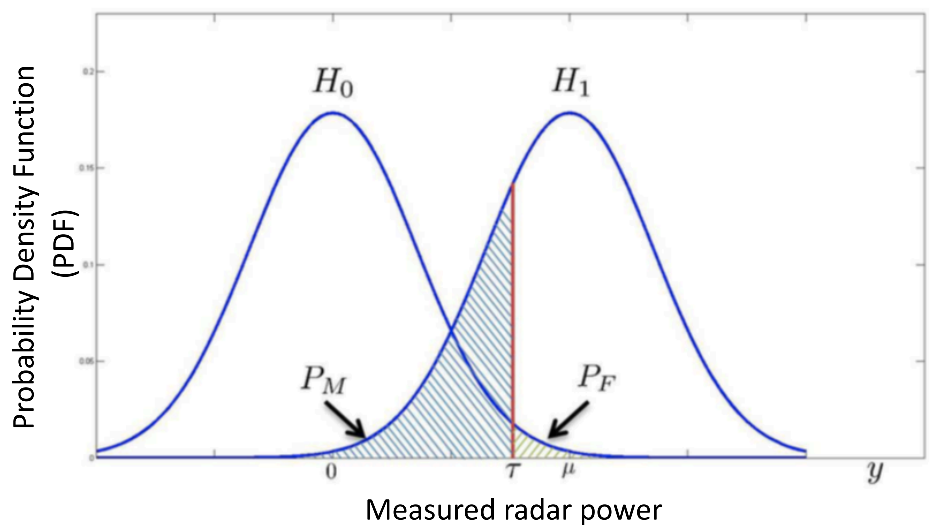


- Because the decision is based on random data, **it is also random**
  - Evaluate probability of detection, false alarm, etc.

$$\text{Measurement: y}$$

$$\text{if } \frac{p_1(y)}{p_0(y)} > \tau, \text{ decide } H_1$$

$$\text{else decide } H_0$$



Binary Hypothesis testing

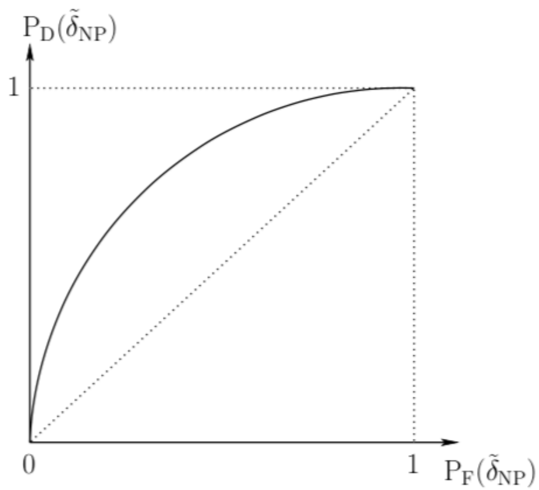
$$\text{Measurement: } y$$

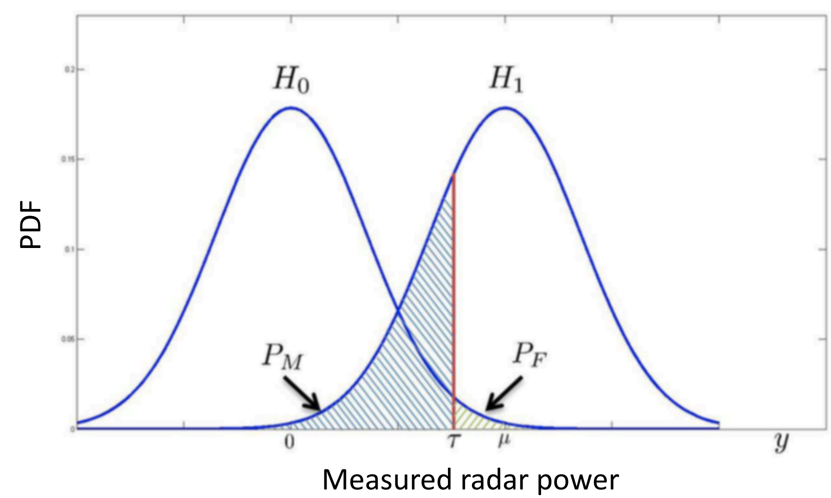
$$\text{if } \frac{p_1(y)}{p_0(y)} > \tau, \text{ decide } H_1$$

$$\text{else decide } H_0$$

Ways to choose threshold τ		
Maximum Likelihood (ML)	Choose the hypothesis that has a larger PDF at y	$\tau = 1$
Maximum a Posteriori (MAP)	Incorporate prior knowledge	$\tau = p(H_1)/p(H_0)$
Neyman-Pearson (NP)	Useful if you don't know $p_1(y)$	Choose $P_{\text{false alarm}}$ and solve for τ

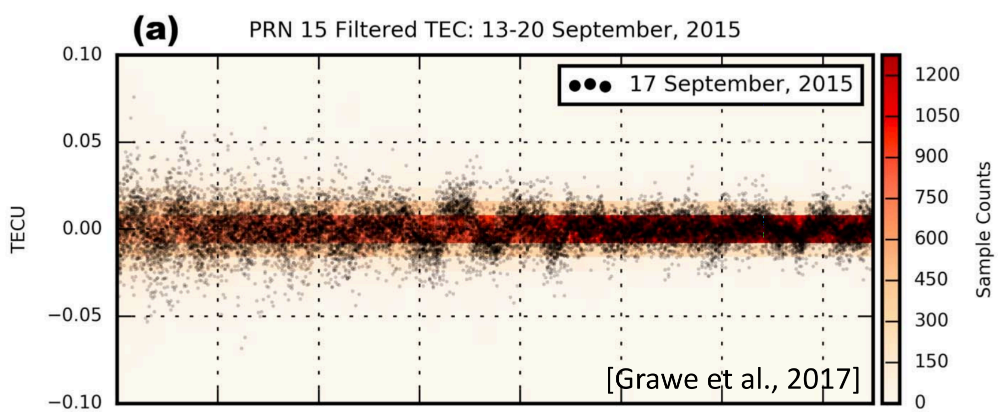
- Receiver Operating Characteristic (ROC) curve
  - ROC characterizes all thresholds

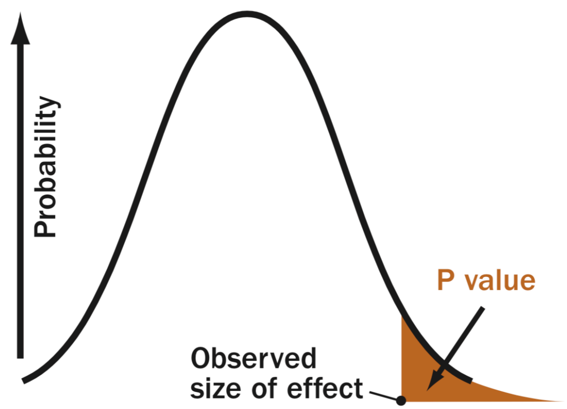




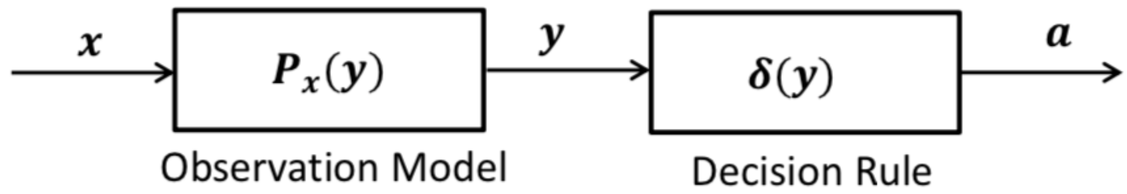
Connection with statistical tests

- Most (all?) statistical tests can be understood in this framework
  - t-test
  - Wilcoxon
  - ANOVA
- P-value is the probability of false alarm
  - i.e., of deciding an effect is real when it is not actually real
  - "statistically significant at the p=0.05 level"
  - NOT "95% significant"
  - NOT "a large effect"
  - NOT "95% chance of H1 being correct"





A P value is the probability of an observed (or more extreme) result arising only from chance.



Making decisions using data

Discrete Decisions

- Is there an MSTID in my data?
- Is there a relationship between atmospheric tides and electric fields?
- Did an earthquake cause an ionospheric effect?

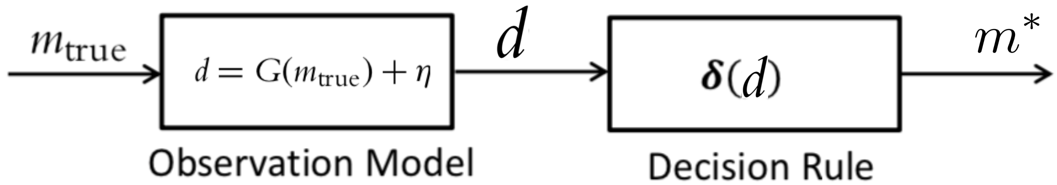
Continuous Decisions

- What is the plasma velocity?
- What percentage of ionospheric variability can be attributed to the neutral atmosphere?
- What is a meteor's mass, based on its plasma trail?

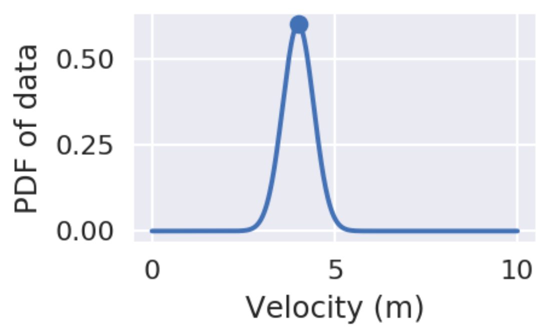
Interpreting Decisions

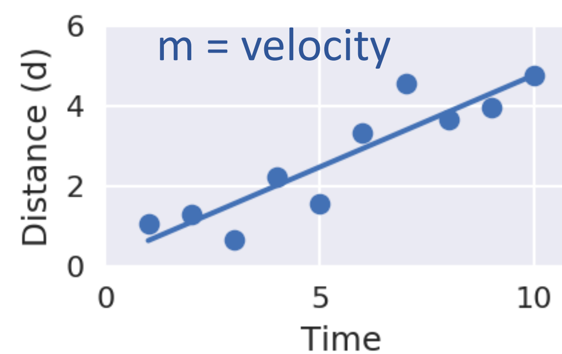
- Accuracy vs Precision
- Correlation vs Causation
- Monte Carlo simulation

Estimation theory



- Estimate velocity using **Maximum Likelihood (ML) Estimation**
- Under Gaussian, uncorrelated noise, ML is Least Squares!
  - If that's not true, Least Squares may not be the best choice for parameter estimation

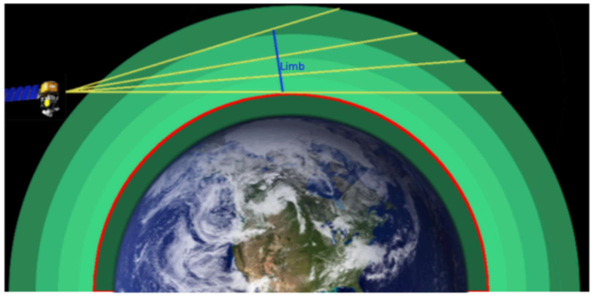


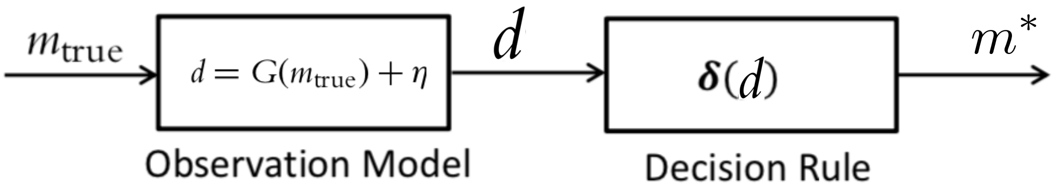


$$d = Gm + \text{noise}$$

$$f_i(d_i|\mathbf{m}) = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}(d_i-(\mathbf{Gm})_i)^2/\sigma_i^2} \longrightarrow \min \sum_{i=1}^{m} \frac{(d_i - (\mathbf{Gm})_i)^2}{\sigma_i^2}.$$

Estimation theory (n-dimensional model)





- Example: Radio occultation
- Generally: Fredholm integral $$d(x) = \int_a^b g(x, \xi) m(\xi) d\xi$$
- ML or Least-squares requires us to restrict solution space
- Often done implicitly by limiting degrees of freedom in m^*
    - Fit Chapman profile
    - Fit spherical harmonics
    - Assume equilibrium conditions
- No ability to track error of these assumptions

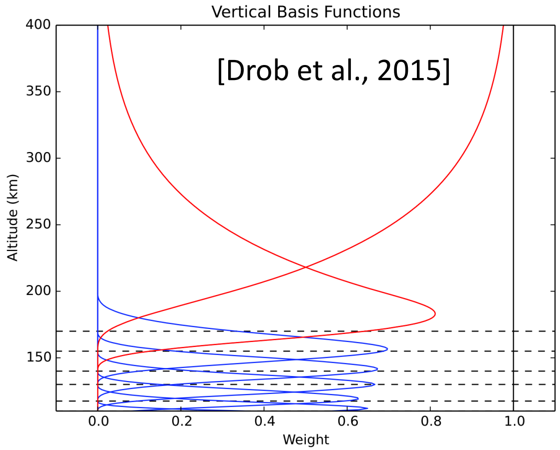
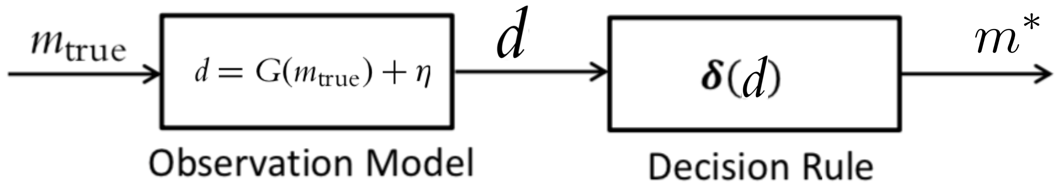


Figure 1. Vertical cubic B-spline basis functions β_i (red, blue) and corresponding data intervals δ_i (dashed) for the new HWM model. The last two basis functions (red) are constructed to approach either 0 or 1, subject to continuity and derivative constraints with the remaining functions.

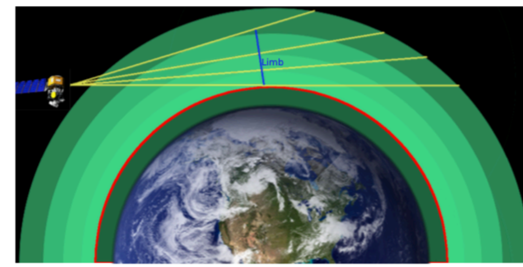
Estimation theory (n-dimensional data)



- Don’t restrict solution space – write PDF and see where it takes you
- Fredholm integrals

$$d(x) = \int_a^b g(x,\ \xi) m(\xi) d\xi$$

$$\mathbf{Gm} = \mathbf{d}$$

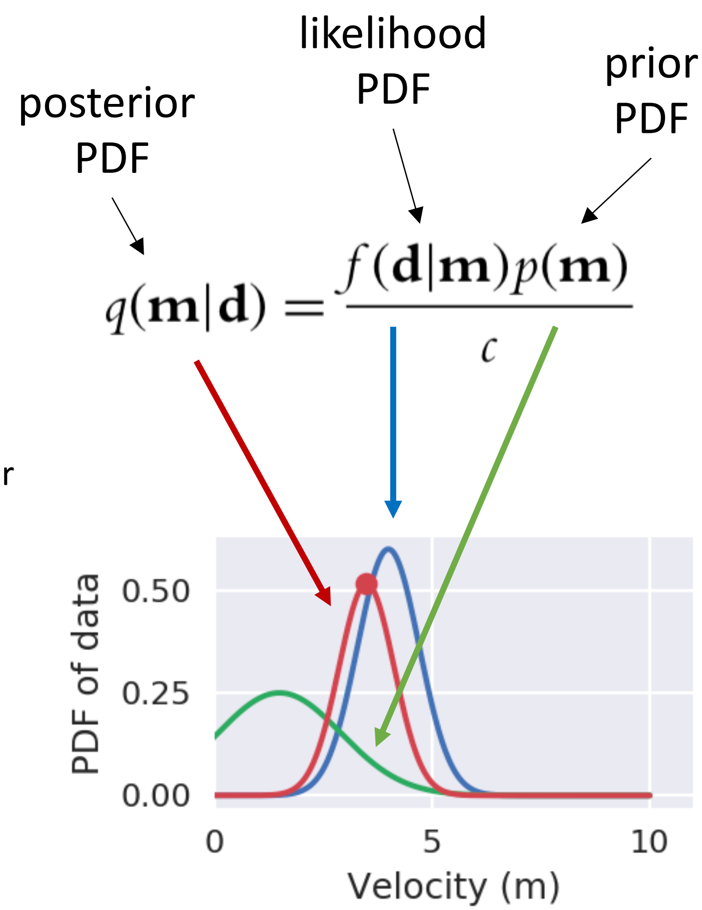


- Tempting to take inverse (or least-squares), but only valid if:
  - G is full rank
  - # data points ≥ # unknowns
  - Errors are Gaussian and uncorrelated
- Even if these are satisfied, result might be too noisy, or physically unrealistic.
- Solution: incorporate prior information

Bayes' Theorem

- Data updates a prior probability/belief
- **Maximum a posteriori (MAP) estimation**
- Example: Gaussian prior with mean $\mathbf{m}_{prior}$ and stddev α
  - Takes form of "cost function" to be minimized

$$\min \ (1/\sigma)^2 \|(\mathbf{Gm} - \mathbf{d})\|_2^2 + (1/\alpha)^2 \|\mathbf{m} - \mathbf{m}_{\text{prior}}\|_2^2,$$



- Equal to ML if prior is constant (i.e., uninformative)
- Prior may seem as arbitrary as fitting pre-determined functions, but:
  - Priors can be learned
  - Priors allow characterization of errors

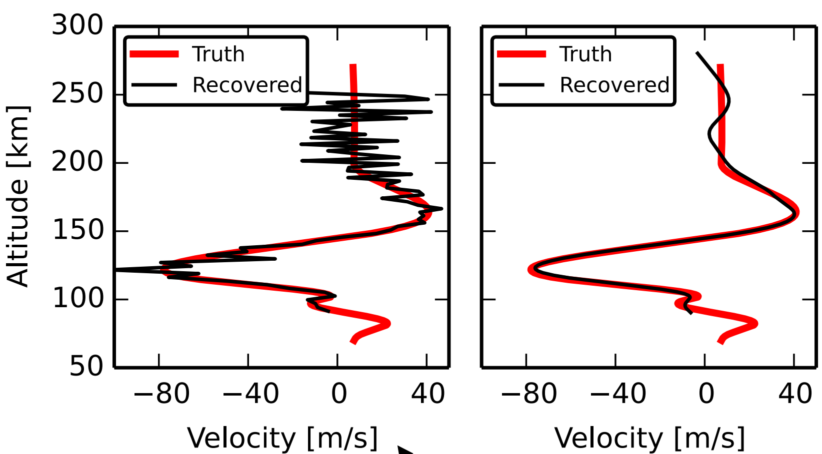
Regularization

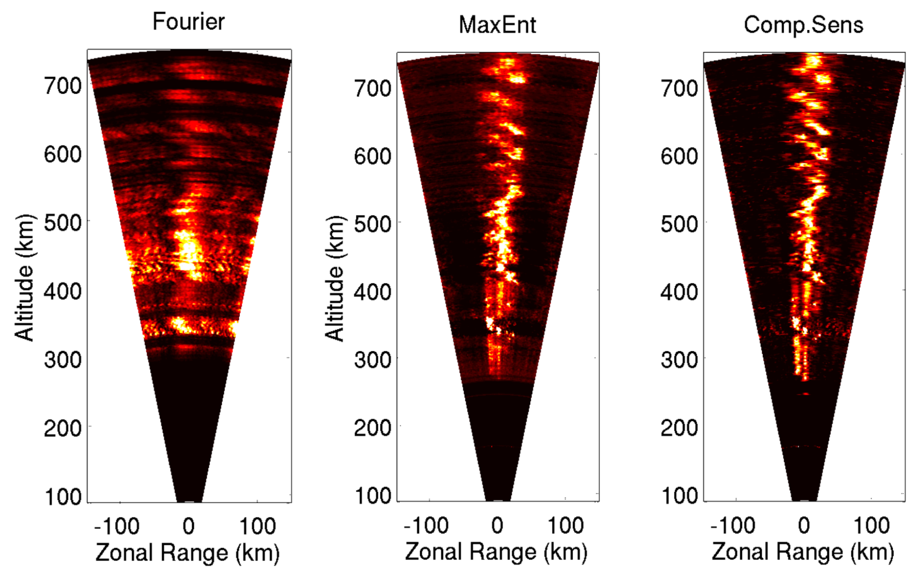
- This "cost function" approach is often useful even when no prior is explicit

$$\min \|\mathbf{Gm} - \mathbf{d}\|_2^2 + \lambda \|\mathbf{Lm}\|_2^2$$

$$\min \|\mathbf{Gm} - \mathbf{d}\|_2^2 + \lambda \|\mathbf{Sm}\|_1$$

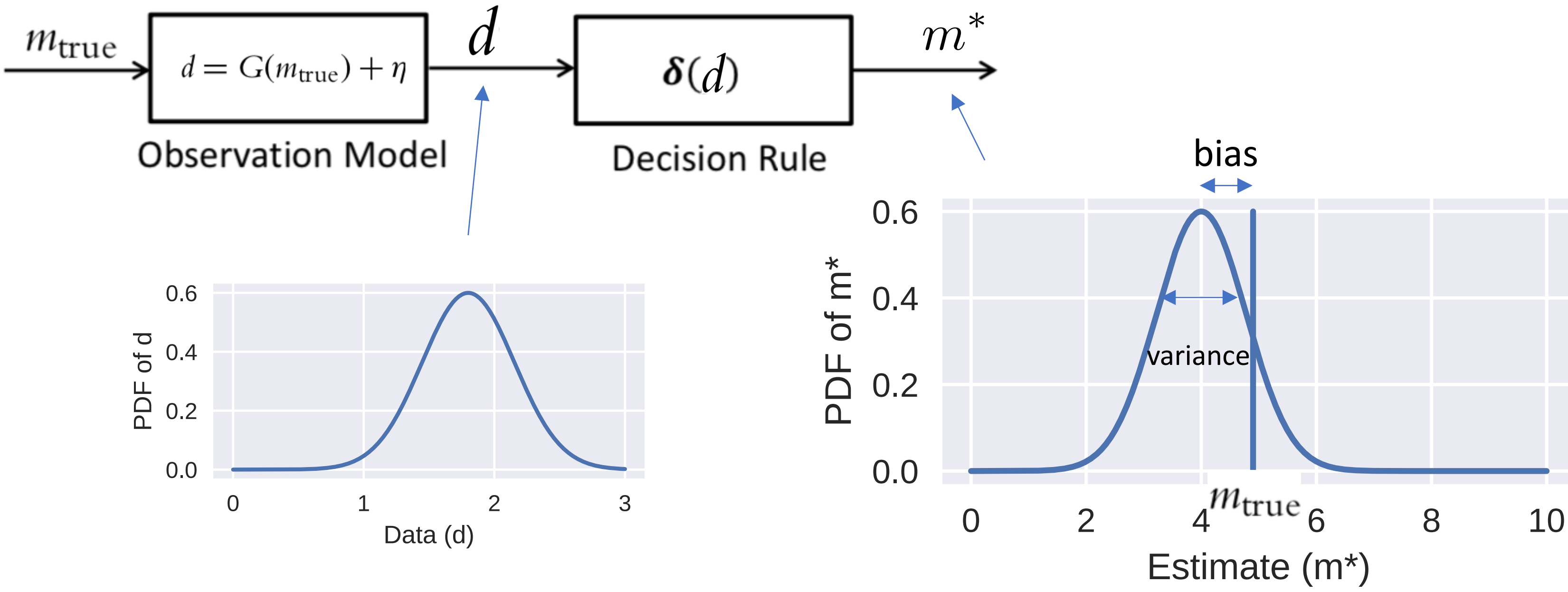
$$\min \|\mathbf{Gm} - \mathbf{d}\|_2^2 + \lambda \cdot \mathrm{entropy}(\mathbf{m})$$





Don't just invert your observation equation, or blindly use least squares!

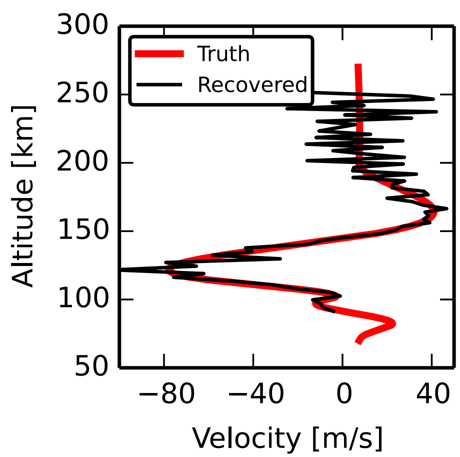
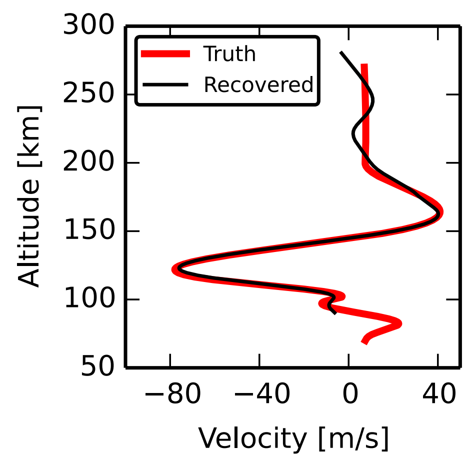
Error propagation

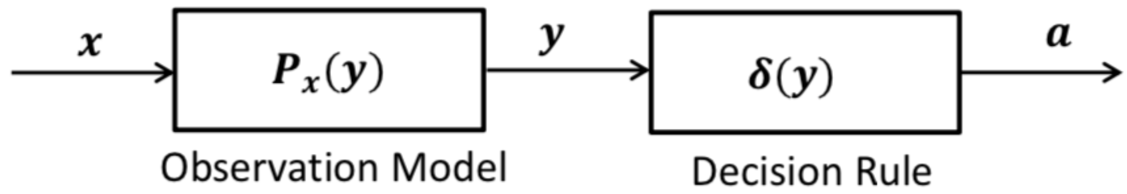


- The estimate is **also** a random variable, with mean and variance (or covariance matrix)
- Even if raw data are uncorrelated, the resulting estimate is often correlated

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2k} & \cdots & \sigma_{kk} \end{bmatrix}$$

Bias-Variance Tradeoff

	Without Regularization	With Regularization
Example Inversion		
Variance	**Large**	**Small**
Mean	**Accurate**	**Less Accurate (Smoothed)**
Inversion	**Easy**	**Hard**



Making decisions using data

Discrete Decisions

- Is there an MSTID in my data?
- Is there a relationship between atmospheric tides and electric fields?
- Did an earthquake cause an ionospheric effect?

Continuous Decisions

- Data Assimiliation

Interpreting Decisions

- Accuracy vs Precision
- Correlation vs Causation
- Monte Carlo simulation

General State-Space Signal Model

The general hidden Markov model (HMM):

Initial prior:	$p_{\boldsymbol{x}_1}(\boldsymbol{x}_1)$	(1)
Measurement/forward model:	$h_i(\boldsymbol{y}_i \vert \boldsymbol{x}_i)$	(2)
State-transition model:	$f_i(\boldsymbol{x}_{i+1} \vert \boldsymbol{x}_i)$	(3)

$$\dim(\boldsymbol{x}_i) = N \qquad \dim(\boldsymbol{y}_i) = M$$

Goal: Compute minimum mean square error (MMSE) estimates of the unknown state $\boldsymbol{x}_i$ given the measurements $\boldsymbol{y}_{1:j} \triangleq \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_j\}$.

$$\widehat{\boldsymbol{x}}_{i|j} \triangleq \mathbb{E}[\boldsymbol{x}_i|\boldsymbol{y}_{1:j}] = \int \boldsymbol{x}_i \, p(\boldsymbol{x}_i|\boldsymbol{y}_{1:j}) \, d\boldsymbol{x}_i \qquad (4)$$

Linear Additive-Noise State-Space Signal Model (Linear Gaussian Model)

$$\text{Initial prior:} \quad \mathbb{E}[\boldsymbol{x}_1] = \boldsymbol{\mu}_1,\ \text{Cov}(\boldsymbol{x}_1) = \boldsymbol{\Pi}_1 \tag{5}$$

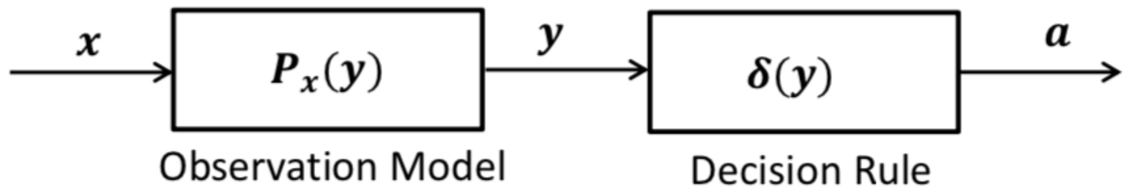
$$\text{Measurement/forward model:} \quad \boldsymbol{y}_i = \boldsymbol{H}_i \boldsymbol{x}_i + \boldsymbol{v}_i \tag{6}$$

$$\text{State-transition model:} \quad \boldsymbol{x}_{i+1} = \boldsymbol{F}_i \boldsymbol{x}_i + \boldsymbol{u}_i \tag{7}$$

- The first and second order statistics of the zero mean state $(\boldsymbol{u}_i)$ and measurement $(\boldsymbol{v}_i)$ noise are given: $\text{Cov}(\boldsymbol{u}_i) = \boldsymbol{Q}_i$ and $\text{Cov}(\boldsymbol{v}_i) = \boldsymbol{R}_i$.

Goal: Compute linear minimum mean square error (LMMSE) estimates of the unknown state $\boldsymbol{x}_i$ given the measurements $\boldsymbol{y}_{1:j}$.

→ Kalman filter



Making decisions using data

Discrete Decisions

- Is there an MSTID in my data?
- Is there a relationship between atmospheric tides and electric fields?
- Did an earthquake cause an ionospheric effect?

Continuous Decisions

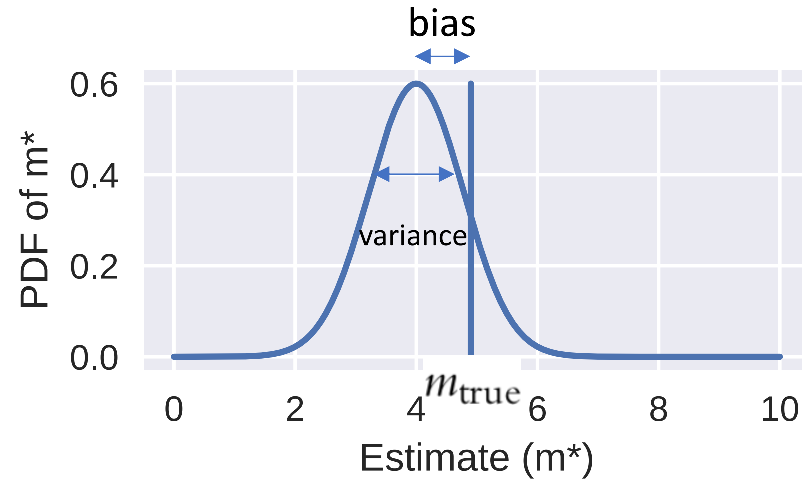
- What is the plasma velocity?
- What percentage of ionospheric variability can be attributed to the neutral atmosphere?
- What is a meteor's mass, based on its plasma trail?

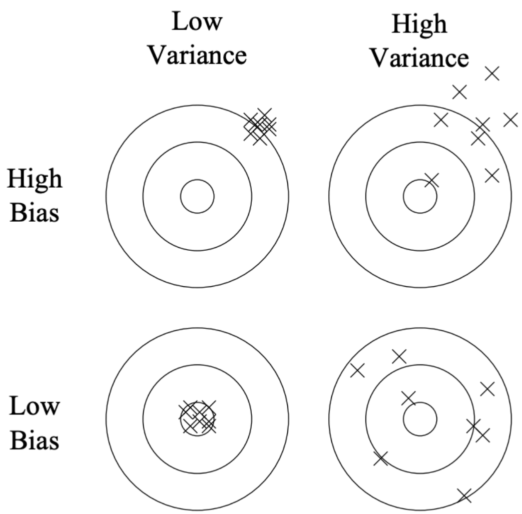
Interpreting Decisions

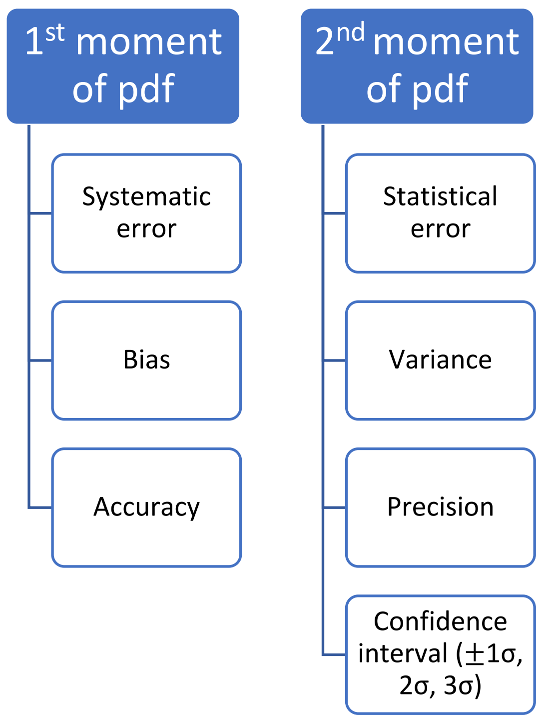
- Accuracy vs Precision
- Correlation vs Causation
- Monte Carlo simulation

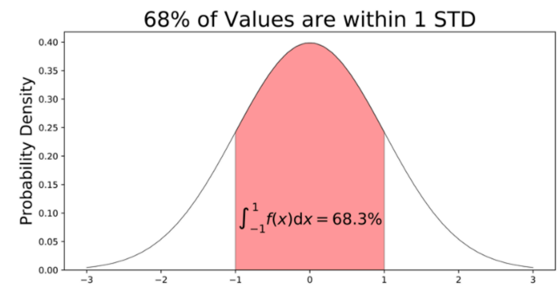
Error/Uncertainty

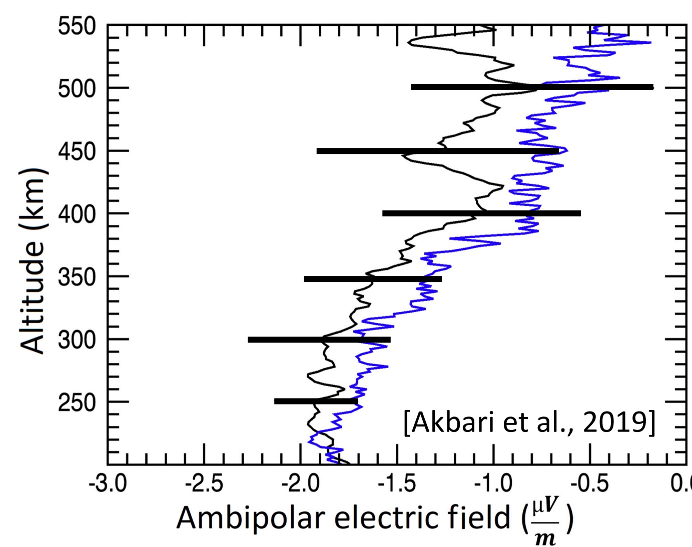
- Important to understand what error bars mean
  - Bias (e.g., calibration error)
  - Variance (e.g., noise)
- Data providers rarely report 1st moment
  - Critical for assimilation and data fusion







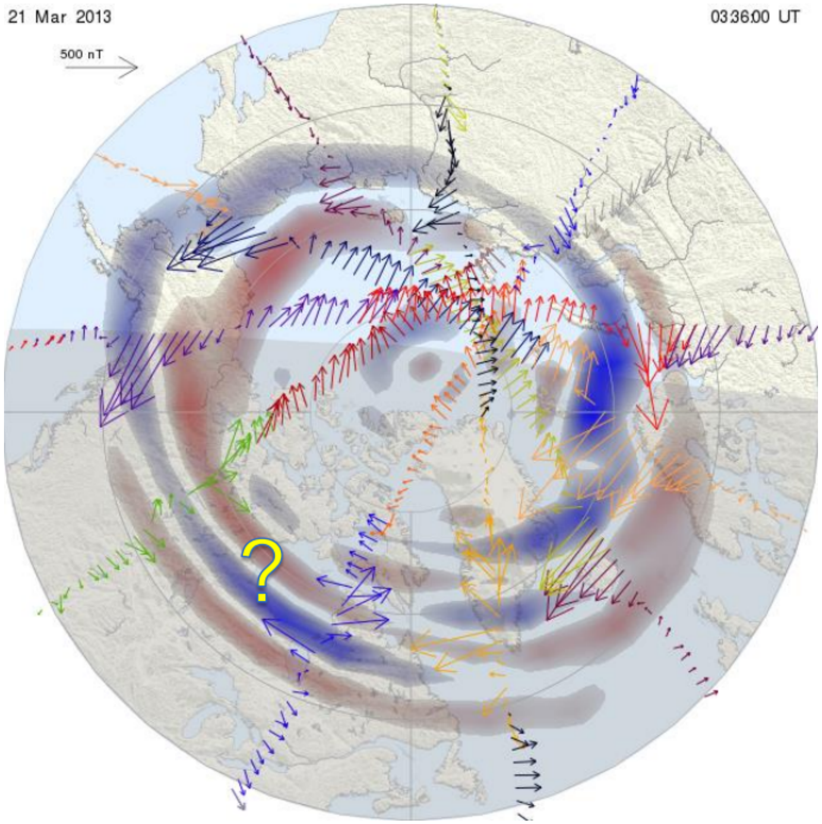




Bias and Resolution

- Geophysical data often have a bias towards “smoothness”
- Can quantify with resolution matrix:

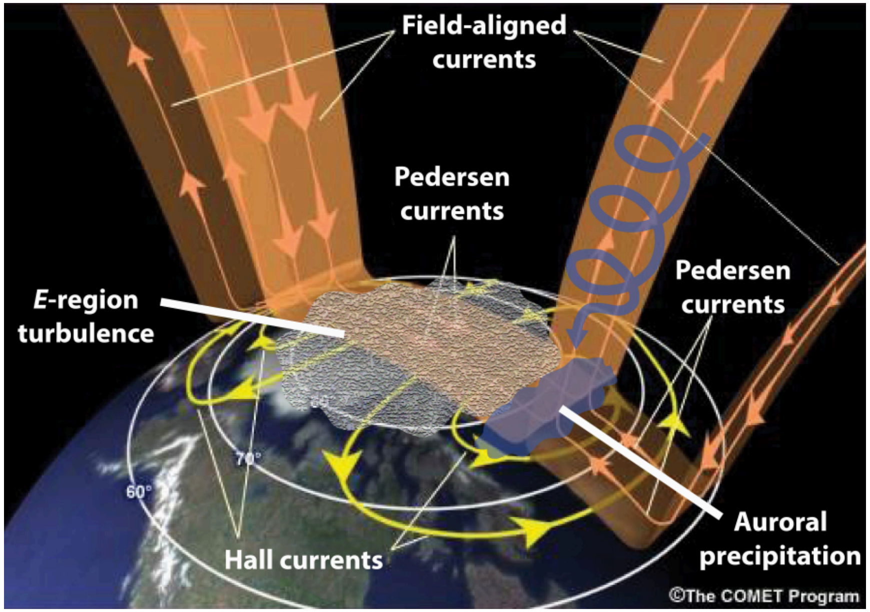
$$\text{if } d = Gm$$
$$\text{and } m^* = G^p d$$
$$\text{then } R = G^p G$$

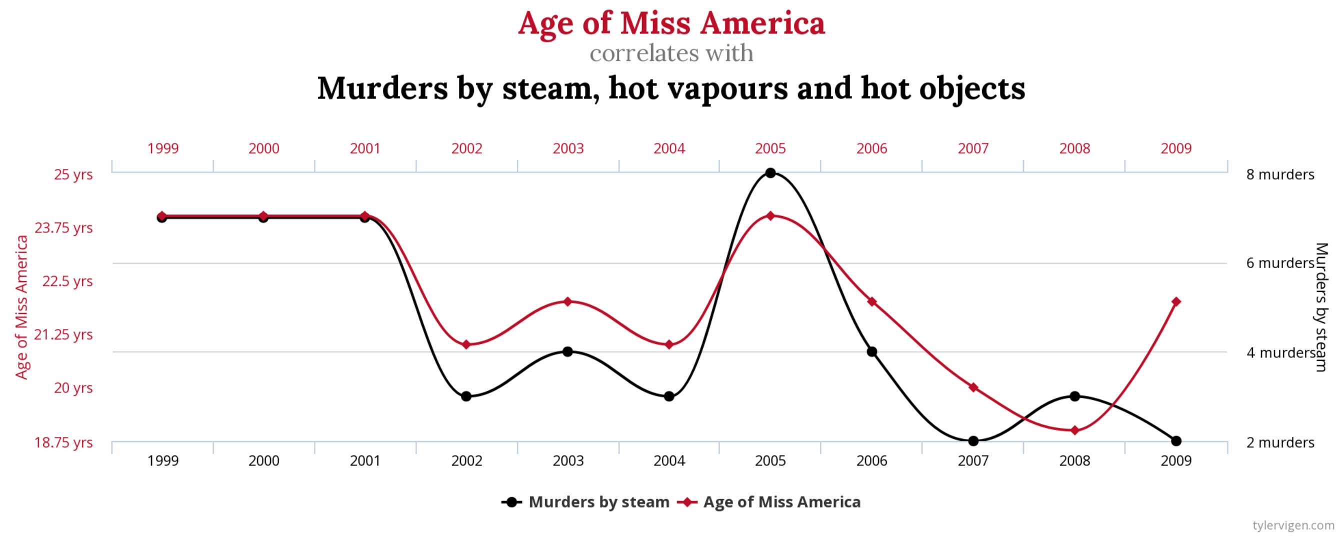


[J. Gjerloev, CEDAR Prize Lecture 2016]

Correlation vs Causation

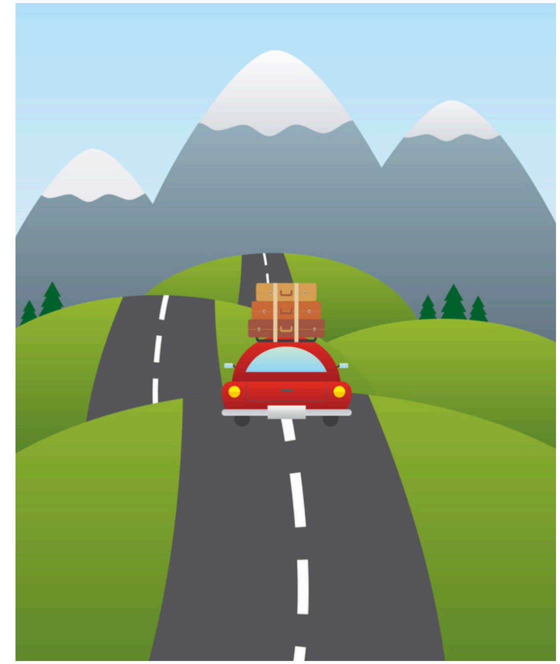
- All machine learning techniques are fueled by correlation
- Coincidental correlation
  - Multiple comparisons
  - p=0.05 → 1 in 20 studies are wrong
- Bidirectional causation
  - Predator-prey
  - Magnetosphere-ionosphere coupling

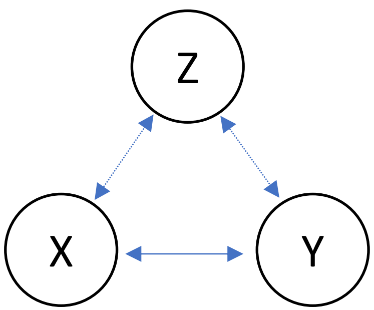


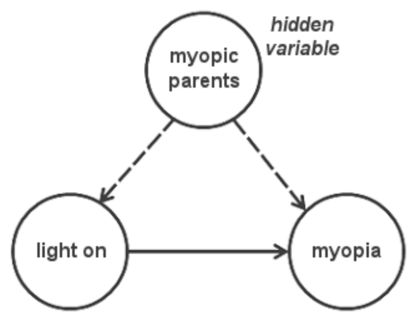


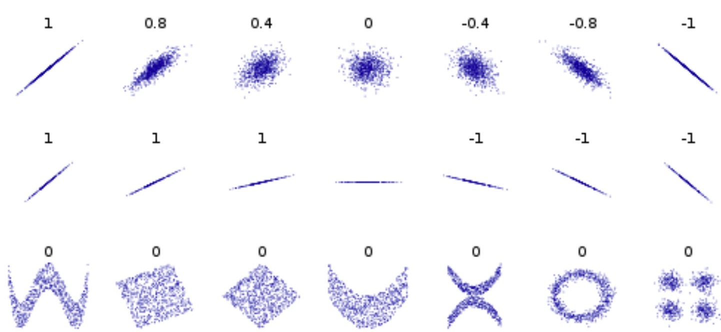
Correlation vs Causation

- Hidden variable
  - Milton Friedman's thermostat
  - How I wasted 6 months in grad school
- Controlled studies are usually the answer, but CEDAR science is largely observational.
  - Use first-principles modeling

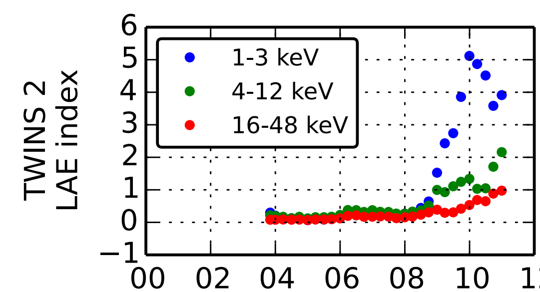


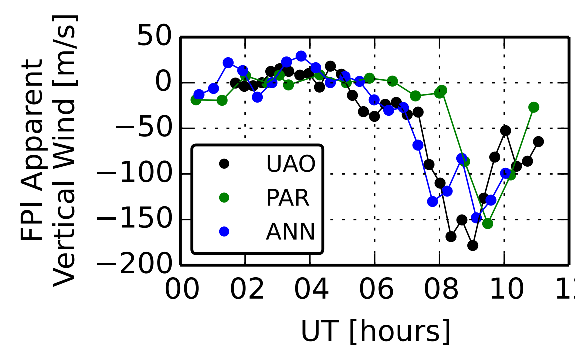


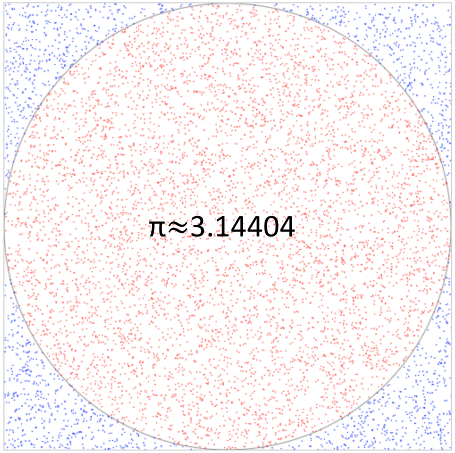


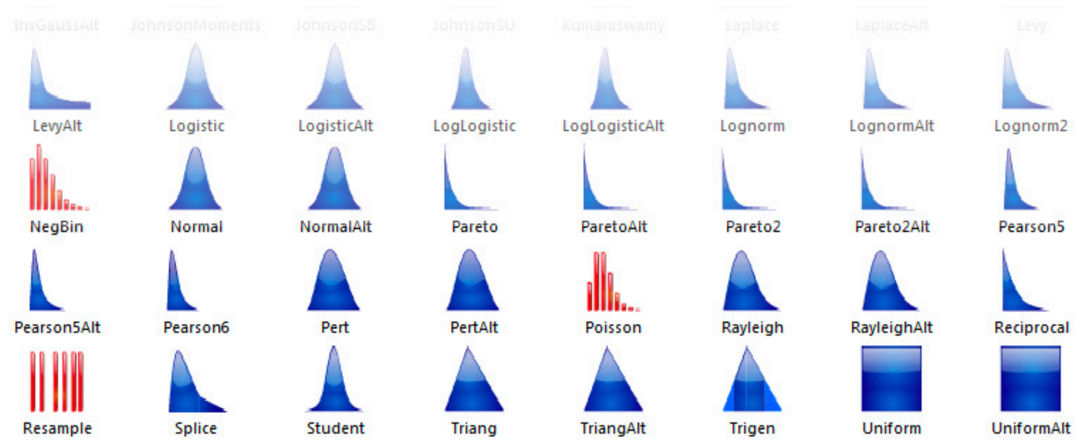


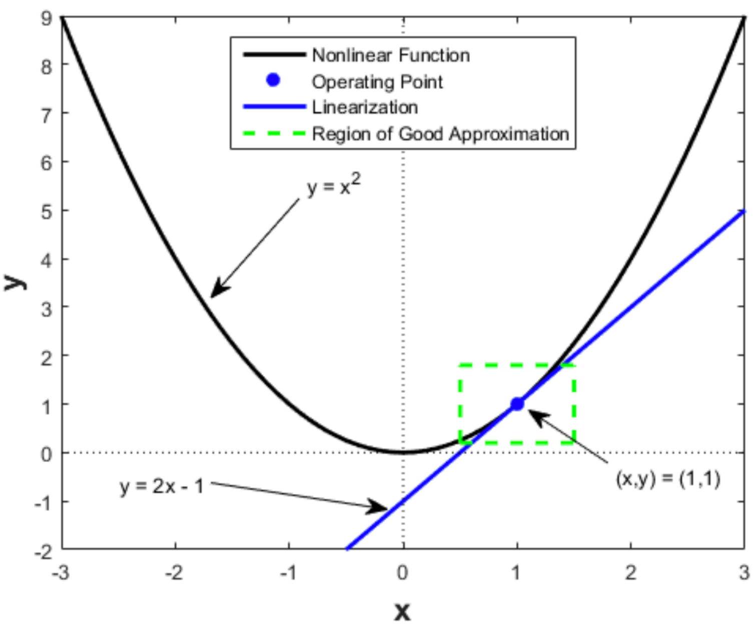
Causation doesn't imply correlation either

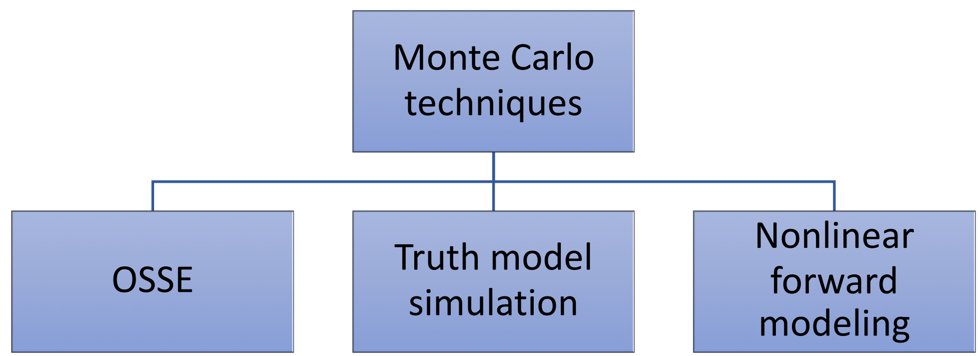




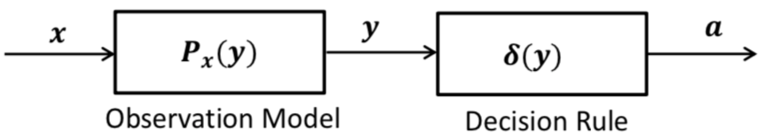

π≈3.14404


LevyAlt
Logistic
LogisticAlt
LogLogistic
LogLogisticAlt
Lognorm
LognormAlt
Lognorm2
NegBin
Normal
NormalAlt
Pareto
ParetoAlt
Pareto2
Pareto2Alt
Pearson5
Pearson5Alt
Pearson6
Pert
PertAlt
Poisson
Rayleigh
RayleighAlt
Reciprocal
Resample
Splice
Student
Triang
TriangAlt
Trigen
Uniform
UniformAlt

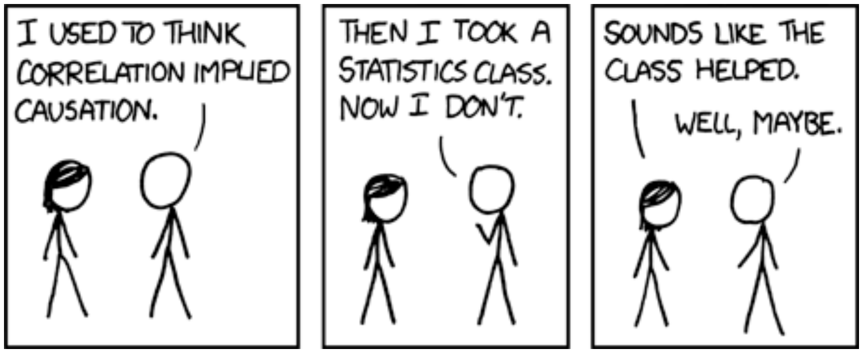

Nonlinear Function
Operating Point
Linearization
Region of Good Approximation
y = x2
y = 2x - 1
(x,y) = (1,1)
y
x


Monte Carlo techniques
OSSE
Truth model simulation
Nonlinear forward modeling

Takeaways



1. Think of all variables as **random**
2. **Don't just invert** your observation equation
   Decision and estimation theory might be able to help
3. **First-order** (systematic) errors are just as important as **second-order** (statistical) errors, especially in geoscience
   All error bars are not created equal
4. Correlation can be misleading



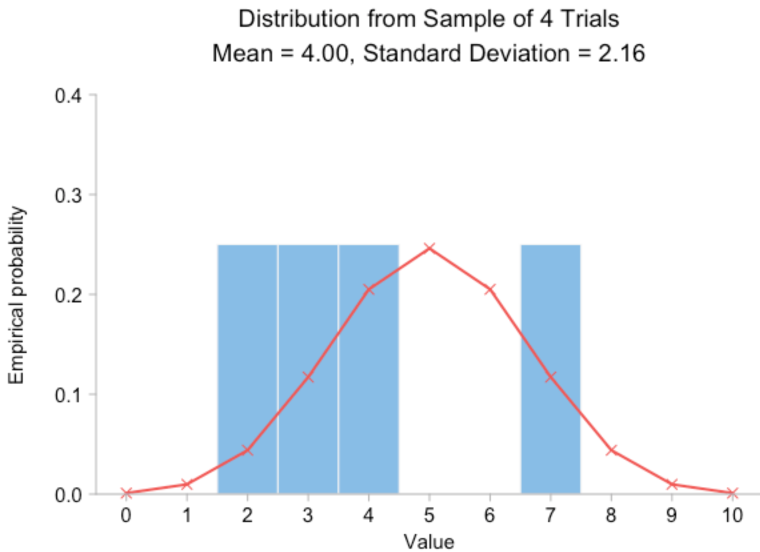
Sources

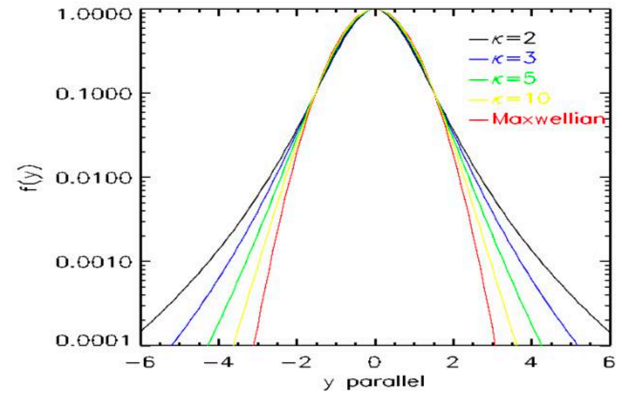
- **https://doi.org/10.1002/scin.5591770721**
- Statistical Inference for Engineers and Data Scientists, Moulin and Veeravali
- https://ccmc.gsfc.nasa.gov/models/exo.php
- Aster, R., Borchers, B., & Thurber, C. H. (2013). *Parameter Estimation and Inverse Problems.*
- https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf
- Maximum entropy: doi: 10.1029/96RS02334

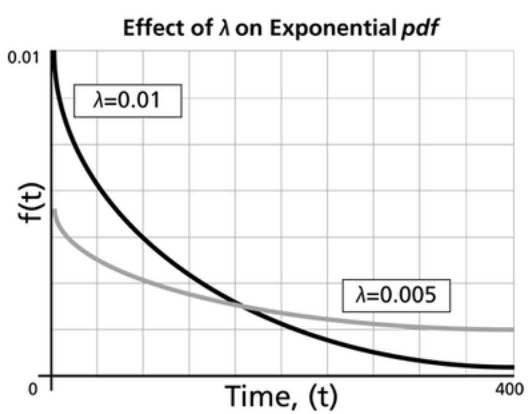
DELETED SLIDES

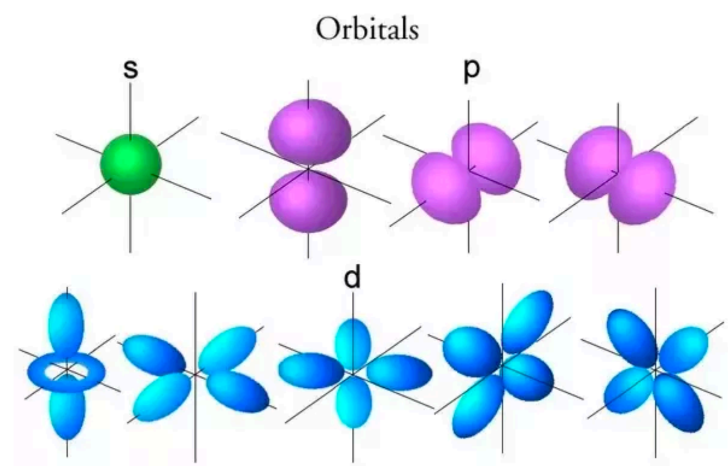
Probability Density Functions (PDFs)

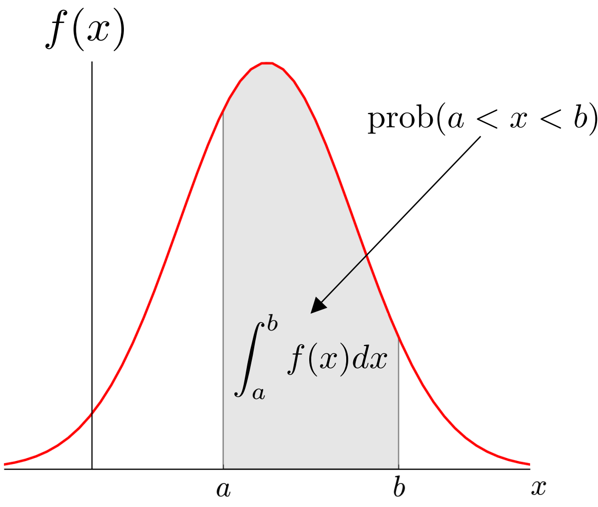
Drop 10 coins and count the heads





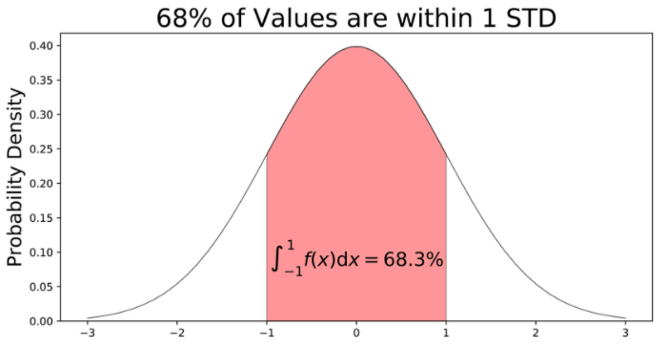






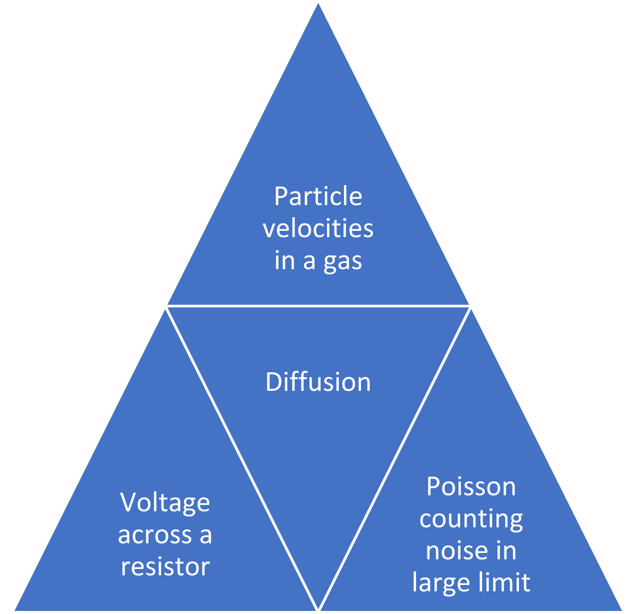
Properties of PDFs

- Integrates to 1
- The probability of any outcome is an integral over the appropriate range
- Maximum → mode, most likely value
- First moment (center of gravity) → mean, expected value
- Second moment → standard deviation, variability



Why are Gaussians used?

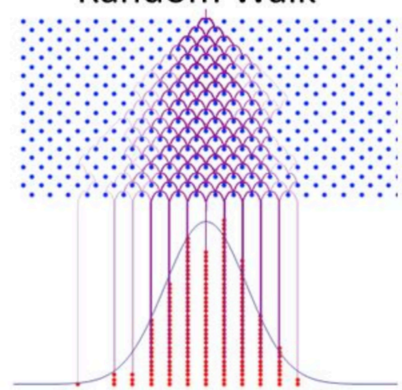
- Central Limit Theorem
- Maximum entropy for given mean & stddev
- Because it makes the math easy



$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

"Plinko" Game

Random Walk



Multivariate PDFs

- Generalize to multi-dimensional data
- Covariance matrix is important – geophysical data often have correlated errors
  - Not often reported
  - Diagonal covariance matrix often assumed – this lets you write PDF as product of individual PDFs

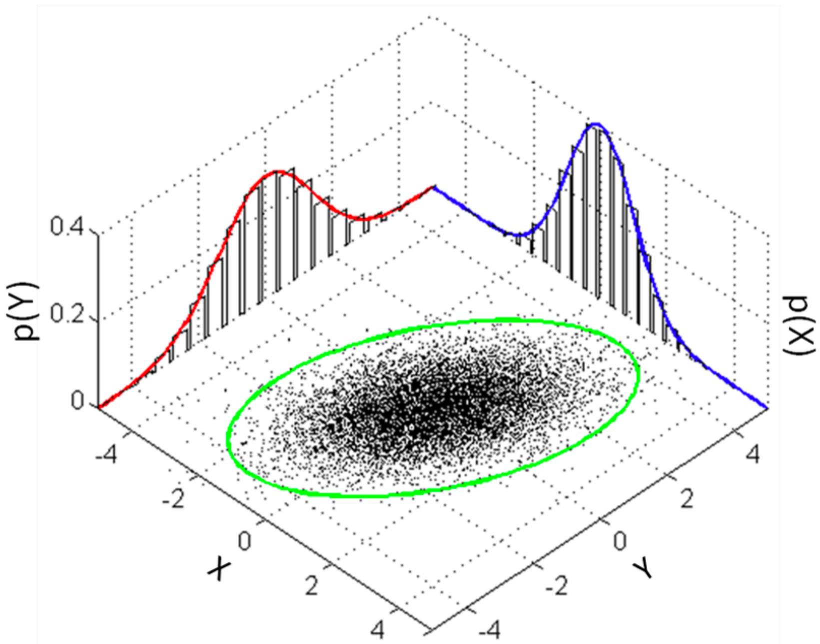
$$f(x_1,\ldots,x_k) = f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2k} & \cdots & \sigma_{kk} \end{bmatrix}$$

the variables $X_1, X_2, \ldots, X_k$ are called **mutually independent** if

$$f(x_1,\ldots,x_k) = f_1(x_1)\,f_2(x_2)\ldots f_k(x_k)$$



Multivariate PDFs

- If Gaussian, mean and covariance matrix are all you need to know
- If not, it's complicated
  - Uncorrelated vs independent

$$f\left(x_1,\ldots,x_k\right) = f\left(\mathbf{x}\right) = \frac{1}{\left(2\pi\right)^{k/2}\left|\Sigma\right|^{1/2}} e^{-\frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}\right)'\Sigma^{-1}\left(\mathbf{x}-\boldsymbol{\mu}\right)}$$

where

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2k} & \cdots & \sigma_{kk} \end{bmatrix}$$