

Fundamentals of Statistics: From Data to Knowledge to Decision-Making

Victor M. Zavala

Department of Chemical and Biological Engineering
University of Wisconsin-Madison

victor.zavala@wisc.edu





Outline

- ① Introduction
- ② Models of Random Variables
- ③ Estimation Techniques
- ④ Multivariate Statistics
- ⑤ Data-Driven Modeling
- ⑥ Statistical Learning
- ⑦ Decision-Making Under Uncertainty



Motivation

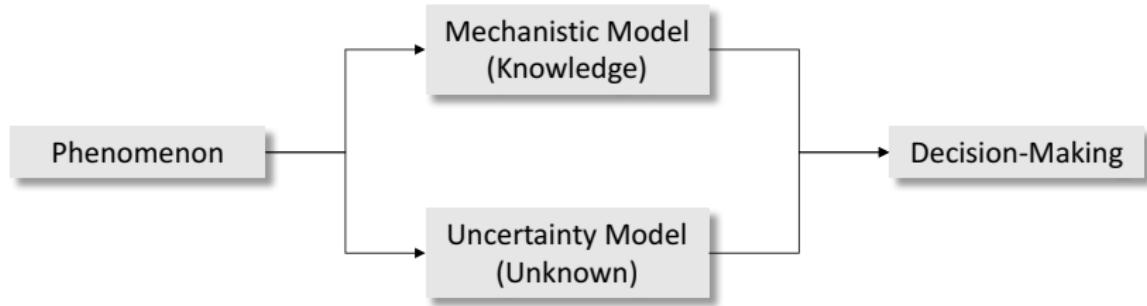
As engineers, we often use *laws* of physics and chemistry to make *decisions*:

- The discovery of these governing laws has been the result of extensive collection and analysis of observations (data)
- A governing law is often expressed in the form of a mechanistic model
- A mechanistic model provides a concise summary of observations (knowledge) that allow us to predict and generalize

These laws are powerful but only provide limited descriptions of phenomena:

- Laws are applicable under specific settings (e.g., continuum vs. atomistic)
- Discovering laws and new mechanistic models might be challenging or cost-prohibitive (e.g., climate)
- Mechanistic predictions will *always* face a certain degree of *uncertainty* (due to our limited knowledge of the world).
- Despite these limitations, we still want to be able to *make decisions*. In fact, we (as humans) make many decisions in our daily lives with limited use of mechanistic models and (somehow) by accounting for uncertainty.

Motivation





Motivation

Statistics is the branch of mathematics that offers tools to:

- Collect, analyze, and extract knowledge (models) from data
- Characterize and model the unknown (uncertainty)
- Systematically make decisions in the face of uncertainty

For an engineering perspective, *statistics* aids the discovery and development of mechanistic models and provides complementary (data-driven) modeling capabilities.

From a scientific perspective, *statistics* provides a framework for thinking about the world that can help us understand how humans naturally process data to extract knowledge and to ultimately make decisions.



Random Variables

In statistics, we use random variables (RVs) to *model* uncertain behavior. An RV (denoted as X) does not have a known value and exhibits *variability* and has the following properties:

- An RV is characterized by a realization set $\omega \in \Omega$ with associated values $x_\omega \in \mathcal{D}_X$ (a.k.a. realizations of X). Here, \mathcal{D}_X is the domain of X .
- An RV is characterized by a measure $\mathbb{P} : \Omega \rightarrow [0, 1]$, which assigns probability to events (combinations of realizations); e.g., $\mathbb{P}(X \in \mathcal{A})$ for some $\mathcal{A} \subseteq \mathcal{D}_X$.



Random Variables

- The measure \mathbb{P} has an associated cumulative density function (cdf) $F_X : \mathcal{D}_X \rightarrow [0, 1]$. The cdf assigns a probability to the event that X is below a certain threshold value x ; i.e., $F_X(x) = \mathbb{P}(X \leq x)$.
- The cdf has an associated probability density function (pdf) $f_X : \mathcal{D}_X \rightarrow [0, 1]$. The pdf assigns a probability to the event that X takes a specific value x ; i.e., $f_X(x) = \mathbb{P}(X = x)$.

An RV that has a unique and known value (exhibits no variability) is called a *deterministic variable*.

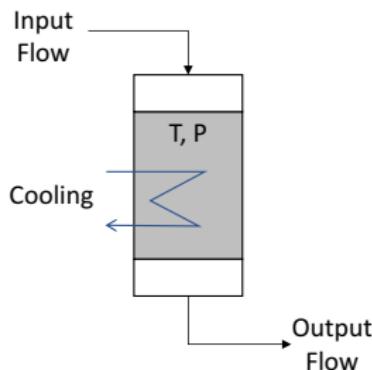


Random Variables

Don't Forget: A random variable is a *model* of a unknown phenomenon.

Example: Gibbs Reactor

- Consider a reactor under which the reaction $CO + 2H_2 \leftrightarrow CH_3OH$ takes place
- Equilibrium is favored (achieves higher conversion C) by high pressure (P) and low temperature (T)
- Control system maintains P and T at desired conditions



$$\mu_{out}^k = \mu_{in}^k + \gamma^k C, \quad k \in K$$

$$\mu_{tot} = \sum_{k \in K} \mu_{out}^k$$

$$a^k = \left(P \frac{\mu_{out}^k}{\mu_{tot}} \right)^{\gamma^k}, \quad k \in K$$

$$K_{eq}(T) = \prod_{k \in K} a^k$$

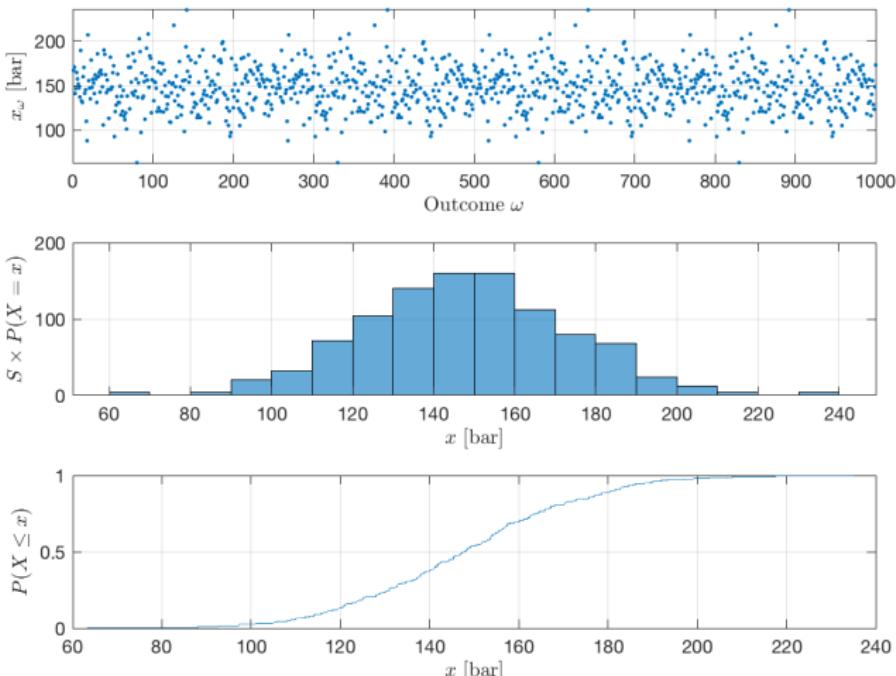
$$\log K_{eq}(T) = -\frac{\Delta H}{RT} + \frac{\Delta S}{R}$$

- What is unknown (what are sources of uncertainty)?



Example: Gibbs Reactor

- Assume pressure (P) varies due to malfunction of control and model this as an RV
- Outcomes, pdf, and cdf of RV are shown below. How do we interpret these?





Types of Random Variables

RVs are categorized as multivariate vs. univariate and continuous vs. discrete:

- A *multivariate* RV $X = (X_1, X_2, \dots, X_n)$ has realizations that are vector values $x_\omega = (x_{\omega,1}, x_{\omega,2}, \dots, x_{\omega,n}) \in \mathbb{R}^n$; e.g., temperature and reaction conversion.
- A *univariate* RV X is a multivariate with $n = 1$ and has realizations that are scalar values $x_\omega \in \mathbb{R}$; e.g., temperature.
- A *continuous* RV X is that in which the domain \mathcal{D}_X is continuous; e.g., $X = (X_1, X_2)$ has realizations satisfying $0 \leq x_{\omega,1} \leq 1$ and $0 \leq x_{\omega,2} \leq 1$.
- A *discrete* RV X is that in which the domain \mathcal{D}_X is discrete; e.g., $X = (X_1, X_2)$ has realizations satisfying $x_{\omega,1} \in \{0, 1\}$ and $x_{\omega,2} \in \{0, 1\}$.

There is a wide range of models of random variables that apply to the different categories (e.g., Gaussian is for continuous and Poisson for discrete). We will explore these later.



Probability Density of Discrete and Continuous RVs

A discrete RV X has a discrete domain \mathcal{D}_X and, as such, its pdf $f_X(x)$ is not a continuous function. Consequently:

$$f(x) \geq 0, \quad x \in \mathcal{D}_X$$

$$\sum_{x \in \mathcal{D}_X} f(x) = 1$$

$$\mathbb{P}(X \in \mathcal{A}) = \sum_{x \in \mathcal{A}} f(x), \quad \mathcal{A} \subseteq \mathcal{D}_X.$$

A continuous RV X has a continuous domain \mathcal{D}_X and its pdf $f_X(x)$ is a continuous function. Consequently:

$$f(x) \geq 0, \quad x \in \mathcal{D}_X$$

$$\int_{x \in \mathcal{D}_X} f(x) dx = 1$$

$$\mathbb{P}(X \in \mathcal{A}) = \int_{x \in \mathcal{A}} f(x) dx, \quad \mathcal{A} \subseteq \mathcal{D}_X.$$



Probability Density of Discrete and Continuous RVs

- A discrete RV is easy to handle computationally (involves summations):
 - If \mathcal{D}_X is discrete then $\mathbb{P}(X \leq a) = F_X(a) = \sum_{x \in \mathcal{D}_X} f_X(x) \mathbf{1}[x \leq a]$.
Here, we use the indicator function $\mathbf{1}[x \leq a] = 1$ if $x \leq a$ and $\mathbf{1}[x \leq a] = 0$ if $x > a$.
- A continuous RV is difficult to handle computationally (involves integrals) but has useful properties that facilitate analysis:
 - If \mathcal{D}_X continuous then $\mathbb{P}(X \leq a) = F_X(a) = \int_{x \in \mathcal{D}_X} f_X(x) dx$.
 - The cdf and pdf are related as $\frac{dF_X(x)}{dx} = f_X(x)$ and thus $\int_{x \in \mathcal{A}} f_X(x) dx = \int_{x \in \mathcal{A}} dF_X(x)$.
- Continuous RVs are often approximated using discrete RVs (discretization). This is analogous to discrete time and continuous time.



From Data to Random Variables

We begin our discussion by considering RVs that are *univariate*.

- In practice, we often count with observations (data) for a given variable of interest $x_\omega, \omega \in \mathcal{S}$. Our objective is to use the data to model this variable as an RV.
- If we assume that \mathcal{S} (a.k.a. sample set) is a subset of the realization set Ω (which is usually extremely large), we can construct a data-driven approximation (a.k.a. empirical or sample approximation) of the domain, cdf, and pdf of X .
- Empirical domain \hat{D}_X is domain covered by observations $x_\omega, \omega \in \mathcal{S}$.
- The “theoretical” pdf is approximated using empirical pdf:

$$\hat{f}_X(x) = \frac{1}{S} \sum_{\omega \in \mathcal{S}} \mathbf{1}[x_\omega = x], \quad x \in \hat{D}_X$$

i.e., this is the frequency at which X takes a value x normalized by $S = |\mathcal{S}|$.

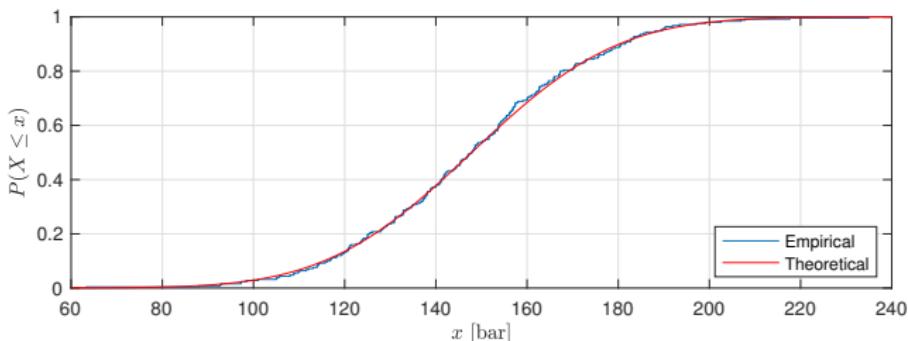
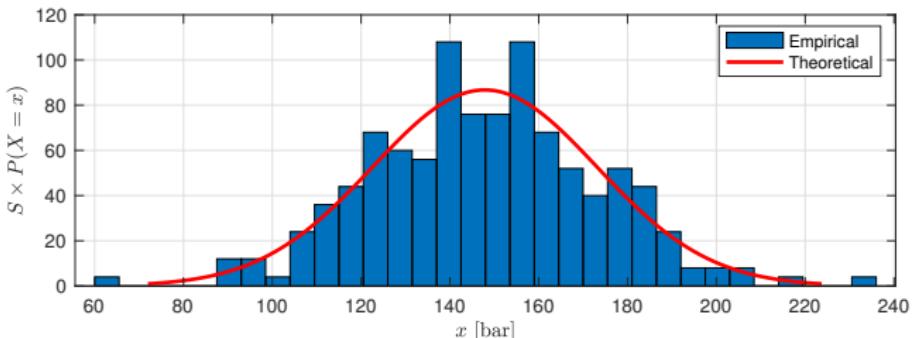
- The “theoretical” cdf is approximated using empirical cdf:

$$\hat{F}_X(x) = \frac{1}{S} \sum_{\omega \in \mathcal{S}} \mathbf{1}[x_\omega \leq x], \quad x \in \hat{D}_X$$

i.e., this is the frequency at which X takes a value below x normalized by $S = |\mathcal{S}|$.

Example: Gibbs Reactor

- Here is a comparison of empirical and theoretical pdfs and cdfs for pressure





Summarizing Statistics (Basic)

- The pdf and cdf are *functions* that fully characterize an RV X . However, in practice, we might be interested in using values (and not functions) to describe X .
- This is done by using *summarizing statistics* (a.k.a. descriptive statistics).
- Popular summarizing statistics are the expected value and variance.

For a discrete RV we have:

- Expected Value (measure of magnitude):* $\mathbb{E}_X := \sum_{x \in \mathcal{D}_X} x f_X(x)$
- Variance and Standard Deviation (measure of variability):*

$$\mathbb{V}_X := \sum_{x \in \mathcal{D}_X} f_X(x)(x - \mathbb{E}_X)^2, \quad \mathbb{SD}_X := \sqrt{\mathbb{V}_X}$$

For a continuous RV we have:

- Expected Value (measure of magnitude):* $\mathbb{E}_X := \int_{x \in \mathcal{D}_X} x f_X(x) dx$
- Variance and Standard Deviation (measure of variability):*

$$\mathbb{V}_X := \int_{x \in \mathcal{D}_X} f_X(x)(x - \mathbb{E}_X)^2 dx, \quad \mathbb{SD}_X := \sqrt{\mathbb{V}_X}$$

When convenient, we will use notation $\mathbb{E}[X]$, $\mathbb{V}[X]$, $\mathbb{SD}[X]$, and so on.



Summarizing Statistics (Sample Approximations)

If we only have a finite set of data for X available, we can approximate the summarizing statistics using their sample approximations:

- *Sample Mean (measure of magnitude):*

$$\hat{\mathbb{E}}_X := \sum_{x \in \hat{\mathcal{D}}_X} x \hat{f}_X(x) = \frac{1}{S} \sum_{\omega \in \mathcal{S}} x_\omega$$

- *Sample Variance and Standard Deviation (measure of variability):*

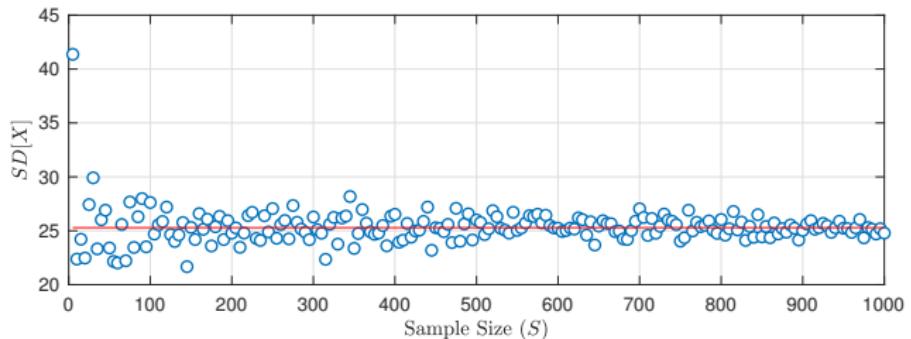
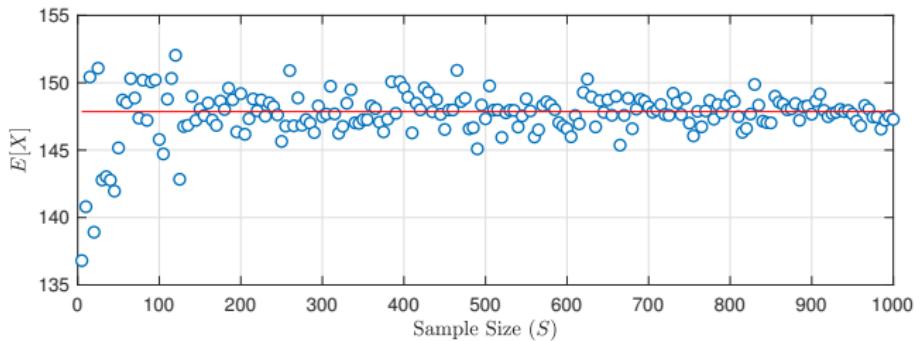
$$\hat{\mathbb{V}}_X := \sum_{x \in \hat{\mathcal{D}}_X} (x - \hat{\mathbb{E}}_X)^2 \hat{f}_X(x) = \frac{1}{S} \sum_{\omega \in \mathcal{S}} (x_\omega - \hat{\mathbb{E}}_X)^2, \quad \hat{\mathbb{SD}}_X = \sqrt{\hat{\mathbb{V}}_X}$$

Intuition tells us that these become better approximations as we accumulate data of X (i.e., as \mathcal{S} approaches Ω). We will see later on that this is indeed the case.



Example: Gibbs Reactor

- Here is behavior of sample mean and standard deviation as we increase sample size S





Summarizing Statistics (Quantiles)

An important family of summarizing statistics are the quantiles (a.k.a. percentiles).

- The quantile is the inverse function of the cdf and, as such, it might be easier to explain it from this perspective. Consider the following equation for some $\alpha \in [0, 1]$:

$$F_X(x) = \mathbb{P}(X \leq x) = \alpha$$

- A value x that satisfies this equation is the α -quantile of the random variable X and is denoted as $Q_X(\alpha)$. This means that we can express the quantile as:

$$Q_X(\alpha) = F_X^{-1}(\alpha)$$



Summarizing Statistics (Quantiles)

Some important observations about quantiles:

- Since the cdf can have a “staircase” form, there might be multiple values of x satisfying $F_X(x) = \alpha$. Consequently, the α -quantile might be not be unique. Typically, the definition of the quantile is refined by looking for the smallest or center values of x satisfying $F_X(x) \geq \alpha$.
- Quantiles are related to other summarizing statistics for interest. For instance:
 - $Q_X(1/2)$ is the *center value* of X (a.k.a. the median and denoted as M_X)
 - $Q_X(1) = \max_{x \in \mathcal{D}_X} x$ is the maximum value of X
 - $Q_X(0) = \min_{x \in \mathcal{D}_X} x$ is the minimum value of X
- We can use empirical cdf $\hat{F}_X(x)$ to estimate empirical quantiles $\hat{Q}(\alpha)$.



Summarizing Statistics (Moments)

- *Central moments* are an important family of summarizing statistics
- The moments of X with pdf $f(x, \theta)$ are given by:

$$m_k := \mathbb{E}[(X - \mathbb{E}[X])^k], \quad k = 1, 2, 3, 4, \dots,$$

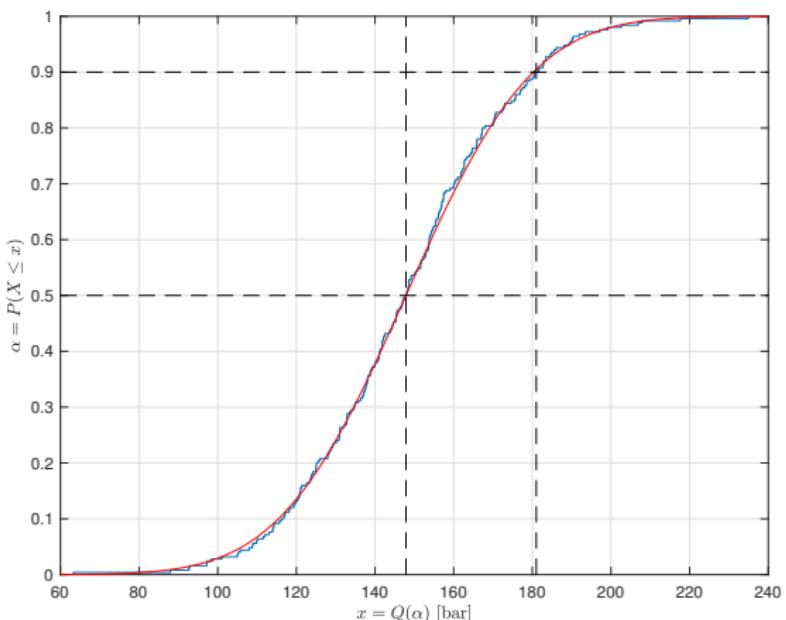
- The normalized moments of X with pdf $f(x, \theta)$ are given by:

$$m_k := \frac{\mathbb{E}[(X - \mathbb{E}[X])^k]}{SD[X]^k}, \quad k = 1, 2, 3, 4, \dots,$$

- First moment is simply $m_1 = 0$, second moment $m_2 = \mathbb{V}[X]$ is variance, third moment m_3 is known as skewness, and fourth moment m_4 as kurtosis.
- As with expectation and variance, we can use data to construct sample approximations for moments \hat{m}_k .

Example: Gibbs Reactor

- Sample moment approximations with $S = 1000$ are $\hat{m}_1 = 0, \hat{m}_2 = 638, \hat{m}_3 = 1539$.
- Theoretical are $\hat{m}_1 = 0, \hat{m}_2 = 639, \hat{m}_3 = 0$. What is going on with third moment?
- Empirical and theoretical quantiles are shown below. How do we interpret this?





Uncertainty Propagation and Mitigation

The uncertainty (and variability) associated with random variables propagate through systems in complex ways. Fortunately, we often have the ability to manipulate a system (through design or control) in order to mitigate the effects of this variability.

Consider propagation of X through system $\varphi(X, u)$, where $u \in \mathcal{U}$ is an action (decision):

$$Y = \varphi(X, u)$$

We make the following observations:

- The output Y is an RV if the input X is an RV.
- The nature of Y (its cdf, pdf, and domain) depends on the system function φ . Some systems might magnify uncertainty and variability while others might damp it.
- The nature of Y depends on the decision u . We can implement actions on the system φ to control the uncertainty and variability of Y .



Uncertainty Propagation and Mitigation

Having data x_ω , $\omega \in \mathcal{S}$ and a system φ , we can characterize the cdf, pdf, domain, and summarizing statistics of Y using the following procedure.

- For a given u , perform simulations of the form:

$$y_\omega = \varphi(x_\omega, u), \omega \in \mathcal{S}$$

- We use y_ω to compute approximations of quantities of interest for Y such as:
 - Sample mean:

$$\hat{\mathbb{E}}_Y = \frac{1}{S} \sum_{\omega \in \mathcal{S}} y_\omega = \frac{1}{S} \sum_{\omega \in \mathcal{S}} \varphi(x_\omega, u)$$

- Sample variance:

$$\hat{\mathbb{V}}_Y = \frac{1}{S} \sum_{\omega \in \mathcal{S}} (y_\omega - \hat{\mathbb{E}}_Y)^2$$

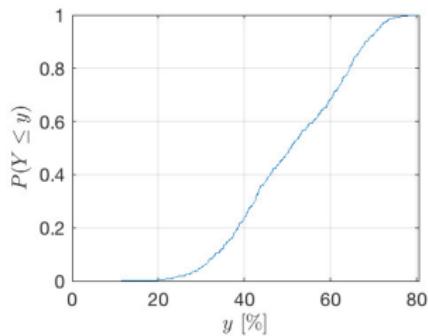
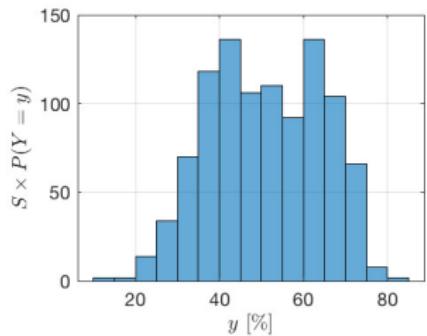
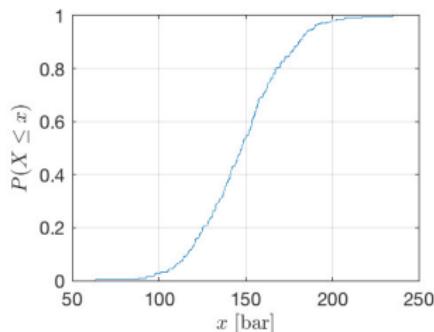
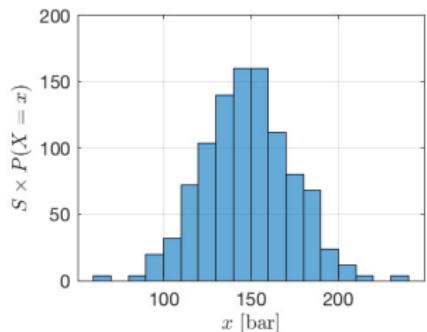
- Empirical cdf:

$$\hat{F}_Y(y) = \frac{1}{S} \sum_{\omega \in \mathcal{S}} \mathbf{1}[y_\omega \leq y]$$

and quantiles $\hat{Q}_Y(\alpha)$.

Example: Gibbs Reactor

- Empirical pdf and cdf for pressure (X) and conversion (Y). What do you notice?





Decision-Making under Uncertainty

Consider now that we would like to find a decision $u \in \mathcal{U}$ that controls $Y(u) = \varphi(X, u)$ in some desirable way. This gives rise to a couple of questions:

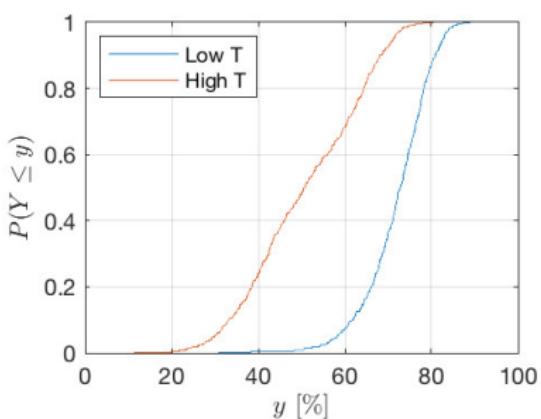
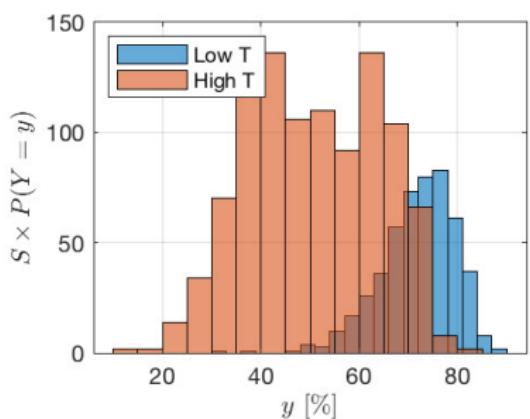
- If we have a couple of competing decisions u and u' giving rise to random outputs $Y(u)$ and $Y(u')$. How can we tell which one is better?
- How can we find the best possible decision u ?

Some observations:

- If we assume a *deterministic setting* with no uncertainty, then $Y(u)$ and $Y(u')$ will each take a single value and one would select, *unambiguously*, the one with larger (or smaller) value. Mathematically, one would select u if $Y(u) \geq Y(u')$.
- In a *setting under uncertainty* this is no longer possible because $Y(u)$ and $Y'(u)$ have multiple possible outcomes and with different probabilities.
- The concept of “better” under uncertainty is ambiguous and the mathematical statement $Y(u) \leq Y(u')$ does not even make sense. Does $Y(u) \leq Y(u')$ mean that all the outcomes of $Y(u)$ are lower than those $Y(u')$? Does it mean that only a subset of outcomes is lower?

Example: Gibbs Reactor

- Can counteract variability in pressure X by operating at low or high temperature u
- Compare empirical pdf and cdf for conversion at low $Y(u)$ and high temp $Y(u')$
- Should we operate at low or high temperature?





Estimation

Given data $x_\omega, y_\omega, \omega \in \mathcal{S}$ available, we now place our attention to the question:

Is the data following a particular pattern (trend)? If there is a pattern, can we model it?

In this context, by a model, we mean two things:

- If we have empirical statistics (e.g., cdf, pdf, mean, variance) obtained from data x_ω, y_ω , do these match the statistics of a *known* RV?
- If we do not know the system model φ that relates x_ω and y_ω , can we determine this by using input-output data?

Estimation is task of determining models from data. Having model will allow us:

- Determine if the available data is sufficient to say something meaningful about events that have not been observed (e.g., need more data to make a decision?).
- Make predictions about other possible events and their respective probabilities (e.g., how likely is an extreme event from happening?)
- Extract trends that help us summarize the data available (from data to knowledge).
- Conduct uncertainty propagation and ultimately make decisions.

Our first step is to postulate an RV model and see if this fits the data.



Outline

- 1 Introduction
- 2 Models of Random Variables
- 3 Estimation Techniques
- 4 Multivariate Statistics
- 5 Data-Driven Modeling
- 6 Statistical Learning
- 7 Decision-Making Under Uncertainty



Model of a Gaussian RV

- A wide range of RV models have been developed over the years based on identification of common patterns that emerge in real-life phenomena.
- Many phenomena follow the behavior of a normal RV (a.k.a. Gaussian RV).

A Gaussian RV is continuous and has an associated pdf:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad x \in \mathcal{D}_X = \{-\infty \leq x \leq \infty\}$$

- The scalar values $\mu \in \mathbb{R}, \sigma \in \mathbb{R}$ are hyperparameters that are specific to the application of interest. These are the values that we will seek to tune to match the model to the available data.
- We express the fact that an RV X is Gaussian as $X \sim \mathcal{N}(\mu, \sigma^2)$.

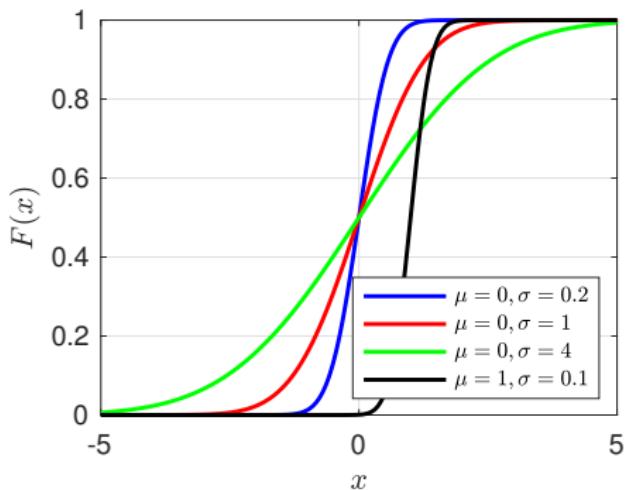
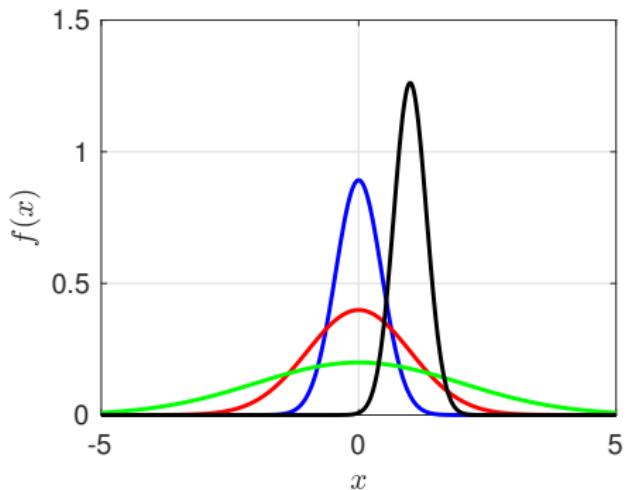


Model of a Gaussian RV

- The pdf tells us the behavior of the Gaussian RV:
 - the probability of an outcome x decays exponentially fast as we move from μ
 - the outcome of maximum probability (most likely outcome) is μ
 - the speed of the decay is dictated by σ
 - the decay in probability is symmetric around μ
- Gaussian model assumes that an outcome x can take any value in $(-\infty, \infty)$. This introduces complications, as many phenomena involves variables that cannot take negative values (e.g., mass) or infinite values (e.g., temperatures).
- Gaussian RVs can model a wide range of phenomena (e.g., diffusion). Moreover, many phenomena have the Gaussian RV as a limiting case (we will show this later).

Model of a Gaussian RV

- Here are the pdfs for $\mathcal{N}(\mu, \sigma)$ for different values of μ and σ .
- What do you observe?
- What happens when $\sigma^2 \rightarrow 0$?





Example: Diffusion Phenomena

- Gaussian RV naturally emerges in diffusion phenomena
- Consider question: What is probability of finding a particle in a particular location x in the spatial domain $[-\infty, \infty]$ and at a given time $t \in [0, T]$?
- One can show that such probability, denotes as $f(x, t)$ solves the diffusion equation:

$$\frac{\partial f(x, t)}{\partial t} = D \frac{\partial^2 f(x, t)}{\partial x^2}$$

with boundary conditions:

$$f(x, t) = 0 \quad x \in [-\infty, \infty]$$

$$\int_{-\infty}^{\infty} f(x, t) dx = 1, \quad t \in [0, T]$$

- The solution is:

$$f(x, t) = \frac{1}{\sqrt{4\pi Dt}} e^{-\frac{x^2}{4Dt}}, \quad x \in [-\infty, \infty], \quad t \in [0, T]$$

- Define dispersion coefficient $\sigma^2 = 2Dt$ and write:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}, \quad x \in [-\infty, \infty]$$

- That is position X is random and given by $X \sim \mathcal{N}(0, \sigma^2)$
- How does probability change with diffusivity D ?



Properties of a Gaussian RV

A Gaussian RV $X \sim \mathcal{N}(\mu, \sigma^2)$ has many useful properties. For instance:

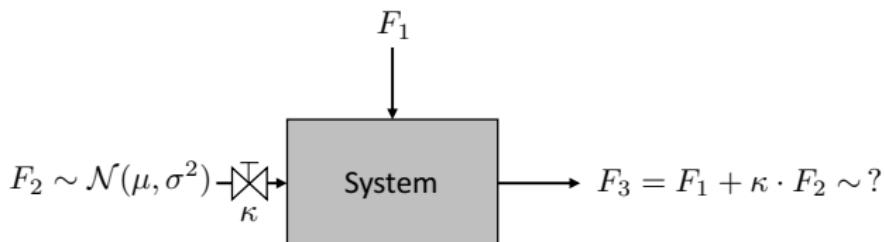
- Its expected value and variance are $\mathbb{E}_X = \mu$ and $\mathbb{V}[X] = \sigma^2$.
- Any linear transformation $Y = a + bX$ yields a Gaussian RV $Y \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$. This implies that $\mathbb{E}_Y = a + b\mathbb{E}_X$ and $\mathbb{V}_Y = b^2\mathbb{V}_X$.
- The cdf of $Y = a + bX$ satisfies $F_Y(y) = F_X(x)$ for all $y = a + bx$.

Think about the implications of the above properties from an estimation and uncertainty propagation perspective:

- We can estimate μ and σ from data simply as $\hat{\mu} = \hat{\mathbb{E}}_X$ and $\hat{\sigma}^2 = \hat{\mathbb{V}}[X]$. This is sufficient to create our empirical Gaussian model.
- Any linear system $\varphi(X) = a + bX$ will generate a Gaussian RV as output if the input X is linear. Moreover, the system will shrink the variability of X if $b < 1$ and will magnify if $b > 1$.

Example: Mixing Problem

- We have an input flow $F_1 = 10$ (gpm) that can be measured with high accuracy to the point that it is OK to assume this to be deterministic.
- We have another input flow F_2 (gpm) that cannot be measured with high accuracy and is thus modeled as an RV $\mathcal{N}(20, 1)$. This flow can be controlled using a valve with coefficient $\kappa \in [0, 1]$.
- What type of RV is the output flow $F_3 = F_1 + \kappa \cdot F_2$? What is its mean and std dev?
- How does uncertainty in F_3 change $\kappa \rightarrow 0$ and $\kappa \rightarrow 1$? Why?



- We have that $F_3 = 10 + \kappa F_2$ and is thus F_3 a linear transformation of F_2
- We thus have that $F_3 \sim \mathcal{N}(10 + \kappa \cdot 20, \kappa^2 \cdot 1)$



Properties of a Gaussian RV

The cdf of a Gaussian RV is given by:

$$F_X(x) = \frac{1}{2} \left(1 + \operatorname{erf} \left(\frac{(x - \mu)/\sigma}{\sqrt{2}} \right) \right)$$

where $\operatorname{erf} : \mathbb{R} \rightarrow \mathbb{R}$ is the error function:

$$\operatorname{erf} \left(\frac{(x - \mu)/\sigma}{\sqrt{2}} \right) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{(x - \mu)/\sigma}{\sqrt{2}}} e^{-t^2} dt$$

Computing the cdf involves evaluating an integral that depends on μ and σ .



Properties of a Gaussian RV

- Fortunately, one can exploit properties of Gaussian RVs to avoid this issue.
- Note that $Z = (X - \mu)/\sigma$ is a linear transformation of $X \sim \mathcal{N}(\mu, \sigma^2)$ and thus $Z \sim \mathcal{N}(0, 1)$.
- Now note that the pdf and cdf of Z are simply:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$$F_Z(z) = \frac{1}{2} \left(1 + \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right) \right), \quad \operatorname{erf}\left(\frac{z}{\sqrt{2}}\right) = \frac{2}{\sqrt{\pi}} \int_0^{\frac{z}{\sqrt{2}}} e^{-t^2} dt$$

which do not depend on hyperparameters; Z is known as the standard normal RV.



Properties of a Gaussian RV

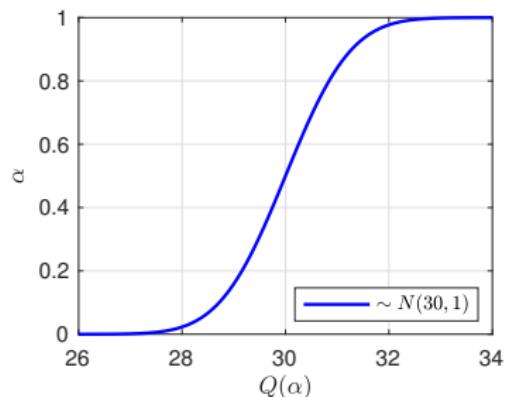
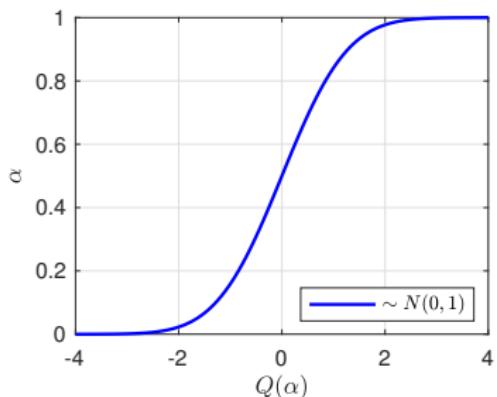
- From linear transformation properties we have that $F_X(x) = F_Z(z)$ holds for any $z = (x - \mu)/\sigma$ and thus we can evaluate $F_X(x)$ at a given value x by transforming x into z and then evaluate $F_Z(z)$.
- Since $F_Z(z)$ does not depend on any parameters, it can be precomputed (values of $F_Z(z)$ are available in software packages).
- If we want to compute $Q_X(\alpha) = F_X^{-1}(\alpha)$. As with cdf, we can compute this by using pre-computed quantiles of Z , which we denote as $z_\alpha := F_Z^{-1}(\alpha)$.
- Relationship between the quantiles of X and Z is obtained directly from the linear transformation $x = \mu + \sigma z$:

$$Q_X(\alpha) = \mu + \sigma z_\alpha.$$

The values z_α are known as the critical values of the standard normal. As with the cdf, these values have been precomputed and are available in software packages.

Example: Mixing Problem

- Consider $\kappa = 1$ and thus flow $F_3 \sim \mathcal{N}(30, 1)$
- Quantile functions for $\mathcal{N}(0, 1)$ and $\mathcal{N}(30, 1)$ are shown below



- Quantile of $\mathcal{N}(0, 1)$ at $\alpha = 0.977$ is $Q(\alpha) = 2$
- Quantile of $\mathcal{N}(\mu, \sigma)$ should be $\mu + 2 \cdot \sigma = 32$. Confirm this is true from plot



Properties of a Gaussian RV

- Standardization allows us to easily determine probability that X is in specific ranges.
- Imagine that you precomputed $F_Z(k) = \mathbb{P}(Z \leq k)$ for $k = 0, 1, 2, 3, \dots$. We have:

$$\mathbb{P}(Z \leq 0) = 50.0\% \iff \mathbb{P}(X \leq \mu) = 50.0\%$$

$$\mathbb{P}(Z \leq 1) = 84.1\% \iff \mathbb{P}(X \leq \mu + \sigma) = 84.1\%$$

$$\mathbb{P}(Z \leq 2) = 97.7\% \iff \mathbb{P}(X \leq \mu + 2\sigma) = 97.7\%$$

$$\mathbb{P}(Z \leq 3) = 99.9\% \iff \mathbb{P}(X \leq \mu + 3\sigma) = 99.9\%$$

- i.e., the probability that X is below its mean μ plus one σ is always 84.1%, the probability that it is below μ plus three σ is always 99.9% and so on.



Properties of a Gaussian RV

- Standardization allows us to easily determine *confidence regions* for X .
- Assume a probability level $\alpha \in [0, 1]$; we can show that critical value z satisfying

$$\mathbb{P}(-z \leq Z \leq z) = 1 - \alpha$$

is $z = F_Z^{-1}(1 - \frac{\alpha}{2})$ (we denote this as $z_{1-\frac{\alpha}{2}}$).

- In other words, we have that:

$$\mathbb{P}(-z_{1-\frac{\alpha}{2}} \leq Z \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha$$

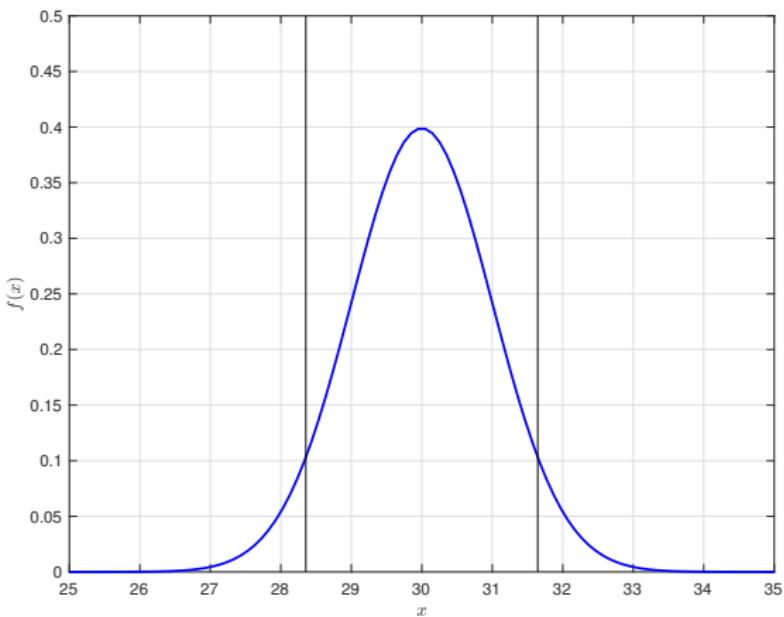
- Using the linear transformation property we obtain:

$$\mathbb{P}(\mu - z_{1-\frac{\alpha}{2}}\sigma \leq X \leq \mu + z_{1-\frac{\alpha}{2}}\sigma) = 1 - \alpha.$$

- i.e.; the probability of finding $X \sim \mathcal{N}(\mu, \sigma^2)$ in region $\mu \pm z_{1-\frac{\alpha}{2}}\sigma$ is $1 - \alpha$.
- This gives an idea of how confident we are of finding X in a given region; conversely, can also determine region under which we can find X with a desired confidence.
- The concept of the confidence region is important in many topics of statistics.

Example: Mixing Problem

- In what interval do we expect flow $F_3 \sim \mathcal{N}(30, 1)$ to be with 90% probability?



- We have $1 - \alpha = 0.9$ and thus $\alpha = 0.1$ and quantile for $Q(1 - \frac{\alpha}{2}) = 1.644$.
- We thus have $F_3 \in [30 - 1.64 \cdot 1, 30 + 1.64 \cdot 1] = [28.36, 31.64]$ with 90% prob.



Model of an Exponential RV

Another RV that often appears in applications is the exponential random variable.

An exponential RV is continuous and has a pdf of the form:

$$f_X(x) = \frac{1}{\beta} e^{-x/\beta}, \quad x \in \mathcal{D}_X = \{0 \leq x \leq \infty\}.$$

- The only hyperparameter of this model is $\beta \in \mathbb{R}_+$ (a.k.a. scale value).
- The reciprocal $\eta = 1/\beta$ is known as the intensity and thus the pdf can also be written as $f_X(x) = \eta e^{-\eta \cdot x}$.
- We express the fact that X is exponential as $X \sim \text{Exp}(\beta)$.



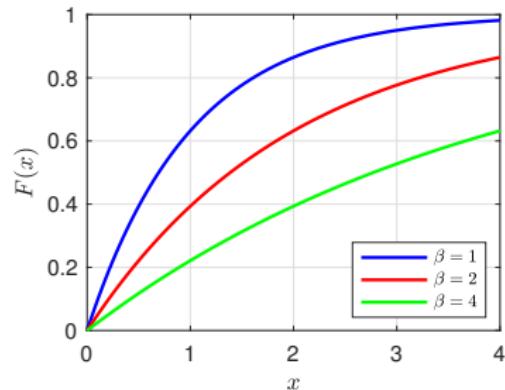
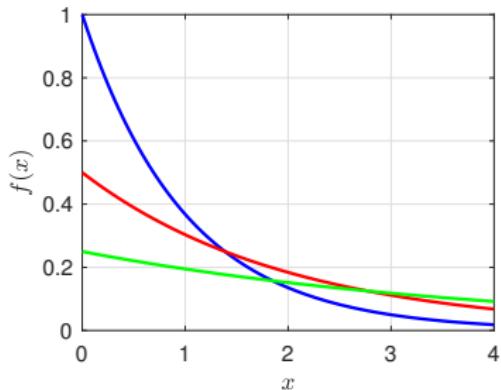
Model of an Exponential RV

- Pdf of the exponential RV tell us that:
 - Probability of finding x away from zero decays exponentially fast at rate η
 - There is zero probability of finding X below zero (pdf is asymmetric). One can think of an exponential RV as one side of a Gaussian RV.
- Cdf is $F_X(x) = 1 - e^{-x/\beta}$, expected value and variance are $\mathbb{E}_X = \beta$ and $\mathbb{V}_X = \beta^2$.
- This RV is often used to model time phenomena associated with *failures*.
- For instance, X can be used to model the amount of time that we have to wait until we observe the first occurrence of an event (e.g., engine fails). In this context, we know the average time that we have to wait ($\mathbb{E}_X = \beta$) but the actual time is random (unknown).



Model of an Exponential RV

- Below are the pdfs and cdfs for $\text{Exp}(\beta)$ for different values of β .

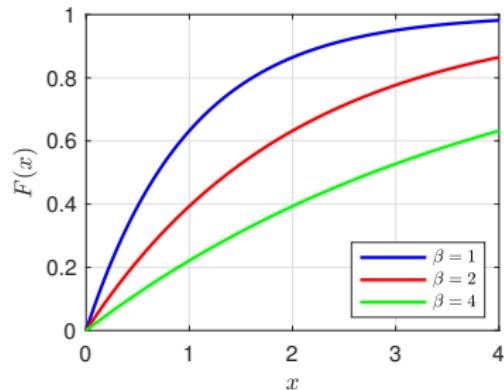
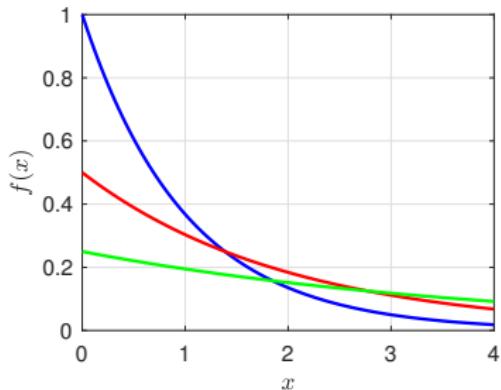


How to determine β from cdf?



Model of an Exponential RV

- Below are the pdfs and cdfs for $\text{Exp}(\beta)$ for different values of β .



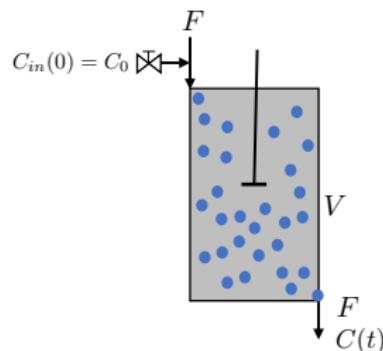
How to determine β from cdf?

- Note that $F_X(x) = 1 - e^{-1} = 0.63$ when $x = \beta$.

Example: Residence Time in Mixing System

- Mixed system with volume V with input and output flow F
- At time $t = 0$ we inject particles so that input concentration is C_0 .
- The particle concentration in system at time t is $C(t)$

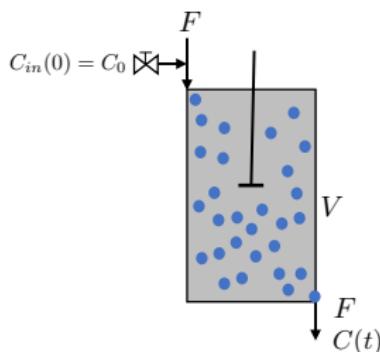
For how long will a particle reside in the system? What factors influence this time?



Example: Residence Time in Mixing System

- Mixed system with volume V with input and output flow F
- At time $t = 0$ we inject particles so that input concentration is C_0 .
- The particle concentration in system at time t is $C(t)$

For how long will a particle reside in the system? What factors influence this time?



- A material balance for the system reveals that:

$$f_T(t) = \frac{C(t)}{C_0} = \frac{1}{\tau} e^{-t/\tau} \quad \text{with } \tau = V/F \quad (1)$$

- Interpret T as time required for particle to exit system (residence time)
- Pdf $f_T(t)$ is interpreted as fraction of particles exiting at time $T = t$



Example: Residence Time in Mixing System

$$f_T(t) = \frac{C(t)}{C_0} = \frac{1}{\tau} e^{-t/\tau} \text{ with } \tau = V/F \quad (2)$$

- Balance suggests that $T \sim \text{Exp}(\tau)$
- We thus have mean of residence time is $\mathbb{E}[T] = \tau = V/F$ and variance is $\mathbb{V}[T] = (V/F)^2$. What is effect of V and F ?
- Fraction of particles that have exited up to time t is $\mathbb{P}(T \leq t) = F_T(t)$.
- Note that $F_T(t) = \int_0^t (1/\tau) e^{-t/\tau} dt = (1 - e^{-t/\tau})$ and thus $F(0) = 0$ and $F(\infty) = 1$.
- Fraction of particles that are still in the system at time t is $\mathbb{P}(T > t) = 1 - F_T(t)$.
This function is known as the survival function.



Model of a Gamma RV

The Gamma RV is a generalization of the exponential RV that has a pdf of the form:

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha-1}, \quad x \in \mathcal{D}_X = \{0 \leq x \leq \infty\}.$$

- The pdf has two hyperparameters $\alpha, \beta \in \mathbb{R}_+$
- $\Gamma(\alpha)$ is the gamma function (for integer $\alpha \geq 1$ we have $\Gamma(\alpha) = (\alpha - 1)!$).
- We express the fact that X is a gamma RV as $X \sim \text{Gamma}(\alpha, \beta)$

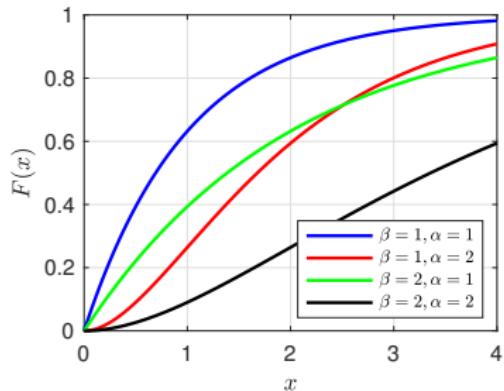
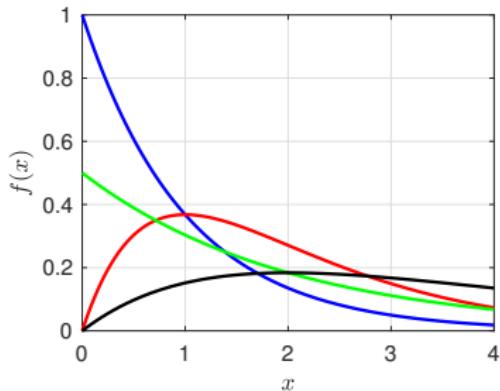


Model of a Gamma RV

- The pdf tells us that one recovers an exponential RV when $\alpha = 1$.
- The term $x^{\alpha-1}$ introduces a competing (opposite) effect for the exponential decay, giving rise to peak in the pdf. The location of this peak is $x^* = (\alpha - 1)\beta$ and is the mode of the RV (point that maximizes $\mathbb{P}(X = x)$).
- The expected value and variance are $\mathbb{E}_X = \alpha\beta$ and $\mathbb{V}_X = \alpha\beta^2$ (i.e., the hyperparameters can be estimated from data by solving a set of two equations).
- In the context of time phenomena, this RV generalizes the exponential in that it models the amount of time that we have to wait until we observe the α -th occurrence of an event. Consequently, $\alpha=1$ means the first event (as in the exponential RV).
- This RV has applications not only in temporal but also in spatial phenomena. For instance, it can be used to model the distance until we find the α -th occurrence of certain type of atom in a molecule.

Model of an Gamma RV

- Below are the pdfs and cdfs for $\text{Gamma}(\alpha, \beta)$ for different values of α, β .



- Note same as exponential for $\alpha = 1$.
- Note emergence of peaks due to competing effects for $\alpha = 2$.



Model of a Chi-Squared RV

The Chi-Square RV has a pdf of the form:

$$f_X(x) = \frac{1}{2^{r/2}\Gamma(r/2)} e^{-x/2} x^{r/2-1}, \quad x \in \mathcal{D}_X = \{0 \leq x \leq \infty\}.$$

- The pdf has only one hyperparameter $r \in \mathbb{Z}_+$ (a.k.a degrees of freedom).
- We express the fact that X is a gamma RV as $X \sim \chi^2(r)$
- Note that this is a Gamma RV with $\beta = 2$ and $\alpha = r/2$ (for a positive integer r).
- The expected value and variance can be derived from those of the Gamma RV.
- A crucial property of a Chi-Squared RV is that it is related to the standard normal RV. In particular, one can show that:

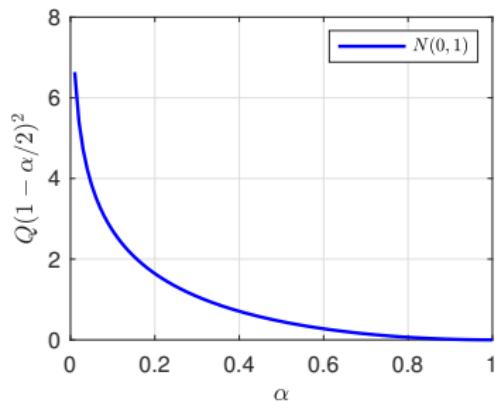
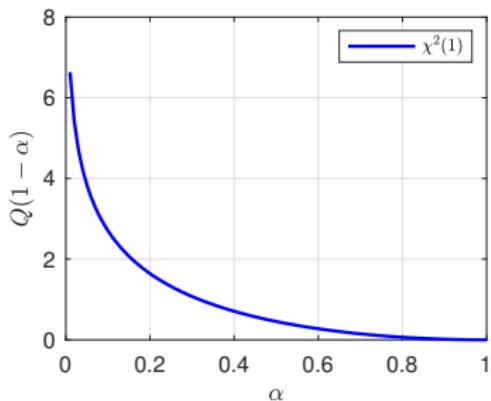
$$\sum_{i=1}^r X_i^2 \sim \chi^2(r)$$

if $X_i \sim \mathcal{N}(0, 1)$. This property will be useful later on.

- From relationship with $\mathcal{N}(0, 1)$ we can show that $\mathbb{Q}(1 - \alpha)$ of $\chi^2(1)$ is equal to $\mathbb{Q}(1 - \alpha/2)^2$ of $\mathcal{N}(0, 1)$. This is because $\mathbb{P}(-z \leq Z \leq z) = \mathbb{P}(Z^2 \leq z)$.

Model of an Chi-Squared RV

- Compare quantile functions of $\mathbb{Q}(1 - \alpha)$ of $\chi^2(1)$ and $\mathbb{Q}(1 - \alpha/2)^2$ of $\mathcal{N}(0, 1)$.





Model of a Weibull RV

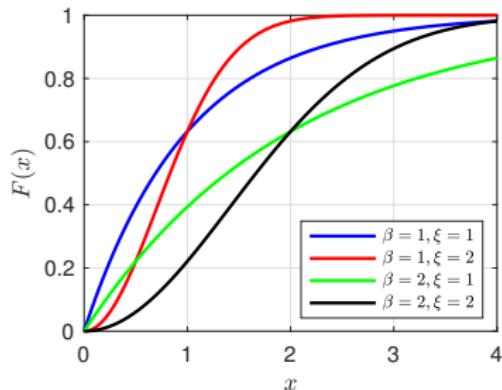
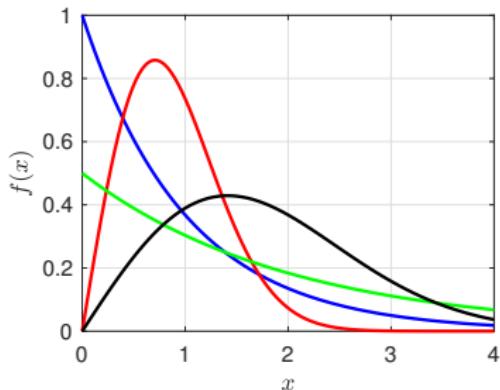
The Weibull RV is a generalization of the exponential RV that has a pdf:

$$f_X(x) = \frac{\xi}{\beta} \left(\frac{x}{\beta}\right)^{\xi-1} \exp\left[-\left(\frac{x}{\beta}\right)^\xi\right], \quad x \in \mathcal{D}_X = \{0 \leq x \leq \infty\}.$$

- The pdf has two hyperparameters $\xi, \beta \in \mathbb{R}_+$ (a.k.a scale and shape).
- We express the fact that X is a gamma RV as $X \sim \text{Weibull}(\xi, \beta)$.
- Note that one recovers an exponential RV when $\xi = 1$.
- The expected value and variance are $\mathbb{E}_X = \beta\Gamma(1 + 1/\xi)$ and $\mathbb{V}_X = \beta^2 (\Gamma(1 + 2/\xi) + \Gamma(1 + 1/\xi)^2)$. The dependence on the gamma function makes it difficult to estimate ξ, β from these relationships.
- cdf has a nice form $F_X(x) = 1 - e^{(-x/\beta)^\xi}$ (ξ, β are often inferred from the cdf). Note that $\mathbb{P}(X \leq \beta) = 0.632$ for any ξ .
- Weibull RV is the *de facto* model used in failure analysis and was discovered by Fischer and Tippett.
- Turns out that, as in the case of the Gaussian RV, many phenomena have the Weibull RV as a limiting case (we will show this later).

Model of Weibull RV

- Below at the pdfs and cdfs for $\text{Weibull}(\beta, \xi)$ for different values of β, ξ .



- Note same as exponential for $\xi = 1$.
- Note emergence of peaks due to competing effects for $\xi = 2$.
- Note $\mathbb{P}(X \leq \beta) = 0.632$ for any ξ (this differentiates it from Gamma RV).



Families of Random Variables

- The exponential, Gamma, Chi-Squared, and Weibull RVs are interrelated. In fact, these are captured by the generalized gamma model with pdf:

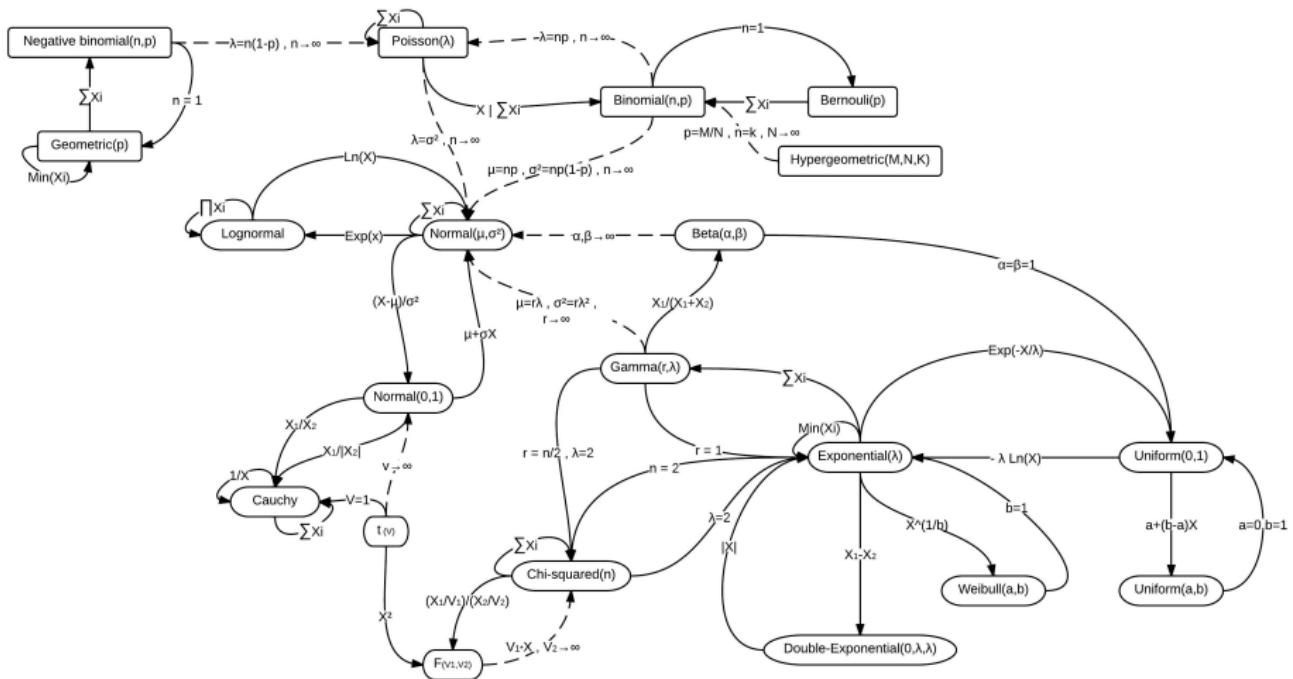
$$F_X(x) = \frac{1}{\beta^{\alpha\xi}\Gamma(\alpha)} \exp\left[-\left(\frac{x-\delta}{\beta}\right)^{\xi}\right] \xi(x-\delta)^{\alpha\xi-1}, \quad x \in \mathcal{D}_X = \{0 \leq x \leq \infty\}.$$

- These RVs are known as the Gamma family.
- There are three major families of continuous RVs:
 - Gaussian family (includes Gaussian, LogNormal, and Rayleigh)
 - Gamma family (includes exponential, Gamma, Chi-Squared, and Weibull)
 - Ratio family (includes Cauchy, Uniform, Beta, Fisher, Student)
- There is one family for discrete RVs, which includes Uniform (discrete), Bernoulli, Binomial, and Poisson.
- Each family models different type of phenomena and there exist relationships between RVs within families and across families.

A detailed discussion of the modeling properties of RVs is beyond our scope. Here, we have only discussed the RVs that will be relevant in our subsequent discussion.



Families of Random Variables





Outline

- ① Introduction
- ② Models of Random Variables
- ③ Estimation Techniques
- ④ Multivariate Statistics
- ⑤ Data-Driven Modeling
- ⑥ Statistical Learning
- ⑦ Decision-Making Under Uncertainty



Estimation Techniques

- Now that we have a basic idea of the types of RV models available we proceed to develop procedures to estimate an RV model from available data.
- We will define an RV model in terms of its $f_X(x|\theta)$ (or $F_X(x|\theta)$), where θ are the hyperparameters.
- By estimating an RV model we mean that we seek to find θ that best fits the data.

We will explore a couple of estimation methods:

- Point Estimation (Method of Moments and Least-Squares Method)
 - Maximum Likelihood Estimation
-
- Our first step will be to explore our data and postulate a specific RV model (e.g., Gaussian or Exponential) based on any patterns exposed by the data.
 - Our second step would be to tune θ for the given postulated RV model to see if this fits the data satisfactorily. If the fit is not adequate, we postulate another model.



Method of Moments

- Recall the moments of X (with pdf $f_X(x|\theta)$ and hyperparameters θ) are given by:

$$m_k(\theta) := \mathbb{E}[(X - \mathbb{E}[X])^k], \quad k = 1, 2, \dots, N$$

- Here, we highlight the dependence of the moments on the hyperparameters θ .
- Method of moments uses data $x_\omega, \omega \in \mathcal{S}$ to obtain sample approximations:

$$M_k = \frac{1}{S} \sum_{\omega \in \mathcal{S}} (x_\omega - m)^k, \quad k = 1, 2, \dots, N$$

where $m = \frac{1}{S} \sum_{\omega \in \mathcal{S}} x_\omega$ is the sample mean.



Method of Moments

- Our objective is to find the hyperparameters θ that solve the matching equations:

$$m_k(\theta) = M_k, \quad k = 1, 2, \dots, N$$

- In other words, we want to find θ that matches model and data moments.
- A solution to the equations can also be found by solving the minimization problem:

$$\min_{\theta} \frac{1}{2} \sum_{k=1}^N (m_k(\theta) - M_k)^2.$$

This problem is known as a least-squares minimization problem (seeks to minimize the discrepancy between the model and data moments).

- Least-squares is preferred when the matching equations do not have a solution. In this case, the minimization problem finds θ that is most compatible with the data.



Example: Method of Moments

We illustrate how to estimate $\theta = (\mu, \sigma^2)$ using moment matching for $\mathcal{N}(0, 1)$.

- We are given observations $x_\omega, \omega \in \mathcal{S}$.
- Sample mean is $\hat{E}_X = \frac{1}{S} \sum_{s \in \mathcal{S}} x_\omega$ and variance $\hat{V}_X = \frac{1}{S} \sum_{s \in \mathcal{S}} (x_\omega - \hat{E}_X)^2$.
- We have that $m_1(\theta) = \mathbb{E}[X - \mathbb{E}_X] = \mathbb{E}_X - \mu = 0$ and thus we estimate $\mu = \hat{\mathbb{E}}_X$.
- We have that $m_2(\theta) = \mathbb{V}_X = \sigma^2$ and thus we estimate $\sigma^2 = \hat{\mathbb{V}}_X$

Now consider we wish to estimate $\theta = (\xi, \beta)$ using moment matching for Weibull(ξ, β).

- We are given observations $x_\omega, \omega \in \mathcal{S}$.
- Sample mean is $\hat{E}_X = \frac{1}{S} \sum_{s \in \mathcal{S}} x_\omega$ and variance $\hat{V}_X = \frac{1}{S} \sum_{s \in \mathcal{S}} (x_\omega - \hat{E}_X)^2$.
- We find β, ξ that match $m_1(\theta)$ and $m_2(\theta)$ by solving the nonlinear equations:

$$\hat{\mathbb{E}}_X = \beta \Gamma(1 + 1/\xi)$$

$$\hat{\mathbb{V}}_X = \beta^2 \left(\Gamma(1 + 2/\xi) + \Gamma(1 + 1/\xi)^2 \right)$$

- This is computationally challenging due to Γ function. Is there another way?



Least-Squares Method

- Functional form of moments $m_k(\theta)$ might be too complex for some RVs (e.g., Weibull).
- In such cases, we can use $F_X(t|\theta)$ to find the hyperparameters. For example, for Weibull we have $F_X(t|\theta) = (1 - e^{-(t/\beta)^\xi})$ with $\theta = (\xi, \beta)$.
- In least-squares method, we find θ that best matches the empirical cdf (obtained from data x_ω , $\omega \in \mathcal{S}$).
- Here, we propose a set of threshold values t_k , $k = 1, 2, \dots, N$ and compute:

$$\hat{F}_k = \frac{1}{S} \sum_{\omega \in \mathcal{S}} \mathbf{1}[x_\omega \leq t_k], \quad k = 1, 2, \dots, N$$

- We use these approximations to solve the least-squares minimization problem:

$$\min_{\theta} \frac{1}{2} \sum_{k=1}^N (F_X(t_k|\theta) - \hat{F}_k)^2$$

- When cdf has exponential form, it is convenient to use log transformations such as:

$$\min_{\theta} \frac{1}{2} \sum_{k=1}^N (\log F_X(t_k|\theta) - \log \hat{F}_k)^2$$

- Least-squares method is general and can be used to match other statistics of the RV available such as empirical quantiles, and empirical pdfs.



Example: Weibull using Least-Squares

Consider we wish to estimate $\theta = (\xi, \beta)$ using least-squares for Weibull(ξ, β).

- We are given observations x_ω , $\omega \in \mathcal{S}$ and obtain empirical cdfs $\hat{F}(t_k)$ for given t_k , $k = 1, \dots, N$.
- Recall cdf of Weibull is $F_X(t|\theta) = (1 - e^{-(t/\beta)^\xi})$ and notice that:

$$\log(1 - F_X(t|\theta)) = -(t/\beta)^\xi$$

Moreover, note that:

$$\log(-\log(1 - F_X(t|\theta))) = \xi \log t - \xi \log \beta$$

- We have $\log t = a \log(-\log(1 - F_X(t|\theta))) + b$ with $a = 1/xi$ and $b = \log \beta$.
- We want to find a, b that best match equations
 $\log t_k = a \cdot \log(-\log(1 - \hat{F}(t_k))) + b$. This is done by solving:

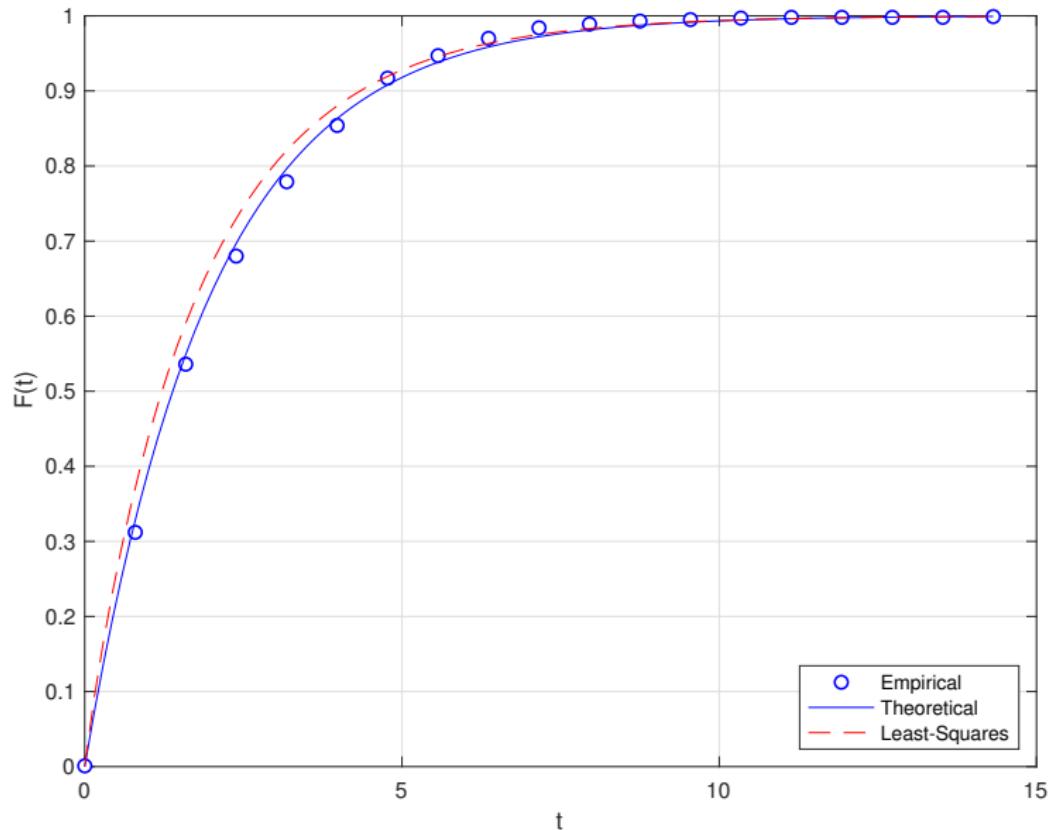
$$\min_{a,b} \sum_{k=1}^N (y_k - (a \cdot x_k + b))^2$$

where $x_k = \log(-\log(1 - \hat{F}(t_k)))$ and $y_k = \log t_k$.

- This is a simple linear optimization problem. We will see later one how to solve this.



Example: Weibull using Least-Squares





Maximum Likelihood Method

The maximum likelihood estimation (MLE) method proceeds as follows:

- Assume we have observations x_ω , $\omega \in \mathcal{S}$ (selected at random).
- We postulate $f_X(x|\theta)$ for the RV X and recall that $f(x_\omega|\theta)$ is the probability (likelihood) that X takes the value of observation x_ω .
- Consequently, we find hyperparameters θ that maximize *joint* probability that X takes observations x_ω , $\omega \in \mathcal{S}$. This is done by solving the maximization problem:

$$\max_{\theta} L(\theta) = \prod_{\omega \in \mathcal{S}} f(x_\omega|\theta).$$

Here, $L(\theta)$ is known as the likelihood function.

- It is often convenient to solve the equivalent problem:

$$\max_{\theta} \log L(\theta) = \sum_{\omega \in \mathcal{S}} \log f(x_\omega|\theta).$$

This problem can be solved by hand when the pdf is simple but requires numerical techniques when complex.

- We will soon discuss what “random observations” and “joint probability” mean.



Example: Maximum Likelihood for Exponential RV

- Want to estimate β for $\text{Exp}(\beta)$ using observations x_{ω} .
- Pdf of the exponential is $f(x|\beta) = \frac{1}{\beta}e^{-x/\beta}$ and thus likelihood function:

$$\begin{aligned}L(\beta) &= \left(\frac{1}{\beta} e^{-x_1/\beta} \right) \left(\frac{1}{\beta} e^{-x_2/\beta} \right) \cdots \left(\frac{1}{\beta} e^{-x_S/\beta} \right) \\&= \frac{1}{\beta^S} \exp \left(-\frac{1}{\beta} \sum_{\omega=1}^S x_{\omega} \right).\end{aligned}$$

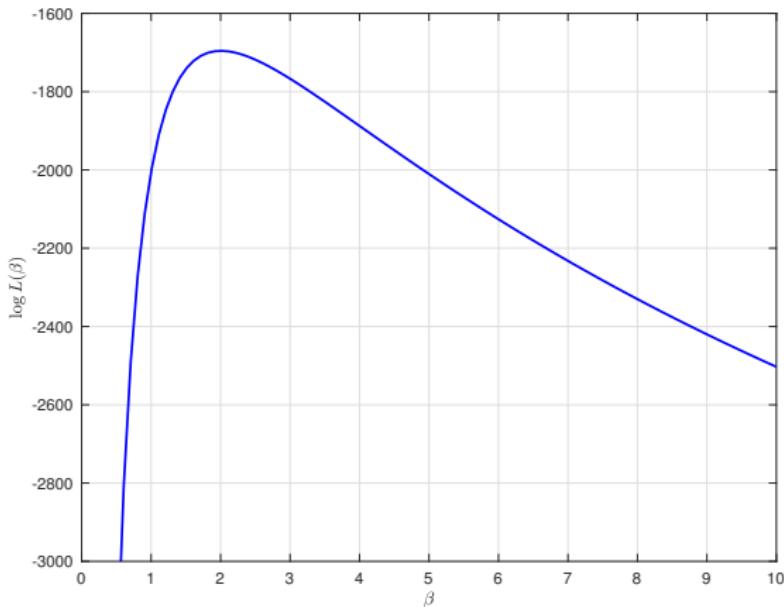
- We find β that maximizes log likelihood $\log L(\beta)$:

$$\max_{\beta} \log L(\beta) = -S \log \beta - \frac{1}{\beta} \sum_{\omega=1}^S x_{\omega}$$

- The solution of this problem indicates that $\hat{\beta} = \frac{1}{S} \sum_{\omega=1}^S x_{\omega}$.
- Best estimate $\hat{\beta}$ is sample mean (which makes sense because $\mathbb{E}[X] = \beta$).

Example: Maximum Likelihood for Exponential RV

- Here we present a plot for $\log L(\beta)$ using data x_ω generated from $\text{Exp}(2)$.
- Value of β that maximizes $\log L(\beta)$ ($\hat{\beta}$) coincides with true value $\beta = 2$.



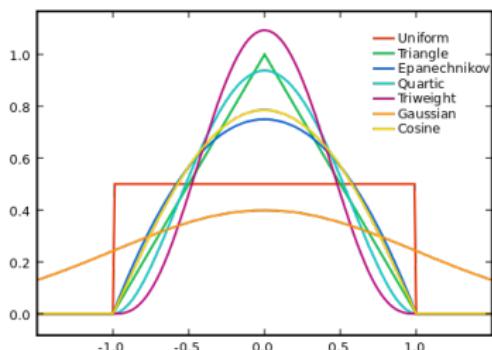
Kernel-Based Techniques

- If data does not fit a known RV model, one can resort to determine the pdf and cdf numerically by using kernel-based (functional) techniques.
- Kernel-based techniques use data points x_ω , $\omega \in \mathcal{S}$ as support points to create a functional approximation:

$$\hat{f}_h(x) = \frac{1}{S} \sum_{\omega=1}^S K_h(x - x_\omega)$$

where h is a tuning parameter known as the bandwidth.

- Here are typical kernel functions:



- We will explore later the principles behind kernel methods.



Sampling and Asymptotic Properties

So far, we have assumed that we have data x_ω , $\omega \in \mathcal{S}$ but we have not discussed yet how this is being collected and generated. Moreover, we want to know how our empirical approximations behave as we keep accumulating data.

- The data sample sequence $x_\omega \in \mathcal{S}$ is a set of observations of X collected from a statistical population Ω by a defined procedure; i.e., sampling is a data collection procedure.
- A data sample sequence $x_\omega \in \mathcal{S}$ is called random if each sample x_ω is drawn from the same underlying pdf $f_X(x)$ and if it is drawn independently from the others. In other words, the samples are independent and identically distributed (i.i.d.).
- If a sample x_ω is selected at random, the sample itself is an RV. Consequently, sometimes we denote the data sample sequence as a sequence of RVs $X_\omega \in \mathcal{S}$.

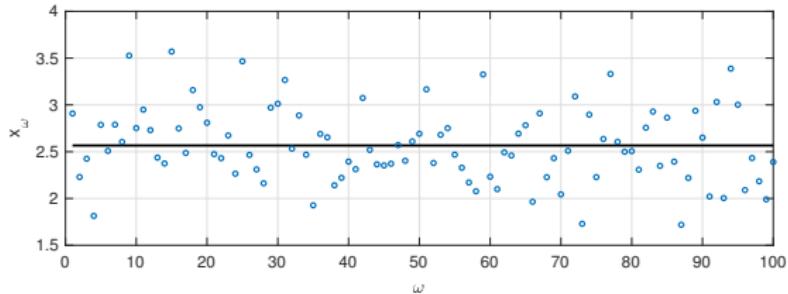
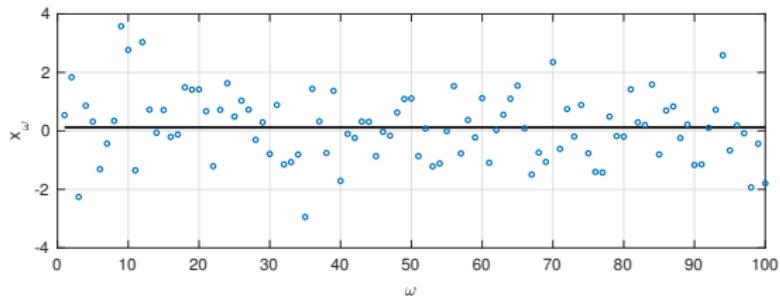


Sampling and Asymptotic Properties

- Random sample X_ω has same probability $1/S$ of being selected and is *unbiased*.
- The lack of bias indicates that $\mathbb{E}[X_\omega] = \mathbb{E}[X]$ (drawing sample many times and averaging the results gives same expected value of the actual RV X).
- Random samples can be used to construct approximation techniques with rather striking asymptotic properties, as we will show next.
- Collecting data at random is not as easy as it sounds, one must ensure that there is no bias in selecting a sample (i.e., there is no hidden mechanism). As humans, it is strikingly difficult to pick something randomly.

Sampling and Asymptotic Properties

- Below we show a random and nonrandom sample sequence for $\mathcal{N}(0, 1)$ and their corresponding long-term averages.
- For nonrandom, we select a particular element of the sample (this introduces a bias).





Monte Carlo Approximations

Monte Carlo (MC) is a set of computational techniques that use *random samples* to infer properties of X and derived quantities (e.g., summarizing statistics). For example, we can use the random sample $x_\omega \in \mathcal{S}$ to compute the sample approximations:

- Expectation of X : $\hat{E}_X^S := \frac{1}{S} \sum_{\omega \in \mathcal{S}} x_\omega \approx \mathbb{E}[X]$
- Variance of X : $\hat{V}_X^S := \frac{1}{S} \sum_{\omega \in \mathcal{S}} (x_\omega - \hat{E}_X^S)^2 \approx \mathbb{V}[X]$
- CDF of X : $\hat{F}_X^S(x) := \frac{1}{S} \sum_{\omega \in \mathcal{S}} \mathbf{1}[x_\omega \leq x] \approx F_X(x) = \mathbb{E}[\mathbf{1}[X \leq x]]$
- Expectation of system $\hat{E}_\varphi^S := \frac{1}{S} \sum_{\omega \in \mathcal{S}} \varphi(x_\omega, u) \approx \mathbb{E}[\varphi(X, u)]$.

Natural questions that emerge here are:

- Do the approximations become exact as $S \rightarrow \infty$?
- How accurate are these approximations for finite S ?
- Most approximations use an expectation function and we thus restrict our discussion to the behavior of the approximation \hat{E}_X^S .
- There are approximation techniques that use systematic (biased) sampling methods (data is not collected at random).



Law of Large Numbers

Lets first assess the quality of the MC approximation as $S \rightarrow \infty$.

- Consider a random sample sequence X_1, X_2, \dots, X_S for RV X . Since samples are identically distributed, they have same underlying pdf with expected value $\mathbb{E}[X]$.
- This implies that $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \dots = \mathbb{E}[X_S] = \mathbb{E}[X]$.
- Consider the MC approximation of $\mathbb{E}[X]$:

$$\hat{\mathbb{E}}_X^S = \frac{1}{S} \sum_{\omega \in \mathcal{S}} X_\omega$$

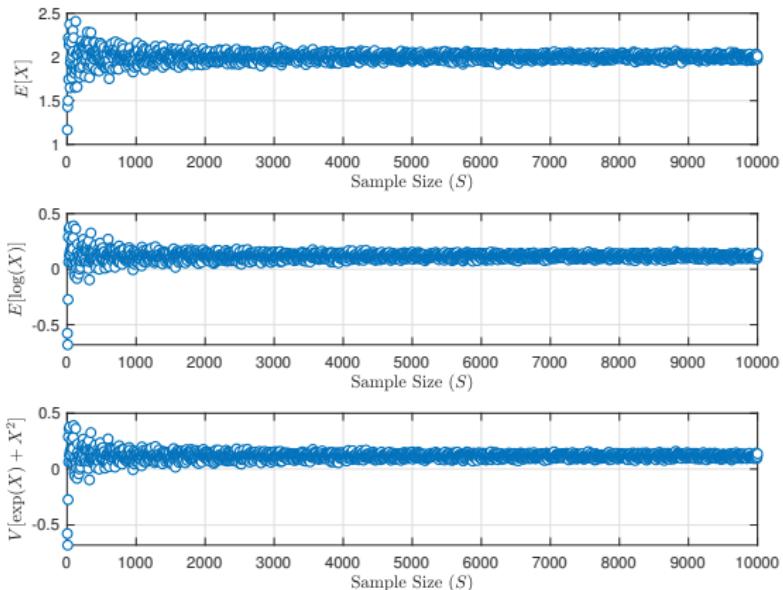
The law of large numbers (LLN) states that:

$$\lim_{S \rightarrow \infty} \hat{\mathbb{E}}_X^S = \mathbb{E}[X]$$

- The LLN is a fundamental result in statistics and is important because it guarantees *stable long-run* behavior of random variables.
- In other words, if the process is truly random, samples fluctuate around $\mathbb{E}[X]$ and average out. Conversely, if there is a systematic bias, the fluctuations will accumulate and the process will drift.
- The LLN implies that Monte Carlo approximations become asymptotically exact as $S \rightarrow \infty$. Importantly, the result holds for any random variable (e.g., discrete, continuous, univariate, multivariate).

Example: Monte Carlo Approximations

- Consider random samples x_ω , $\omega \in \mathcal{S}$ obtained from $X \sim \text{Weibull}(2, 1)$
- MC approximations for $\mathbb{E}[X]$, $\mathbb{E}[\log(X)]$ and $\mathbb{V}[\exp(X) + X^2]$ for different S .





Central Limit Theorem

lets turn our attention to the issue of assessing quality of the MC approximation for finite sample size S .

- This can be addressed by using a powerful result in statistics known as the central limit theorem (CLT).
- Consider that the random sample sequence X_1, X_2, \dots, X_S is i.i.d and has *known* expected value $e = \mathbb{E}[X]$ and variance $v^2 = \mathbb{V}[X]$.
- We know that X_ω are RVs and so is $\hat{\mathbb{E}}_X^S$.

The CLT will answer the following questions:

- What is the distribution of $\hat{\mathbb{E}}_X^S$? If we know it, we can say something about how much variability the MC approximation has.
- What is the distribution of $\hat{\mathbb{E}}_X^S$ as $S \rightarrow \infty$?



Central Limit Theorem

The CLT states that:

$$\lim_{S \rightarrow \infty} \hat{\mathbb{E}}_X^S \sim \mathcal{N}(e, v/\sqrt{S})$$

This result is one of the most surprising and useful results in statistics. Let's explain why:

- CLT says that, *regardless* of the underlying nature of X (e.g., Uniform, Weibull, Exponential), its sample average approximation $\hat{\mathbb{E}}_X^S$ will *always* become a Gaussian RV as S increases.
- CLT also says that the variance of the Gaussian $\hat{\mathbb{E}}_X^S \sim \mathcal{N}(e, v/\sqrt{S})$ shrinks with S . In other words, $\hat{\mathbb{E}}_X^S$ becomes more certain as S increases.
- Implication is that, since we have a cdf and pdf for $\hat{\mathbb{E}}_X^S$, we can compute all quantities of interest for it (e.g., confidence regions):

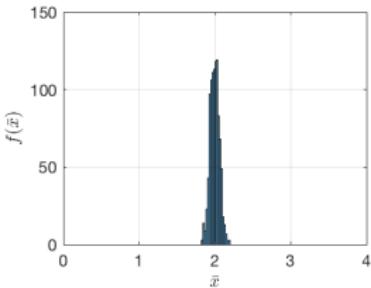
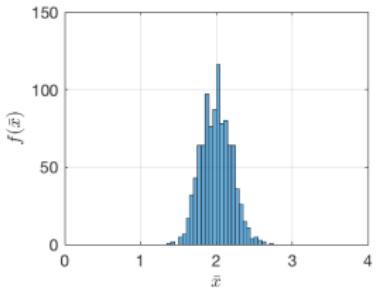
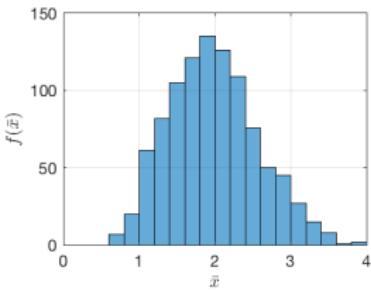
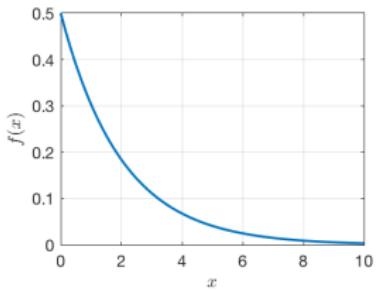
$$\mathbb{P}\left(e - z_{1-\alpha/2}v/\sqrt{S} \leq \hat{\mathbb{E}}_X^S \leq e + z_{1-\alpha/2}v/\sqrt{S}\right) = 1 - \alpha.$$

Consequently, for given S , we can know how confident we are that the approximation $\hat{\mathbb{E}}_X^S$ is in a region.



Example: Central Limit Theorem

- We sample X_1, X_2, \dots, X_S obtained from $X \sim \text{Weibull}(2, 1)$.
- Want distribution of $\bar{X} = \frac{1}{S} \sum_{\omega=1}^S X_\omega$ for different values of S .
- Below we show pdf of Weibull(2, 1) and of \bar{X} for $S = 10, 100, 1000$.





Extreme Value Distribution

- In CLT we are interested in *sample average* $\hat{\mathbb{E}}_X^S = \frac{1}{S} \sum_{\omega \in S} X_\omega$.
- But what if we are interested in a different statistic? For instance, the *sample max*:

$$X_{max}^S = \max\{X_1, X_2, \dots, X_S\}$$

- There exists a result (analogous to CLT) that characterizes pdf of X_{max}^S as $S \rightarrow \infty$. The result is known as the extreme value theorem (EVT).
- Consider, as before, a random sample sequence X_1, X_2, \dots, X_S for RV X .

The EVT states that:

$$\lim_{S \rightarrow \infty} X_{max}^S \sim GEV(a, b, c)$$

where $GEV(a, b, c)$ is the generalized extreme value RV with hyperparameters a, b, c .



Extreme Value Distribution

The generalized extreme value RV has a pdf of the form:

$$f(s; \xi) = \begin{cases} (1 + cs)^{(-1/c)-1} \exp(-(1 + cs)^{-1/c}) & c \neq 0 \\ \exp(-s) \exp(-\exp(-s)) & c = 0 \end{cases}$$

and a cdf of the form:

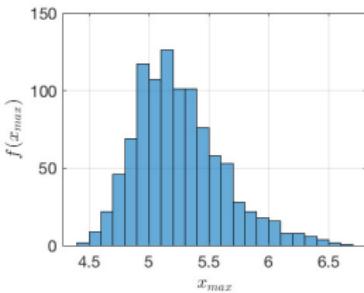
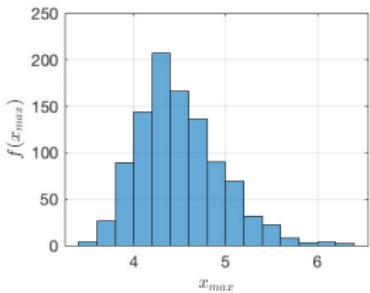
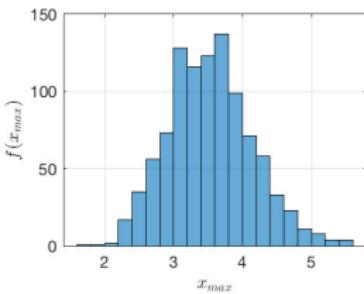
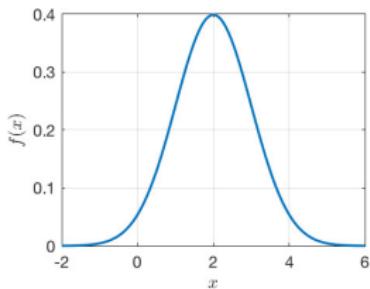
$$F_X(x) = \begin{cases} \exp(-(1 + cs)^{-1/c}) & c \neq 0 \\ \exp(-\exp(-s)) & c = 0 \end{cases}$$

where $s = (x - a)/b$ is a standard variable.

- GEV RV is a generalization that includes Weibull (for $c < 0$), Frechet (for $c > 0$), and Gumbel (for $c = 0$) RVs.
- GEV RV is widely used in failure analysis because max operator characterizes peak (extreme) events.
- As with CLT, EVT does not depend on the underlying nature of X .

Example: Extreme Value Theorem

- We sample X_1, X_2, \dots, X_S obtained from $X \sim \mathcal{N}(2, 1)$.
- Want distribution of $X_{max} = \max\{X_1, X_2, \dots, X_S\}$ for different values of S .
- Below we show pdf of $X \sim \mathcal{N}(2, 1)$. and of \bar{X} for $S = 10, 100, 1000$.





Outline

- ① Introduction
- ② Models of Random Variables
- ③ Estimation Techniques
- ④ Multivariate Statistics**
- ⑤ Data-Driven Modeling
- ⑥ Statistical Learning
- ⑦ Decision-Making Under Uncertainty



Multivariate Statistics

- So far, we have assumed that the RV X is univariate and thus has observations that are scalar values ($x_\omega \in \mathbb{R}$).
- We have also informally mentioned the concept of independent and joint RVs in the context of estimation and sampling. What do these mean?
- Consider a multivariate RV $X = (X_1, X_2, \dots, X_n)$ with observations $x_\omega = (x_{\omega,1}, x_{\omega,2}, \dots, x_{\omega,n}) \in \mathbb{R}^n$; e.g., consider input-output pair $(X, \varphi(X))$ discussed in uncertainty propagation.

Questions that we are interested in answering are:

- Are there any connections between RVs? Is there a pattern that suggests they vary together? Are they independent of one another?
- If there are connected, how strong are connections?
- If there are connected, how does knowledge of one affects uncertainty of the other?
- How to analyze connections between many RVs? (e.g., n is in the hundreds)
- How to generalize results from univariate case to multivariate case?



Joint PDFs and CDFs

For simplicity, we consider bivariate RV $X = (X_1, X_2)$. The concepts presented easily generalize to higher dimensions.

- Observation $\omega \in \Omega$ of RV X generates observation pair $x_\omega = (x_{\omega,1}, x_{\omega,2})$
- We assume that the domain of X is a 2-D box:

$$\mathcal{D}_1 = \{-\infty \leq x_1 \leq \infty\}$$

$$\mathcal{D}_2 = \{-\infty \leq x_2 \leq \infty\}$$

$$\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2 = \{-\infty \leq x_1 \leq \infty, -\infty \leq x_2 \leq \infty\}$$

- The *joint* pdf and cdf of multivariate RV are:

$$f(x_1, x_2) = \mathbb{P}(X_1 = x_1 \& X_2 = x_2), (x_1, x_2) \in \mathcal{D}$$

$$F(x_1, x_2) = \mathbb{P}(X_1 \leq x_1 \& X_2 \leq x_2), (x_1, x_2) \in \mathcal{D}$$

- Note sign $\&$ inside the measure. For example, $f(x_1, x_2)$ is probability of event in which X_1 takes value x_1 *and* X_2 takes value x_2 .
- The pdf must satisfy $f(x_1, x_2) \geq 0, (x_1, x_2) \in \mathcal{D}$.



Joint PDFs and CDFs

Consider a subdomain $\mathcal{A} \subseteq \mathcal{D}$ of the form:

$$\mathcal{A} = \{a_1 \leq x_1 \leq b_1 \& a_2 \leq x_2 \leq b_2\}$$

- For discrete RV we have that pdf is discontinuous and:

$$\begin{aligned}\mathbb{P}(X \in \mathcal{A}) &= \sum_{(x_1, x_2) \in \mathcal{A}} f(x_1, x_2) \\ &= \sum_{\omega \in \Omega} f(x_1, x_2) \mathbf{1}[(x_{\omega,1}, x_{\omega,2}) \in \mathcal{A}].\end{aligned}$$

- For continuous RV we have that pdf is continuous and:

$$\begin{aligned}\mathbb{P}(X \in \mathcal{A}) &= \int_{x \in \mathcal{A}} f(x) dx \\ &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} f(x_1, x_2) dx_1 dx_2 \\ &= \int_{a_1}^{b_1} \int_{a_2}^{b_2} dF(x_1, x_2)\end{aligned}$$

The last expression implies that $\frac{dF(x_1, x_2)}{dx_1 dx_2} = f(x_1, x_2)$.



Example: Control System Reliability

- Reliability of control system depends on lifetime of processor X_1 and actuator X_2 (if one fails, entire system fails).
- Probability that that lifetime of $X_1 = x_1$ and $X_2 = x_2$ is given by joint pdf of $X = (X_1, X_2)$:

$$f(x_1, x_2) = \frac{1}{50} e^{-(0.2x_1 + 0.1x_2)}$$

- What is probability that system lasts more than 2 years?

$$\mathbb{P}(X_1 \geq 2 \& X_2 \geq 2) = \int_2^{\infty} \int_2^{\infty} \frac{1}{50} e^{-(0.2x_1 + 0.1x_2)} dx_1 dx_2 = 0.549$$

- What is probability that processor lasts more than 5 years and actuator lasts more than 10 years?

$$\mathbb{P}(X_1 \geq 5 \& X_2 \geq 10) = \int_5^{\infty} \int_{10}^{\infty} \frac{1}{50} e^{-(0.2x_1 + 0.1x_2)} dx_1 dx_2 = 0.135$$



Conditional PDFs and CDFs

The joint pdf $f(x_1, x_2)$ tells us probability that (X_1, X_2) takes value (x_1, x_2) .

- Imagine now we want to know probability that X_1 takes x_1 given knowledge that X_2 takes value x_2 (or other way around).
- These probabilities are obtained from *conditional pdfs*:

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}, \quad x_1 \in \mathcal{D}_1$$

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}, \quad x_2 \in \mathcal{D}_2$$

- Conditionals represent $f(x_1|x_2) = \mathbb{P}(X_1 = x_1|X_2 = x_2)$ and $f(x_2|x_1) = \mathbb{P}(X_2 = x_2|X_1 = x_1)$.
- These expressions can also be written as:

$$f_2(x_2)f(x_1|x_2) = f(x_1, x_2), \quad x_1 \in \mathcal{D}_1$$

$$f_1(x_1)f(x_2|x_1) = f(x_1, x_2), \quad x_2 \in \mathcal{D}_2$$

- As an example, consider we want $\mathbb{P}(a_1 \leq X_1 \leq b_1|X_2 = x_2)$; we have that:

$$\mathbb{P}(a_1 \leq X_1 \leq b_1|X_2 = x_2) = \int_{a_1}^{b_1} f(x_1|x_2)dx_1$$

- Joint pdfs have associated marginal cdfs $F(x_1|x_2)$ and $F(x_2|x_1)$.



Marginal PDFs

- Imagine now we want to know probability that X_1 takes x_1 *regardless* of what value X_2 takes (or other way around).
- These probabilities are obtained from marginal pdfs. For a continuous RV:

$$f_1(x_1) = \int_{x_2 \in \mathcal{D}_2} f(x_1, x_2) dx_2, \quad x_1 \in \mathcal{D}_1$$

$$f_2(x_2) = \int_{x_1 \in \mathcal{D}_1} f(x_1, x_2) dx_1, \quad x_2 \in \mathcal{D}_2$$

i.e., marginal pdfs integrate out effect of RV we ignore (for discrete RV we sum out)

- Marginals represent:

$$f_1(x_1) = \mathbb{P}(X_1 = x_1 | X_2 \in \mathcal{D}_2) = \mathbb{P}(X_1 = x_1)$$

$$f_2(x_2) = \mathbb{P}(X_2 = x_2 | X_1 \in \mathcal{D}_1) = \mathbb{P}(X_2 = x_2).$$

- As an example, consider we want $\mathbb{P}(a_1 \leq X_1 \leq b_1)$; we have that:

$$\mathbb{P}(a_1 \leq X_1 \leq b_1) = \int_{a_1}^{b_1} \int_{x_2 \in \mathcal{D}_2} f(x_1, x_2) dx_2 dx_1 = \int_{a_1}^{b_1} f_1(x_1) dx_1$$

- Marginal pdfs have associated marginal cdfs $F_1(x_1)$ and $F_2(x_2)$.



Independence

So the conditional pdfs are telling us know knowledge in one RV affects the uncertainty in another (i.e., how much knowledge of one is embedded in the other one).

So how about that knowledge of one does not affect uncertainty of the other?

This gives rise to the concept of *independence*.

- RVs X_1 and X_2 are said to be independent if:

$$f(x_1|x_2) = f_1(x_1), \quad x_1 \in \mathcal{D}_1$$
$$f(x_2|x_1) = f_2(x_2), \quad x_2 \in \mathcal{D}_2$$

- This implies that:

$$f(x_1, x_2) = f_2(x_2)f_1(x_1), \quad (x_1, x_2) \in \mathcal{D}$$

- Equivalently:

$$\mathbb{P}(X_1 = x_1 \& X_2 = x_2) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2)$$



Example: Control System Reliability

- Recall joint pdf of $X = (X_1, X_2)$ is:

$$f(x_1, x_2) = \frac{1}{50} e^{-(0.2x_1 + 0.1x_2)}$$

- Probability that $X_1 = x_1$ regardless of knowledge that $X_2 = x_2$ is given by:

$$f_1(x_1) = \int_0^{\infty} \frac{1}{50} e^{-(0.2x_1 + 0.1x_2)} dx_2 = \frac{1}{5} e^{-0.2x_1}$$

- Probability that $X_2 = x_2$ regardless of knowledge that $X_1 = x_1$ is given by:

$$f_2(x_2) = \int_0^{\infty} \frac{1}{50} e^{-(0.2x_1 + 0.1x_2)} dx_1 = \frac{1}{10} e^{-0.1x_2}$$

- Probability that $X_1 = x_1$ given that we know $X_2 = x_2$ is given by:

$$f(x_1|x_2) = \frac{\frac{1}{50} e^{-(0.2x_1 + 0.1x_2)}}{\frac{1}{10} e^{-0.1x_2}} = \frac{1}{5} e^{-0.2x_1}$$

- Probability that $X_2 = x_2$ given that we know $X_1 = x_1$ is given by:

$$f(x_2|x_1) = \frac{\frac{1}{50} e^{-(0.2x_1 + 0.1x_2)}}{\frac{1}{5} e^{-0.2x_1}} = \frac{1}{10} e^{-0.1x_2}$$

- We thus conclude that lifetime X_1 is independent of lifetime X_2 (and viceversa).



Summarizing Statistics for Multivariate RVs

Computing summarizing statistics in multivariate case is similar to univariate case but there are a few key differences that we highlight:

- Joint expectation of $X = (X_1, X_2)$ is a vector $\mathbb{E}[X] = (\mathbb{E}[X_1], \mathbb{E}[X_2])$ with:

$$\mathbb{E}[X_1] = \int_{x_1 \in \mathcal{D}_1} x_1 f_1(x_1) dx_1, \quad \mathbb{E}[X_2] = \int_{x_2 \in \mathcal{D}_2} x_2 f_2(x_2) dx_2.$$

- Joint expectation of $\varphi(X) = \varphi(X_1, X_2)$ is a scalar:

$$\mathbb{E}[\varphi(X)] = \int_{x_1 \in \mathcal{D}_1} \int_{x_2 \in \mathcal{D}_2} \varphi(x_1, x_2) f(x_1, x_2) dx_1 dx_2$$

- Conditional expectation of X_1 (given knowledge $X_2 = x_2$) is:

$$\mathbb{E}[X_1 | X_2 = x_2] = \int_{x_1 \in \mathcal{D}_1} x_1 f(x_1 | x_2) dx_1$$

- Conditional expectation of $\varphi(X) = \varphi(X_1, X_2)$ (given knowledge $X_2 = x_2$) is:

$$\mathbb{E}[\varphi(X) | X_2 = x_2] = \int_{x_1 \in \mathcal{D}_1} \varphi(x_1, x_2) f(x_1 | x_2) dx_1$$

- Expressions for discrete RVs are analogous (replace integrals for sums).



Summarizing Statistics for Multivariate RVs

In multivariate case, concepts of *covariance* and *correlation* emerge:

- Define the marginal expectations and variances:

$$\mu_1 = \mathbb{E}[X_1]$$

$$\mu_2 = \mathbb{E}[X_2]$$

$$\sigma_1^2 = \mathbb{V}[X_1] = \mathbb{E}[(X_1 - \mu_1)^2]$$

$$\sigma_2^2 = \mathbb{V}[X_2] = \mathbb{E}[(X_2 - \mu_2)^2].$$

- Covariance between X_1 and X_2 is:

$$\begin{aligned}\text{Cov}(X_1, X_2) &= \mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= \int_{x_1 \in \mathcal{D}_1} \int_{x_2 \in \mathcal{D}_2} (X_1 - \mu_1)(X_2 - \mu_2) f(x_1, x_2) dx_1 dx_2.\end{aligned}$$

- Define $\sigma_{i,j} = \text{Cov}(X_i, X_j)$ and note $\sigma_{2,1} = \sigma_{1,2}$, $\text{Cov}(X_1, X_1) = \sigma_1^2$ and $\text{Cov}(X_2, X_2) = \sigma_2^2$.
- Correlation between X_1 and X_2 is:

$$\text{Corr}(X_1, X_2) = \frac{\sigma_{1,2}}{\sigma_1 \sigma_2}$$

- Define $\rho_{i,j} = \text{Corr}(X_i, X_j)$ and note $\rho_{1,2} \in [-1, 1]$, $\rho_{2,1} = \rho_{1,2}$, $\rho_{1,1} = \rho_{2,2} = 1$.



Summarizing Statistics for Multivariate RVs

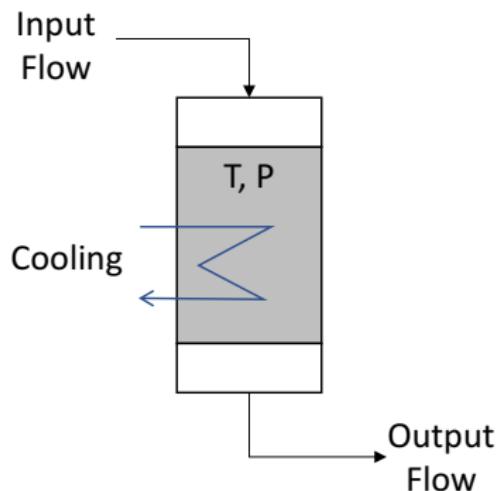
The covariance and correlation tells us how, on average, X_1 varies with X_2 :

- If $\sigma_{1,2} > 0$ ($\rho_{1,2} > 0$) we have *positive correlation*: The event $X_1 > \mu_1$ is, on average, associated with $X_2 > \mu_2$ and event $X_1 < \mu_1$ is, on average, associated with $X_2 < \mu_2$. In other words, X_1 and X_2 move in the same direction (on average).
- If $\sigma_{1,2} < 0$ ($\rho_{1,2} < 0$) we have *negative correlation*: The event $X_1 > \mu_1$ is, on average, associated with $X_2 < \mu_2$ and event $X_1 < \mu_1$ is, on average, associated with $X_2 > \mu_2$. In other words, X_1 and X_2 move in opposite directions (on average).
- If $\sigma_{1,2} = 0$ ($\rho_{1,2} = 0$) we have *no correlation*: On average, changes in $(X_1 - \mu_1)$ are not associated with changes in $(X_2 - \mu_2)$. If X_1 and X_2 are independent then $\sigma_{1,2} = 0$. However, $\sigma_{1,2} = 0$ does not imply independence because variables might move together (but not on average).

Presence of correlation reveals emergent trends. For instance, if X_1 and X_2 are related as $X_2 = \alpha X_1$ then $\text{Cov}(X_1, X_2) = \alpha \mathbb{V}[X_1]$.

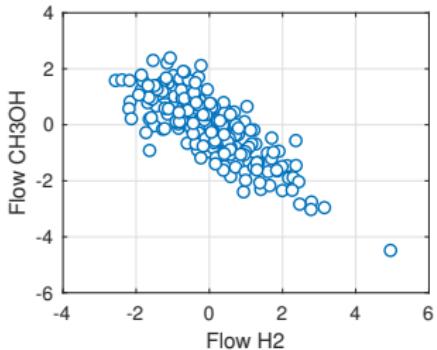
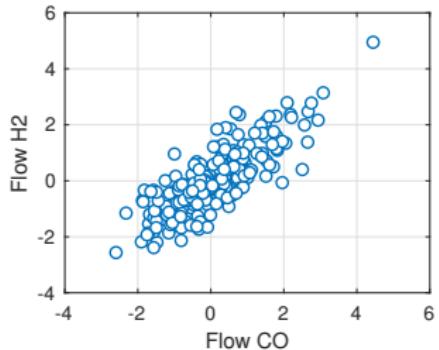
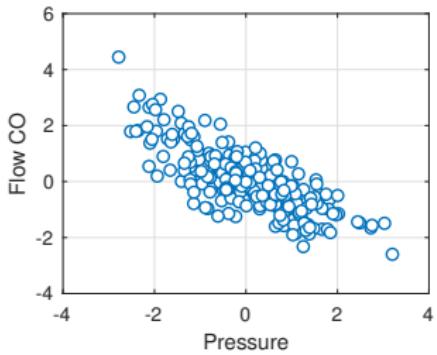
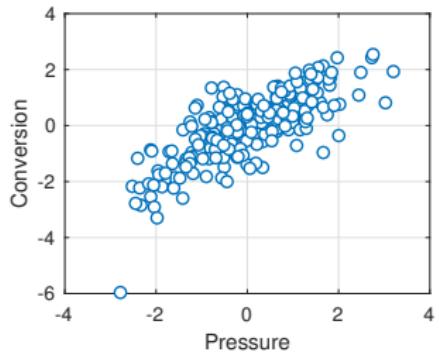
Example: Covariance and Correlation for Reactor

- Consider a reactor under which the reaction $CO + 2H_2 \leftrightarrow CH_3OH$ takes place
- Equilibrium and is favored by high pressure (P) and low temperature (T)
- We have data for pressure, conversion, and output flow of CO , H_2 , and CH_3OH



- Do you expect a positive or negative correlation between conversion and pressure?
- How are output flows for CO , H_2 , and CH_3OH related?

Example: Covariance and Correlation for Reactor





Covariance and Correlation Matrices

- Covariance between variables is often expressed in matrix form as:

$$\text{Cov}[X] = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix}$$

this matrix is symmetric because $\sigma_{1,2} = \sigma_{2,1}$ and has positive eigenvalues (it is positive definite).

- Covariance matrix (for any dimension n) can be computed as:

$$\text{Cov}[X] = \mathbb{E} \left[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T \right]$$

- Correlation between variables is often expressed in matrix form as:

$$\text{Corr}[X] = \begin{bmatrix} 1 & \rho_{1,2} \\ \rho_{2,1} & 1 \end{bmatrix}$$

this matrix is symmetric because $\rho_{1,2} = \rho_{2,1}$ and is positive definite.

- Correlation matrix (for any dimension n) can be computed as:

$$\text{Corr}[X] = D^{-1} \text{Cov}(X) D^{-1}$$

where $D = \text{diag}(\text{Cov}[X])$.

- Sample covariance and correlation matrices can be computed from data.



Example: Covariance and Correlation for Reactor

- Consider a reactor under which the reaction $CO + 2H_2 \leftrightarrow CH_3OH$ takes place
- We have data for pressure, conversion, and output flow of CO , H_2 , and CH_3OH
- Sample covariance and correlation matrices are shown below

$$\hat{\text{Cov}}[X] = \begin{bmatrix} 1.28 & 1.03 & -0.97 & -0.90 & 0.96 \\ 1.03 & 1.38 & -1.05 & -1.00 & 1.04 \\ -0.97 & -1.05 & 1.20 & 0.98 & -0.97 \\ -0.90 & -1.00 & 0.98 & 1.23 & -0.97 \\ 0.96 & 1.04 & -0.97 & -0.97 & 1.21 \end{bmatrix}$$

$$\hat{\text{Corr}}[X] = \begin{bmatrix} 1.00 & 0.77 & -0.78 & -0.71 & 0.77 \\ 0.77 & 1.00 & -0.81 & -0.77 & 0.81 \\ -0.78 & -0.81 & 1.00 & 0.81 & -0.80 \\ -0.71 & -0.77 & 0.81 & 1.00 & -0.79 \\ 0.77 & 0.81 & -0.80 & -0.79 & 1.00 \end{bmatrix}$$

- Do you expect a positive or negative correlation between conversion and pressure?
- How are output flows for CO , H_2 , and CH_3OH related?



Multivariate Gaussian Variables

Surprisingly enough, there are actually few established models for multivariate RVs. The most used model is that of the Gaussian RV, which has a wide range of properties:

- Consider a multivariate RV vector $X = (X_1, X_2, \dots, X_n)$.
- Denote as $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ are hyperparameters.
- RV $X \sim \mathcal{N}(\mu, \Sigma)$ has joint pdf of the form:

$$f_X(x) = f_X(x_1, x_2, \dots, x_n) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

- $|\Sigma|$ is determinant of Σ (product of its eigenvalues) and Σ is positive definite.
- Domain of X is $\mathcal{D} = [-\infty, \infty]^n$.
- Hyperparameters are given by expected value $\mu = \mathbb{E}[X]$ and covariance $\Sigma = \text{Cov}[X]$.



Properties of Multivariate Gaussian RVs

Consider the case with $n = 2$:

- Can show that marginal pdfs of X are Gaussian:

$$f_1(x_1) = \int_{x_2 \in \mathbb{D}_2} f(x_1, x_2) dx_2 = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(\frac{-(x - \mu_1)^2}{2\sigma_1^2}\right)$$

$$f_2(x_2) = \int_{x_1 \in \mathbb{D}_1} f(x_1, x_2) dx_1 = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(\frac{-(x - \mu_2)^2}{2\sigma_2^2}\right)$$

- In other words, $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.



Properties of Multivariate Gaussian RVs

- Can show that conditional pdfs of X are Gaussian:

$$f(x_1|x_2) = \frac{1}{\sqrt{2\pi\sigma_{1|2}^2}} \exp\left(\frac{-(x - \mu_{1|2})^2}{2\sigma_{1|2}^2}\right)$$
$$f(x_2|x_1) = \frac{1}{\sqrt{2\pi\sigma_{2|1}^2}} \exp\left(\frac{-(x - \mu_{2|1})^2}{2\sigma_{2|1}^2}\right).$$

- That is, $X_1|X_2 \sim \mathcal{N}(\mu_{1|2}, \sigma_{1|2})$ and $X_2|X_1 \sim \mathcal{N}(\mu_{2|1}, \sigma_{2|1})$ with hyperparameters:

$$\mu_{1|2} = \mu_1 + \sigma_{1,2}\sigma_{22}^{-1}(x_2 - \mu_2)$$

$$\sigma_{1|2} = \sigma_{1,1} - \sigma_{1,2}\sigma_{2,2}^{-1}\sigma_{2,1}$$

$$\mu_{2|1} = \mu_2 + \sigma_{2,1}\sigma_{11}^{-1}(x_1 - \mu_1)$$

$$\sigma_{2|1} = \sigma_{2,2} - \sigma_{2,1}\sigma_{11}^{-1}\sigma_{1,2}$$

- What if X_1 and X_2 are independent? What if they are strongly correlated?



Example: Uncertainty Reduction in Reactor

- Consider covariance for pressure X_1 and conversion X_2 and assume Gaussian.
- Marginal means are:

$$\mu_1 = 147$$

$$\mu_2 = 0.69$$

- Covariance matrix is:

$$\text{Cov}[X] = \begin{bmatrix} 641.31 & 1.84 \\ 1.84 & 0.24 \end{bmatrix}$$

- Marginal variance for conversion is $\sigma_{2,2} = 0.24$ (we use this as a measure of its uncertainty).
- Since pressure and conversion are correlated, we expect that having knowledge of conversion decreases uncertainty in the conversion. To verify this, we compute the variance of conditional density $f(x_2|x_1)$

$$\begin{aligned}\sigma_{2|1}^2 &= (0.24 - 1.84 \cdot (641.31)^{-1} \cdot 1.84)^2 \\ &= 0.05\end{aligned}$$

- Consequently, uncertainty in conversion is reduced by 80% if we know pressure.
- In other words, since pressure and conversion are correlated, data on pressure carries information on conversion (and the other way around).

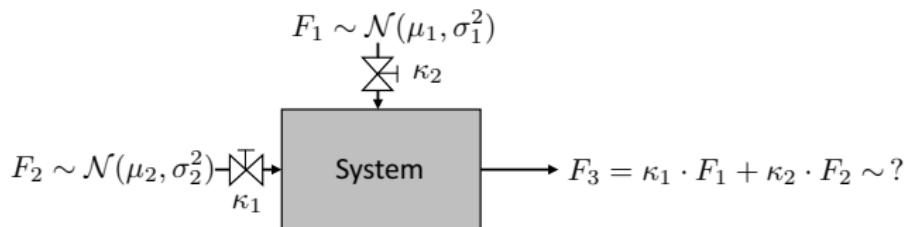


Properties of Multivariate Gaussian RVs

- Linear transformation of a multivariate Gaussian is Gaussian:
 - If $X \sim \mathcal{N}(\mu, \Sigma)$, then $Y = AX + b$ is $Y \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$.
- Linear transformation property is used to establish useful properties:
 - Mixture model:
If $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then $Y = \sum_{i=1}^n X_i$ is Gaussian with $Y \sim \mathcal{N}(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2)$.
 - Multivariate standard normal:
If $X \sim \mathcal{N}(0, I)$ then $Y = \sqrt{\Sigma}X + \mu$ is Gaussian with $Y \sim \mathcal{N}(\mu, \Sigma)$.
- Standardization result used to generate samples of $\mathcal{N}(\mu, \Sigma)$ from samples of $\mathcal{N}(0, I)$.

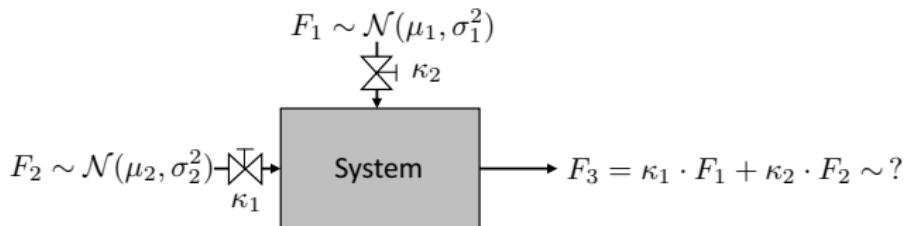
Example: Gaussian Mixture

- Have system with random revenue streams $F_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $F_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- Can control amount of revenue using $\kappa_1 \in [0, 1]$ and $\kappa_2 \in [0, 1]$
- What is uncertainty of total revenue F_3 ? How is this affected by κ_1 and κ_2 ?



Example: Gaussian Mixture

- Have system with random revenue streams $F_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $F_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$
- Can control amount of revenue using $\kappa_1 \in [0, 1]$ and $\kappa_2 \in [0, 1]$
- What is uncertainty of total revenue F_3 ? How is this affected by κ_1 and κ_2 ?



- We can write F_2 as a linear combination of F_1 and F_2 :

$$F_3 = \begin{bmatrix} \kappa_1 & \kappa_2 \end{bmatrix} \begin{bmatrix} F_1 \\ F_2 \end{bmatrix} = \kappa_1 F_1 + \kappa_2 F_2$$

- We thus have that $F_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$ with:

$$\mu_3 = \begin{bmatrix} \kappa_1 & \kappa_2 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} = \kappa_1 \mu_1 + \kappa_2 \mu_2$$

$$\sigma_3^2 = \begin{bmatrix} \kappa_1 & \kappa_2 \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \kappa_1 \\ \kappa_2 \end{bmatrix} = \kappa_1^2 \sigma_1^2 + \kappa_2^2 \sigma_2^2$$



Geometry of Multivariate Gaussian Variables

Understanding geometry of multivariate Gaussian facilitates *data visualization*.

- For $n = 2$ we write joint pdf of $X = (X_1, X_2)$ as:

$$f_X(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp(-U(x_1, x_2))$$

where

$$U(x_1, x_2) = \frac{1}{2(1 - \rho^2)} \left[\frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - 2\rho \frac{(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} \right]$$

- If we fix probability $f_X(x_1, x_2) = p$ with $\alpha \in [0, 1]$, pdf defines an equation in variables (x_1, x_2) . This is equation of ellipse centered at μ_1 and μ_2 . This ellipse is known as the p -level set of pdf.
- Correlation coefficient $\rho = \text{Corr}(X_1, X_2)$ dictates *orientation of ellipse*: if $\rho > 0$ this is tilted to right. if $\rho < 0$ this is tilted to left, and if $\rho = 0$ (X_1 and X_2 are independent) ellipse has no tilt.
- Length of axes* are dictated by σ_1^2 and σ_2^2 (variances of X_1 and X_2).
- Maximum value of $f_X(x_1, x_2)$ is achieved at $x_1 = \mu_1$ and $x_2 = \mu_2$.



Geometry of Multivariate Gaussian Variables

- We are interested in box confidence regions $\mathcal{B}(\alpha)$ satisfying:

$$\mathbb{P}(X \in \mathcal{B}(\alpha)) = 1 - \alpha.$$

- For univariate $X \sim \mathcal{N}(\mu, \sigma^2)$, box region is:

$$\mathcal{B}(\alpha) = \{x \mid x \in [\mu \pm \sqrt{\mathbb{Q}(\alpha)}\sigma]\}$$

where $\mathbb{Q}(\alpha)$ is the α -quantile of $\chi^2(1)$.

- For multivariate $X = (X_1, X_2)$ with $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ we can build a box of the form:

$$\mathcal{B}(\alpha) = \{(x_1, x_2) \mid x_1 \in [\mu_1 \pm \sqrt{\mathbb{Q}(\alpha)}\sigma_1] \& x_2 \in [\mu_2 \pm \sqrt{\mathbb{Q}(\alpha)}\sigma_2]\}.$$

- This box (a.k.a. marginal box) does not capture correlations in X_1 and X_2 .



Geometry of Multivariate Gaussian Variables

- In multivariate Gaussians, observations concentrate in ellipses (p -level sets). We are thus interested in finding *ellipsoidal* confidence region $\mathcal{E}(\alpha)$ satisfying:

$$\mathbb{P}(X \in \mathcal{E}(\alpha)) = 1 - \alpha.$$

- For $X \sim \mathcal{N}(\mu, \Sigma)$, the ellipsoidal region is given by:

$$\mathcal{E}(\alpha) = \{x \mid (x - \mu)^T \Sigma^{-1} (x - \mu)^T \leq \mathbb{Q}(\alpha)\}.$$

where $\mathbb{Q}(\alpha)$ is the α -quantile of $\chi^2(n)$.

- Interpretation of region is:
 - If draw sample from $\mathcal{N}(\mu, \Sigma)$, there is probability α that it will land in $\mathcal{E}(\alpha)$
 - The larger the α , the larger the ellipsoid (more likely it is to land in $\mathcal{E}(\alpha)$)
- Tighest box that encloses $\mathcal{E}(\alpha)$ is:

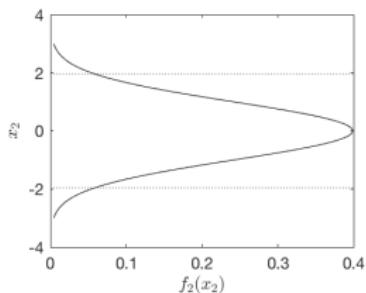
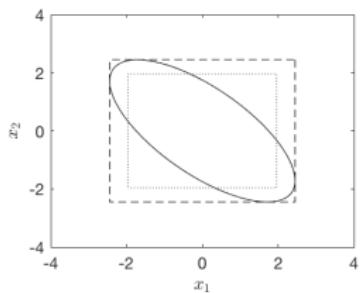
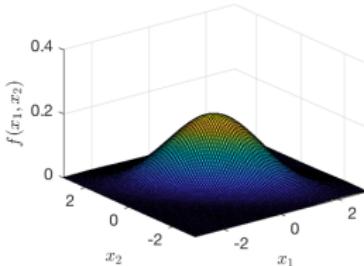
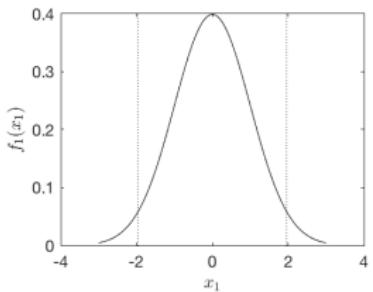
$$\mathcal{B}(\alpha) = \{(x_1, x_2) \mid x_1 \in [\mu_1 \pm \sqrt{\mathbb{Q}(\alpha)}\sigma_1] \text{ } \& \text{ } x_2 \in [\mu_2 \pm \sqrt{\mathbb{Q}(\alpha)}\sigma_2]\}.$$

where $\mathbb{Q}(\alpha)$ is α -quantile of $\chi^2(n)$ (note difference with the marginal box).

- Ellipsoidal and enclosing box capture correlations in X_1 and X_2 .

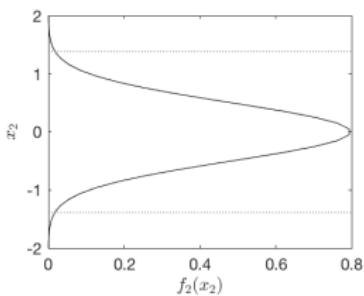
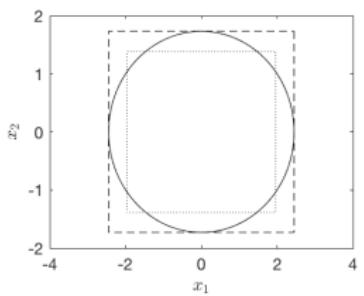
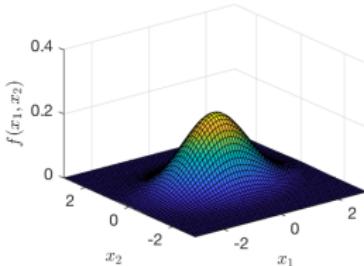
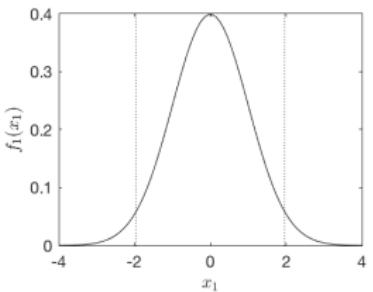
Example: Geometry of Multivariate Gaussian Variables

Marginal pdfs, joint pdf and confidence ellipse and boxes for Gaussian with correlation.



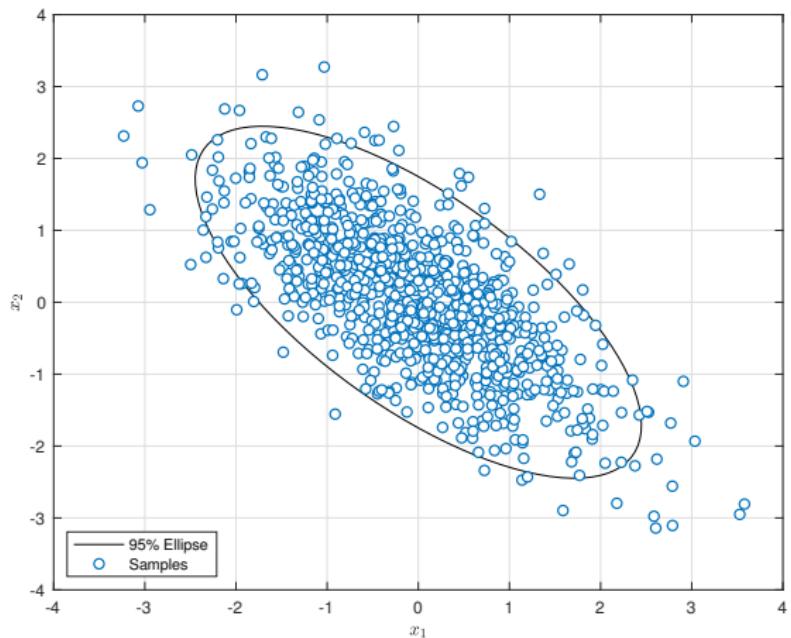
Example: Geometry of Multivariate Gaussian Variables

Marginal pdfs, joint pdf and confidence ellipse and boxes for Gaussian with no correlation.



Example: Geometry of Multivariate Gaussian Variables

- Here are $S = 1000$ observations for Gaussian and 95% confidence ellipsoid
- Total of 942 samples lie inside of ellipsoid





Outline

- 1 Introduction
- 2 Models of Random Variables
- 3 Estimation Techniques
- 4 Multivariate Statistics
- 5 Data-Driven Modeling**
- 6 Statistical Learning
- 7 Decision-Making Under Uncertainty



Data-Driven Modeling

When X, Y are correlated, variations in Y align with those in X (there is a *trend*).

What if connection is deep (e.g., variations in Y are *caused* by variations in X)?

- Consider univariate RVs Y and X and postulate that any behavior in Y is due to a systematic dependence of X :

$$Y = \theta X$$

Here, θ is a parameter that captures a *linear* dependence between X and Y .

- Use random samples pairs (y_ω, x_ω) and, based on postulated model, we assume they are related as:

$$y_\omega = \theta x_\omega + \epsilon_\omega, \omega \in \mathcal{S}.$$

- We introduce hidden RV $\epsilon_\omega \in \mathcal{N}(0, \sigma^2)$ with known σ^2 to capture behavior that cannot be explained by model (the unknown).
- Note y_ω is an RV because ϵ_ω is an RV. Moreover, y_ω is linear transformation of ϵ_ω and thus y_ω is Gaussian with $\mathbb{E}[y_\omega | x_\omega, \theta] = \theta x_\omega$ and $\mathbb{V}[y_\omega | x_\omega, \theta] = \sigma^2$.



Data-Driven Modeling

- Recall $f(y_\omega|x_\omega, \theta)$ is probability that $Y = y_\omega$ given that we know $X = x_\omega$ and θ . This is equivalent to assume that θ and x_ω are *deterministic* (more on this later).
- We use a maximum likelihood approach and seek to find θ that maximizes joint likelihood $\prod_{\omega \in \mathcal{S}} f(y_\omega|x_\omega, \theta)$. This gives:

$$\max_{\theta} \log L(\theta) = \sum_{\omega \in \mathcal{S}} \log f(y_\omega|x_\omega, \theta)$$

- Since $y_\omega \sim \mathcal{N}(\theta x_\omega, \sigma^2)$, we know that:

$$\log f(y_\omega|x_\omega, \theta) = -\log \sqrt{2\pi\sigma^2} - \frac{(y_\omega - \theta x_\omega)^2}{2\sigma^2}$$

- Terms $\log \sqrt{2\pi\sigma^2}$ and $2\sigma^2$ are constants and we thus obtain:

$$\min_{\theta} \frac{1}{2} \sum_{\omega \in \mathcal{S}} (y_\omega - \theta x_\omega)^2$$

This is a least-squares problem and aims to find the estimate $\hat{\theta}$ that minimizes discrepancy between the observed output y_ω and model prediction $\hat{\theta}x_\omega$.



Data-Driven Modeling

We denote best parameter that is *learned* from data as $\hat{\theta}$.

- Recall $\hat{\theta}$ minimizes $S(\theta) = \frac{1}{2} \sum_{\omega \in S} (y_\omega - \theta x_\omega)^2$ and thus satisfies:

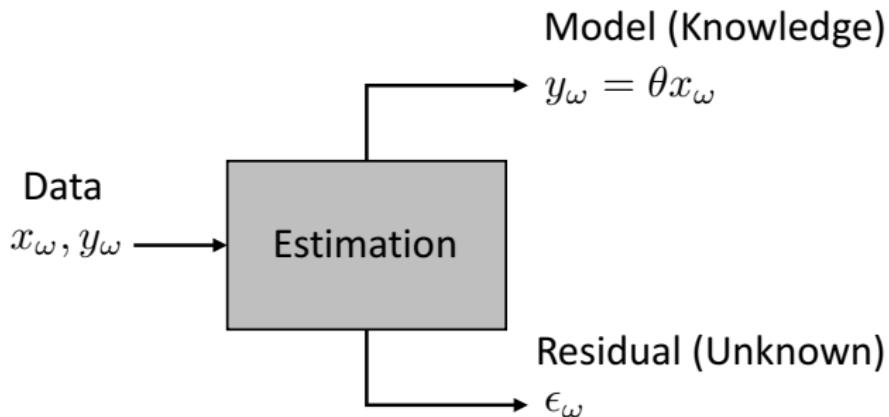
$$\frac{\partial S(\hat{\theta})}{\partial \theta} = - \sum_{\omega \in S} x_\omega (y_\omega - \theta x_\omega) = 0, \quad \frac{\partial^2 S(\hat{\theta})}{\partial \theta^2} > 0$$

These conditions lead to:

$$\hat{\theta} = \frac{\sum_{\omega \in S} x_\omega y_\omega}{\sum_{\omega \in S} x_\omega^2}, \quad \sum_{\omega \in S} x_\omega^2 > 0$$

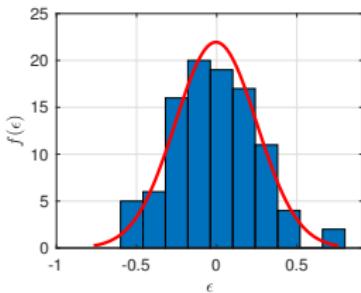
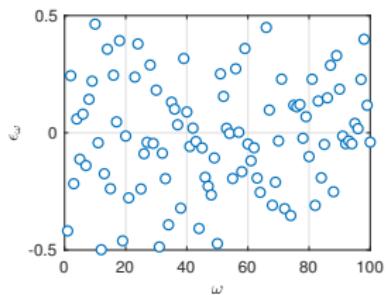
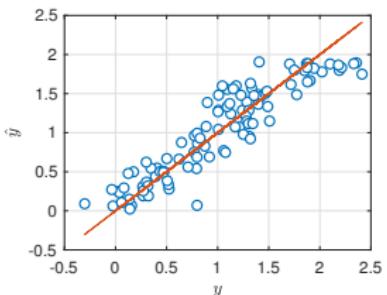
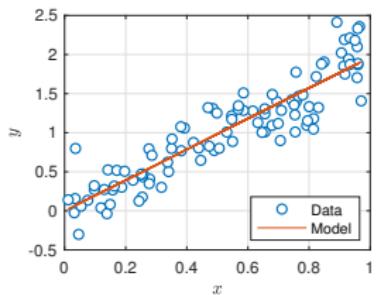
- Estimate $\hat{\theta}$ captures *interactions* in data x_ω, y_ω .
- Estimate $\hat{\theta}$ is *unique* and becomes better defined as we add more data.
- Recall $y_\omega \sim \mathcal{N}(\hat{\theta}x_\omega, \sigma^2)$, implying that prediction $\hat{\theta}x_\omega$ is the most likely outcome and that the larger σ^2 , the more uncertainty we have in y_ω .
- Estimate gives *residual noise estimates* $\hat{\epsilon}_\omega = y_\omega - \hat{\theta}x_\omega$. If these estimates follow our assumption $\mathcal{N}(0, \sigma^2)$ then the available data and postulated model is satisfactory. If not, more data or another model is needed (e.g., nonlinear).
- From $y_\omega = \hat{\theta}x_\omega + \epsilon_\omega$ we note that model $\hat{\theta}x_\omega$ represents what we know about y_ω while ϵ_ω represents the unknown. Estimation problem thus seeks to extract maximum knowledge from data (minimize the unknown).

Data-Driven Modeling



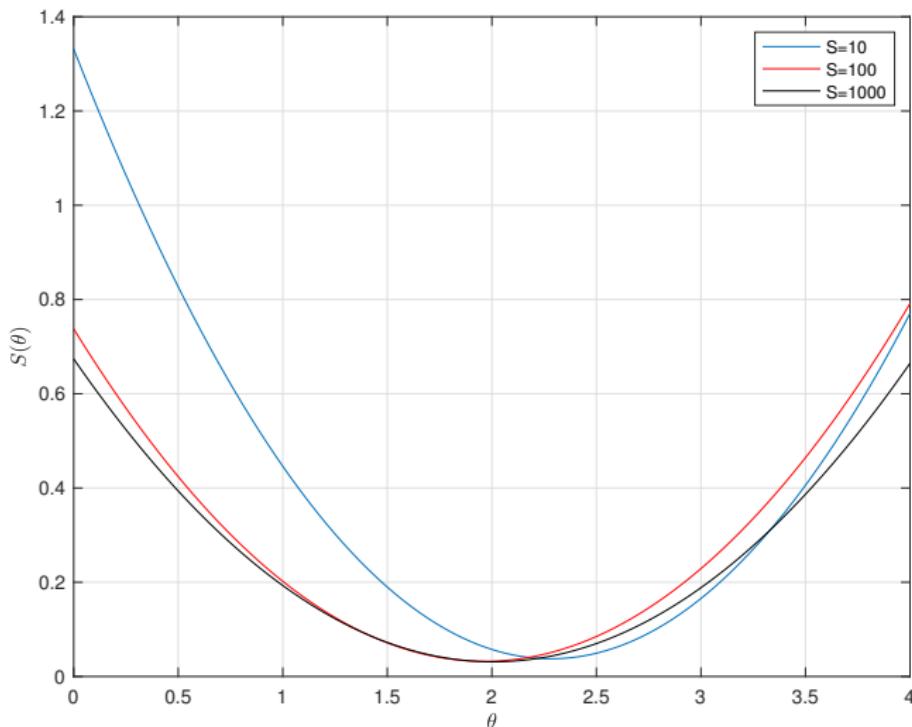
Example: Data-Driven Modeling

- Consider input out pair data (Y, X) with true relationship $Y = \theta X$ and $\theta = 2$
- Observations (y_ω, x_ω) are corrupted by noise $\epsilon_\omega \sim \mathcal{N}(0, 0.25)$
- If we use $S = 100$ observations, best estimate is $\hat{\theta} = 1.96$



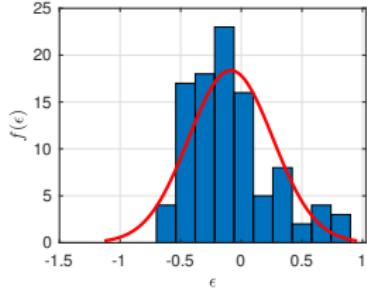
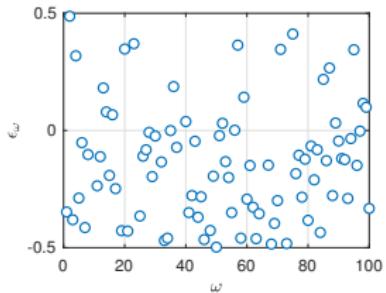
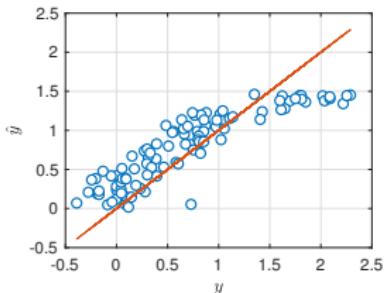
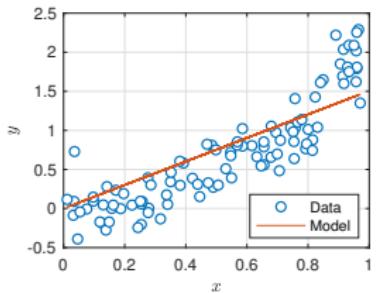
Example: Data-Driven Modeling

- Below we show function $S(\theta)$ for different amounts of data $S = 10, 100, 1000$
- Note how surface becomes better defined as we add data



Example: Data-Driven Modeling

- Assume now true relationship $Y = \theta X^2$ and $\theta = 2$
- Below we show model predictions if we postulate wrong model $Y = \theta X$



- Residual statistics play key role in determining appropriateness of postulated model.



Data-Driven Modeling

Now generalize linear model to account for influence of multiple inputs. We postulate:

$$Y = \theta_0 + \sum_{i=1}^n \theta_i X_i$$

Use samples (data) pairs (y_ω, x_ω) and, based on postulated model, we have that:

$$y_\omega = \theta_0 + \sum_{i=1}^n \theta_i x_{i,\omega} + \epsilon_\omega, \quad \omega \in \mathcal{S}.$$

These set of equations can be expressed compactly using matrix notation:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

where:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_S \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & & & \ddots & \\ 1 & x_{S,1} & x_{S,2} & \dots & x_{S,n} \end{bmatrix} \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_S \end{bmatrix}$$

We assume unknown noise is $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{\omega,\omega} = \sigma^2$ for $\omega \in \mathcal{S}$.



Data-Driven Modeling

Best estimate $\hat{\theta}$ is found as the solution of max likelihood problem:

$$\min_{\theta} \frac{1}{2}(\mathbf{y} - \mathbf{X}\theta)^T(\mathbf{y} - \mathbf{X}\theta)$$

Problem can also be written as:

$$\min_{\theta} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\theta\|^2$$

Solution of this problem must satisfy:

$$\begin{aligned}\frac{\partial S(\theta)}{\partial \theta} &= -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\theta) = 0 \\ \left| \frac{\partial^2 S(\theta)}{\partial \theta^2} \right| &= \left| \mathbf{X}^T \mathbf{X} \right| > 0\end{aligned}$$

First condition yields:

$$\hat{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Second condition indicates that $\hat{\theta}$ is unique if matrix $\mathbf{X}^T \mathbf{X}$ is positive definite.



Data-Driven Modeling

Now note that $\mathbf{y} = \mathbf{X}\theta + \epsilon$ is an RV (θ is true parameter). Consequently:

$$\begin{aligned}\hat{\theta} &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{X}\theta + \epsilon) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{X})\theta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \epsilon \\ &= \theta + (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X} \epsilon\end{aligned}$$

Estimate $\hat{\theta}$ is thus an RV (a linear transformation of noise $\epsilon \sim \mathcal{N}(0, \Sigma)$) and thus:

$$\hat{\theta} \sim \mathcal{N}(\theta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$$

Some observations:

- Expected value of estimate is $\mathbb{E}[\hat{\theta}] = \theta$ (estimate is unbiased).
- Covariance of estimate is $\text{Cov}[\hat{\theta}] = (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2$. As variance σ^2 increase so does variance of $\hat{\theta}$ (covariance is sensitivity of estimates to noise).



Data-Driven Modeling

How much of the variability in the data y can be explained by our gained knowledge (model $\hat{y} = \mathbf{X}^T \hat{\theta}$) and how much of it cannot be explained (unknown ϵ)?

This can be addressed using *analysis of variance* (ANOVA). Total variability of data is:

$$S_y = \sum_{\omega \in S} (y_\omega - \bar{y})^2 \quad \text{with} \quad \bar{y} = \frac{1}{S} \sum_{\omega \in S} y_\omega$$

Total variability can be decomposed into contributions as:

$$S_y = \underbrace{\sum_{\omega \in S} (\hat{y}_\omega - \bar{y})^2}_{S_m} + \underbrace{\sum_{\omega \in S} (y_\omega - \hat{y}_\omega)^2}_{S_e}$$

Here, S_m is known as model sum of squares and S_e is known as sum of squared errors. Based on these quantities we define index:

$$R^2 = \frac{S_m}{S_y} = 1 - \frac{S_e}{S_y}$$

This represents fraction of variability captured by model. The fraction that is left unexplained is $1 - R^2 = S_e/S_y$. Consequently, $R^2 \rightarrow 1$ indicates model fully explains data variability.



Data-Driven Modeling

How confident are we in estimate $\hat{\theta}$ and in model predictions $\hat{\mathbf{y}} = \mathbf{X}\hat{\theta}$?

- We have established that $\hat{\theta} \sim \mathcal{N}(\theta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$. We thus have that true θ lies in region $\mathcal{E}(\alpha)$ defined by:

$$(\hat{\theta} - \theta)^T \left(\frac{\mathbf{X}^T \mathbf{X}}{\sigma^2} \right) (\hat{\theta} - \theta) \leq \mathbb{Q}(\alpha)$$

where $\mathbb{Q}(\alpha)$ is α -quantile of $\chi^2(n+1)$.

- In most scientific literature, confidence in estimates is reported using marginals:

$$\hat{\theta}_i \sim \mathcal{N}(\theta_i, \sigma^2 (\mathbf{X} \mathbf{X})_{ii}^{-1}), \quad i = 0, \dots, n$$

as:

$$\theta_i = \hat{\theta}_i \pm m_i \quad m_i = \sqrt{\mathbb{Q}(\alpha) \sigma^2 (\mathbf{X} \mathbf{X})_{ii}^{-1}}$$

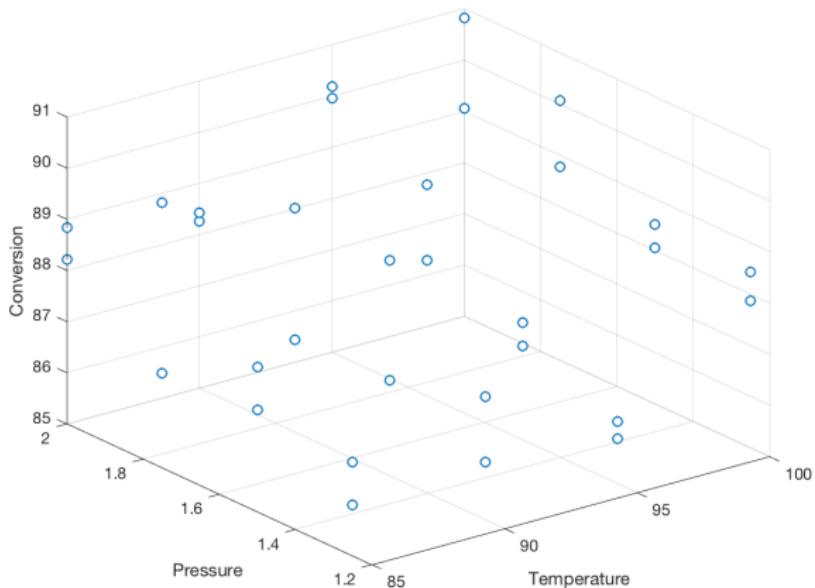
where $\mathbb{Q}(\alpha)$ is α -quantile of $\chi^2(1)$. This disregards correlations in parameters.

- We can construct confidence intervals for model predictions by noticing that $\hat{\mathbf{y}} = \mathbf{X}\hat{\theta}$ and $\hat{\theta} \sim \mathcal{N}(\theta, (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2)$ and thus:

$$\hat{\mathbf{y}} \sim \mathcal{N}(\mathbf{y}, \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2)$$

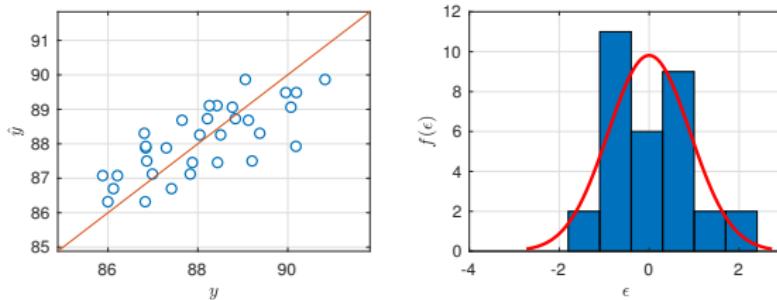
Example: Catalytic Reactor

- Given experimental data for pressure X_1 , temperature X_2 , and conversion Y
- Total of $S = 32$ observations



Example: Catalytic Reactor

- We postulate model $Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2$ with noise model $\epsilon \sim \mathcal{N}(0, \sigma)$
- Model fit and residual pdf is shown below



- $R^2 = 0.55$ (55% of total variability is explained by model)
- Parameters 95% confidence intervals (marginals) are:

$$\theta_0 \in [69.8858, 81.8434]$$

$$\theta_1 \in [0.0149, 0.1366]$$

$$\theta_2 \in [1.9946, 4.4299]$$

Model seems satisfactory, low R^2 suggests that measurements are inaccurate.



Data-Driven Modeling

Is the available data sufficient (or too much) to construct model?

Some observations on having sufficient data:

- Matrix $\mathbf{X}^T \mathbf{X}$ plays a fundamental role as it contains all input data and determines sharpness of minimum and variance of estimate $\hat{\theta}$.
 - If one or more eigenvalues of $\mathbf{X}^T \mathbf{X}$ are large, $\hat{\theta}$ is well-defined by the data and its variance is small. This manifests as a sharp minimum $S(\hat{\theta})$.
 - If one or more eigenvalues of $\mathbf{X}^T \mathbf{X}$ are close to zero, $\hat{\theta}$ is ill-defined by the data and variance is large. This manifests as a flat minimum $S(\hat{\theta})$.
 - If one eigenvalue of $\mathbf{X}^T \mathbf{X}$ is zero, $\hat{\theta}$ cannot be resolved uniquely from data.
- Volume of data is not sufficient, we also require *quality of data*.
 - Observations do not provide information if redundant ($\mathbf{X}^T \mathbf{X}$ has dependent rows).
 - If selected inputs X do not explain output y estimates $\hat{\theta}$ might exhibit high variability (regardless of number of observations).
 - Using knowledge of application to select input variables is important.
- Selection of input variables and observations is a topic of *design of experiments*.
- Inputs X are a.k.a. *regressor variables*, *explanatory variables*, *features*, or *descriptors*.



Data-Driven Modeling

Is the available data sufficient (or too much) to construct model?

Some observations on using excessive data:

- A common issue is that we use too many inputs X to explain output Y . This can result in a large number of parameters and *overfitting*.
- Check that $\hat{\theta}$ obtained with observations (y_ω, x_ω) , $\omega \in \mathcal{S}$ predicts well in an independent set of observations (y_ω, x_ω) , $\omega \in \mathcal{T}$. This procedure is called *cross-validation* or *out-of-sample testing*.
- Cross validation will ensure that model is *generalizable*.
- In linear models, an adjusted R^2 index is used to account for number of parameters:

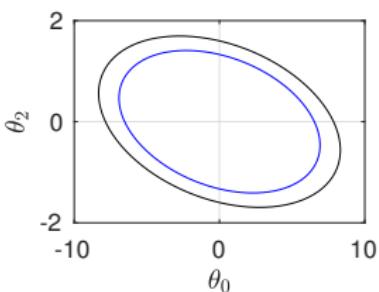
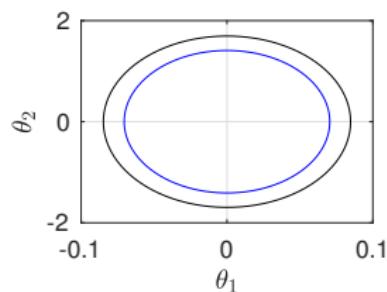
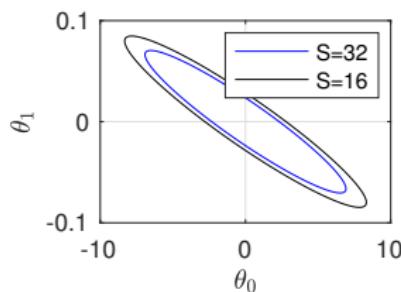
$$R_{adj}^2 = 1 - \frac{S_e}{S_y} \frac{(S-1)}{(S-n)}$$

As number of parameters n increases, we have that R_{adj}^2 decreases.

- Note that the size of confidence ellipsoid $\mathcal{E}(\theta)$ depends on n .
- A strategy to deal with many parameters is *regularization* (will be covered later). Regularization seeks to embed *prior* knowledge in estimation problem.

Example: Catalytic Reactor

- $R^2 = 0.55$, $R_{adj}^2 = 0.52$ (model does not seem overparameterized)
- Below we show confidence ellipses for $S = 32$ (all data) and $S = 16$ (reduced data). This tells us that there is a bit of redundancy in data.





Data-Driven Modeling (Nonlinear)

Now generalize the linear model to a general model of the form:

$$y_\omega = g(\theta, x_\omega) + \epsilon_\omega, \quad \omega \in \mathcal{S}.$$

where $g : \mathbb{R}^n \times \mathbb{R}^S \rightarrow \mathbb{R}$ is a model function of the parameters and inputs.

- The model function can be nonlinear and capture mechanistic relationships between the inputs, parameters, and outputs.
- In linear case we have $g(\theta, x_\omega) = \theta_0 + \sum_{i=1}^n \theta_i x_{i,\omega} + \epsilon_\omega$.
- Use an MLE framework to estimate θ :

$$\min_{\theta} S(\theta) = \frac{1}{2} \sum_{\omega \in \mathcal{S}} (y_\omega - m_\omega(\theta))^2$$

where $m_\omega(\theta) = g(\theta, x_\omega)$. Problem can be expressed in vector form:

$$\min_{\theta} \frac{1}{2} \|\mathbf{y} - \mathbf{m}(\theta)\|^2$$

- Solution $\hat{\theta}$ satisfies following set of n nonlinear equations (a.k.a. score functions):

$$\nabla_{\theta} S(\theta) = 0 \quad \iff \quad \nabla_{\theta} m(\theta)^T (\mathbf{y} - \mathbf{m}(\theta)) = 0$$



Data-Driven Modeling (Nonlinear)

- Solution $\hat{\theta}$ also satisfies $|\mathbf{H}(\theta)| > 0$ where:

$$\mathbf{H}(\theta) = \frac{\partial^2 S(\theta)}{\partial \theta^2} = \begin{bmatrix} \frac{\partial^2 S(\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 S(\theta)}{\partial \theta_1 \partial \theta_2} & \cdots & \frac{\partial^2 S(\theta)}{\partial \theta_1 \partial \theta_n} \\ \frac{\partial^2 S(\theta)}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 S(\theta)}{\partial \theta_2 \partial \theta_2} & \cdots & \frac{\partial^2 S(\theta)}{\partial \theta_n \partial \theta_2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial^2 S(\theta)}{\partial \theta_1 \partial \theta_n} & \frac{\partial^2 S(\theta)}{\partial \theta_2 \partial \theta_n} & \cdots & \frac{\partial^2 S(\theta)}{\partial \theta_n \partial \theta_n} \end{bmatrix}$$

This matrix is known as the *Hessian* matrix.

- For linear models the Hessian is:

$$\mathbf{H}(\theta) = X^T X$$

because $\nabla_{\theta} m(\theta) = X$.

- As in linear case, eigenvalues of Hessian matrix encode information about data quality and quantity.



Nonlinear Estimation

What is different about nonlinear estimation?

- Difficult to obtain pdf of $\hat{\theta}$. Typically, this is approximated as:

$$\hat{\theta} \sim \mathcal{N}(\theta, \mathbf{H}(\hat{\theta})^{-1}\sigma^2)$$

The approximation is accurate if nonlinearity of model $m(\theta)$ is not too strong.

- Function $S(\theta)$ might have multiple points satisfying optimality conditions.
- Problems are often solved using local search algorithms (based on Newton's method) and thus the initial guess of $\hat{\theta}$ influences estimate found. One can also resort to using global search algorithms.
- Hessian can be difficult to compute, one can approximate it as $H(\theta) \approx \nabla_{\theta}m(\theta)^T \nabla_{\theta}m(\theta)$ (a.k.a. Gauss-Newton approximation).
- There are modern languages that compute exact Hessians (e.g., JuMP, Casadi).
- Surprisingly, standard tools such as Matlab do not provide efficient capabilities to compute Hessians.



Example: Hessian of Heat Capacity Problem

- You would like to create a model to predict heat capacity as a function of temperature of the form:

$$C = \theta_0 + \theta_1 T$$

- Assume that you have an experimental data point available (T_1, C_1)

Can θ_0, θ_1 be estimated *uniquely* from (T_1, C_1) ? Why?

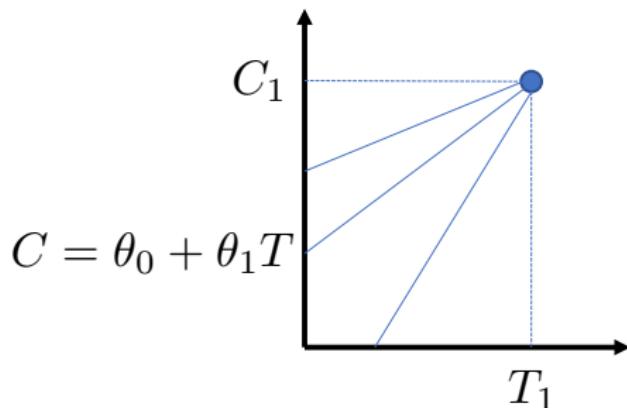
Example: Hessian of Heat Capacity Problem

- You would like to create a model to predict heat capacity as a function of temperature of the form:

$$C = \theta_0 + \theta_1 T$$

- Assume that you have an experimental data point available (T_1, C_1)

Can θ_0, θ_1 be estimated *uniquely* from (T_1, C_1) ? Why?





Example: Hessian of Heat Capacity Problem

- Lets analyze how to reach a conclusion by analyzing the Hessian matrix. Estimation problem is:

$$\min_{\theta_0, \theta_1} S(\theta_0, \theta_1) = \frac{1}{2}(C_1 - \theta_0 - \theta_1 T_1)^2$$

- The first-order derivatives of $S(\theta)$ are:

$$\frac{\partial S}{\partial \theta_0} = -(C_1 - \theta_0 - \theta_1 T_1)$$

$$\frac{\partial S}{\partial \theta_1} = -T_1(C_1 - \theta_0 - \theta_1 T_1)$$

- The second derivatives of $S(\theta)$ are:

$$\frac{\partial^2 S}{\partial \theta_0 \partial \theta_0} = 1, \quad \frac{\partial^2 S}{\partial \theta_0 \partial \theta_1} = T_1$$

$$\frac{\partial^2 S}{\partial \theta_1 \partial \theta_0} = T_1, \quad \frac{\partial^2 S}{\partial \theta_1 \partial \theta_1} = T_1^2.$$



Example: Hessian of Heat Capacity Problem

- The Hessian matrix is thus:

$$H(\theta) = \begin{bmatrix} 1 & T_1 \\ T_1 & T_1^2 \end{bmatrix}$$

- The determinant of this matrix is:

$$|H(\theta)| = T_1^2 - T_1^2 = 0$$

- Consequently, the matrix is singular (it has eigenvalues that are zero)
- Parameters θ_0, θ_1 cannot be estimated uniquely, as expected.



Example: Hessian of Heat Capacity Problem

- Now imagine that you get another experimental data point (T_2, C_2) to estimate parameters. Estimation problem becomes:

$$\min_{\theta_0, \theta_1} S(\theta_0, \theta_1) = \frac{1}{2}(C_1 - \theta_0 - \theta_1 T_1)^2 + \frac{1}{2}(C_2 - \theta_0 - \theta_1 T_2)^2$$

- Hessian of this problem is:

$$H(\theta) = \begin{bmatrix} 2 & T_1 + T_2 \\ T_1 + T_2 & T_1^2 + T_2^2 \end{bmatrix}$$

- Determinant of this matrix is:

$$\begin{aligned}|H(\theta)| &= 2(T_1^2 + T_2^2) - (T_1 + T_2)^2 \\&= T_1^2 - 2T_1T_2 + T_2^2 \\&= (T_1 - T_2)^2\end{aligned}$$

- Under what conditions are parameters unique?



Example: Hessian of Heat Capacity Problem

- Now imagine that you get another experimental data point (T_2, C_2) to estimate parameters. Estimation problem becomes:

$$\min_{\theta_0, \theta_1} S(\theta_0, \theta_1) = \frac{1}{2}(C_1 - \theta_0 - \theta_1 T_1)^2 + \frac{1}{2}(C_2 - \theta_0 - \theta_1 T_2)^2$$

- Hessian of this problem is:

$$H(\theta) = \begin{bmatrix} 2 & T_1 + T_2 \\ T_1 + T_2 & T_1^2 + T_2^2 \end{bmatrix}$$

- Determinant of this matrix is:

$$\begin{aligned}|H(\theta)| &= 2(T_1^2 + T_2^2) - (T_1 + T_2)^2 \\&= T_1^2 - 2T_1T_2 + T_2^2 \\&= (T_1 - T_2)^2\end{aligned}$$

- Under what conditions are parameters unique?
- Parameters are unique as long as $T_1 \neq T_2$ (data is not redundant).



Example: Hougen-Watson Reaction

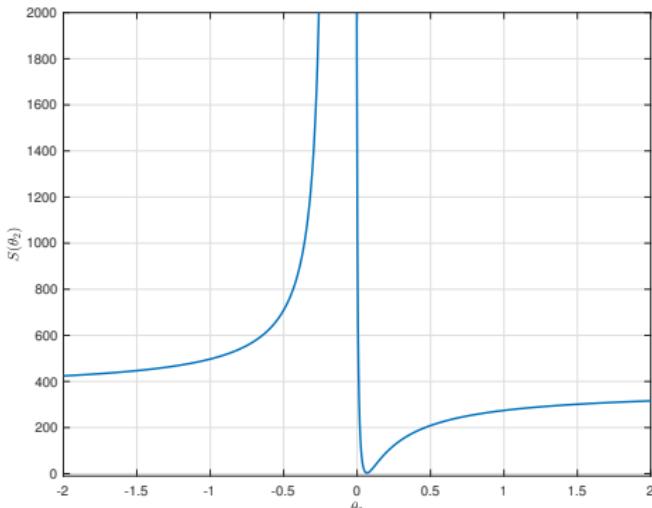
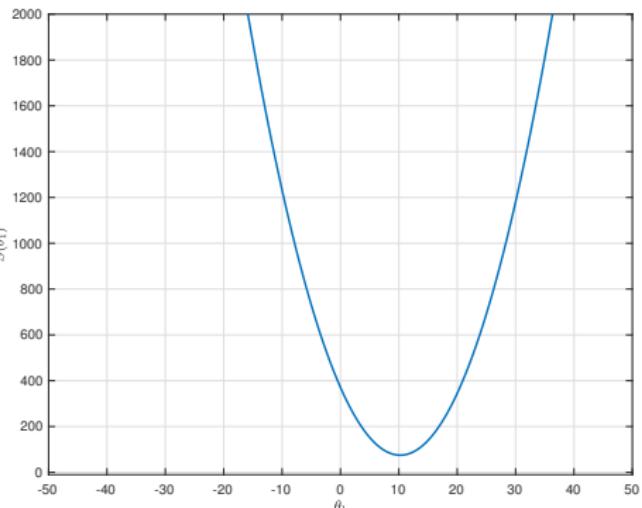
- You would like to find parameters for the Hougen-Watson function:

$$Y = \frac{(\theta_0 X_2 - X_3 / \theta_4)}{(1 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3)}$$

- Model captures effect of competitive reaction and adsorption rates, Y is production rate and X are species concentrations.
- Parameters θ have physical meaning (represent adsorption and reaction rates).
- We have $S = 13$ data pairs (y_ω, x_ω) available and $m = 5$ parameters to estimate.

Example: Hougen-Watson Reaction

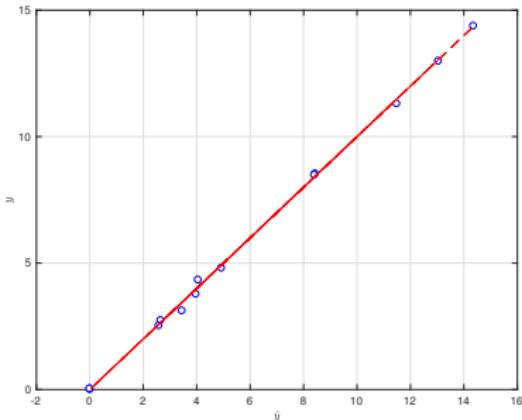
- The model function is highly nonlinear and this can result in high nonlinearity of $S(\theta)$.
- Below we show the behavior of $S(\theta)$ as we span a range for θ_1 and θ_2 .



- Note that, in nonlinear models, we might see emergence of maxima and minima in $S(\theta)$.

Example: Hougen-Watson Reaction

- Solution of problem with Matlab built-in function *fitnlm* gives following fit:



- The parameters 95% confidence intervals are given by:

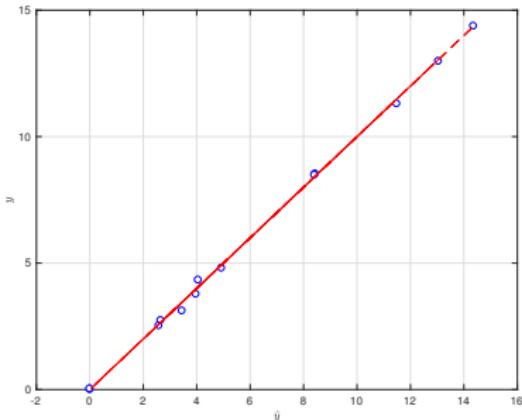
$$\theta_0 \in [-0.75, 3.25], \quad \theta_1 \in [-0.04, 0.16]$$

$$\theta_2 \in [-0.03, 0.11], \quad \theta_3 \in [-0.06, 0.29], \quad \theta_4 \in [-0.74, 3.12].$$

- What is wrong with these?

Example: Hougen-Watson Reaction

- Solution of problem with Matlab built-in function *fitnlm* gives following fit:



- The parameters 95% confidence intervals are given by:

$$\theta_0 \in [-0.75, 3.25], \quad \theta_1 \in [-0.04, 0.16]$$

$$\theta_2 \in [-0.03, 0.11], \quad \theta_3 \in [-0.06, 0.29], \quad \theta_4 \in [-0.74, 3.12].$$

- What is wrong with these?
- if we analyze output of Matlab, we realize that the algorithm does not report the Hessian matrix. Because of this, it is impossible to conclude if estimated parameters are unique.
- Check also what happens if you provide a bad initial guess to parameters.



Prior Knowledge

How can we incorporate prior (expert, physical) knowledge about parameters in the estimation problem?

- Prior knowledge help us eliminate spurious estimates $\hat{\theta}$ (e.g., avoid estimates with no physical meaning or large parameters).
- Prior knowledge help us reduce number of meaningful parameters or narrow space over we search on.

We can incorporate prior knowledge in estimation (a.k.a. regularization) by using:

- Bounds on parameters (e.g., kinetic parameters are positive):

$$\min_{\theta} \frac{1}{2} \|\mathbf{y} - \mathbf{m}(\theta)\|^2$$

$$\text{s.t. } \theta_L \leq \theta \leq \theta_U$$

- Constraints to fix parameters (e.g., sum of parameters must be equal to some value):

$$\min_{\theta} \frac{1}{2} \|\mathbf{y} - \mathbf{m}(\theta)\|^2$$

$$\text{s.t. } \Pi\theta = r$$



Prior Knowledge

- Penalty term to control parameter behavior (e.g., penalize movements away from reference value):

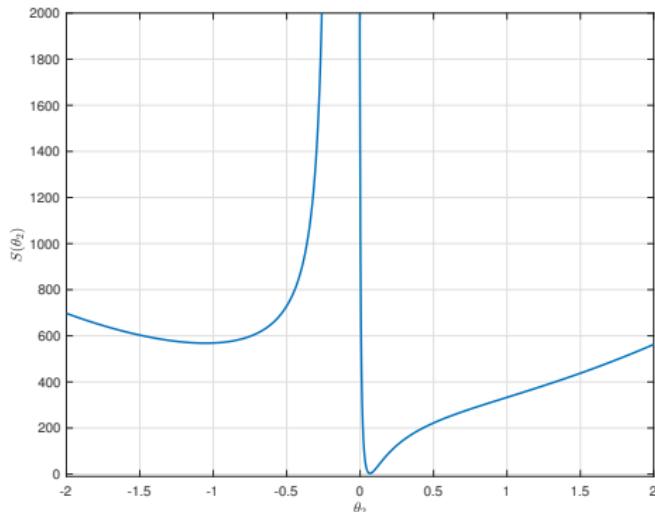
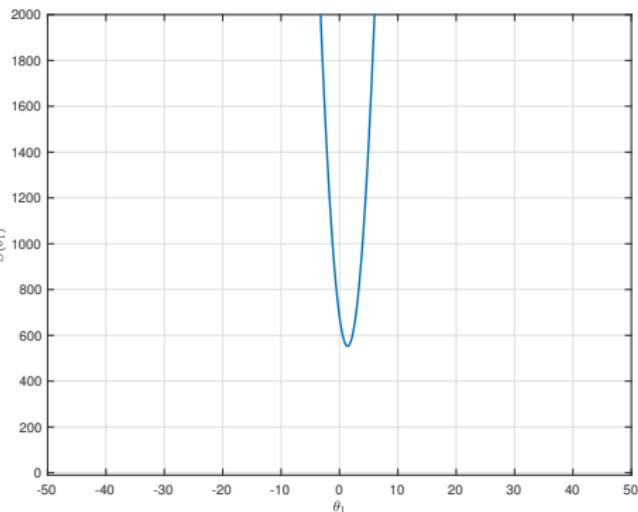
$$\min_{\theta} \frac{1}{2} \|\mathbf{y} - \mathbf{m}(\theta)\|^2 + \kappa \cdot \rho(\theta)$$

Where $\kappa \geq 0$ is constant that trades-off fit and allowed movement.

- Common choices for penalty function $\rho(\theta)$ are:
 - ℓ_2 norm (a.k.a ridge or Tikhonov penalty): $\rho(\theta) = \frac{1}{2} \|\theta - \bar{\theta}\|_2^2 = \frac{1}{2} \sum_{i=1}^m (\theta_i - \bar{\theta}_i)^2$
 - ℓ_1 norm (a.k.a. lasso penalty): $\rho(\theta) = \|\theta - \bar{\theta}\|_1 = \sum_{i=1}^m |\theta_i - \bar{\theta}_i|$
 - Bayes penalty: $\rho(\theta) = \frac{1}{2} (\theta - \bar{\theta})^T \Sigma_{\theta}^{-1} (\theta - \bar{\theta})$
- Penalties (a.k.a. regularizers) induce different behavior and some of them can be derived from statistical principles. There are many regularizers in the literature (seeking to embed different type of knowledge).
- Statistics help us to understand the type of *prior knowledge* that regularizers convey.
- Understanding statistical principles of constraints is a more difficult question but one can often reformulate constraints as penalty functions.

Example: Hougen-Watson Reaction

- Below we show the behavior of $S(\theta) + \frac{1}{2}(\theta - \bar{\theta})^T \Sigma_{\theta}^{-1}(\theta - \bar{\theta})$



- Bayes regularizer helps better define minima but in this case does not avoid presence of non-physical regions (this can only be achieved with constraints).
- The presence of non-physical regions can introduce numerical instability (nothing prevents solvers from visiting such regions).



Example: Hougen-Watson Reaction

- Instead, we filter out parameters with non-physical meaning by directly imposing constraints on parameters. We formulate problem:

$$\begin{aligned} \min_{\theta} \frac{1}{2} \|\mathbf{y} - \mathbf{m}(\theta)\|^2 \\ \text{s.t. } \theta_L \leq \theta \leq \theta_U \end{aligned}$$

where

$$m_\omega(\theta) = \frac{(\theta_0 x_{2,\omega} - x_{3,\omega}/\theta_4)}{(1 + \theta_1 x_{1,\omega} + \theta_2 x_{2,\omega} + \theta_3 x_{3,\omega})}$$

and we set $\theta_L = (0, 0, 0, 0, 0)$ and $\theta_U = (2, 2, 2, 2, 2)$.

- This problem can be solved using Matlab's built-in tool lsqnonlin.
- This tool also provides $\nabla_{\theta} m(\theta)$, which we can use to approximate $H(\theta) \approx \nabla_{\theta} m(\theta)^T \nabla_{\theta} m(\theta)$.



Bayesian Estimation

Bayes theorem provides a solid statistical basis to derive a wide range of estimation formulations (e.g., that include prior knowledge).

- In the context of our estimation problem of interest, Bayes theorem states that:

$$f(\theta|y) = \frac{f(y|\theta)f(\theta)}{f(y)}$$

- $f(\theta|y)$ is probability that parameters take value θ given knowledge that output takes value y (a.k.a. posterior pdf)
- $f(y|\theta)$ is probability that output takes value y given knowledge that parameters take value θ
- $f(\theta)$ is marginal probability of parameters (a.k.a prior pdf)
- $f(y)$ is marginal probability of outputs (this is irrelevant as it does not carry knowledge on parameters)



Bayesian Estimation

- Goal in Bayesian estimation is to maximize probability of parameters $f(\theta|y)$ (not of outputs as in MLE). Bayes theorem tells us that:

$$f(\theta|y) \propto f(y|\theta)f(\theta)$$

- This approach carries prior knowledge of θ . Recall that in MLE we find estimate $\hat{\theta}$ that maximizes $f(y|\theta)$ (it is assumed that θ is deterministic).
- We thus find estimate $\hat{\theta}$ by solving:

$$\max_{\theta} \log f(y|\theta) + \log f(\theta)$$

This problem is equivalent to that of MLE but we incorporate prior term $\log f(\theta)$.

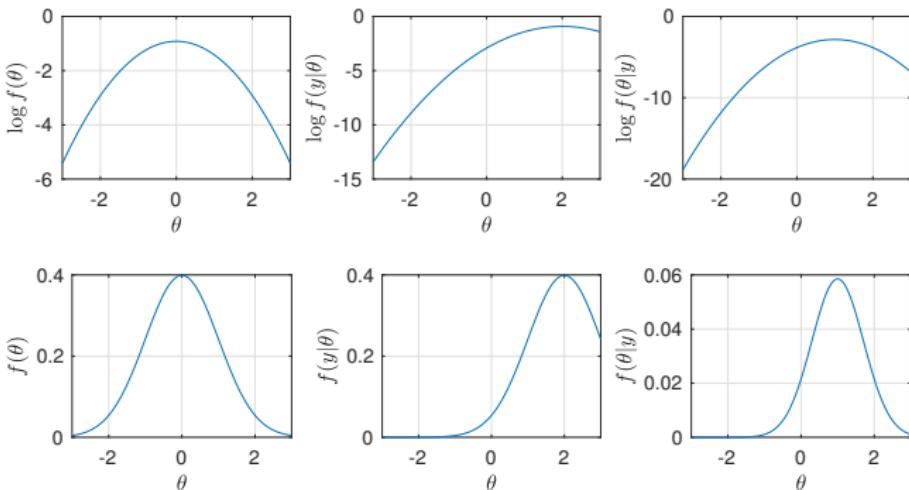
- If we have prior knowledge that $\theta \sim \mathcal{N}(\bar{\theta}, \Sigma_\theta)$, then:

$$\min_{\theta} \frac{1}{2}\|y - m(\theta)\|^2 + \frac{\kappa}{2}(\theta - \bar{\theta})\Sigma_\theta^{-1}(\theta - \bar{\theta})$$

- Gaussian prior $f(\theta)$ achieves its maximum at $\theta = \bar{\theta}$ while $f(y|\theta)$ achieves its maximum when θ fits data.
- Estimation problem seeks to find *balance* between what we previously knew about θ and new knowledge gained through observations y .
- If ignore prior knowledge, all that we know about θ is through y (which might lead to ambiguity if data is insufficient).
- Adding prior knowledge reduces *ambiguity*.

Example: Bayesian Estimation

- Consider estimation problem for model $Y = \theta X$ and assume single observation (x, y)
- Evidence provided by prior knowledge is in $\mathcal{N}(0, 1)$ with density $f(\theta)$
- Evidence provided by data (x, y) is $\mathcal{N}(y - \theta x, 2)$ with density $f(y|\theta)$
- Combined evidence is density $f(\theta|y) \propto f(y|\theta)f(\theta)$
- Below we show the pdfs in log and original scale for the pdfs



- According to prior evidence $f(\theta)$, $\theta = 1$ is most likely value
- According to data evidence $f(y|\theta)$, $\theta = 2$ is most likely value
- According to combined evidence $f(\theta|y)$, $\theta = 1$ is most likely value



Outline

- ① Introduction
- ② Models of Random Variables
- ③ Estimation Techniques
- ④ Multivariate Statistics
- ⑤ Data-Driven Modeling
- ⑥ Statistical Learning**
- ⑦ Decision-Making Under Uncertainty



Statistical Learning

Machine learning is fast growing field that combines techniques from diverse branches of science and engineering to perform tasks such as:

- Data Analysis (e.g., dimension reduction, clustering, visualization, signal recognition, computer vision)
- Data-Driven Modeling (e.g., neural nets, kriging, support vector machines)
- Artificial Intelligence (e.g., data collection, experimentation, learning, control)

Statistical learning is a subset of machine learning that provides tools derived from *statistical principles*.

- Some tools of machine learning are derived from other mathematical principles (e.g., geometry, topology, optimization, linear algebra).
- Our focus here is not to provide an extensive review of all tools. Instead, we focus on general statistical principles behind such tools.



Data Analysis

- How can I interpret and extract knowledge from high-dimensional data?
- How can I reduce (compress) my data?



Analysis of Multivariate Data

Consider the following problem:

- You have multiple input RVs $X = (X_1, X_2, \dots, X_n)$ entering a system.
- You want to create a product that is a *mixture* (blend) of these RVs:

$$t = \sum_{i=1}^n w_i X_i = w^T X$$

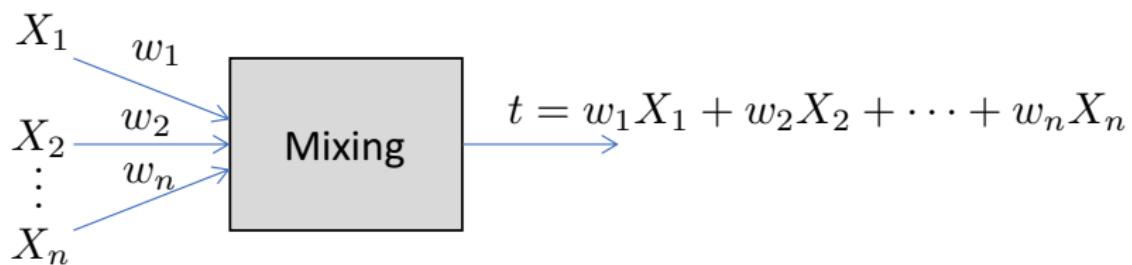
where $w_i \in \mathbb{R}$ are mixture proportions satisfying $\|w\| = 1$, with $w = (w_1, w_2, \dots, w_n)$.

- What proportions w give a product that contains maximum information about X ?
- What proportions w give a product that contains the second most information about X ?
- What proportions w give a product that contains minimum information about X ?

Think about analogy of this problem with a physical blending (mixing) process:

We want to mix a set of input flows in a way that product is as valuable as possible.

Analysis of Multivariate Data





Analysis of Multivariate Data

Data mixing problem can be solved using *principal component analysis* (PCA).

- Collect observations for $x_\omega \in \mathbb{R}^n$, $\omega \in 1, \dots, S$ for the RV $X = (X_1, X_2, \dots, X_n)$.
- Store the observations in $S \times n$ data matrix:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{S,1} & x_{S,2} & \cdots & x_{S,n} \end{bmatrix}$$

- Normalize columns in such a way that $\frac{1}{S} \sum_{i=1}^S \mathbf{X}_{i,j} = 0$ for all $j = 1, \dots, n$. This centers the data around zero.
- After normalization, the sample covariance matrix of X is $\text{Cov}(X) = \mathbf{X}^T \mathbf{X}$ (denote this as Σ and note this is $n \times n$).
- Covariance matrix contains all information (variance) of inputs X .



Analysis of Multivariate Data

- For mixture $t = w^T X$ we can show that $\hat{\mathbb{V}}[t] = w^T \Sigma w$.
- Consequently, the mix proportions that contain max variance of X can be found as:

$$\max_w w^T \Sigma w \text{ s.t. } \|w\| = 1$$

- Solution of this problem is eigenvector w_1 of Σ associated with largest eigenvalue λ_1 . Consequently, proportions w_1 give mixture $t_1 = w_1^T X$ that contains maximum information about X .
- The mixture that contains minimum information about X is found by solving:

$$\min_w w^T \Sigma w \text{ s.t. } \|w\| = 1$$

this gives eigenvector w_n associated with smallest eigenvalue $\lambda_n = w_n^T \Sigma w_n$

- We can find and rank mixtures based on information content by performing an eigendecomposition:

$$\Sigma = W \Lambda W^T = \lambda_1 w_1 w_1^T + \lambda_2 w_2 w_2^T + \cdots + \lambda_n w_n w_n^T$$

- Eigenvalues are arranged as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Consequently, mix $t_1 = w_1^T X$ contains most info while $t_2 = w_2^T X$ contains second most info and so on.



Analysis of Multivariate Data

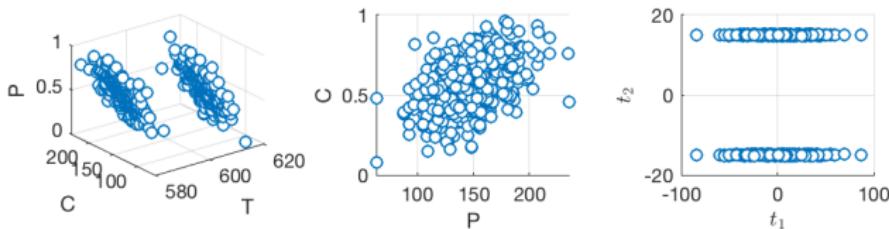
- Mixtures (a.k.a. principal components) t_1, t_2, \dots, t_n contain information about X . Consequently, select a few of them (typically two or three) to visualize data X in a small dimensional space (PCA is a dimensionality reduction technique).
- Truncating the eigendecomposition series enables compression of data matrix Σ .
- Key property of principal components is that they retain structure of high-dimensional data. Visualizing data by dropping variables destroys original structure.
- We can show that principal components are uncorrelated and thus $\text{Cov}(t_i, t_j) = 0$ for all $i \neq j$ (i.e., mixtures contain complementary types of information).
- We can use eigenvector matrix $W = [w_1 | w_2 | \cdots | w_n] \in \mathbb{R}^{n \times n}$, to can project data matrix \mathbf{X} to the principal component (information) space as:

$$\mathbf{T} = \mathbf{X}W$$

where $\mathbf{T} \in \mathbb{R}^{S \times n}$ is a matrix with entries $t_{i,j}$, $i = 1, \dots, S$ $j = 1, \dots, n$.

Example: Gibbs Reactor

- Have data set with pressure, temperature, and conversion ($P_\omega, T_\omega, C_\omega$).
- If we visualize data in a 3D (P, T, C) space, there are clearly two clusters revealed (one is reactor operating under failure mode and one under normal mode).
- Imagine we drop T data because we believe this is unimportant and visualize (P, C) space alone. We can see that the failure clusters cannot be spotted. This illustrated how improper reduction of data can miss significant information.
- By reducing data using PCA projection, we retain failure structure and can visualize it in 2D space (t_1, t_2) .



- The eigenvalues obtained with PCA are $\lambda = (3.19 \times 10^5, 1.12 \times 10^5, 4)$.
- This indicates that one of the principal components carries small information (there is a redundant dimension). This is because of the high correlation between P and C (i.e., if you know P , you know C and viceversa).



Singular Value Decomposition

- PCA decomposes covariance $\Sigma = \mathbf{X}^T \mathbf{X}$ by using eigendecomposition $\Sigma = W \Lambda W^T$
- One can also implement PCA by decomposing the data matrix $\mathbf{X} \in \mathbb{R}^{S \times n}$ directly.
This is done using a technique called singular value decomposition (SVD):

$$\mathbf{X} = U S V^T$$

- Here, $U \in \mathbb{R}^{S \times S}$ is left eigenvector matrix, $V \in \mathbb{R}^{n \times n}$ is the right eigenvector matrix and $S \in \mathbb{R}^{S \times n}$ is the singular value matrix (a diagonal matrix).
- Matrices U, V are orthogonal and thus satisfy $U^T U = I, UV = 0$, and $V^T V = I$:

$$\begin{aligned}\mathbf{X}^T \mathbf{X} &= (U S V^T)^T (U S V^T) \\ &= V S^T S V^T\end{aligned}$$

- By defining $W = V$ and $\Lambda = S^T S$ we obtain standard PCA.
- The SVD of \mathbf{X} can be written as an expansion of the form:

$$\mathbf{X} = \sum_{j=1}^n s_j u_j v_j^T$$

- where $s_j = S_{j,j}$ and u_j and v_j are the columns of U and V , respectively.
- Why would you prefer to decompose \mathbf{X} over decomposing Σ ?

Example: Image Compression

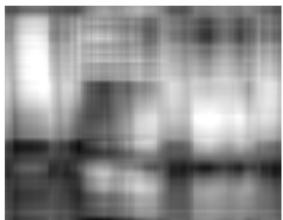
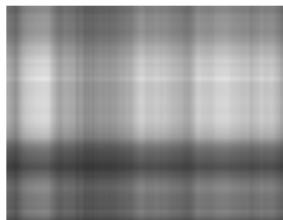
- An image is a matrix \mathbf{X} (each entry represents a pixel and the number the intensity)
- We decompose the image using SVD as $\mathbf{X} = \sum_{j=1}^n s_j u_j v_j^T$ (elements of the series are arranged in information content of the image).
- Large components tend to contain broad (global) features of the image while small components contain granular (local) features.
- We can truncate the series to compress an image (i.e., ignore granular behavior).



- This image is a matrix \mathbf{X} of dimension 747×941

Example: Image Compression

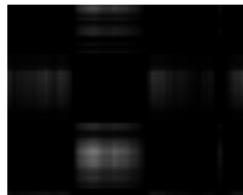
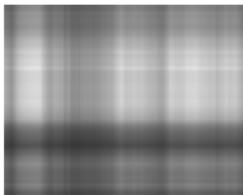
- Approximate images (matrices) obtained using truncated series $\mathbf{X}_k = \sum_{j=1}^k s_j u_j v_j^T$ for $k = 1, 3, 10, 100$:



- Note how coarse features develop quickly while granular develop slowly
- Each element $s_j u_j v_j^T$ encodes different “features” of the image
- Nearly perfect image obtained with $k = 100$ (13% of original data)
- Truncating series cuts down non-informative features

Example: Image Compression

- Here are first nine elements $s_j u_j v_j^T$ of series $\mathbf{X} = \sum_{j=1}^n s_j u_j v_j^T$.





Analysis of Space-Time Data

- PCA extract patterns from observations x_ω , $\omega \in \mathcal{S}$ and assumes them to be independent
- But what if there is correlation between snapshots x_ω and $x_{\omega'}$?
- For instance, what if there is a temporal correlation between snapshots (i.e., x_ω , $x_{\omega+1}$ are correlated)?
- Consider $x_\omega \in \mathbb{R}^n$ are observations at time t_ω and consider mixture:

$$t_\omega = w^T x_\omega, \quad \omega \in \mathcal{S}$$

- This forms a time series t_1, t_2, \dots, t_S and we assume temporal model of the form:

$$\hat{t}_k = \beta_1 t_{k-1} + \beta_2 t_k + \cdots + \beta_r t_{k-r}$$

- This is known as an autoregressive time-series model (a whole topic on its own).



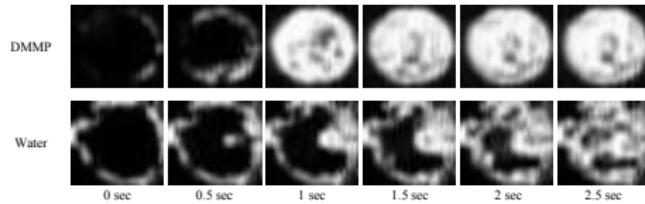
Analysis of Space-Time Data

- Our goal is to find parameters w and β such that t_ω capture most information of x_ω .
- This can be done by solving the following optimization problem:

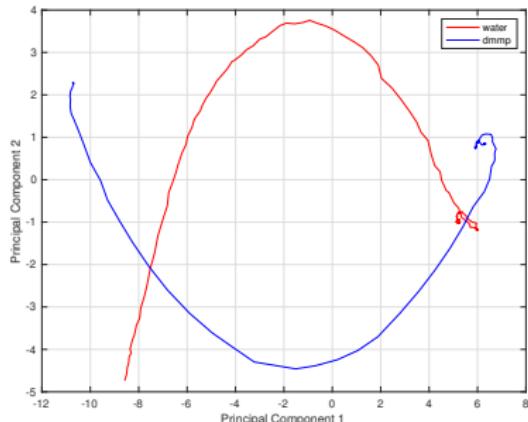
$$\begin{aligned} \max_{w, \beta} \quad & \sum_{\omega=1}^S t_k \hat{t}_\omega \\ \text{s.t.} \quad & t_\omega = w^T x_\omega, \omega = 1, \dots, S \\ & \hat{t}_\omega = \beta_1 t_{\omega-1} + \beta_2 t_\omega + \dots + \beta_r t_{\omega-r}, \omega = 1, \dots, S \end{aligned}$$

- This problem can be interpreted as finding w, β that maximize predictability of \hat{t}_ω given $t_{\omega-1}, \dots, t_{\omega-S}$.
- One can show that the above problem is a special type of an eigenvalue problem that captures static and temporal variance of x_ω .
- The time series \hat{t}_ω , $\omega \in \mathcal{S}$ thus contains maximum information about x_ω (these are known as dynamic principal components).
- As in PCA, we find a time series that contains second most information and so on.
- This provides powerful approach to visualize dynamics of high-dimensional data.

Example: Dynamic Image Analysis



- Chemical sensor responds to different contaminants, seek to spot differences in space-dynamic response to identify contaminant type
- Each observation x_ω is an image and we have a sequence $\omega = 1, \dots, S$
- Below we show dynamics for first two principal components extracted from data





Data-Driven Modeling (Classification)

Consider following problem:

- Imagine you have input RVs $X = (X_1, X_2, \dots, X_n)$ with observations $x_\omega \in \mathbb{R}^n$, $\omega \in \mathcal{S}$ and domain $\mathcal{D} = \mathbb{R}^n$ (X can be discrete or continuous) and an output RV Y with observations $y_\omega \in \mathbb{R}$ with domain $\mathcal{D}_Y = \{0, 1\}$ (discrete).
- We postulate that there exists a relationship between X and Y of the form:

$$Y = g(X, \theta)$$

where $\theta \in \mathbb{R}^n$ are parameters and $g(X, \theta)$ is model.

- Our objective is to find model that best explains observations:

$$y_\omega = g(x_\omega, \theta) + \epsilon_\omega$$

- This estimation problem is known as a *classification* problem. In this context, inputs X are known as features or descriptors while Y are known as labels or classes.
- What is different (and difficult) here is the binary nature of output Y . In standard estimation problems, Y is assumed to be continuous.

Think about analogy of this problem with a chemical classification problem:

Given a set of features for a chemical, can we predict if this is toxic or not?



Classification

- To solve problem, we postulate a model function of the form:

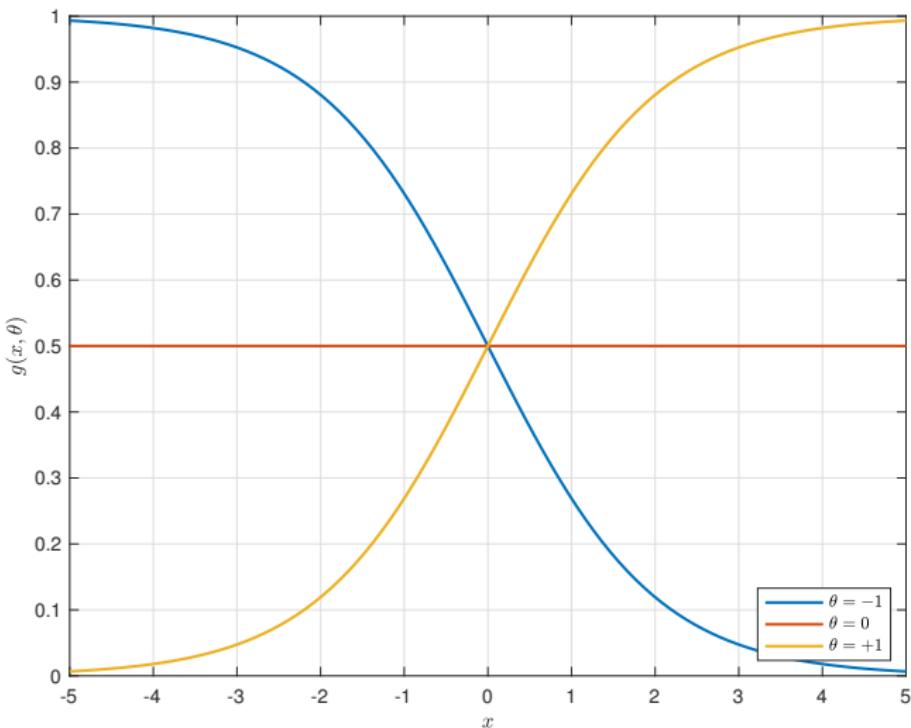
$$g(x, \theta) = \frac{1}{1 + e^{-\theta^T x}}$$

- Here, the mix $\theta^T x = \sum_{i=1}^n \theta_i x_i$ is known as *evidence*. The parameters reflect weights that we place on different features.
- Postulated model is known as the logistic function and seeks to capture 0-1 logic.
- Logistic function is a sigmoidal function that satisfies $g(x, \theta) \rightarrow 1$ as $\theta^T x \rightarrow \infty$, $g(x, \theta) \rightarrow 0$ as $\theta^T x \rightarrow -\infty$, and $g(x, \theta) = \frac{1}{2}$ if $\theta^T x = 0$.



Classification

Here is a visualization of the logistic function for different values of θ and for scalar x .





Classification

- Logistic function $g(x, \theta)$ can be used to model probabilities $\mathbf{P}(Y = 1 | x, \theta)$ and $\mathbf{P}(Y = 0 | x, \theta) = 1 - \mathbf{P}(Y = 1 | x, \theta)$:
 - If evidence is strongly positive, there is a high probability that $Y = 1$
 - If evidence is strongly negative, there is a high probability that $Y = 0$
 - If evidence is weak, there is ambiguity (it is equally probable that $Y = 0$ or $Y = 1$)
- This logic can be captured by defining a conditional pdf of the form:

$$f(y|x, \theta) = \mathbb{P}(Y = y | x, \theta) = g(x, \theta)^y (1 - g(x, \theta))^{1-y}$$

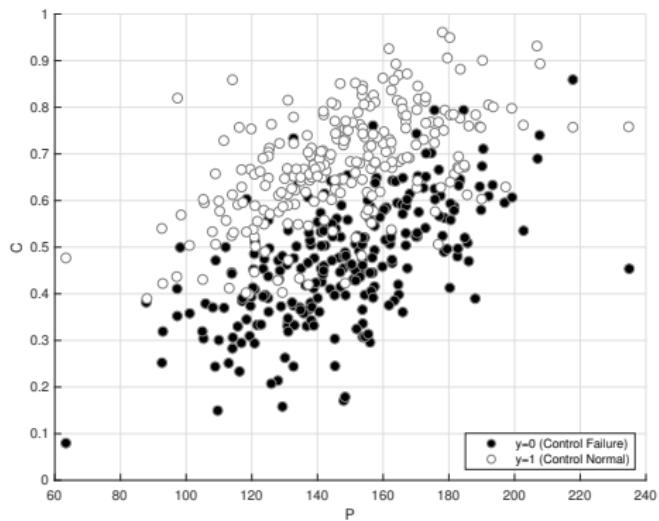
- Goal is to find estimate $\hat{\theta}$ that maximizes joint probability $\prod_{\omega \in S} f(y_\omega, |x_\omega, \theta)$:

$$\max_{\theta} \sum_{\omega \in S} y_\omega \log(g(x_\omega, \theta)) + (1 - y_\omega) \log(1 - g(x_\omega, \theta))$$

- Having estimate $\hat{\theta}$ we use model $g(x, \hat{\theta})$ to predict class of input with features x .
- This approach does not minimize sum of squared errors (and is standard estimation). The objective function is called the *log loss*.

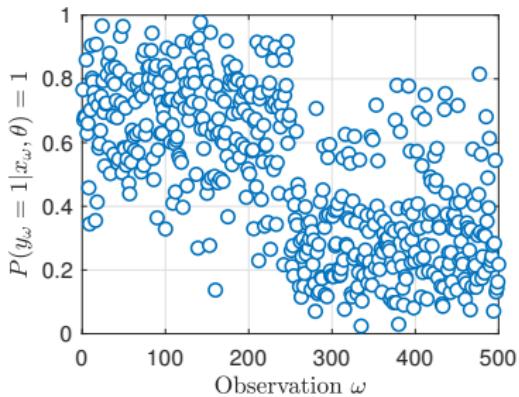
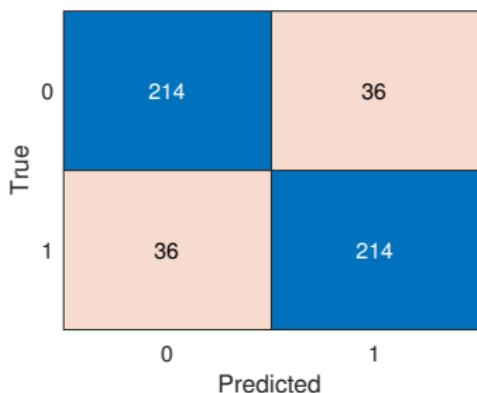
Example: Gibbs Reactor

- Have data $(P_\omega, C_\omega, T_\omega)$ for reactor.
- Reactor is in normal mode when T_ω is low ($y_\omega = 1$) and in failure mode when T_ω is high ($y_\omega = 0$).
- We seek to develop a model that can predict failure mode by using data $x_\omega = (P_\omega, C_\omega)$ alone.
- Data tells us that this might be achievable due to strong cluster separation.



Example: Gibbs Reactor

- After training, we find that model correctly classifies failure 85% of the time.
- Logistic classification also gives us an estimate of the probability of an observation to be in a given class and which observations are in the “fuzzy” region.
- If data noise is increased, cluster separation is not as strong and accuracy decreases.





Data-Driven Modeling (Kernel Methods)

- A flexible way of constructing models is to *mix* different types of models.
- Assume that Y and $X = (X_1, \dots, X_n)$ follow a relationship of the form:

$$Y = \sum_{j=1}^m \theta_j \phi_j(X) = \theta^T \phi(X)$$

- $\phi_j(X)$ $j = 1, \dots, m$ is collection of basic models (a.k.a. basis functions). We define vector $\phi(X) = (\phi_1(X), \dots, \phi_m(X))$.
- Basis function $\phi_j(X)$ can be nonlinear (e.g., polynomial, sigmoidal, exponential) or linear (in which case $\phi(X) = X$).
- Parameters $\theta_j \in \mathbb{R}$ are mixing coefficients of basis functions.



Data-Driven Modeling (Kernel Methods)

- We determine estimate $\hat{\theta}$ by solving regularized MLE problem:

$$\min_{\theta} S(\theta) = \frac{1}{2} \sum_{\omega \in \mathcal{S}} (y_{\omega} - \theta^T \phi(x_{\omega}))^2 + \lambda \theta^T \theta$$

In vector form:

$$\min_{\theta} S(\theta) = \frac{1}{2} (\Phi \theta - \mathbf{y})^T (\Phi \theta - \mathbf{y}) + \lambda \theta^T \theta$$

where $\Phi \in \mathbb{R}^{S \times n}$ is input data matrix with entries $\Phi_{\omega,j} = \phi_j(x_{\omega})$ and \mathbf{y} is output data vector.

- If basis functions are linear then $\Phi = X$.



Kernel Methods

- Optimality conditions indicate that best estimate satisfies:

$$\theta = \frac{1}{\lambda} \Phi^T \mathbf{r}$$

where we define residuals (mismatch errors) $\mathbf{r} = (\Phi\theta - \mathbf{y})$. By substituting in $S(\theta)$ we obtain:

$$S(\theta) = \frac{1}{2} \mathbf{r}^T \Phi \Phi^T \Phi \Phi^T \mathbf{r} - \mathbf{r}^T \Phi \Phi^T \mathbf{y} + \frac{1}{2} \mathbf{y}^T \mathbf{y} + \frac{\lambda}{2} \mathbf{r}^T \Phi \Phi^T \mathbf{r}$$

We define matrix $K = \Phi \Phi^T \in \mathbb{R}^{S \times S}$ and simplify:

$$S(\theta) = \frac{1}{2} \mathbf{r}^T K K \mathbf{r} - \mathbf{r}^T K \mathbf{y} + \frac{1}{2} \mathbf{y}^T \mathbf{y} + \frac{\lambda}{2} \mathbf{r}^T K \mathbf{r}$$

- Note that function $S(\theta)$ can be defined entirely in terms of \mathbf{r} .



Kernel Methods

- This suggests an alternative strategy to find estimate. Here, we solve problem:

$$\min_{\mathbf{r}} \frac{1}{2} \mathbf{r}^T K K \mathbf{r} - \mathbf{r}^T K \mathbf{y} + \frac{1}{2} \mathbf{y}^T \mathbf{y} + \frac{\lambda}{2} \mathbf{r}^T K \mathbf{r}$$

to find an optimal $\hat{\mathbf{r}}$ and then recover $\hat{\theta} = \Phi^T \hat{\mathbf{r}}$.

- But it turns out that parameters $\hat{\theta}$ are *not needed at all*. To see this, note that $\hat{\mathbf{r}}$ satisfies $\mathbf{r} = (K + \lambda I)^{-1} \mathbf{y}$ and thus optimal model prediction is:

$$\begin{aligned}\hat{\mathbf{y}} &= K \hat{\mathbf{r}} \\ &= K(K + \lambda I)^{-1} \mathbf{y}\end{aligned}$$

- Optimal prediction only depends on input data (in K) and output data (in \mathbf{y}).
- This implies that parameters θ are not needed (these are just intermediary variables).



Kernel Methods

- Matrix K is known as kernel matrix, which captures interactions in input variables.
- Given input data x_ω , winS , a kernel matrix can be constructed as:

$$K_{i,j} = k(x_i, x_j) \quad i, j \in \mathcal{S}$$

where $k(x_i, x_j)$ is known as the *kernel function*. This provides a general approach for estimation.

- The case for linear estimation corresponds to defining kernel function:

$$k(x_i, x_j) = x_i^T x_j$$

In this case, note that $K = \text{Cov}(X)$ (after normalization).

- The case of basis functions corresponds to defining a kernel function:

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

- Typically, kernel function is chosen to be the radial basis function (RBF):

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$$

where γ is a hyperparameter. This kernel is also known as the Gaussian kernel.



Kernel Methods

The RBF kernel is simple but captures a wide range of nonlinear behavior. To see this:

Consider scalar case (x_ω is a scalar) and notice that:

$$\begin{aligned}\exp(-\gamma(x_i - x_j)^2) &= \exp(-\gamma x_i^2 + 2\gamma x_i x_j - \gamma x_j^2) \\ &= \exp(-\gamma x_i^2 - \gamma x_j^2) \left(1 + \frac{2\gamma x_i x_j}{1!} + \frac{(2\gamma x_i x_j)^2}{2!} + \dots\right)\end{aligned}$$

We thus have that radial basis function can be written as:

$$\exp(-\gamma(x_i - x_j)^2) = \phi(x_i)^T \phi(x_j)$$

where $\phi(x)$ is an *infinite* collection of polynomial basis functions:

$$\phi(x) = e^{-\gamma x} (1, \sqrt{2\gamma/1!}, \sqrt{(2\gamma)^2/2!}, \sqrt{(2\gamma)^3/3!}, \dots).$$

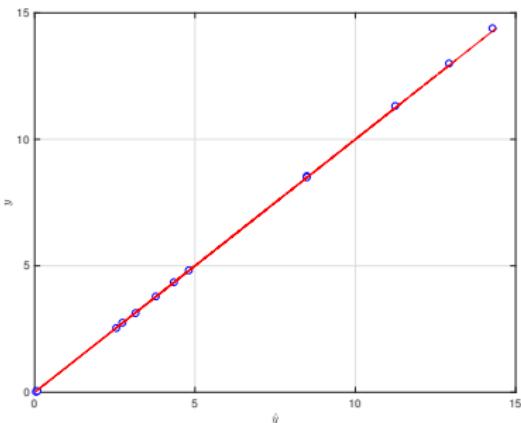
- In kernel methods we do not need to postulate an input-output model and estimate its parameters θ .
- Instead, we postulate a kernel function and estimate its hyperparameters (e.g., γ).
- The number of parameters of a kernel function is often small (typically less than ten). This is a key advantage over standard estimation approaches.

Example: Hougen-Watson Reaction

- Recall the mechanistic model is highly nonlinear and has a total of five parameters:

$$Y = \frac{(\theta_0 X_2 - X_3 / \theta_4)}{(1 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3)}$$

- Instead of using this model, we directly use data (x_ω, y_ω) to construct model.
- We use Gaussian kernel $k(x_\omega, x_{\omega'}) = \exp(-\gamma \|x_\omega - x_{\omega'}\|)$ with $\gamma = 0.01$.
- Note that this model only requires one parameter (γ).
- Below we show the fit of the data using the kernel model.





Data-Driven Modeling (Kriging)

How do we estimate kernel function hyperparameters?

This can be done by using a technique called kriging.

- In kriging we postulate a model of the form:

$$y_\omega = g(x_\omega) + \epsilon_\omega, \omega \in \mathcal{S}$$

In vector form:

$$\mathbf{y} = \mathbf{g} + \boldsymbol{\epsilon}$$

- If we assume that $\epsilon_\omega \sim \mathcal{N}(0, \sigma)$ then we have that $f(\mathbf{y}|\mathbf{g})$ corresponds to $\mathcal{N}(\mathbf{g}, \sigma\mathbb{I})$.
- Note the absence of parameters in postulated model.



Kriging

- What is unique about kriging is that \mathbf{g} is treated as a random function (non-parametric).
- In the techniques that we have covered, we defined parametric functions $\mathbf{g}(\theta)$.
- We assume that random function \mathbf{g} has pdf $f(\mathbf{g}|\gamma)$ corresponding to $\mathcal{N}(0, K(X, \gamma))$.
- Here, $\text{Cov}(\mathbf{g}) = K(\gamma)$ is a covariance function and γ are its hyperparameters.
- Think about $K(\gamma)$ as a function that generates samples of the random model \mathbf{g} . This matrix has entries:

$$K_{i,j}(\gamma) = k(x_i, x_j, \gamma) \quad (3)$$

where $k(x_i, x_j, \gamma)$ is a kernel function.



Kriging

- One can show that marginal of \mathbf{y} is:

$$f(\mathbf{y}|\gamma) = \int f(\mathbf{y}|\mathbf{g})f(\mathbf{g}|\gamma)d\mathbf{g}$$

- This corresponds to pdf of $\mathcal{N}(0, C(\gamma))$ with entries:

$$C(x_i, x_j, \gamma) = K_{i,j}(\gamma) + \sigma\delta_{i,j} \quad i, j \in \mathcal{S}$$

where $\delta_{i,j} = 1$ if $x_i = x_j$ and zero otherwise.

- One approach to estimate $\hat{\gamma}$ consists of maximizing $\log f(\mathbf{y}|\gamma)$:

$$\max_{\gamma} -\frac{1}{2} \log |C(\gamma)| - \frac{1}{2} \mathbf{y}^T C(\gamma) \mathbf{y} - \frac{S}{2} \log 2\pi$$

- As a kernel method, in kriging one computes predictions by using kernel function (which only depends on $\hat{\gamma}$ and input data).
- This is done by using conditional pdf $f(y|\mathbf{y})$ where y is predicted output at new point x and \mathbf{y} are observations used to determine $\hat{\gamma}$.



Kriging

- One can show that the conditional pdf $f(y|\mathbf{y})$ is Gaussian $\mathcal{N}(m(x), \sigma^2(x))$ with:

$$m(x) = \mathbf{k}^T C^{-1} \mathbf{y}$$

$$\sigma^2(x) = c - \mathbf{k}^T C \mathbf{k}$$

and:

$$C = C(\hat{\gamma})$$

$$c = k(x, x, \hat{\gamma})$$

$$\mathbf{k}_i = k(x_i, x, \hat{\gamma}), i = 1, \dots, S$$

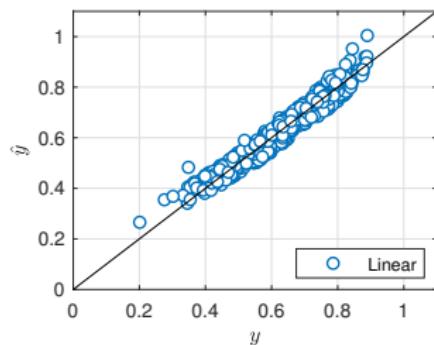
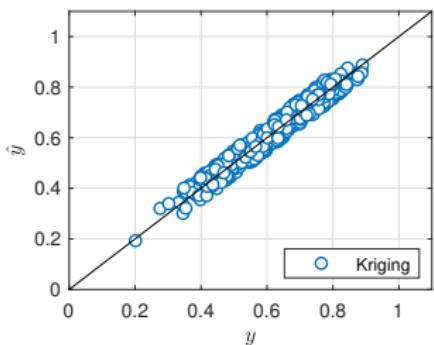
- Mean prediction of y is $m(x)$ and confidence intervals can be constructed using $\sigma^2(x)$.
- There is a wide range of kernel functions that can be used in kriging. The generalized Gaussian kernel resembles the radial basis function and takes the form:

$$k(x_i, x_k, \gamma) = \gamma_0 \exp\left(-\frac{\gamma_1}{2} \|x_i - x_k\|^2\right) + \gamma_2$$

- It is possible to combine parametric and non-parametric estimation within kriging. This can be done by defining $f(\mathbf{y}|\mathbf{g}, \theta)$ as $\mathcal{N}(\mathbf{g} + \Phi\theta, \sigma\mathbb{I})$ where Φ is an input data matrix with entries $\Phi_{\omega,j} = \phi_j(x_\omega)$ and $\phi_j(x)$ is a collection of basis functions.

Example: Gibbs Reactor

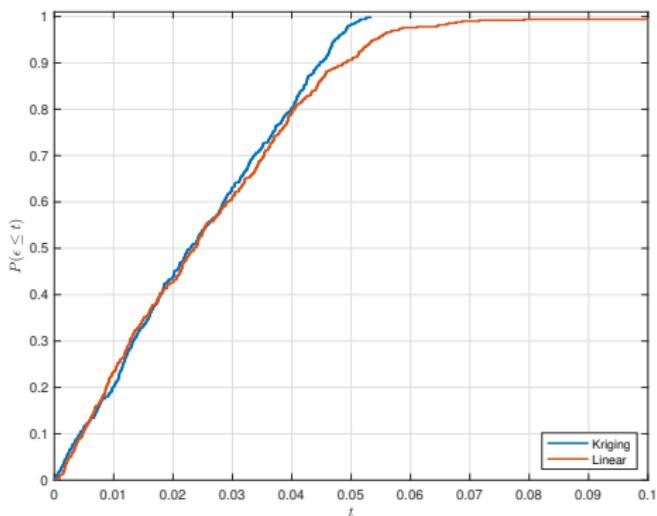
- We compare the performance of kriging (nonlinear) and linear models in Gibbs reactor
- We construct models of the form $Y = g(X, \theta)$ with inputs $X = (P, T)$ and output $Y = C$
- Below we show the fit obtained with both models



- From these it is rather obvious that the kriging model can predict better over the entire domain
- Linear model predicts accurately in the middle of the domain but fails on the extremes (where nonlinearity is more marked)

Example: Gibbs Reactor

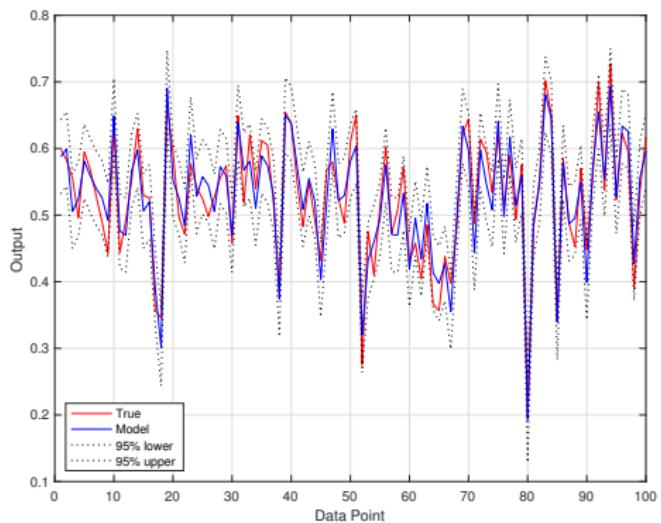
- An interesting and effective way to benchmark models is to look at the distributions of their residuals
- Below we show empirical cdfs for residuals $|\epsilon| = |y - \hat{y}|$ for kriging and linear models



- Residuals of kriging never exceed a threshold of $t = 0.06$ while those of linear reach levels of $t = 0.1$
- How do you interpret the above graphs?

Example: Gibbs Reactor

- An important feature of kriging is that it provides uncertainty information on predictions





Data-Driven Modeling (Neural Networks)

How do humans learn? How do they establish corrections between variables to make predictions and decisions?

- Neural networks (NNs) are nature-inspired parametric models that seek to establish causal models of between $X = (X_1, \dots, X_n)$ and Y of the form:

$$Y = g(X, \theta)$$

- What is different about NNs is that the model function $g(X, \theta)$ is automatically constructed by using a set of activation (basis) functions that are mixed (combined) in a hierarchical manner. This seeks to mimic how the brain works.
- NNs provide a flexible approach to capture virtually any type of relationship between X and Y . A key advantage of this is that we do not need to postulate a model (e.g., nonlinear, linear, logistic, mechanistic).
- As with estimation and classification the objective is to build the NN model (the parameters θ) that best explain the observations:

$$y_\omega = g(x_\omega, \theta) + \epsilon_\omega, \quad \omega \in \mathcal{S}$$

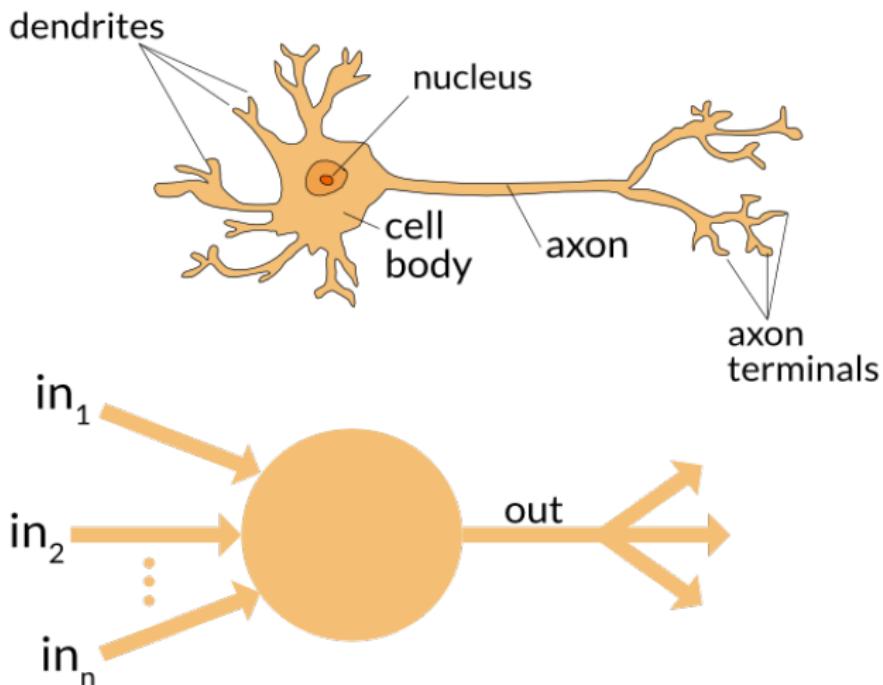
- The determination of θ from data mimics *learning* process of a human.



Neural Networks

- NNs seek to mimic how brain responds when exposed to data and how we learn and accumulate knowledge.
- Brain is a highly sophisticated network that has neurons as basic processing (decision) units that interact with one another through signals.
- Mathematical description of a neuron is called a *perceptron*. Perceptron generates a signal if evidence $\sum_{i=1}^n \theta_i x_i + \theta_0$ is strong enough or does not generate signal if evidence is not strong.
- Binary behavior of perceptron is similar to that logistic classification. In fact, perceptrons are often modeled using logistic functions. The additional parameter θ_0 in the evidence is called the *bias*.
- At basic level, a perceptron is a decision unit that decides to "fire" or not when exposed to data (evidence). How to respond to different pieces of evidence is captured by parameters.
- The bias captures situations in which we make decisions by taking into account not only evidence but also our inherent biases.
- An NN is a hierarchical architecture of perceptrons that constructs *complex logic*.

Neural Networks





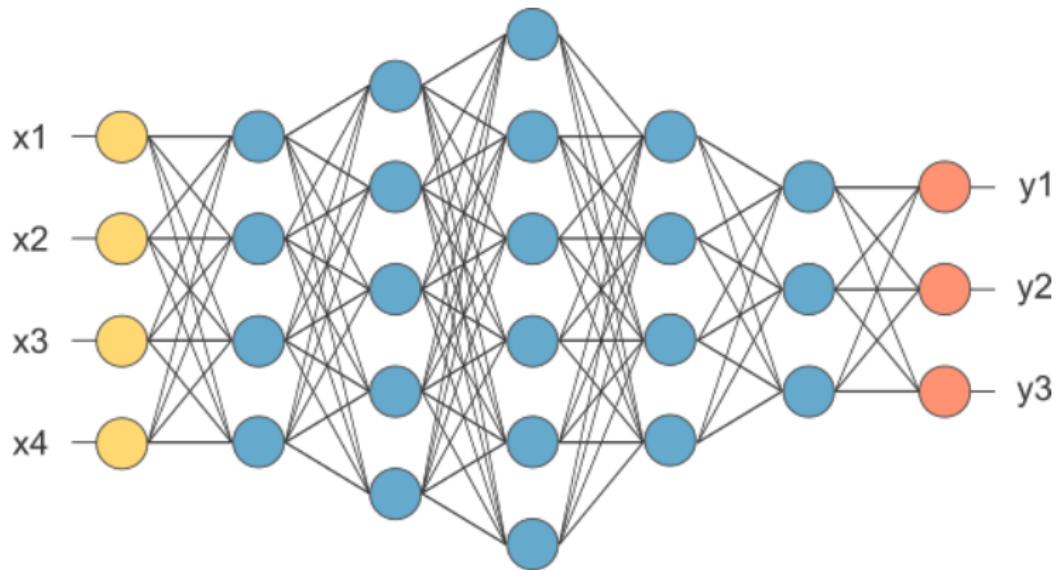
Neural Networks

Components of NN hierarchy are: input layers, hidden layers, and output layer.

- Input layer contains perceptrons that take input data X to generate signals. This layer captures basic logic (i.e., I decided this because of that).
- Hidden layer contains perceptrons that take signals from input layer to generate signals. This layer captures abstract logic. This layer captures aspects that are difficult for humans to rationalize (i.e., why did I decide that?).
- Output layer contains perceptrons that take signals from hidden layer to generate a final output signal Y . Signal can be continuous or discrete.

We now proceed to show how to train NNs. For simplicity, we will assume a NN architecture with one input layer, one hidden layer, and one output layer.

Neural Networks





Neural Networks

- *Input layer* is composed of $j = 1, \dots, n_I$ perceptrons. Evidence in perceptron j is :

$$a_j = \sum_{i=1}^n \theta_{j,i}^I x_i + \theta_{j,0}^I$$

Here, $\theta_{j,i}^I$ are parameters of perceptron ($\theta_{j,0}$ is bias) and x_i is input data (for a given observation ω). Given evidence, perceptron j generates output signal:

$$z_j = h(a_j)$$

here, $h(a_j)$ is known as activation function and is often modeled using a sigmoidal function (tanh and max functions are also used).

- *Hidden layer* is composed of $k = 1, \dots, n_H$ perceptrons. Evidence in perceptron k is:

$$a_k = \sum_{j=1}^{n_I} \theta_{k,j}^H z_j + \theta_{k,0}^H$$

Given evidence, output signal is:

$$w_k = h(a_k)$$



Neural Networks

- *Output layer* takes signals of hidden layer and can have one or multiple perceptrons (depending on how many outputs Y we have). If there is one output, evidence takes the form:

$$a = \sum_{k=1}^{n_H} \theta_k^O w_k + \theta_0^O$$

Final output (model prediction) is $m = \sigma(a)$.

where $\sigma(a)$ is a sigmoidal function if output Y is binary (e.g., classification) or $\sigma(a) = a$ if output is continuous (e.g., regression).

- Given parameters $\theta = (\theta^I, \theta^H, \theta^O)$, NN propagates input x_ω into output $m_\omega(\theta)$. This forward propagation can be written as:

$$m_\omega(\theta) = \sigma \left(\sum_{k=1}^{n_H} \theta_k^O h \left(\sum_{j=1}^n g \left(\sum_{i=1}^n \theta_{j,i}^I x_i^\omega + \theta_{j,0}^I \right) + \theta_{k,0}^H \right) + \theta_0^O \right), \quad \omega \in \mathcal{S}$$

- Derivatives $\partial m_\omega(\theta) / \partial \theta$ are computed using a technique called *back propagation*.



Neural Networks

- We can use MLE framework to estimate θ . For regression problems we solve:

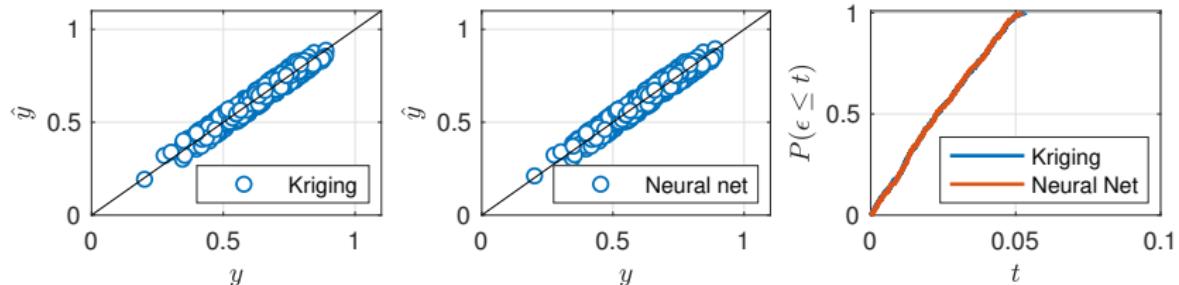
$$\min_{\theta} S(\theta) = \frac{1}{2} \sum_{\omega \in \mathcal{S}} (y_{\omega} - m_{\omega}(\theta))^2$$

In the context of NNs, $S(\theta)$ is known as the loss function.

- For classification problems we find θ that maximizes the log loss.
- Main advantage of NNs is that they are completely data-driven (no mechanistic understanding is needed). However, this is also the main disadvantage of NNs.
- Parameters and hidden variables often have no physical meaning and it is thus difficult to convey prior knowledge.
- This often manifests as a need for large amounts of data to determine parameters which are also many (on the order of thousands to millions). In the architecture discussed with have $n_I \cdot n + n_H \cdot n_I + n_H$ parameters.
- As a result, a typical issue with NNs is that of *overfitting*. Regularization terms (e.g., $\rho(\theta) = \lambda \theta^T \theta$) are often added to the loss function but there is often no mechanistic basis to construct more sophisticated strategies (e.g., Bayesian).
- Conducting cross-validation in NNs is particularly critical.

Example: Gibbs Reactor

- We now compare the performance of NN against kriging
- NN contains one hidden layer (with two perceptrons)



- We see that performance is virtually the same (NN can capture high nonlinearity)
- Reducing perceptrons of NN to one worsens predictions (but not by much)

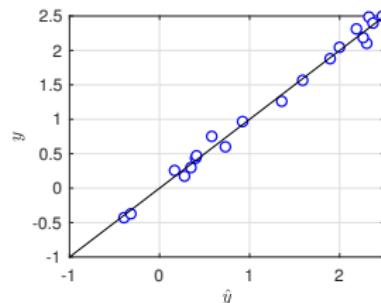
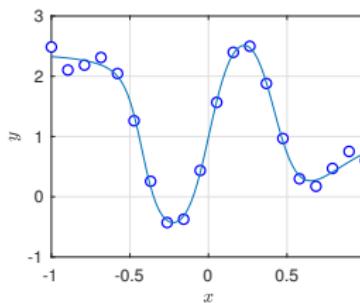
Example: Sigmoids as Basis Functions

- So why are NNs so effective at capturing nonlinear behavior?
- NNs operate as a sophisticated basis function model (the activation functions)
- To see this, construct a simple NN model with sigmoidal activation functions:

$$\phi_k(z) = \frac{1}{1 + e^{-z}}, \quad k = 1, \dots, N$$

where $z = \theta_k \cdot x + b_k$

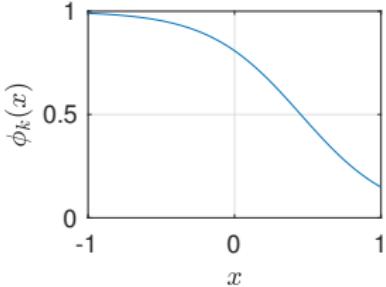
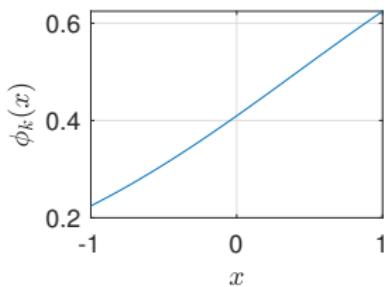
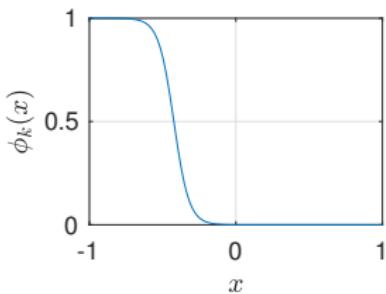
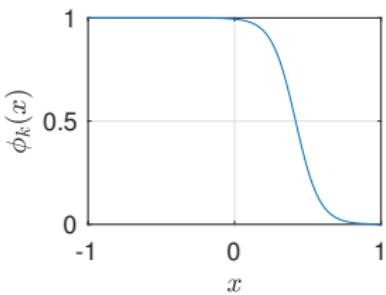
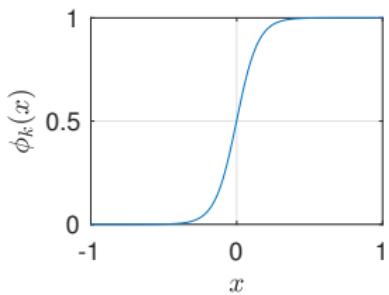
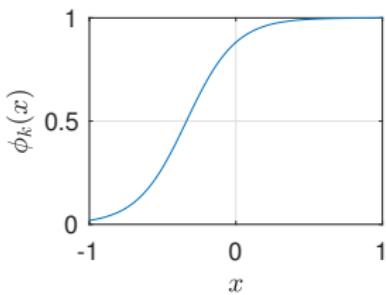
- Output is $y = \sum_{k=1}^N w_k \phi_k(z)$ (note there are no hidden layers).
- Fit of NN model to function $y = \sin(4x) + \sin(8x) + \exp(-x)$ is shown below.



- Hessian $H(\hat{\theta})$ has five eigenvalues below 1×10^{-5} (surface of $S(\theta)$ is flat).

Example: Sigmoids as Basis Functions

- Basis functions (with parameters obtained from training the model) are shown below.





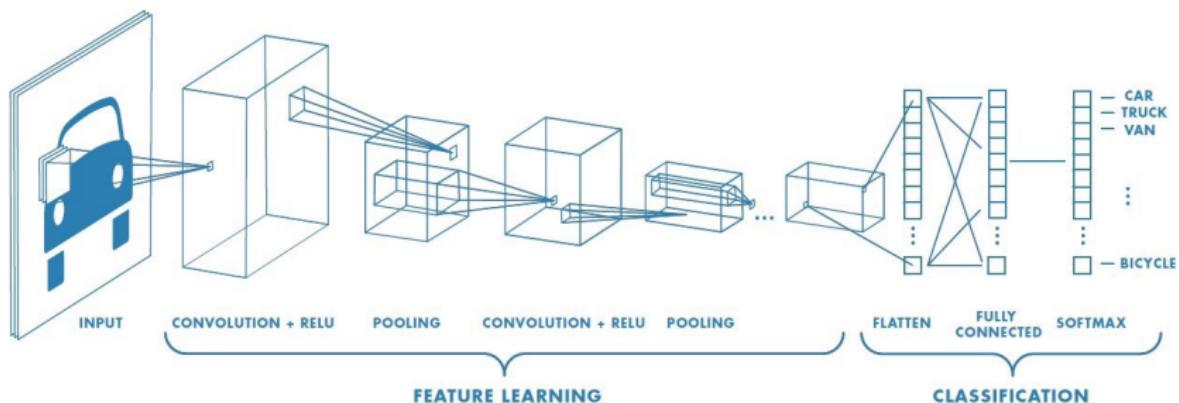
Convolutional Neural Networks

- Convolutional Neural Nets (CNNs) are NNs that are specialized to take images (matrices \mathbf{X}_ω) as inputs to make predictions y_ω (e.g., classify or predict value).
- In principle, we could create a large vector $\text{vec}(\mathbf{X}_\omega)$ and feed this into a regular NN to make a prediction. However, this approach can lead to an extremely large number of parameters.
- Instead, CNNs seek to summarize (compress) images by extracting relevant features. Such features are extracted by using filters (a filter is a small matrix with a given pattern \mathbf{W}).
- The extracted features are then used as evidence in the perceptrons that generate activation signals (also matrices).
- In a CNN, the parameters to learn from data are the filters.
- Different filters extract different features (patterns) of an image (analogous to SVD).



Convolutional Neural Networks

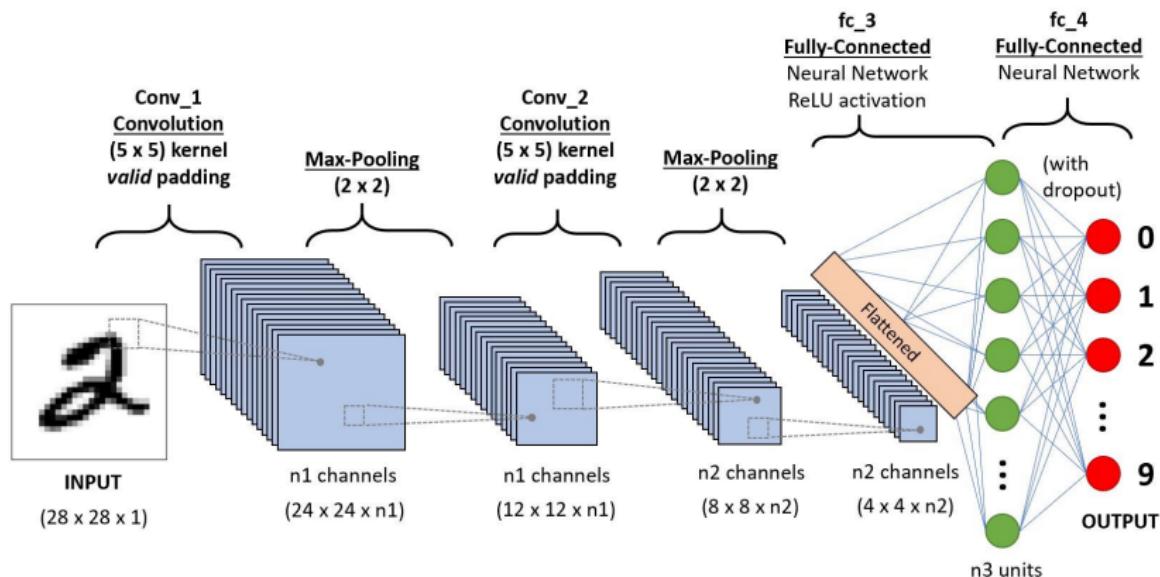
- Here is a typical architecture of a CNN:





Convolutional Neural Networks

- Images can be expressed by multiple matrices (color channels).



Convolutional Neural Networks

- Convolution is a pattern matching operation
- In signal processing, convolution is used to identify if a signal looks like another reference signal (filter)
- Procedure generates a filtered image (convolved feature)

1 $\times 1$	1 $\times 0$	1 $\times 1$	0	0
0 $\times 0$	1 $\times 1$	1 $\times 0$	1	0
0 $\times 1$	0 $\times 0$	1 $\times 1$	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

Convolved Feature



Convolutional Neural Networks

0	0	0	0	0	0	...
0	156	155	156	158	158	...
0	153	154	157	159	159	...
0	149	151	155	158	159	...
0	146	146	149	153	158	...
0	145	143	143	148	158	...
...

Input Channel #1 (Red)

-1	-1	1
0	1	-1
0	1	1

Kernel Channel #1



308

+

0	0	0	0	0	0	...
0	167	166	167	169	169	...
0	164	165	168	170	170	...
0	160	162	166	169	170	...
0	156	156	159	163	168	...
0	155	153	153	158	168	...
...

Input Channel #2 (Green)

1	0	0
1	-1	-1
1	0	-1

Kernel Channel #2



-498

0	0	0	0	0	0	...
0	163	162	163	165	165	...
0	160	161	164	166	166	...
0	156	158	162	165	166	...
0	155	155	158	162	167	...
0	154	152	152	157	167	...
...

Input Channel #3 (Blue)

0	1	1
0	1	0
1	-1	1

Kernel Channel #3



164

+

+ 1 = -25

Bias = 1

Output

-25				...
				...
				...
				...
...



Convolutional Neural Networks

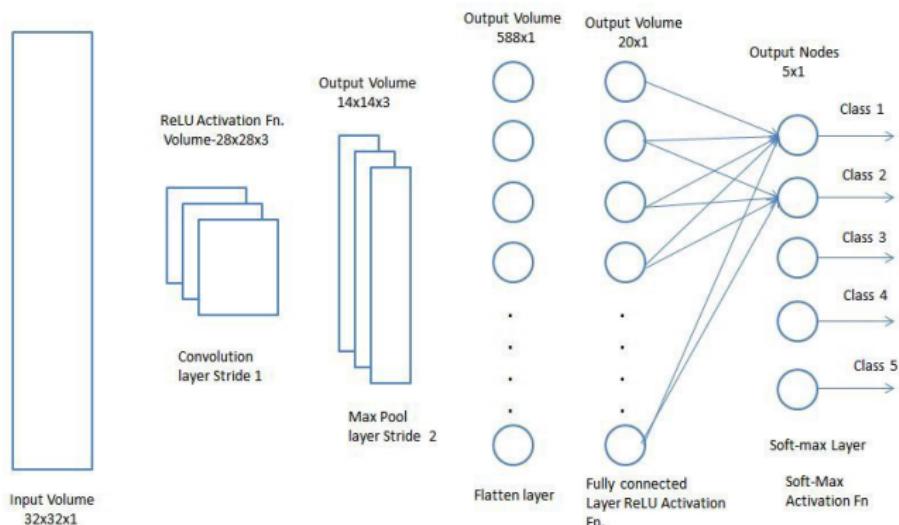
- Pooling is a procedure to summarize filtered images into smaller sizes

3.0	3.0	3.0
3.0	3.0	3.0
3.0	2.0	3.0

3	3	2	1	0
0	0	1	3	1
3	1	2	2	3
2	0	0	2	2
2	0	0	0	1

Convolutional Neural Networks

- Subsequent convolutions of the image generate output features that are used by a standard NN to perform predictions
- This NN layer is called the fully connected layer



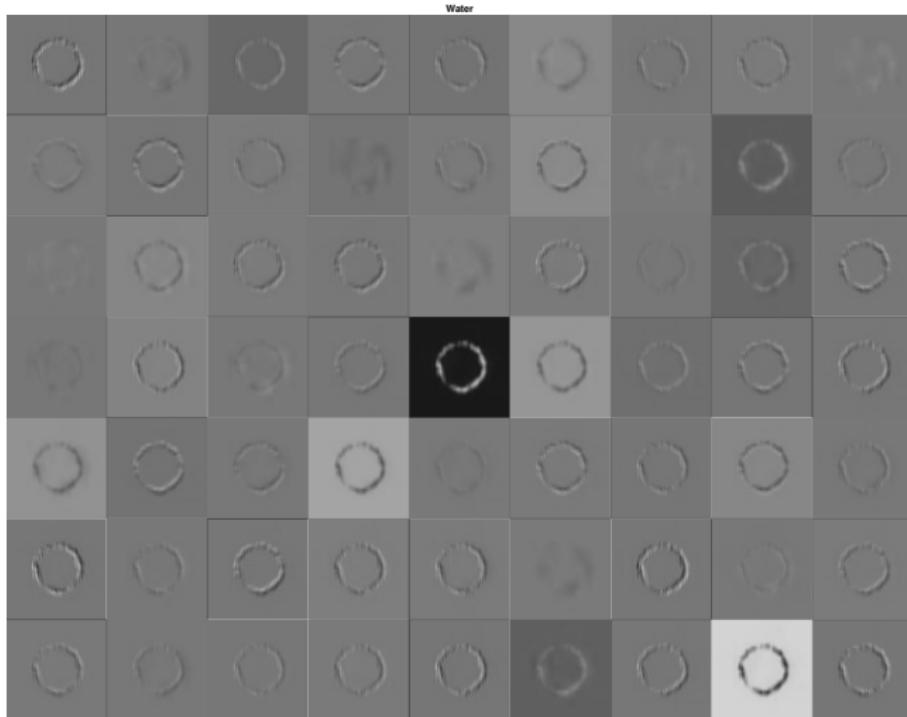
Example: Classifying Contaminants

- Here are some example images of a sensor exposed to DMMP and Water
- Have a total set of $S = 500$ for training



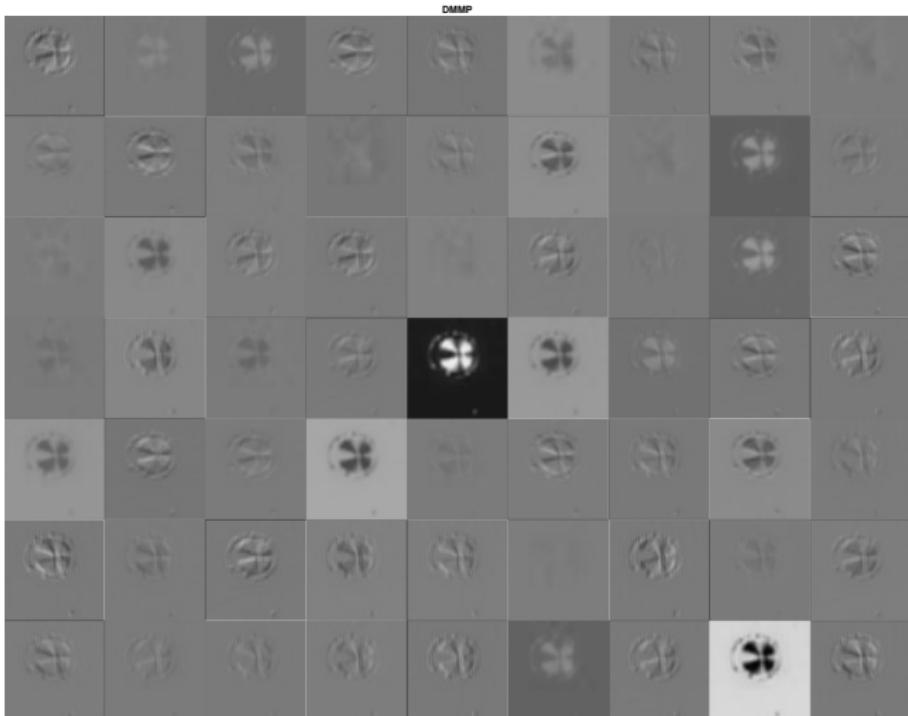
Example: Classifying Contaminants

- Here are the filtered images (activations) obtained from filters for DMMP



Example: Classifying Contaminants

- Here are the filtered images (activations) obtained from filters for Water



Example: Classifying Contaminants

- Here is the confusion matrix obtained from classification for an independent set of images (accuracy of 77%)

		Predicted	
		DMMP	Water
True	DMMP	20	20
	Water	3	57



Outline

- ① Introduction
- ② Models of Random Variables
- ③ Estimation Techniques
- ④ Multivariate Statistics
- ⑤ Data-Driven Modeling
- ⑥ Statistical Learning
- ⑦ Decision-Making Under Uncertainty



Decision-Making under Uncertainty

- We have learned to use data to model uncertainty as RVs and to build models to capture connections between RVs.
- We have also learned how to characterize uncertainty for functions of RVs using Monte Carlo simulations and RV transformations.
- We now shift our attention on how to use these capabilities to *make decisions*.
- Fundamental issue with making decisions under uncertainty is that humans take different attitudes towards risk and tend to severely under/overestimate uncertainty. Also, “risk” means different things to different people.
- Consider setting in which we would like to choose between decisions u and u' with associated univariate RVs $Y(u)$ and $Y(u')$ (e.g., cost).
- Another setting is that in which we want to find a decision u that manipulates RV $Y(u)$ in a particular way (e.g., it maximizes it or reduces its variability).



Decision-Making under Uncertainty

Consider now the following questions:

- What is risk? How do I measure risk?
- How can I make a decision that withstands uncertainty?
- How can I take optimal proactive actions in the face of uncertainty?

Concepts and techniques that answer these questions are studied in the area of *stochastic optimization* (a.k.a. stochastic programming).



Defining and Measuring Risk

At this point, we have all fundamentals of statistics needed to properly define risk and figure out how to measure it.

- Consider an input X with pdf $f_X(x)$ and cdf $F_X(x)$ and utility function $\varphi(X, u)$ that depends on the input RV and a decision u .
- Since X is an RV, utility is also an RV that we represent as univariate RV $Y(u) = \varphi(X, u)$ with pdf $f_{Y(u)}(y)$ and cdf $F_{Y(u)}(y)$.
- The pdf and cdf of $Y(u)$ depend on the decision u ; consequently, u can be used to manipulate these and to manipulate the statistics of $Y(u)$.
- In popular culture, risk is associated with a decision or situation (e.g., “this is a risky investment”, “the reactor is operating at risky conditions”) and is associated with extreme events or large/catastrophic losses.



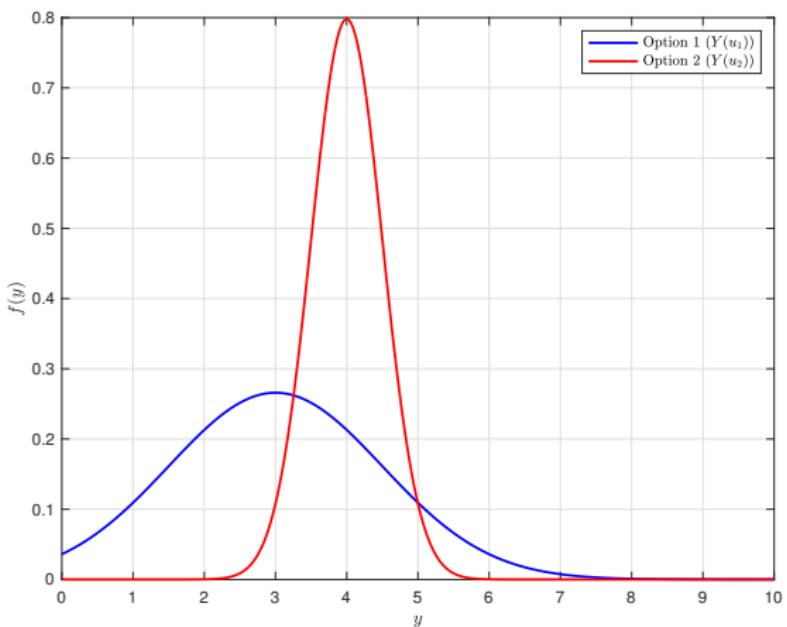
Defining and Measuring Risk

- When making decisions, however, we want precise measures that tell us exactly what constitutes a risky decision and how risky it is.
- Unfortunately, there is no unique mathematical definition of risk because humans tend to value different aspects of uncertainty (e.g., probability of failure vs. magnitude of failure vs. worst-case failure).
- Moreover, humans tend to disagree on what probability levels are acceptable (e.g., what seems risky to me might not seem risky to you).
- These disagreements arise because decision-makers (DMs) take different attitudes towards risk.
- As a result, what we want to explore here is not what definition of risk exist but rather what *definitions* of risk exist. Ultimately, establishing a definition of risk for a particular situation at hand should be based on consensus between DMs.



Example: Defining and Measuring Risk

- Consider options $Y(u_1)$ and $Y(u_2)$
 - $Y(u_1)$ has low mean cost but high variance
 - $Y(u_2)$ has high mean cost and but low variance
- Which one is better?





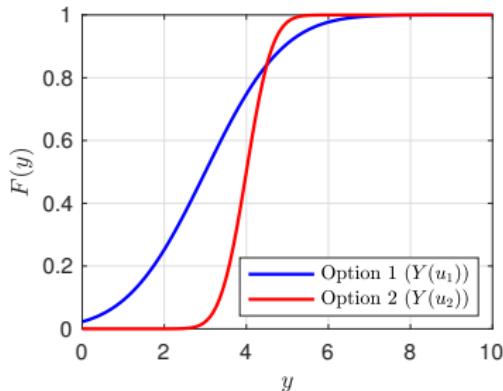
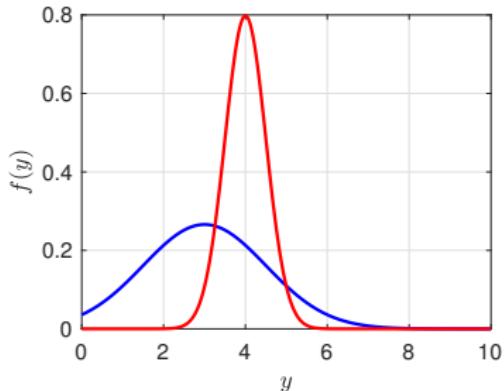
Defining and Measuring Risk

To motivate potential definitions of risk that we might consider, we explore different attitudes towards risk that DMs might take:

- *Risk-Neutral*: DM prefers $Y(u)$ over $Y(u')$ if $\mathbb{E}_{Y(u)} \leq \mathbb{E}_{Y(u')}$. This DM worries about performance on average and is not concerned with the fact that $Y(u)$ might have outcomes with large values compared to those of $Y'(u)$.
- *Risk-Conscious*: DM prefers $Y(u)$ over $Y(u')$ if $\mathbb{P}(Y(u) \leq y) \geq \mathbb{P}(Y(u') \leq y)$. A risk-conscious DM wants a decision that will likely achieve lower outcomes. This type of DM might also prefer $\mathbb{V}_{Y(u)} \leq \mathbb{V}_{Y(u')}$ because u has lower variability than u' .
- *Risk-Averse*: DM prefers $Y(u)$ over $Y(u')$ if $\max Y(u) \leq \max Y(u')$. This DM only worries about the worst possible outcome of u and u' .
- *Risk-Taker*: DM prefers $Y(u)$ over $Y(u')$ if $\min Y(u) \leq \min Y(u')$. This DM only worries about the best possible outcome of u and u' .

Example: Defining and Measuring Risk

- Which option to select if you are risk-neutral, risk-conscious, risk-averse, and risk-taker?





Defining and Measuring Risk

So, how do we measure risk?

A risk measure is a function $\rho(Y(u))$ that maps a univariate RV $Y(u)$ to a scalar quantity (it is a summarizing statistic). Common risk measures used in practice are:

- Expected Value: $\mathbb{E}[Y(u)]$
- Variance: $\mathbb{V}[Y(u)] = \mathbb{E}[(Y(u) - \mathbb{E}[Y(u)])^2]$
- Mean-Variance: $\text{MV}_\kappa[Y(u)] = \mathbb{E}[Y(u)] + \kappa\mathbb{V}[Y(u)]$ (for some κ)
- Probability of Loss: $\mathbb{P}(Y(u) > y)$ or $\mathbb{P}(Y(u) \leq y)$ (a.k.a. probability of failure)
- Value-at-Risk: $\mathbb{Q}_{Y(u)}(\alpha)$ (also written as VaR_α)
- Conditional Value-at-Risk: $\mathbb{E}[Y(u)|Y(u) \geq \text{VaR}_\alpha]$ (a.k.a. expected short fall)
- Mean Deviation: $\mathbb{E}[|Y(u) - \mathbb{E}[Y(u)]|]$
- Worst/Best-Case: $\max Y(u)$ or $\min Y(u)$



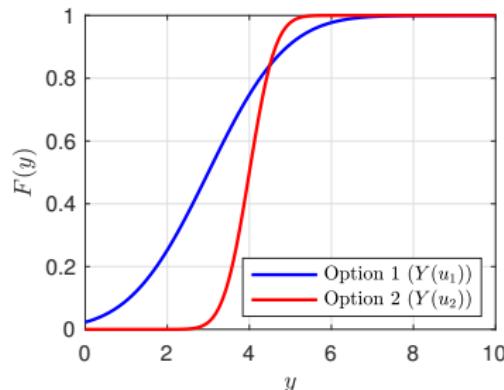
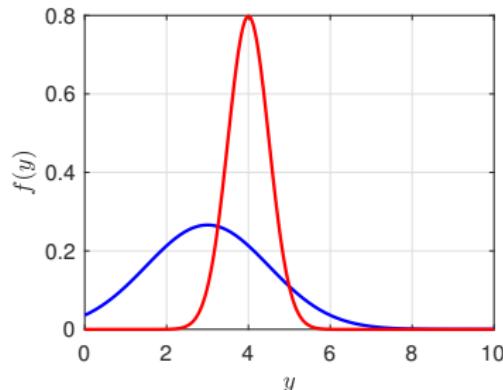
Defining and Measuring Risk

Some observations:

- Different measures are used to model different risk attitudes.
- Probability of loss is what is colloquially known as risk because it has a natural interpretation that DMs find useful in practice. We will see that this measure has its caveats.
- In principle, we have freedom of using any statistic $Y(u)$ of interest (e.g., moments, entropy) as risk measure but, as expected, not all measures are expected to be adequate.

Example: Defining and Measuring Risk

For previous example we have:



- Expected Value: $\mathbb{E}[Y(u_1)] = 3$, $\mathbb{E}[Y(u_2)] = 4$
- Variance: $\mathbb{V}[Y(u_1)] = 2.2$, $\mathbb{V}[Y(u_2)] = 0.2$
- Mean-Variance: $MV_\kappa[Y(u_1)] = 5.2$, $MV_\kappa[Y(u_2)] = 4.2$ for $\kappa = 1$
- Probability of Loss:
 - $\mathbb{P}(Y(u_1) > y) = 9\%$, $\mathbb{P}(Y(u_2) > y) = 2\%$ for $y = 5$
 - $\mathbb{P}(Y(u_1) > y) = 2\%$, $\mathbb{P}(Y(u_2) > y) = 5\%$ for $y = 4$
- Value-at-Risk: $\mathbb{Q}_{Y(u_1)}(\alpha) = 4.92$, $\mathbb{Q}_{Y(u_2)}(\alpha) = 4.64$ for $\alpha = 0.9$
- Conditional Value-at-Risk: $CVaR_\alpha[Y(u_1)] = 5$, $CVaR_\alpha[Y(u_2)] = 4.9$ for $\alpha = 0.9$



Stochastic Dominance

So, what are the desirable features of a risk measure?

- To answer this question, it is necessary to first discuss what makes an RV better than another RV (how do we compare RVs?).
- An important concept that arises here is that of *stochastic dominance* (SD).

We say that Y dominates Y' (written as $Y \preceq Y'$) if:

$$\mathbb{P}(Y \leq y) \geq \mathbb{P}(Y' \leq y) \text{ for all } y \in \mathcal{D}$$

where \mathcal{D} is the domain of Y and Y' .

Some observations:

- SD says that Y is better (dominates) Y' if probability of Y taking a value less than a threshold y is higher than probability of Y' being below same threshold. Moreover, the probability is higher or equal for any threshold value.
- Can also express SD as $F_Y(y) \geq F_{Y'}(y)$ for all $y \in \mathcal{D}$. We thus have that cdf curve of Y is always above that of Y' .



Stochastic Dominance

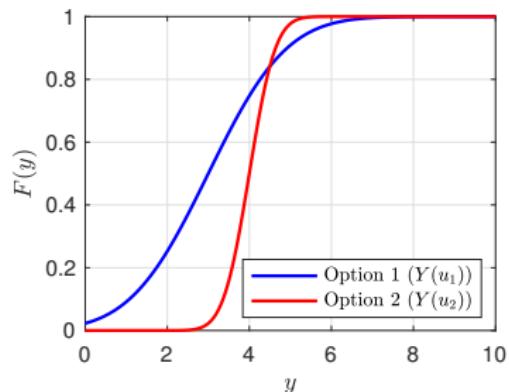
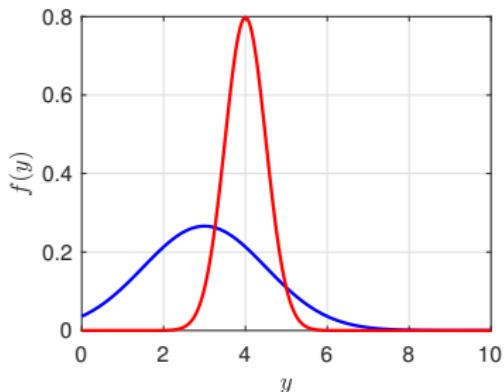
- Note how SD is different than the traditional dominance concept we are familiar with (a decision dominates another one if it is *always* better). In statistical terms, this would imply that outcomes of Y are always lower or equal than those of Y' :

$$y_\omega \leq y'_\omega \quad \omega \in \Omega$$

- This requirement is strict and it is unlikely to occur in practice. Consequently, SD is a more flexible requirement.
- A risk-conscious DM would typically require that $\mathbb{P}(Y \leq y) \geq \mathbb{P}(Y' \leq y)$ holds *or a single y*. Note that SD is a stricter requirement because it requires $\mathbb{P}(Y \leq y) \geq \mathbb{P}(Y' \leq y)$ to hold *for all possible y*. As such, there are weakened (relaxed) versions of SD that are also often used in practice (e.g., holds *for some y*).

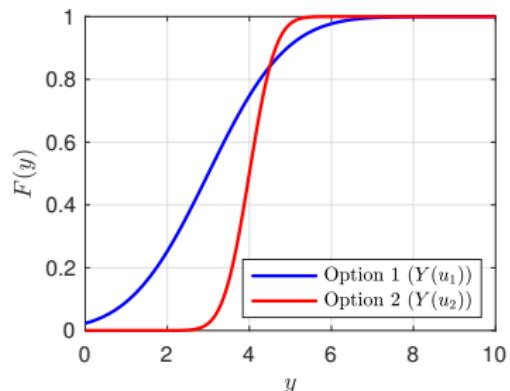
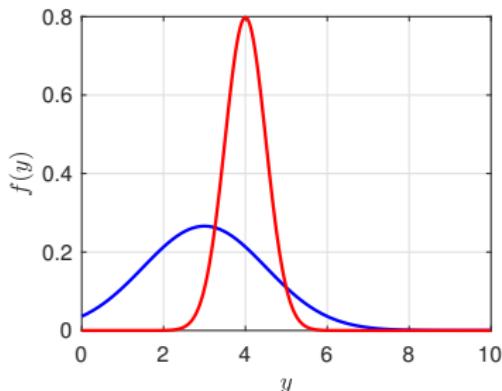
Example: Stochastic Dominance

- Does $Y(u_1)$ dominate $Y(u_2)$ (i.e., $Y(u_1) \preceq Y(u_2)$)?



Example: Stochastic Dominance

- Does $Y(u_1)$ dominate $Y(u_2)$ (i.e., $Y(u_1) \preceq Y(u_2)$)?



- No, $Y(u_1)$ dominates $Y(u_2)$ only for some values of threshold y .



Coherency Properties of Risk Measures

There are a number of fundamental properties (a.k.a. axioms) that a proper (a.k.a. coherent) risk measure should satisfy. These properties have been proposed based on their usefulness in actual practical applications and based on mathematical consistency.

- Translation Invariance: $\rho(Y + c) = \rho(Y) + c$ for $c \in \mathbb{R}$. This indicates that adding a constant value to an RV should result in adding same constant to the risk measure (i.e., adding a constant does not alter inherent properties of RV).
- Subadditivity: $\rho(Y + Y') \leq \rho(Y) + \rho(Y')$. This indicates that the risk of a combined pair of RVs cannot exceed the summation of their individual risks.
- Monotonicity: If $Y \preceq Y'$ then $\rho(Y) \leq \rho(Y')$. This indicates that, if Y stochastically dominates Y' , then its risk should also be lower. This indicates that risk measure reflects dominance.
- Positive Homogeneity: $\rho(c \cdot Y) = c\rho(Y)$ for $c \in \mathbb{R}_+$. This indicates that scaling RV by a constant does not affect inherent properties of RV.

We will see later that some of these properties resemble those of vector norms.



Risk Measures as Norms

In fact, it turns out that the risk measure of an RV is analogous to a norm of a vector. To see this, let's begin with the following question:

How can we say that vector $\mathbf{y} = (y_1, y_2, \dots, y_S)$ is better than $\mathbf{y}' = (y'_1, y'_2, \dots, y'_S)$?

This reveals similar difficulties that arise when comparing RVs. Consider following possibilities:

- We can say that \mathbf{y} is better than \mathbf{y}' if the total magnitude of its entries is lower ($\sum_{i=1}^S y_i \leq \sum_{i=1}^S y'_i$). This can be written as $(1/S) \sum_{i=1}^S y_i \leq (1/S) \sum_{i=1}^S y'_i$ and is analogous to expected value.
- We can say that \mathbf{y} is better than \mathbf{y}' if $\max_i y_i \leq \max_i y'_i$. This is analogous to worst-case risk measure.
- We can say that \mathbf{y} is better than \mathbf{y}' if $y_i \leq y'_i$ for all $i = 1, \dots, S$. This is analogous to say that Y is always better than Y' . Note, however, that the entries are not arranged in order so maybe an entry of \mathbf{y} is much larger than any entry of \mathbf{y}' .
- Consequently, perhaps a better approach would be arranging entries in decreasing order and then compare the rearranged entries. This is analogous to stochastic dominance (we establish a threshold on the magnitude of the entries) and count how many entries of the vectors are below that threshold value.



Risk Measures as Norms

- Recall that norm $\rho(\mathbf{y})$ of a vector \mathbf{y} is a measure of its size. Norms are used to compare vectors and to establish bounding properties.
- Recall that most used vector norm is ℓ -p norm:

$$\|\mathbf{y}\|_p = \left(\sum_{i=1}^S |y_i|^p \right)^{1/p}.$$

This norm has the following special cases:

- For $p = 1$ we have that $\|\mathbf{y}\|_1 = \sum_{i=1}^S |y_i|$ and note that this is S times average magnitude of the entries (analogous of expected value).
- For $p = \infty$ we have that $\|\mathbf{y}\|_\infty = \max_i |y_i|$ and note this is largest value (analogous of worst-case).
- As in the case of risk measures, one can define many types of norms to measure and compare vectors in different forms.
- For instance, an interesting norm is k -max norm. This norm is sum of k -largest elements of vector \mathbf{y} . This norm is analogous to expected shortfall.



Risk Measures as Norms

As with risk measures, one must ensure that any norm that we define satisfy basic properties. A *proper* norm of a vector must satisfy the following properties:

- Homogeneity: $\rho(c \cdot \mathbf{y}) = c \cdot \rho(\mathbf{y})$ for $c \in \mathbb{R}_+$
- Subadditivity (Triangle Inequality): $\rho(\mathbf{y} + \mathbf{y}') \leq \rho(\mathbf{y}) + \rho(\mathbf{y}')$
- Normalization: $\rho(0) = 0$.

Homogeneity property is analogous to that of a proper risk measure. Moreover, the triangle inequality is analogous to the subadditivity property.



Example: Risk Measures as Norms

- Consider vectors $y_1 = (100, 70, 50, 20, 10)$ and $y_2 = (110, 80, 40, 10, 0)$
- What is mean and worst-case for such vectors?
 - Mean is $(1/5) \cdot (100 + 70 + 50 + 20 + 10) = 50$ and
 $(1/5) \cdot (110 + 80 + 40 + 10 + 0) = 48$
 - Worst-Case is 100 and 110
- What is CVaR for such vectors?
 - If $\alpha = 5/5$ then $(1/5) \cdot (100 + 70 + 50 + 20 + 10) = 50$ and
 $(1/5) \cdot (110 + 80 + 40 + 10 + 0) = 48$
 - If $\alpha = 4/5$ then $(1/4) \cdot (100 + 70 + 50 + 20) = 60$ and $(1/4) \cdot (110 + 80 + 40 + 10) = 60$
 - If $\alpha = 3/5$ then $(1/3) \cdot (100 + 70 + 50) = 73$ and $(1/3) \cdot (110 + 80 + 40) = 76$
 - If $\alpha = 1/5$ then $(1/1) \cdot (100) = 100$ and $(1/1) \cdot (110) = 60$
- We thus have that CVaR has mean and worst-case as extremes.



Expected Value Properties

We now have all elements to judge whether a risk measure is adequate or not. We begin by discussing the properties of expected value $\mathbb{E}[Y]$.

- Expected value is a measure of the magnitude of Y (i.e., $\mathbb{E}[Y] \leq \mathbb{E}[Y']$ indicates that Y is, on average, smaller than Y') and therefore has a natural intuitive interpretation. Because of this, this is a common measure used in practice.
- Expected value satisfies all properties of coherency.
- Expected value ignores outcomes that can result in high values. For instance, even if $\mathbb{E}[Y] \leq \mathbb{E}[Y']$ holds, Y can have outcomes of large magnitude that Y' does not have.



Variance Properties

We now discuss the properties of variance $\mathbb{V}[Y] = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$.

- Variance is a measure of variability (i.e., $\mathbb{V}[Y] \leq \mathbb{V}[Y']$ indicates that Y has less variability than Y') and therefore has a natural and intuitive interpretation.
- An issue with the variance is that it does not capture magnitude of RV. To remedy this issue, variance is often used in conjunction with expected value by using mean-variance measure $\mathbb{E}[Y] + \kappa\mathbb{V}[Y]$ where κ is a weighting vector that helps span a range of attitudes towards risk (from risk neutral to risk conscious).
- Mean-variance was proposed by Harry Markowitz in the 1950s and became the standard in the finance industry for many years (Markowitz earned a Nobel prize).
- Variance takes into consideration outcomes of large values (in upper tail of pdf) but this connection is not direct and this results in several important inefficiencies. Specifically, variance is centered around the expected value and penalizes deviations from it symmetrically.
- Symmetry is undesirable in applications because we are often only interested in accounting for outcomes of large magnitude (and not with those of low magnitude). Moreover, in many applications, RV does not have a symmetric pdf and therefore variance can fail to properly capture tail effects.



Variance Properties

- Importantly, variance is not a proper risk measure. Specifically, it does not satisfy monotonicity (it is not consistent with SD).
- The variance and expected value do not have same units. Consequently, it is often preferred to use the standard deviation $SV[Y] = \sqrt{V[Y]}$, which has the same units as the expected value.
- Variance remains widely used in industry because it is an easily-interpretable measure of variability. This measure, however, has important deficiencies.

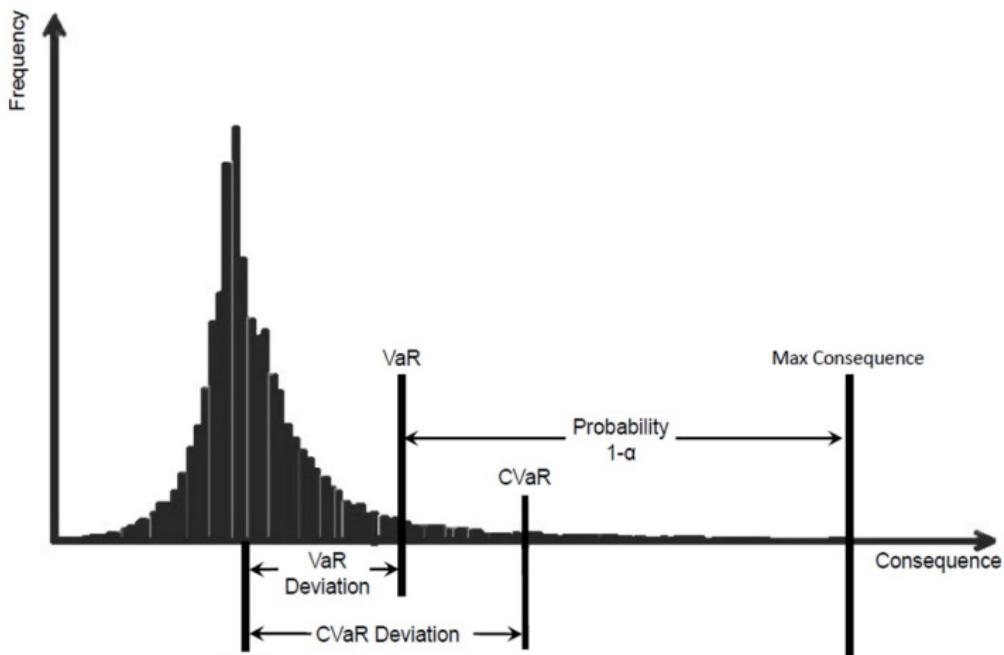


Expected Shortfall Properties

The expected shortfall overcomes many of deficiencies of the variance and has recently become the standard risk measure. This measure has many important properties and connections with other risk measures that are worth highlighting.

- Recall that expected shortfall of Y at probability α is $\mathbb{E}[Y|Y \geq Q(\alpha)]$, where $Q(\alpha)$ is the α -quantile of Y .
- Expected loss takes expected value losses above quantile $Q(\alpha)$. The α -quantile is threshold value t at which $\mathbb{P}(Y \leq t) = \alpha$.
- Expected loss captures magnitude of outcomes of high value while ignoring those of small magnitude (it is an asymmetric risk measure). This becomes obvious if we write the expected loss as $\mathbb{E}[(Y - Q(\alpha))_+]$.

Example: Expected Shortfall Properties





Expected Shortfall Properties

- If we set $\alpha = 0$, quantile $Q(\alpha)$ is minimum value of Y and therefore the expected loss is the expected value $E[Y]$.
- If we set $\alpha = 1$, quantile is maximum value and therefore the expected loss is the worst-case value $\max Y$.
- Consequently, expected loss captures a range of risk attitudes (from risk neutral to conscious to averse).
- Expected loss is a coherent risk measure.
- A caveat of the expected loss is that it offers no direct control on the probability of loss, which is a measure of interest to many DMs. In other words, reducing the expected loss does not necessarily imply reducing the probability of loss.
- Moreover, DM needs to specify α and this selection can lead to disagreement.



Probability of Loss Properties

Probability of loss is one of most widely used measures of risk.

- Recall that probability of loss is simply $\mathbb{P}(Y > y)$ and this is also often expressed as probability of no loss as $\mathbb{P}(Y \leq y) = 1 - \mathbb{P}(Y > y)$.
- What constitutes a loss is defined by threshold value y . Consequently, it is more appropriate to call this measure the “probability of unacceptable loss”.
- Meaning of probability of loss is intuitive.
- Biggest caveat of probability of loss is that it says nothing about actual magnitude of the losses. For example, even if $\mathbb{P}(Y > y) \leq \mathbb{P}(Y' > y)$, one might still prefer Y' because losses incurred are not that large compared to those of Y . In other words, Y' is less catastrophic than Y .
- Probability of loss is not coherent risk measure.
- Finally, one needs to specify a threshold value to express what constitutes an unacceptable loss and this selection often leads to disagreement of DMs.



Risk Measures

The different advantages and disadvantages encountered with risk measures highlight the difficulty in controlling and comparing RVs. In particular:

- Risk measures are often *conflicting* (e.g., reducing one measure often results in increasing another measure).
- Measuring risk leads to *ambiguity* (expressing mathematically what an DM might be looking for is challenging). This is analogous to measuring happiness or fairness.

Despite these limitations, risk measures are an essential component of decision-making under uncertainty.

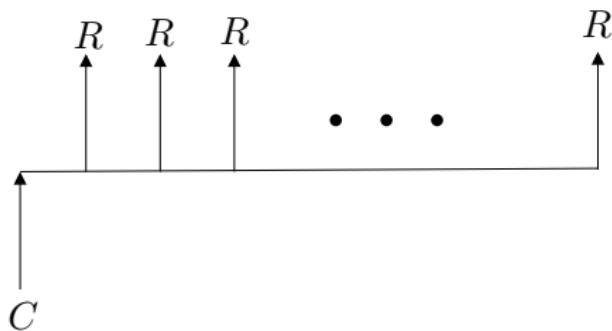


Example: Investment Options

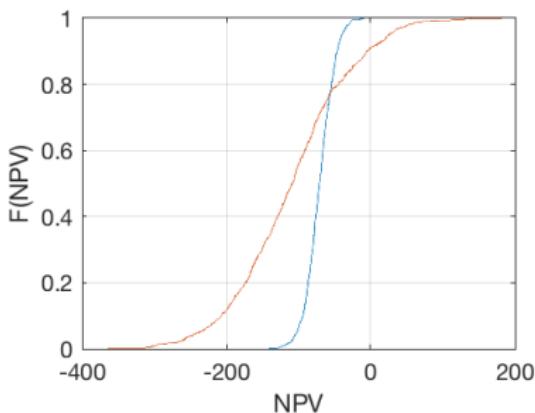
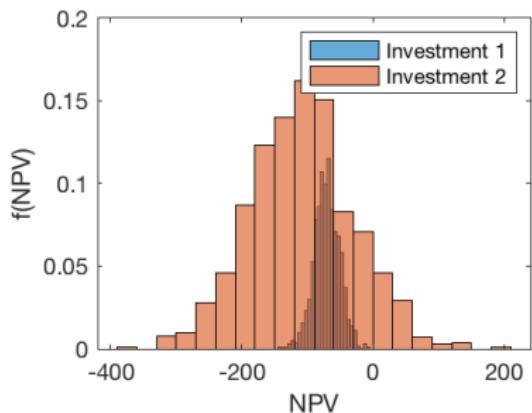
- We have investment options:
 - Initial capital $C_1 = \$25,000$ with annual rate of return $R_1 \sim \mathcal{N}(0.5, 0.1)$
 - Initial capital $C_2 = \$50,000$ with annual rate of return $R_2 \sim \mathcal{N}(0.4, 0.2)$
- We compare net present value for both options (10 yr, interest rate of 5%):

$$NPV_1 = C_1 - R_1 \cdot C_1 \cdot \frac{(1 - (1 + i)^{-10})}{i}$$
$$NPV_2 = C_2 - R_2 \cdot C_2 \cdot \frac{(1 - (1 + i)^{-10})}{i}.$$

- Note we want NPV to be as *negative* as possible.



Example: Investment Options





Making Optimal Decisions

Instead of comparing decisions, a DM might also want to directly find best decision possible. Such a decision is also influenced by the attitude towards risk and gives rise to different optimization problems (depending on the risk measure used):

- Expected value and variance:

$$u^* = \operatorname{argmin}_u \mathbb{E}[Y(u)] + \kappa \mathbb{V}[Y(u)]$$

- Probability of loss

$$u^* = \operatorname{argmin}_u \mathbb{P}(Y(u) > y)$$

- Expected shortfall:

$$u^* = \operatorname{argmin}_u \mathbb{E}[Y(u) | Y(u) \geq \mathbb{Q}(\alpha)]$$

Another possibility is to find a decision that dominates a given benchmark. In other words, we seek u such that $Y(u) \preceq Y'$ for a given Y' . An issue with this approach is that the problem might have no solution.



Deterministic Decision-Making

DMs often make decisions in real-life by ignoring uncertainty all-together. This *deterministic* DM approach follows the logic:

- Assume input X takes a single value $x = \mathbb{E}[X]$ (i.e., average historical value). This assumes that input is deterministic and thus output Y is deterministic with value:

$$y = \varphi(x, u)$$

- Based on this assumed behavior, we can obtain a decision:

$$u_D^* = \underset{u}{\operatorname{argmin}} \varphi(x, u)$$

- The assumed deterministic behavior simplifies the decision-making process (removes ambiguity) but fails to ignore the inherent variability of X seen in real-life.
- Decision u_D^* might not be capable of controlling real output $Y(u_D^*)$, which is characterized by outcomes $y_\omega = \varphi(x_\omega, u_D^*)$, $\omega \in \mathcal{S}$.
- As a result, decision u_D^* might be vulnerable to uncertainty and incur large losses.

Example: Deterministic vs. Stochastic Solutions

- We consider the minimization of the cost function:

$$\varphi(u, X) = (u - X)^2 - e^{\alpha X}$$

with input uncertainty $X \in \mathcal{N}(2, 1)$

- We obtain solution using deterministic approach, which finds decision by solving:

$$u_D^* = \operatorname{argmin}_u \varphi(\mathbb{E}[X], u)$$

- We obtain solution using stochastic approach, which finds decision by solving:

$$u^* = \operatorname{argmin}_u \mathbb{Q}_{\varphi(u, X)}(\alpha)$$

- Pdfs and cdfs for costs $\varphi(u_D^*, X)$ and $\varphi(u^*, X)$ are shown below:

