

# MA/ST 810

# Mathematical-Statistical Modeling and Analysis of Complex Systems

---

## Integrating Mathematical and Statistical Models

- Recap of mathematical models
- Models and data
- Statistical models and sources of variation

# Recap of mathematical models

---

**Deterministic models:** For the purposes of this discussion, we will consider *mathematical models* of the following form

- *System*      $\dot{x}(t) = g\{t, x(t), \theta\}$
- *Solution*      $y = x(t, \theta)$
- $x(t)$  is the vector whose elements correspond to *states* of a system of interest at time  $t$
- $\dot{x}(t)$  is the vector of derivatives of the elements of  $x(t)$  (ODEs)
- Often arise from a *compartmental model* of the system (a gross simplification)

**Some examples...**

# Recap of mathematical models

---

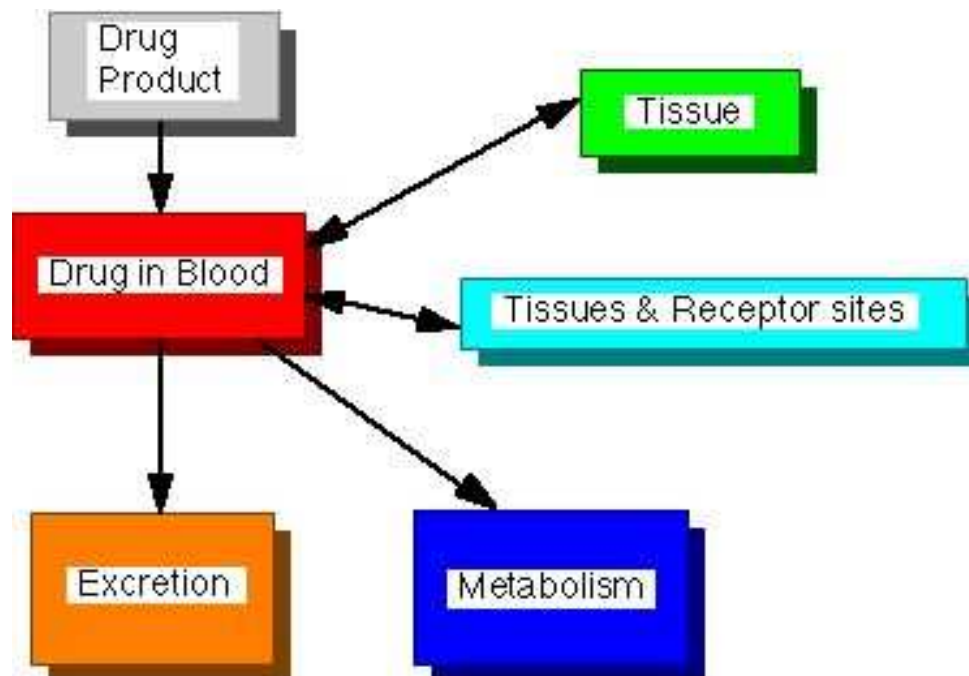
## Example 1: *Pharmacokinetics (PK)* of a drug

- *Pharmacokinetics*: “What the body does to the drug”
- *ADME processes*: *absorption*, *distribution*, *metabolism*, *excretion* (metabolism + excretion = *elimination*) – dictate the *concentrations of drug* in the body
- Critical to understand (*quantify*) ADME processes in the development of *dosing recommendations*
- *Premise*: Measure concentrations of drug in the blood over time and use these to learn about ADME processes

# Recap of mathematical models

---

## ADME:



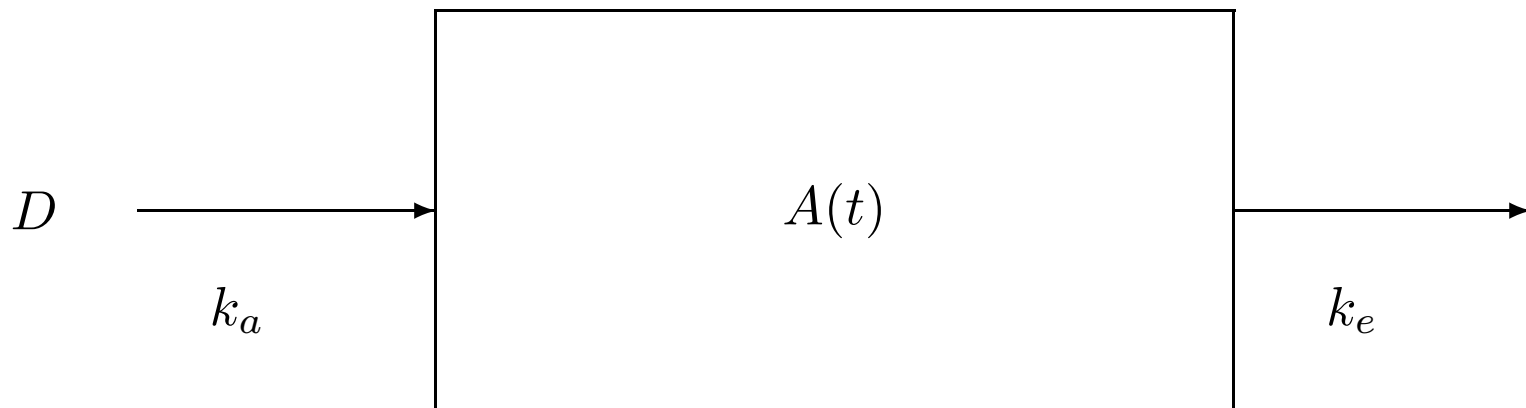
**Routes of drug administration:** *Intravenously, orally, intramuscularly, subcutaneously, . . .*

# Recap of mathematical models

---

## Pharmacokinetics of theophylline: Anti-asthmatic agent

- Common deterministic model: *One compartment model with first-order absorption and elimination* following *oral* dose  $D$  at  $t = 0$
- $A(t)$  = amount of drug in the “*blood compartment*” at time  $t$  (“*well-mixed*”)



- *Assumption*:  $A(t) = VC(t)$  (constant relationship between drug concentration  $C(t)$  and amount of drug in body  $A(t)$  for all  $t$ )

# Recap of mathematical models

---

**System:** Letting  $A_a(t)$  be the amount of drug at the absorption site at time  $t$

$$\begin{aligned}\dot{A}(t) &= k_a A_a(t) - k_e A(t) \\ \dot{A}_a(t) &= -k_a A_a(t)\end{aligned}$$

with initial conditions  $A_a(0) = A_{a0} = FD$ ,  $A(0) = A_0 = 0$ , where  $F$  is the fraction available (take  $F \equiv 1$  for simplicity)

- $x(t) = \{A(t), A_a(t)\}^T$ ,  $\dot{x}(t) = \{\dot{A}(t), \dot{A}_a(t)\}^T = g\{t, x(t), \theta\}$

$$g\{t, x(t), \theta\} = \begin{pmatrix} k_a A_a(t) - k_e A(t) \\ -k_a A_a(t) \end{pmatrix}, \quad \theta = (k_a, k_e)^T$$

**Solution:** Expression for  $A(t)$  [and  $A_a(t)$ ] may be found analytically in a *closed form*

# Recap of mathematical models

---

**Laplace transform of  $A(t)$ :**  $\mathcal{L} A = \int_0^{\infty} e^{-st} A(t) dt$

$$s\mathcal{L} A - A_0 = k_a\mathcal{L} A_a - k_e\mathcal{L} A \quad (1)$$

$$s\mathcal{L} A_a - A_{a0} = -k_a\mathcal{L} A_a \quad (2)$$

- Solve (2) for  $\mathcal{L} X_a$  and substitute in (1) to obtain

$$\mathcal{L} A = \frac{k_a F D}{(s + k_e)(s + k_a)}$$

- From a table of Laplace transforms, we find immediately that

$$A(t) = \frac{k_a F D}{k_a - k_e} \{e^{-k_e t} - e^{-k_a t}\}$$

so that (divide by  $V$ )

$$C(t) = \frac{k_a F D}{V(k_a - k_e)} \{e^{-k_e t} - e^{-k_a t}\}$$

# Recap of mathematical models

---

**Result:** If the model is *perfectly correct*, the relationship

$$C(t) = \frac{k_a F D}{V(k_a - k_e)} \{e^{-k_e t} - e^{-k_a t}\}$$

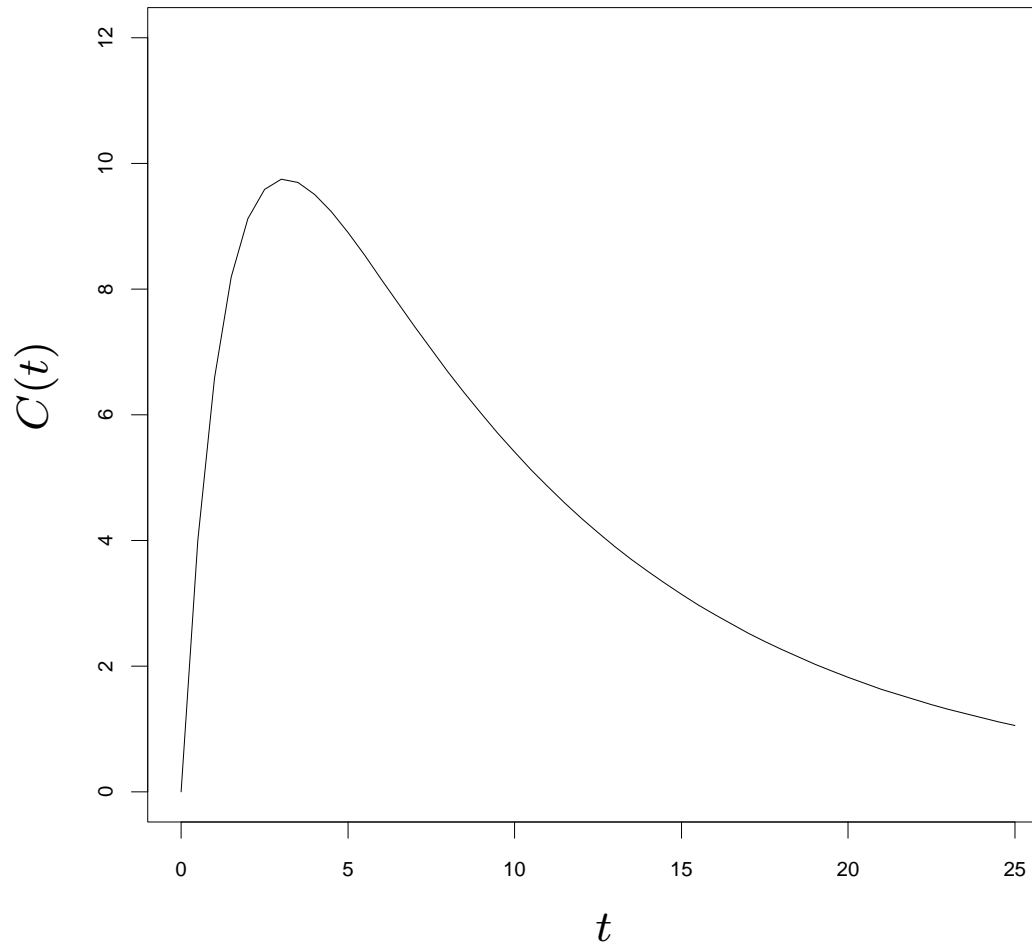
should describe concentration of drug at time  $t$  *within a single subject* to whom oral dose  $D$  was administered at time  $t = 0$

- The expression for  $C(t)$  involves three *parameters*  $\theta = (k_a, k_e, V)^T$ .
- If we *knew*  $\theta$ , could predict concentration at any time  $t$  following any oral dose  $D$  for this subject (*dosing recommendations*)
- Based on *data* (concentrations over time), could learn about *unknown*  $\theta$  for this subject
- More shortly. . .

# Recap of mathematical models

---

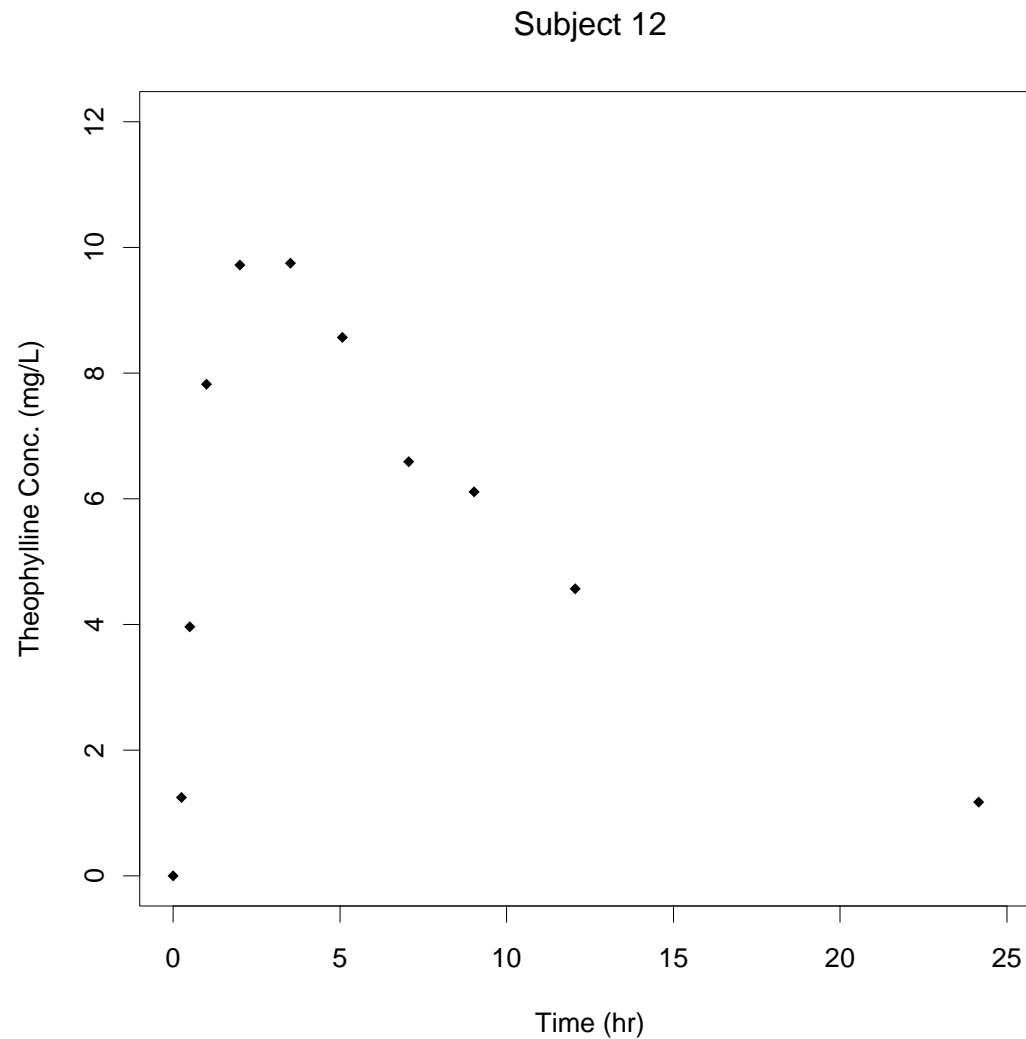
**For example:** With  $k_a = 0.7$ ,  $k_e = 1.1$ ,  $V = 0.4$



# Recap of mathematical models

---

**Data:** From a single subject



# Recap of mathematical models

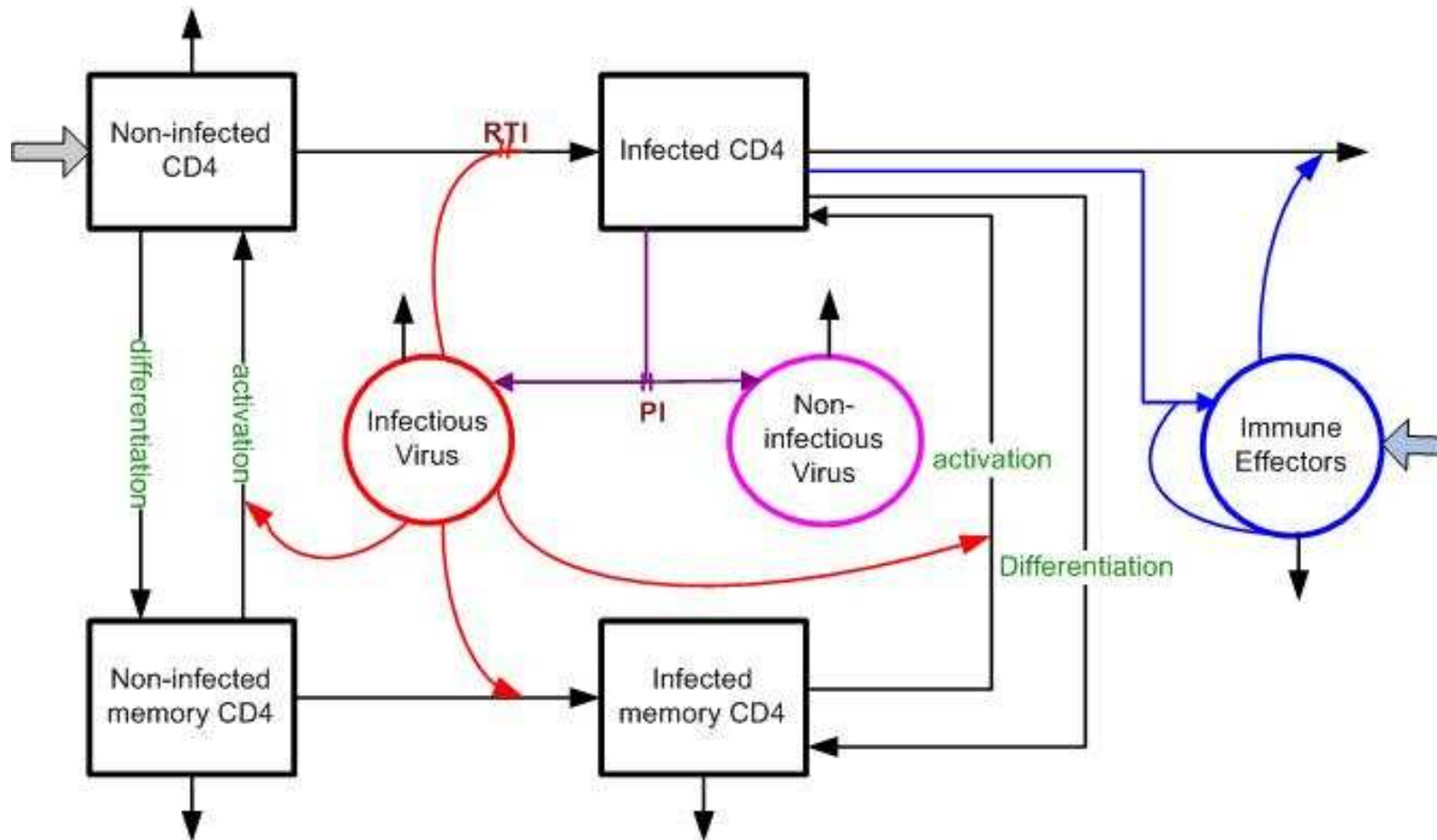
---

**Example 2:** Dynamics of HIV under *antiretroviral* therapy (ARV)

- HIV (*Human immunodeficiency virus*) infects *target cells* in the immune system and uses them to produce *more virus* (that then infects more cells. . . )
- The immune system *responds* to remove infected cells
- A model describes the *interplay* between HIV and immune system taking place *within* a single subject over time
- Can use to *predict* progression of infection under different ARV regimens
- *Reverse transcriptase inhibitor* (*RTI*) *blocks* infectious virus from infecting target cells
- *Protease inhibitor* (*PI*) causes infected cells to produce *non-infectious virus*

# Recap of mathematical models

Possible model for within-subject dynamics:



# Recap of mathematical models

---

## Typical model components:

$T_1, T_1^*$	Type 1 target cells (e.g, CD4 <sup>+</sup> T cells), uninfected, infected
$T_2, T_2^*$	Type 2 target cells (e.g., macrophages), uninfected, infected
$V_I, V_{NI}$	infectious and non-infectious free virus
$E$	cytotoxic T-lymphocytes

- Uninfected cells ( $T_j$ ) become infected ( $T_j^*$ ) via encounters with infectious virus ( $V_I$ ); infection rates  $k_j$
- Uninfected cell source rates  $\lambda_j$  and natural death rates  $d_j$
- Infected cell death rate  $\delta$ ; each cell produces  $N_T$  virions
- Virus natural death rate  $c$
- $\epsilon_1, f\epsilon_2$  govern efficacy of *RTI* in blocking new infections of  $T_1, T_2$
- $\epsilon_2, f\epsilon_2$  govern efficacy of *PI* in causing  $T_1^*, T_2^*$  to produce non-infectious virus

# Recap of mathematical models

---

**Mathematical model:** 7 compartments (states)

$$\begin{aligned}\dot{T}_1 &= \lambda_1 - d_1 T_1 - \{1 - \epsilon_1 u(t)\} k_1 V_I T_1 \\ \dot{T}_2 &= \lambda_2 - d_2 T_2 - \{1 - f \epsilon_1 u(t)\} k_2 V_I T_2 \\ \dot{T}_1^* &= \{1 - \epsilon_1 u(t)\} k_1 V_I T_1 - \delta T_1^* - m_2 E T_1^* \\ \dot{T}_2^* &= \{1 - f \epsilon_1 u(t)\} k_2 V_I T_2 - \delta T_2^* - m_2 E T_2^* \\ \dot{V}_I &= \{1 - \epsilon_2 u(t)\} 10^3 N_T \delta (T_1^* + T_2^*) - c V_I - \{1 - \epsilon_1 u(t)\} \rho_1 10^3 k_1 T_1 V_I \\ &\quad - \{1 - f \epsilon_1 u(t)\} \rho_2 10^3 k_2 T_2 V_I \\ \dot{V}_{NI} &= \epsilon_2 u(t) 10^3 N_T \delta (T_1^* + T_2^*) - c V_{NI} \\ \dot{E} &= \lambda_E + \frac{b_E (T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_b} E - \frac{d_E (T_1^* + T_2^*)}{(T_1^* + T_2^*) + K_d} E - \delta_E E\end{aligned}$$

- $\theta = (\lambda_1, d_1, \epsilon_1, k_1, \dots)$  plus initial conditions

$$\{T_1(0), T_2(0), T_1^*(0), T_2^*(0), V_I(0), V_{NI}(0), E(0)\}$$

- $u(t) =$  ARV input at  $t$  ( $0 \leq u(t) \leq 1$ ,  $0 =$  off,  $1 =$  on)

# Recap of mathematical models

---

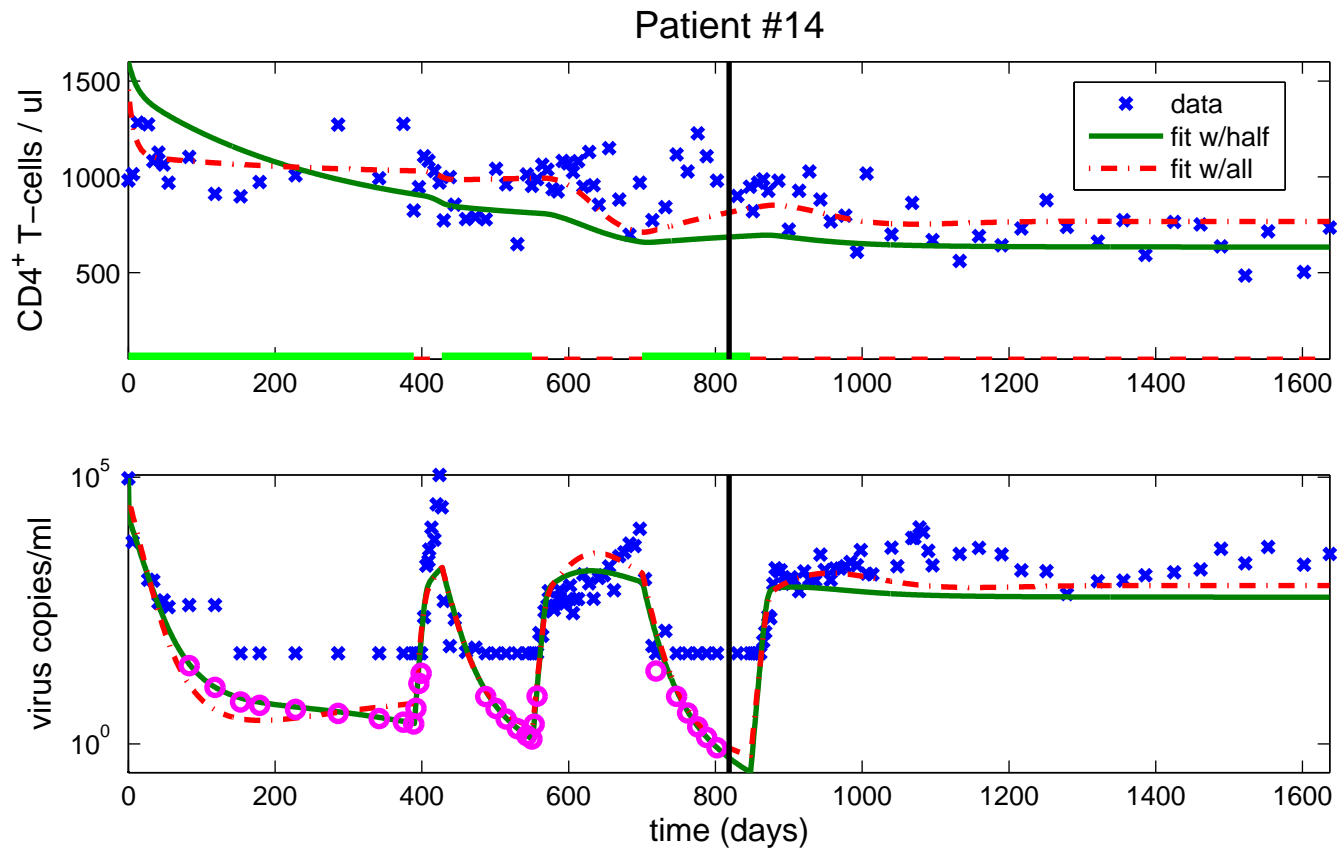
**Thus:**

$$x(t) = \{T_1(t), T_2(t), T_1^*(t), T_2^*(t), V_I(t), V_{NI}(t), E(t)\}^T$$

$$\dot{x}(t) = g\{t, x(t), \theta\} \text{ as on previous slide}$$

- Usual data: *CD4<sup>+</sup> T cell count* =  $T_1 + T_1^*$ , *viral load* =  $V_I + V_{NI}$
- So only *observe* some of the *states* (and do not observe any state explicitly)
- More later...

# Recap of mathematical models



# Recap of mathematical models

---

## Again:

- If we knew  $\theta$ , could predict *viral load*, *CD4 count*, etc, at any time  $t$  under different ARV regimens
- If we had *data* on CD4 counts, viral load under a known regimen followed by the subject, could learn about *unknown*  $\theta$  for this subject (we hope)
- Could use this framework to *design* ARV regimens
- In particular, *adaptive regimens* that use what is known about the subject so far to dictate the best thing to do next
- More later in the course...

# Models and data

---

## Goals of (mathematical) modeling:

- Describe (*quantitatively*) known and hypothesized mechanisms governing behavior of a system of interest. . .
- . . . recognizing that the model is a *simplified depiction* of the *real system* that we would like to understand
- Use the models with *data* (i.e., observations on the real system) to learn about underlying mechanisms

**Forward solution:** Given  $\theta$ , predict observations on the system over time

- Need to *solve* to obtain expressions for the states at any time  $t$
- *Simple models:* A *closed form* expression available for all states at any  $t$
- *More complex models:* Solution at any time  $t$  must be obtained *numerically*

# Models and data

---

**Inverse problem:** Given observations on the system over time, determine the (*unknown*)  $\theta$  governing them

- I.e., observe *data* on the system over time: record  $y_1, \dots, y_n$  at times  $0 \leq t_1 < \dots < t_n$  on  $x(t_1), \dots, x(t_n)$
- Statisticians call this *parameter estimation*
- Primary focus in much of this course is the *inverse problem*
- *Ideally*: Observe *all states* over time, as above ( $y_j$  are *vectors*)
- *In reality*: Observe only some states (PK example) or functions of some states (HIV example)

**For now:** Will start with simplest case of PK example, where  $y_1, \dots, y_n$  are *scalar* observations on a single state, then *generalize*

# Models and data

---

**Critical issue:** *Data* are subject to *variation* and *uncertainty*

- Indeed, observations usually *do not* track exactly on  $y = x(t, \theta)$
- Must recognize this and take it into appropriate account
- *Statistical inference*

**Notation:**

- Solution to the entire system =  $x(t, \theta)$  (all states)
- Solution for the part of the system that is observed is denoted  $f(t, \theta)$ , defined as

$$f(t, \theta) = \mathcal{O}x(t, \theta)$$

for “*observation matrix*”  $\mathcal{O}$  with  $\#$  columns =  $\#$  states

- E.g., for the *PK example* (2 states)  $\mathcal{O} = (1, 0)^T$  ( $1 \times 2$ ) (and divide by  $V$ )

# Models and data

---

**Theophylline study:** PK in *humans* following oral dose

- 12 “*healthy volunteers*” each given dose  $D$  (mg/kg) at time  $t = 0$
- Blood samples drawn at 10 subsequent time points over the next 25 hours for each subject
- Samples *assayed* for theophylline concentration
- *Observe*  $y_1, \dots, y_{10}$  at times  $t_1, \dots, t_{10}$  on each subject

**Objectives:**

- For a *specific subject*, learn about absorption, distribution, and elimination by determining  $k_a, k_e, V \Rightarrow$  dosing recommendations for this subject – we tackle this first
- Learn about how absorption, distribution, and elimination differ from subject to subject  $\Rightarrow$  dosing recommendations for the *population* of likely subjects – later in the course

# Models and data

---

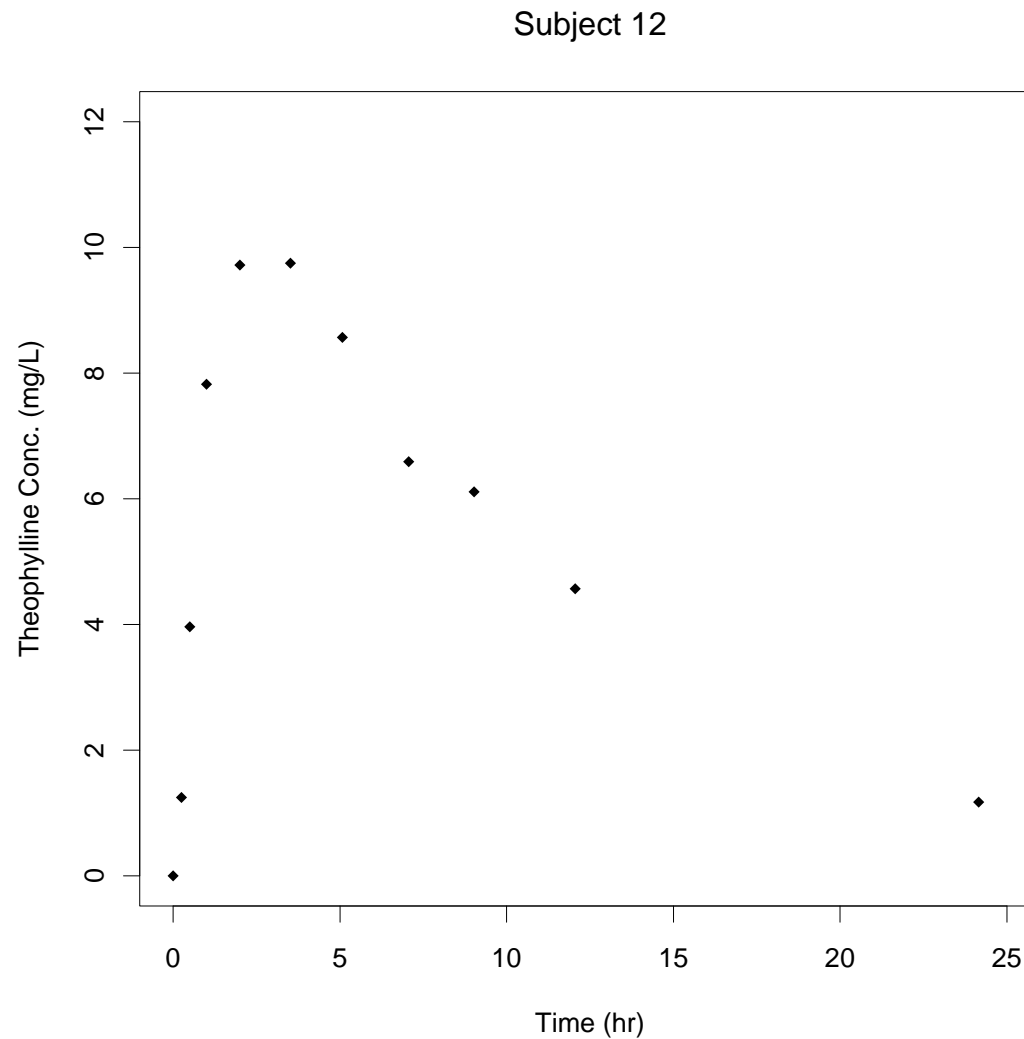
**Model-based expression for theophylline concentration  $C(t)$  at time  $t$  for a single subject:**

$$f(t, \theta) = \frac{k_a F D}{V(k_a - k_e)} \{e^{-k_e t} - e^{-k_a t}\}, \quad \theta = (k_a, k_e, V)^T$$

- We take  $F \equiv 1$
- Note that  $f(t, \theta)$  *also* depends on the dose  $D$ ; we suppress this dependence in the notation for now

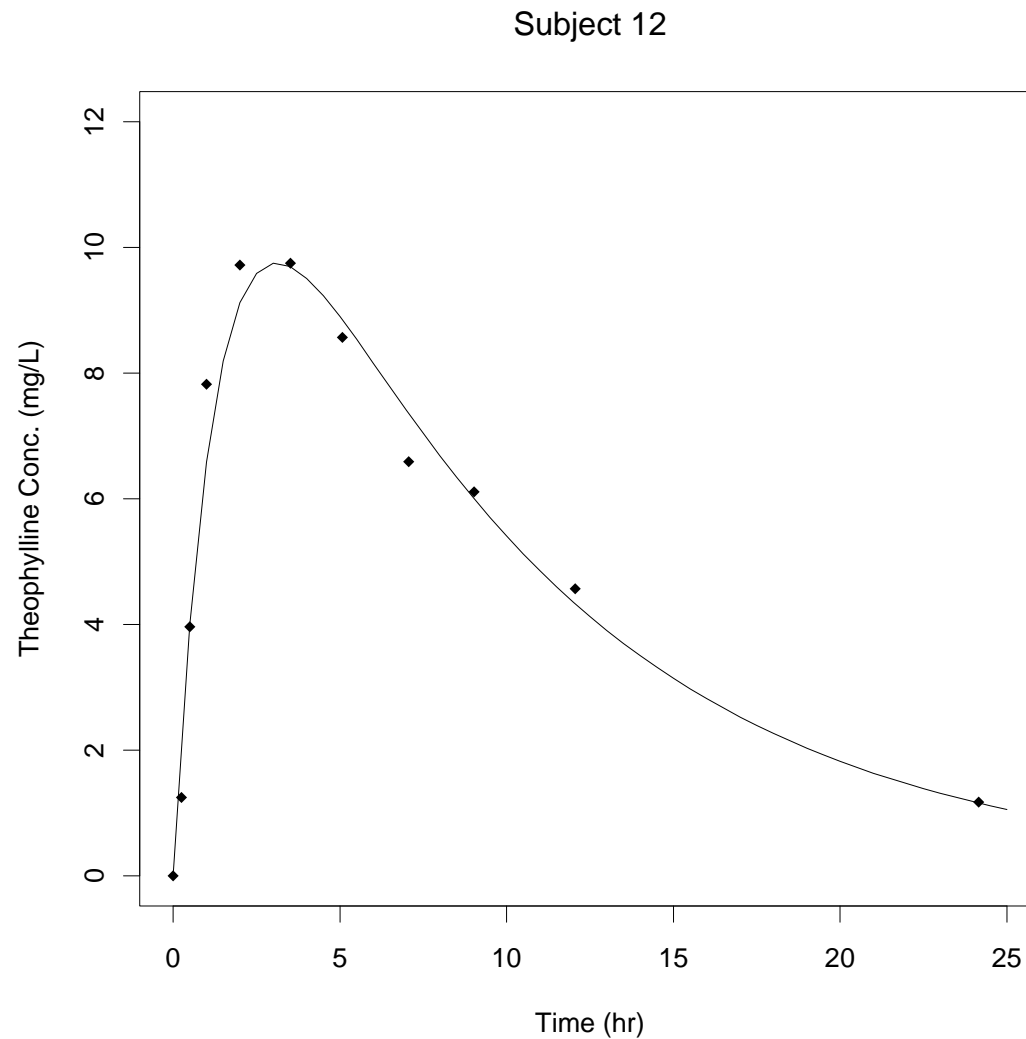
# Models and data

**Data for subject 12:** Plot of concentration vs. time



# Models and data

**Data for subject 12:** With “fitted model” superimposed



# Models and data

---

## Remarks:

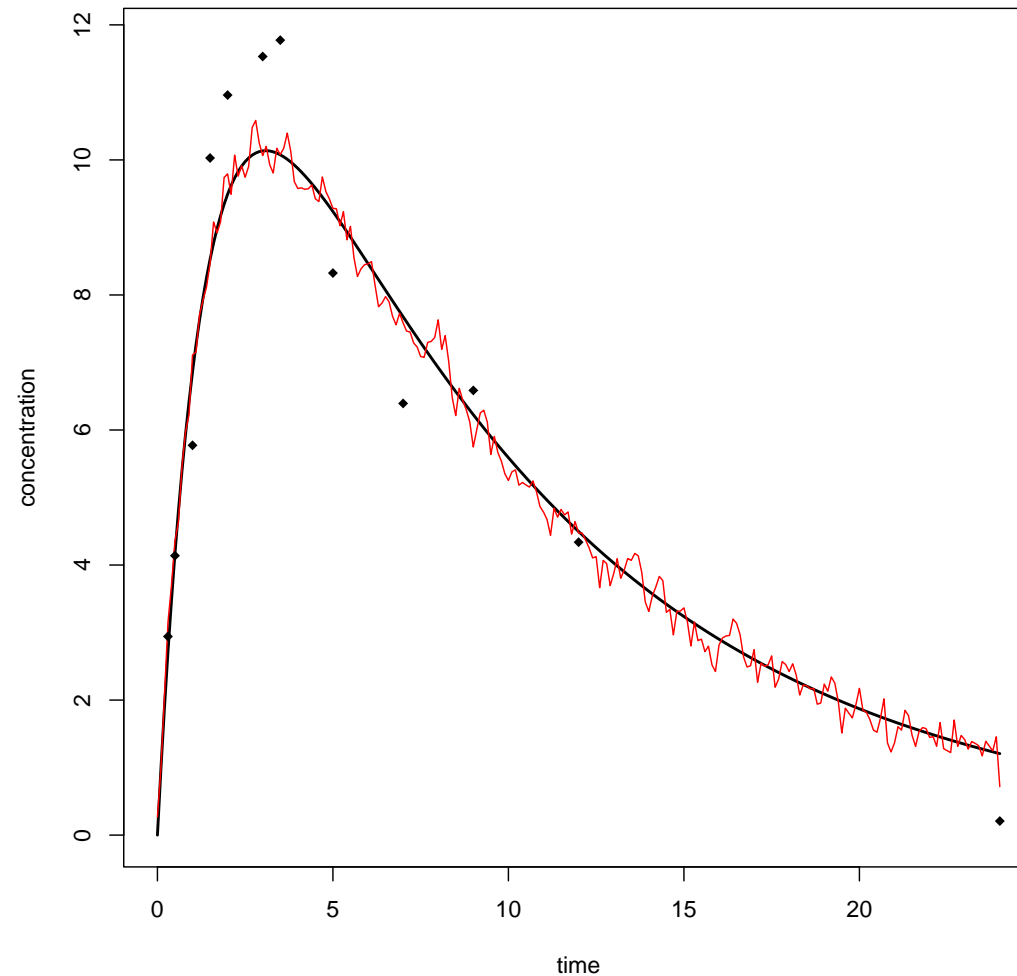
- Observed concentrations appear to trace out a pattern over time quite similar to that dictated by the one compartment model
- But they do not lie *exactly* on a smooth trajectory
- “*Observation error*”

## Why?

- One obvious reason: Assay is not perfect, cannot measure concentration *exactly* (measurement error)
- Other reasons?

# Models and data

**Approximation:** Model is an *idealized* representation of a more complicated biological process



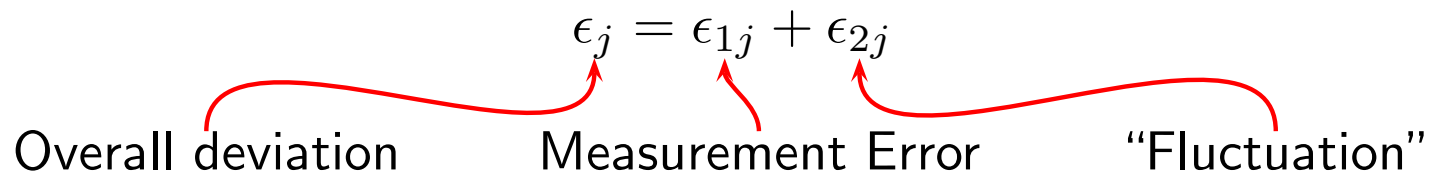
# Models and data

---

**Standard conceptualization:** Think of what we *observe* as

$$y_j = f(t_j, \theta) + \epsilon_j$$

- $f(t, \theta) = C(t)$ , a function of  $\theta = (k_a, k_e, V)^T$
- $\epsilon_j$  is the *deviation* between what the (deterministic) model dictates we would see at  $t_j$  and what we actually observe
- Here,  $\epsilon_j$  represents deviations from  $f(t_j, \theta)$  due to *measurement error*, “*biological fluctuations*”



# Models and data

---

**Thought experiment:** Consider measurement error

- A particular blood sample has a “*true*” concentration of theophylline
- When we measure this concentration, an error is committed, which causes “*observed*” to deviate from “*true*” by an amount that is negative or positive
- Suppose we were to measure the same sample *over and over* (zillions of times) – each time, a possibly different error is committed
- So all such observations would turn out *differently*, even though, *ideally*, they should be all the *same* (measuring the *same thing*)

# Models and data

---

**Result:** Measurement error is a *source of variation* that leads to *uncertainty* in what we *observe*

- In actuality, we measure the concentration only *once*
- The error that results may be thought of as drawn from a “*population*” of *all possible errors* that could be committed when measuring concentration
- ⇒ *UNCERTAINTY* – the observation could have turned out *differently*
- Errors, and hence, observations, are *variable*

# Models and data

---

**Recall:** Would like to determine  $\theta$  from the pairs  $(y_j, t_j)$ ,  $j = 1, \dots, n$

- Any determination of  $\theta$  we try to make from these observations will be subject to *uncertainty*
- That is, if we *estimate*  $\theta$  from data subject to measurement error (and *other sources of variation*), the estimate *could have turned out differently*

**We will see:**

- Failure to acknowledge this can lead to *erroneous conclusions*
- Acknowledging this requires a *formal* way to describe and assess uncertainty, and thus limitations of what can be learned from *data*

# Models and data

---

**Deterministic models:** Representation of the “ideal relationship” between model states and time

**Statistical models:** Representation of the “actual relationship” between *observations* on model states and time

- Incorporate sources of *uncertainty* (variation)
- Framework for *formalizing* assumptions about the effects of variation
- Main tool: *probability*

**Statistical model for observed theophylline concentrations:**

$$Y_j = f(t_j, \theta) + \epsilon_j, \quad j = 1, \dots, n$$

- Think of  $\epsilon_j$  as a *random variable* with a *probability distribution* that characterizes “*populations*” of possible values of phenomena like measurement errors, fluctuations that might occur at  $t_j$

# Models and data

---

## Statistical model for observed theophylline concentrations:

$$Y_j = f(t_j, \theta) + \epsilon_j, \quad j = 1, \dots, n$$

- If  $\epsilon_j$  is a random variable, then so is what we observe
- $\Rightarrow Y_j$  is a random variable with a *probability distribution* that characterizes how observations at  $t_j$  on this subject may vary because of measurement error, fluctuations, etc.
- The model describes pairs  $(Y_j, t_j)$ ,  $j = 1, \dots, n$ , we might see; i.e., the model describes the *mechanism* by which data are thought to arise
- *Data* we observe are realizations of  $Y_j$ ,  $j = 1, \dots, n$ :  $y_1, \dots, y_n$
- The *mechanism* is characterized by assumptions on the *probability distribution* of  $\epsilon_j$  (so, equivalently, on that of  $Y_j$ )

# Statistical models and sources of variation

---

**In general:** *Random variables* and *probability distributions* are the building blocks of *statistical models*

- A *statistical model* is a representation of the *mechanism* by which *data* are assumed to arise
- Phenomena that are *subject to variation* and hence give rise to *uncertainty* in the way data may “*turn out*” are represented by *random variables*
- Assumptions on the nature of *probability distributions* for random variables in statistical models represent assumptions on the nature and extent of such variation
- *Mathematical models* that describe the “idealized” relationship need to be *embedded* in such a model in an appropriate way and given an appropriate interpretation
- Continue with the *theophylline example* for a demonstration...

# Statistical models and sources of variation

---

**Recap:** Theophylline PK on a single subject

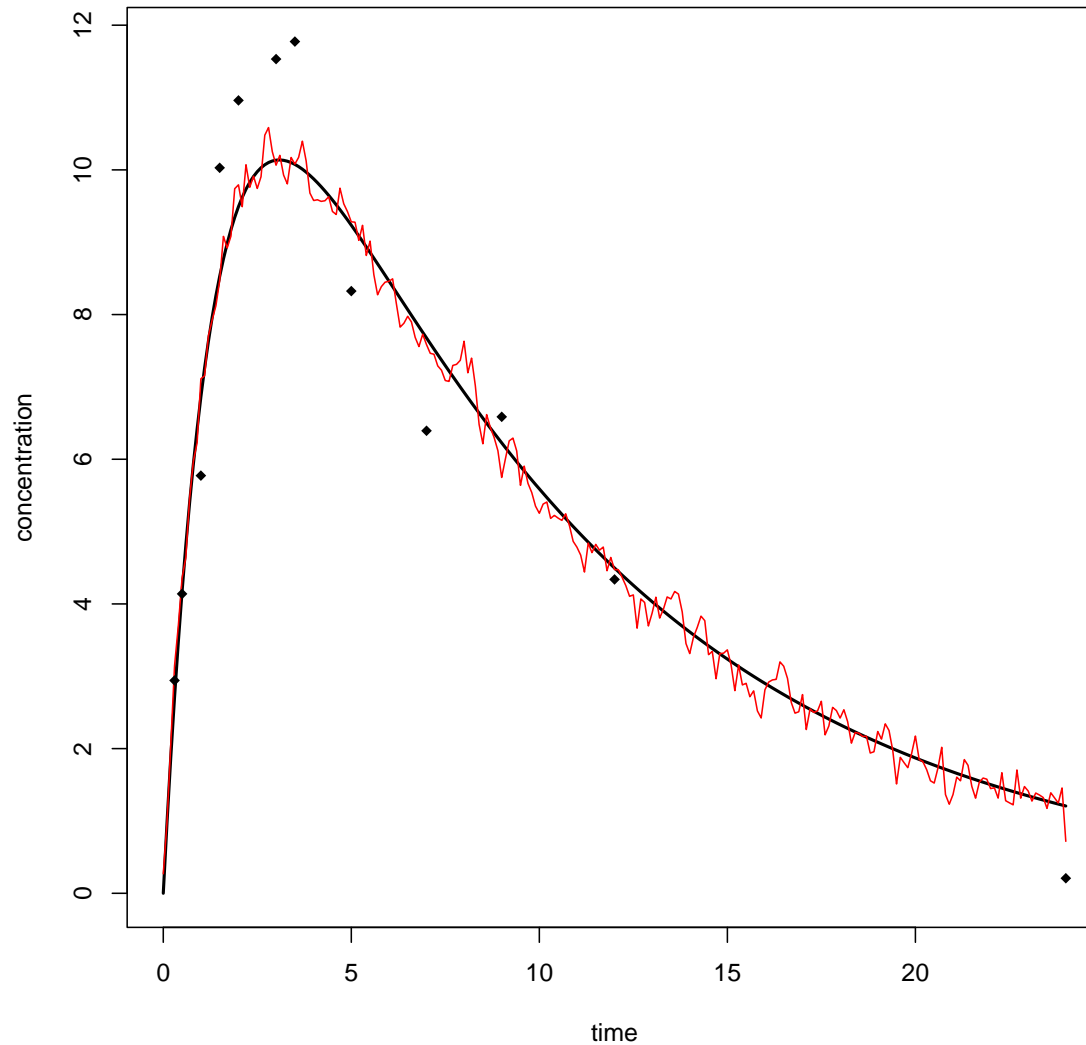
$$Y_j = f(t_j, \theta) + \epsilon_j, \quad j = 1, \dots, n$$

$$f(t, \theta) = \frac{k_a F D}{V(k_a - k_e)} \{e^{-k_e t} - e^{-k_a t}\}, \quad \theta = (k_a, k_e, V)^T$$

- $f(t, \theta)$  derived from the *deterministic* compartment model; also depends on other stuff (*dose D* at  $t = 0$ )
- $\epsilon_j$  represents *deviation* that causes observations to not fall exactly on the smooth path  $f(t, \theta)$
- Aggregate effects of *measurement error*, “*biological fluctuations*,” other phenomena
- $\Rightarrow \epsilon_j$  is a *random variable* whose *probability distribution* reflects assumed features of these phenomena
- And  $Y_j$  is also a *random variable* (transformation of  $\epsilon_j$ )

# Statistical models and sources of variation

## Conceptual representation:



# Statistical models and sources of variation

---

**More formal:** In principle, the PK *process* could be observed at *any time*

- $Y(t)$  is observed concentration that would be seen at time  $t$  and  $\epsilon(t)$  is the corresponding deviation under conditions  $U$  ( $U = \text{dose } D$  at  $t = 0$ )

$$Y(t) = f(t, U, \theta) + \epsilon(t), \quad t \geq 0$$

- Could have  $U = U(t)$  (e.g., HIV example); suppress for now
- Represents the assumed data generating mechanism at *any time*  $t$
- $Y(t)$ ,  $\epsilon(t)$  are *stochastic processes* – *random function* of time with sample space of possible values (functions); e.g.,  $y(t)$ ,  $t \geq 0$  (*sample paths*)
- For a *fixed set* of times  $t_1 < \dots < t_n$  write

$$Y_j = Y(t_j) = f(t_j, U, \theta) + \epsilon(t_j)$$

to represent observations that would be seen at these times under conditions  $U$

- $Y_j = Y(t_j)$ ,  $\epsilon_j = \epsilon(t_j)$

# Statistical models and sources of variation

---

$$Y(t_j) = f(t_j, U, \theta) + \epsilon(t_j)$$

**Thus:**

- $\{\epsilon(t_1), \epsilon(t_2), \dots, \epsilon(t_n)\}^T = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$  and  $\{Y(t_1), Y(t_2), \dots, Y(t_n)\}^T = (Y_1, Y_2, \dots, Y_n)^T$  are *random vectors*
- Can consider the *joint probability distribution* of  $(\epsilon_1, \epsilon_2, \dots, \epsilon_n)^T$  [and thus of  $(Y_1, Y_2, \dots, Y_n)^T$ ]
- *Technical point 1*: Probability distribution for a stochastic process arises from thinking of *all possible* such vectors and their joint distributions for all  $n$  (infinitely many)
- *Technical point 2*: Probability distributions here depend on conditions  $U$  and are hence *conditional probability distributions* (conditional on  $U$ )

# Statistical models and sources of variation

---

$$Y(t_j) = f(t_j, U, \theta) + \epsilon(t_j)$$

**Nature of  $\epsilon(t)$ :** “*Biological fluctuations,*” *measurement error*

$$\epsilon(t) = \epsilon_1(t) + \epsilon_2(t)$$

- $\epsilon_1(t_j) = \epsilon_{1j}$  represents *measurement error* that could be committed at fixed time  $t_j$
- $\epsilon_2(t_j) = \epsilon_{2j}$  represents “*fluctuation*” that might occur at  $t_j$
- These random variables are *continuous* – concentrations *in principle* can take on *any value* (although we may be limited in what we may actually observe due to limits on resolution of measurement)
- Write  $\{\epsilon_1(t_1), \dots, \epsilon_1(t_n)\}^T = (\epsilon_{11}, \dots, \epsilon_{1n})^T$  and  $\{\epsilon_2(t_1), \dots, \epsilon_2(t_n)\}^T = (\epsilon_{21}, \dots, \epsilon_{2n})^T$  – *random vectors*

# Statistical models and sources of variation

---

**Measurement error:** “*Reasonable*” assumptions on  $\epsilon_1(t)$  and hence on aspects of the *joint probability distribution* of  $(\epsilon_{11}, \dots, \epsilon_{1n})^T$  (conditional on  $U$ )

- Measuring device is *unbiased* – does not systematically err in one direction

$$E\{\epsilon_1(t)|U\} = 0 \quad \forall t \Rightarrow E(\epsilon_{1j}|U) = 0 \quad \text{for each } j = 1 \dots, n$$

(All possible errors for measuring concentration for the sample taken at any  $t_j$  “average out” to zero)

- In fact, negative or positive errors are *equally likely*  $\Rightarrow$  the *marginal conditional probability density* of  $\epsilon_{1j}$  is *symmetric* for each  $j$
- Measurement errors at any two times are “*unrelated*”

$$\epsilon_1(t) \perp\!\!\!\perp \epsilon_1(s) | U \quad \forall t, s \Rightarrow \epsilon_{1j} \perp\!\!\!\perp \epsilon_{1j'} | U \Rightarrow \text{cov}(\epsilon_{1j}, \epsilon_{1j'} | U) = 0$$

# Statistical models and sources of variation

---

**Measurement error:** “*Reasonable*” assumptions on  $\epsilon_1(t)$  and hence on aspects of the *joint probability distribution* of  $(\epsilon_{11}, \dots, \epsilon_{1n})^T$  (conditional on  $U$ )

- *Variation* among all errors that might occur at any time is of the *same* magnitude

$$\text{var}\{\epsilon_1(t)|U\} = \sigma_1^2 \quad \forall t \quad \Rightarrow \quad \text{var}(\epsilon_{1j}|U) = \sigma_1^2$$

for all  $j$  (unaffected by time or “actual concentration” in the sample at  $t_j$ ) – *is this realistic?*

- In many cases, *NO* – measurement error variance tends to *increase* with increasing magnitude of the true concentration being measured – approximate by

$$\text{var}\{\epsilon_1(t)|U\} \text{ is a function of } f(t, U, \theta)$$

# Statistical models and sources of variation

---

**“Biological fluctuations”**: “*Reasonable*” assumptions on aspects of the *joint probability distribution* of  $(\epsilon_{21}, \dots, \epsilon_{2n})^T$  (conditional on  $U$ )

- Fluctuations tend to “track” the smooth trajectory  $f(t, U, \theta)$  over time (sample path) but can be “above” or “below” at any time

$$E\{\epsilon_2(t)|U\} = 0 \quad \forall t \quad \Rightarrow \quad E(\epsilon_{2j}|U) = 0$$

(All possible fluctuations at any time “average out” to zero)

- In fact, negative or positive fluctuations at a particular time are *equally likely*  $\Rightarrow$  the *marginal conditional probability density* of  $\epsilon_{2j}$  is *symmetric* for each  $j$
- *Variation* among fluctuations that might occur at any time is *same* at all times

$$\text{var}\{\epsilon_2(t)|U\} = \sigma_2^2 \quad \forall t \quad \Rightarrow \quad \text{var}(\epsilon_{2j}|U) = \sigma_2^2 \quad \forall j$$

# Statistical models and sources of variation

---

**“Biological fluctuations”**: “*Reasonable*” assumptions on aspects of the *joint probability distribution* of  $(\epsilon_{21}, \dots, \epsilon_{2n})^T$  (conditional on  $U$ )

- Fluctuations “*close together*” in time tend to behave “*similarly*,” with extent of “*similarity*” decreasing as the times grow more apart

$$\text{cov}\{\epsilon_2(t), \epsilon_2(s)|U\} = C(|t-s|) \quad \text{and} \quad \text{corr}\{\epsilon_2(t), \epsilon_2(s)|U\} = c(|t-s|) \quad \forall t, s$$

for decreasing functions  $C(\cdot)$ ,  $c(\cdot)$  with  $C(0) = \sigma_2^2$  and  $c(0) = 1$

- Hence at times  $t_j, t_{j'}$

$$\text{cov}(\epsilon_{2j}, \epsilon_{2j'}|U) = C(|t_j - t_{j'}|) \quad \text{and} \quad \text{corr}(\epsilon_{2j}, \epsilon_{2j'}|U) = c(|t_j - t_{j'}|).$$

# Statistical models and sources of variation

---

## “Biological fluctuations,” continued:

- E.g., for  $\text{corr}(\epsilon_{2j}, \epsilon_{2j'}) = c(|t_j - t_{j'}|)$ ,

$$c(u) = \exp(-\phi u^2)$$

(so correlation between fluctuations at two times is *nonnegative*, reflecting “*similarity*”)

- Extent and direction of measurement error at any time  $t$  is *unrelated* to fluctuations at  $t$  or any other time

$$\epsilon_1(t) \perp\!\!\!\perp \epsilon_2(s) | U \quad \forall t, s \Rightarrow \epsilon_{1j} \perp\!\!\!\perp \epsilon_{2j'}$$

for any  $t_j, t_{j'}, j, j' = 1, \dots, n$

# Statistical models and sources of variation

---

## Remarks:

- The foregoing assumptions are not the *only* assumptions one could make, but exemplify the considerations involved
- The *normal probability distribution* is a natural choice to represent the assumption of *symmetry*

# Statistical models and sources of variation

---

**Recapping the assumptions:**  $\epsilon_j = \epsilon_{1j} + \epsilon_{2j}$

- $E(\epsilon_{1j}|U) = 0, E(\epsilon_{2j}|U) = 0 \Rightarrow E(\epsilon_j|U) = 0$

- $\text{var}(\epsilon_{1j}|U) = \sigma_1^2, \text{var}(\epsilon_{2j}|U) = \sigma_2^2, \text{ and } \epsilon_{1j} \perp\!\!\!\perp \epsilon_{2j}|U \text{ for all } j$

$$\Rightarrow \text{var}(\epsilon_j|U) = \sigma_1^2 + \sigma_2^2$$

- $\text{cov}(\epsilon_{1j}, \epsilon_{1j'}|U) = 0, \text{cov}(\epsilon_{2j}, \epsilon_{2j'}|U) = \sigma_2^2 c(|t_j - t_{j'}|) = \sigma_2^2 e^{-\phi(t_j - t_{j'})^2}$

- Conditional on  $U$ ,  $(\epsilon_{11}, \dots, \epsilon_{1n})^T$  has *mean vector 0* and *covariance matrix*

$$\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_1^2 \end{pmatrix} = \sigma_1^2 I_n$$

and has a *multivariate normal distribution*

# Statistical models and sources of variation

**Recapping the assumptions:**  $\epsilon_j = \epsilon_{1j} + \epsilon_{2j}$

- Conditional on  $U$ ,  $(\epsilon_{21}, \dots, \epsilon_{2n})^T$  has *mean vector 0* and *covariance matrix*

$$\sigma_2^2 \begin{pmatrix} 1 & e^{-\phi(t_1-t_2)^2} & \dots & e^{-\phi(t_1-t_n)^2} \\ e^{-\phi(t_1-t_2)^2} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & e^{-\phi(t_{n-1}-t_n)^2} \\ e^{-\phi(t_1-t_n)^2} & \dots & e^{-\phi(t_{n-1}-t_n)^2} & 1 \end{pmatrix} = \sigma_2^2 \Gamma$$

and has a *multivariate normal distribution*

- So, conditional on  $U$ ,  
 $(\epsilon_1, \dots, \epsilon_n)^T = (\epsilon_{11}, \dots, \epsilon_{1n})^T + (\epsilon_{21}, \dots, \epsilon_{2n})^T$  has mean vector 0  
and covariance matrix

$$\sigma_1^2 I_n + \sigma_2^2 \Gamma$$

# Statistical models and sources of variation

---

**Thus:** Implications for  $Y = (Y_1, \dots, Y_n)^T$

- $E(Y_j|U) = f(t_j, U, \theta) + E(\epsilon_j|U) = f(t_j, U, \theta)$
- Thus, may think of  $f(t, U, \theta)$  as the result of averaging across all possible *sample paths of the fluctuation process* and *measurement errors*, so representing the “*inherent trajectory*” for subject 12
- $\text{var}(Y_j|U) = \text{var}(\epsilon_j|U) = \sigma_1^2 + \sigma_2^2$
- $\text{cov}(Y_j, Y_{j'}|U) = \text{cov}(\epsilon_j, \epsilon_{j'}|U) = \sigma_2^2 \exp\{-\phi(t_j - t_{j'})^2\}$
- $Y_j$  is normally distributed (conditional on  $U$ )
- The *random vector*  $Y = (Y_1, \dots, Y_n)^T$  has a (conditional on  $U$ ) *multivariate normal distribution* with *mean vector* and *covariance matrix*

$$f(U, \theta) = \{f(t_1, U, \theta), \dots, f(t_n, U, \theta)\}^T \quad \text{and} \quad \sigma_1^2 I_n + \sigma_2^2 \Gamma$$

# Statistical models and sources of variation

---

**More succinctly:** We have the *statistical model*

$$Y|U \sim \mathcal{N}_n\{f(U, \theta), \sigma_1^2 I_n + \sigma_2^2 \Gamma\} \quad (3)$$

- Each *marginal* is a normal density, e.g.

$$Y_j|U \sim \mathcal{N}\{f(t_j, U, \theta), \sigma_1^2 + \sigma_2^2\}$$

**Simplifications:** We may be willing to make *simplifying assumptions*

- If the  $t_j$  are *far apart*,  $|t_j - t_{j'}|$  may be *large*, and hence  $\exp\{-\phi(t_j - t_{j'})^2\}$  *close to zero*  $\Rightarrow$  “correlation among fluctuations at  $t_1, \dots, t_n$  is *negligible*”
- *Approximate* by assuming  $\epsilon_{2j} \perp\!\!\!\perp \epsilon_{2j'} | U \Rightarrow \text{cov}(\epsilon_{2j}, \epsilon_{2j'} | U) = 0$  and thus  $\Gamma = I_n$ , which implies

$$Y_j \perp\!\!\!\perp Y_{j'} | U \Rightarrow \text{cov}(Y_j, Y_{j'} | U) = 0,$$

$$\text{and } \text{var}(Y_j | U) = \sigma^2 = \sigma_1^2 + \sigma_2^2$$

# Statistical models and sources of variation

---

**Summarizing:** The *statistical model* becomes

$$Y|U \sim \mathcal{N}_n\{f(U, \theta), \sigma^2 I_n\}, \quad \psi = (\theta^T, \sigma^2)^T \quad (4)$$

- The *statistical model* (4) is the standard one that is typically assumed
- The foregoing development shows the considerations involved in *justifying* this model
- These considerations are almost *never* mentioned in the literature on *inverse problems*; indeed, no reference to a statistical model is typically even made
- *Important*: Just because we assume this statistical model holds doesn't mean we're *correct*
- We might proceed as if the model is correct, but need to worry about the implications if it is *not*; more later

# Statistical models and sources of variation

---

**So far:** For simplicity, we have restricted attention to the situation where  $Y_j$  are *scalars*

- $Y_j$  is theophylline concentration at time  $t_j$
- The data observed at each  $t_j$  may be *multivariate*; i.e.,  $Y_j$  is a *random vector*
- In the *HIV example*, the data observed are *bivariate*; i.e.

$$Y_j = (Y_j^{(1)}, Y_j^{(2)})^T = (\text{CD4 count at } t_j, \text{viral load at } t_j)^T$$

- In terms of the *mathematical model* on slide 14, with the 7 states

$$x(t) = \{T_1(t), T_2(t), T_1^*(t), T_2^*(t), V_I(t), V_{NI}(t), E(t)\}^T$$

we have observations on

$$T_1(t) + T_1^*(t) = \text{CD4 count} \quad \text{and} \quad V_I(t) + V_{NI}(t) = \text{viral load} \quad \text{at } t_1, \dots, t_n$$

# Statistical models and sources of variation

---

## Multivariate observations:

- As on slide 20, a model  $f(t, U, \theta)$  ( $2 \times 1$ ) can be derived from the solution  $x(t, U, \theta)$  to the system  $\dot{x} = g\{t, x(t), \theta\}$  as

$$f(t, U, \theta) = \mathcal{O}x(t, U, \theta),$$

where  $\mathcal{O}$  is the ( $2 \times 7$ ) *observation matrix*

$$\mathcal{O} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

- Thus, we have the bivariate model

$$f(t, U, \theta) = \begin{pmatrix} f^{(1)}(t, U, \theta) \\ f^{(2)}(t, U, \theta) \end{pmatrix}.$$

- *All data*:  $Y = (Y_1^T, \dots, Y_n^T)^T$  ( $2n \times 1$ )

# Statistical models and sources of variation

---

## Bivariate stochastic process:

$$Y(t) = \begin{pmatrix} Y^{(1)}(t) \\ Y^{(2)}(t) \end{pmatrix} = f(t, U, \theta) + \epsilon(t), \quad t \geq 0$$

$$\epsilon(t) = \epsilon_1(t) + \epsilon_2(t) = \begin{pmatrix} \epsilon_1^{(1)}(t) \\ \epsilon_1^{(2)}(t) \end{pmatrix} + \begin{pmatrix} \epsilon_2^{(1)}(t) \\ \epsilon_2^{(2)}(t) \end{pmatrix}$$

- Observations

$$Y_j = \{Y^{(1)}(t_j), Y^{(2)}(t_j)\}^T = (Y_j^{(1)}, Y_j^{(2)})^T, \quad j = 1, \dots, n$$

- *Measurement error* deviations

$$\epsilon_{1j} = \{\epsilon_1^{(1)}(t_j), \epsilon_1^{(2)}(t_j)\}^T = (\epsilon_{1j}^{(1)}, \epsilon_{1j}^{(2)})^T$$

- “*Fluctuation*” deviations

$$\epsilon_{2j} = \{\epsilon_2^{(1)}(t_j), \epsilon_2^{(2)}(t_j)\}^T = (\epsilon_{2j}^{(1)}, \epsilon_{2j}^{(2)})^T$$

# Statistical models and sources of variation

---

**In fact:** Observations on each component of  $Y(t)$  need not even be at the *same times*; for simplicity, we assume that they are here

**Statistical model:** Considerations for *each component* of  $\epsilon(t)$  and hence  $\epsilon_j = \epsilon_{1j} + \epsilon_{2j}$  and  $Y_j = (Y_j^{(1)}, Y_j^{(2)})^T$  are as before

- We might still assume  $\epsilon_{1j} \perp\!\!\!\perp \epsilon_{2j} | U$  and believe

$$E(\epsilon_{1j} | U) = 0, \quad E(\epsilon_{2j} | U) = 0 \Rightarrow E(\epsilon_j | U) = 0$$

based on the same rationale as before, applied to each component *separately*

- We now have to consider our beliefs about *variance and correlation* for *each component* in

$$\epsilon_j = \epsilon_{1j} + \epsilon_{2j} = (\epsilon_{1j}^{(1)}, \epsilon_{1j}^{(2)})^T + (\epsilon_{2j}^{(1)}, \epsilon_{2j}^{(2)})^T$$

- **AND correlations** between  $\epsilon_{kj}^{(1)}$  and  $\epsilon_{kj'}^{(2)}$ , for  $k = 1, 2$  at times  $j, j'$

# Statistical models and sources of variation

---

**Statistical model, continued:** For example, *variances*

- *Measurement error variances*

$$\text{var}(\epsilon_{1j}^{(1)} | U) = \sigma_1^{(1)2}, \quad \text{var}(\epsilon_{1j}^{(2)} | U) = \sigma_1^{(2)2}$$

- “*Fluctuation*” variances

$$\text{var}(\epsilon_{2j}^{(1)} | U) = \sigma_2^{(1)2}, \quad \text{var}(\epsilon_{2j}^{(2)} | U) = \sigma_2^{(2)2}$$

- Thus under the foregoing *independence assumption*

$$\text{var}(\epsilon_j^{(1)} | U) = \sigma_1^{(1)2} + \sigma_2^{(1)2}, \quad \text{var}(\epsilon_j^{(2)} | U) = \sigma_1^{(2)2} + \sigma_2^{(2)2}$$

# Statistical models and sources of variation

---

**Statistical model, continued:** For example, *covariances*

- Between “*fluctuations*” for each  $k = 1, 2$

$$\text{cov}(\epsilon_{2j}^{(k)}, \epsilon_{2j'}^{(k)} | U) = \sigma_2^{(k)2} c^{(k)}(|t_j - t_{j'}|) = \sigma_2^{(k)2} e^{-\phi_k(t_j - t_{j'})^2}$$

(*separate* correlation functions depending on  $\phi_k$  for each  $k = 1, 2$ )

- Between *measurement errors* in the same or different components at *different times*

$$\text{cov}(\epsilon_{1j}^{(k)}, \epsilon_{1j'}^{(k')} | U) = 0, \quad j \neq j', \quad k, k' = 1, 2$$

- Between *measurement errors* in different components at the *same time*

$$\text{cov}(\epsilon_{1j}^{(1)}, \epsilon_{1j}^{(2)} | U) = 0$$

Is this *reasonable*?

# Statistical models and sources of variation

---

**Statistical model, continued:** For example, *covariances*

- Between “*fluctuations*” in “*true*” CD4 counts and viral loads

$$\text{cov}(\epsilon_{2j}^{(1)}, \epsilon_{2j'}^{(2)} | U) = ???$$

Do we believe that the fluctuations in “*true*” CD4 counts and viral loads are *independent*?

# Statistical models and sources of variation

---

**Distributional assumption:** As before

- Could assume  $Y_j|U$  has a *multivariate normal* distribution with whatever *covariance structure* is implied by the assumptions made
- I.e., the *marginal density* for each  $Y_j$  is *bivariate normal*

$$Y_j|U \sim \mathcal{N}_2\{f(t, U, \theta), \mathcal{V}\},$$

where  $\mathcal{V}$  is a  $(2 \times 2)$  covariance matrix

- $\psi = (\theta^T, \xi^T)^T$ , where  $\xi$  is the set of all *covariance parameters* in  $\mathcal{V}$
- Implied distributional model for *all data*  $Y$

# Statistical models and sources of variation

---

**Summarizing:** *Lots* to think about and *make assumptions about* here. . .

**Common practice in inverse problems:** Amounts to assuming a statistical model with *all possible correlations negligible*

- E.g.,  $t_j$  are far apart, each component operates *independently* of the other, etc.
- Implies  $Y_j \perp\!\!\!\perp Y_{j'} | U$ ,  $Y_j^{(1)} \perp\!\!\!\perp Y_j^{(2)} | U$  for all  $j, j'$ , and thus

$$\text{var}(Y_j | U) = \mathcal{V} = \begin{pmatrix} \sigma^{(1)2} & 0 \\ 0 & \sigma^{(2)2} \end{pmatrix},$$

$$\text{var}(Y_j^{(1)} | U) = \sigma^{(1)2} = \sigma_1^{(1)2} + \sigma_2^{(1)2}, \quad \text{var}(Y_j^{(2)} | U) = \sigma^{(2)2} = \sigma_1^{(2)2} + \sigma_2^{(2)2}$$

- *Is this reasonable?*

# Statistical models and sources of variation

---

**Implied statistical model:** The statistical model tacitly assumed in inverse problems is

$$Y_j|U \sim \mathcal{N}_2\{f(t, U, \theta), \mathcal{V}\}, \quad j = 1, \dots, n \quad (5)$$

$$\mathcal{V} = \begin{pmatrix} \sigma^{(1)2} & 0 \\ 0 & \sigma^{(2)2} \end{pmatrix}.$$

- *In fact*, sometimes, the even more restrictive assumption that  $\sigma^{(1)2} = \sigma^{(2)2} = \sigma^2$  is imposed, so that  $\mathcal{V} = \sigma^2 I_2$ !
- Does this make any sense? At the very least, it implies that errors in measurement are of *similar magnitude* for all components of  $Y_j$ , *regardless of different scales of measurement*...

**Clearly:** All of this of course generalizes to observation vectors with  $\geq 2$  components in the obvious way (*even more complicated*)

# Statistical models and sources of variation

---

**Key point:** A *statistical model* like (3), (4), or (5) thus describes *all possible probability distributions* for *random vector*  $Y$  representing the *data generating mechanism* for observations we might see at  $t_1, \dots, t_n$  under conditions  $U$

- E.g., for (3), *possible probability distributions* are specified by different values of the *parameter*  $\psi = (\theta^T, \sigma_1^2, \sigma_2^2, \phi)^T \in \Psi$
- *The big question:* Which value of  $\psi$  truly governs the mechanism?
- We are interested in  $\theta$  (the other components of  $\psi$  are required to describe the model fully, but are a *nuisance* ... more later)

**Objective:** If we *collect data* (so observe *a single realization* of  $Y$ ) under conditions  $U$ , what can we learn about  $\psi$ ?

- ... and how can we account for the fact that things could have turned out *differently* (i.e., a *different realization*)?