Bioinformatics: Network Analysis

Model Fitting

COMP 572 (BIOS 572 / BIOE 564) - Fall 2013

Luay Nakhleh, Rice University

Outline

- Parameter estimation
- Model selection

Parameter Estimation

- Generally speaking, parameter estimation is an optimization task, where the goal is to determine a set of numerical values for the parameters of the model that minimize the difference between experimental data and the model.

- In order to avoid cancelation between positive and negative differences, the objective function to be minimized is usually the sum of the squared differences between data and model.

- This function is commonly called SSE (sum of squared errors), and methods searching for models under SSE are called least-squares methods.

- Parameter estimation for linear systems is relatively easy, whereas parameter estimation of nonlinear systems (that can't be transformed into approximately linear ones) is very hard.

Linear Systems:
Linear Regression Involving a Single Variable

- A single variable y depending on only one other variable x

- A scatter plot gives an immediate impression of whether the relationship might be linear.

- If so, the method of linear regression quickly produces the straight line that optimally describes this relationship.

Linear Systems:
Linear Regression Involving a Single Variable

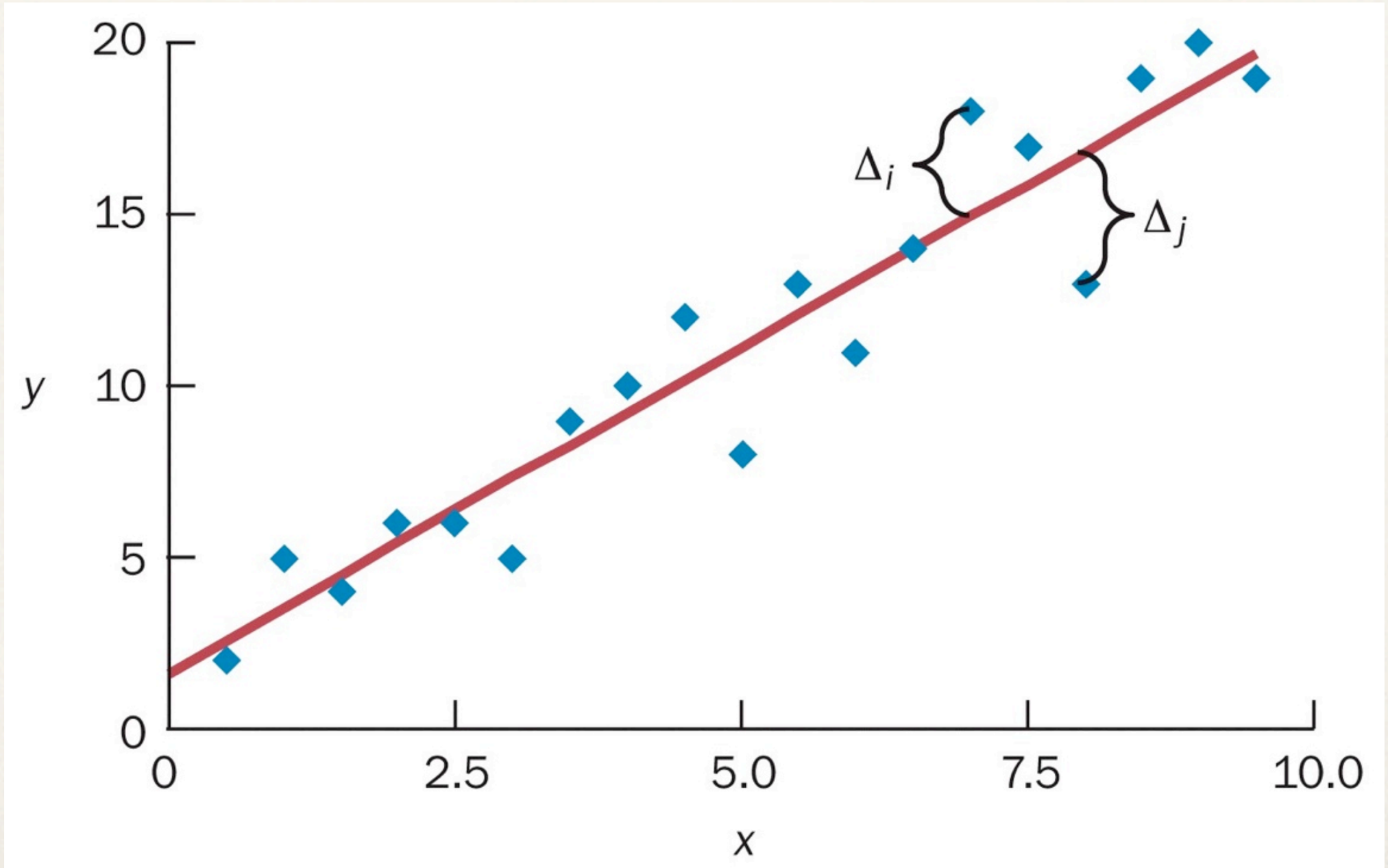


Figure 5.1 A First Course in Systems Biology (© Garland Science 2013)

Linear Systems:
Linear Regression Involving a Single Variable

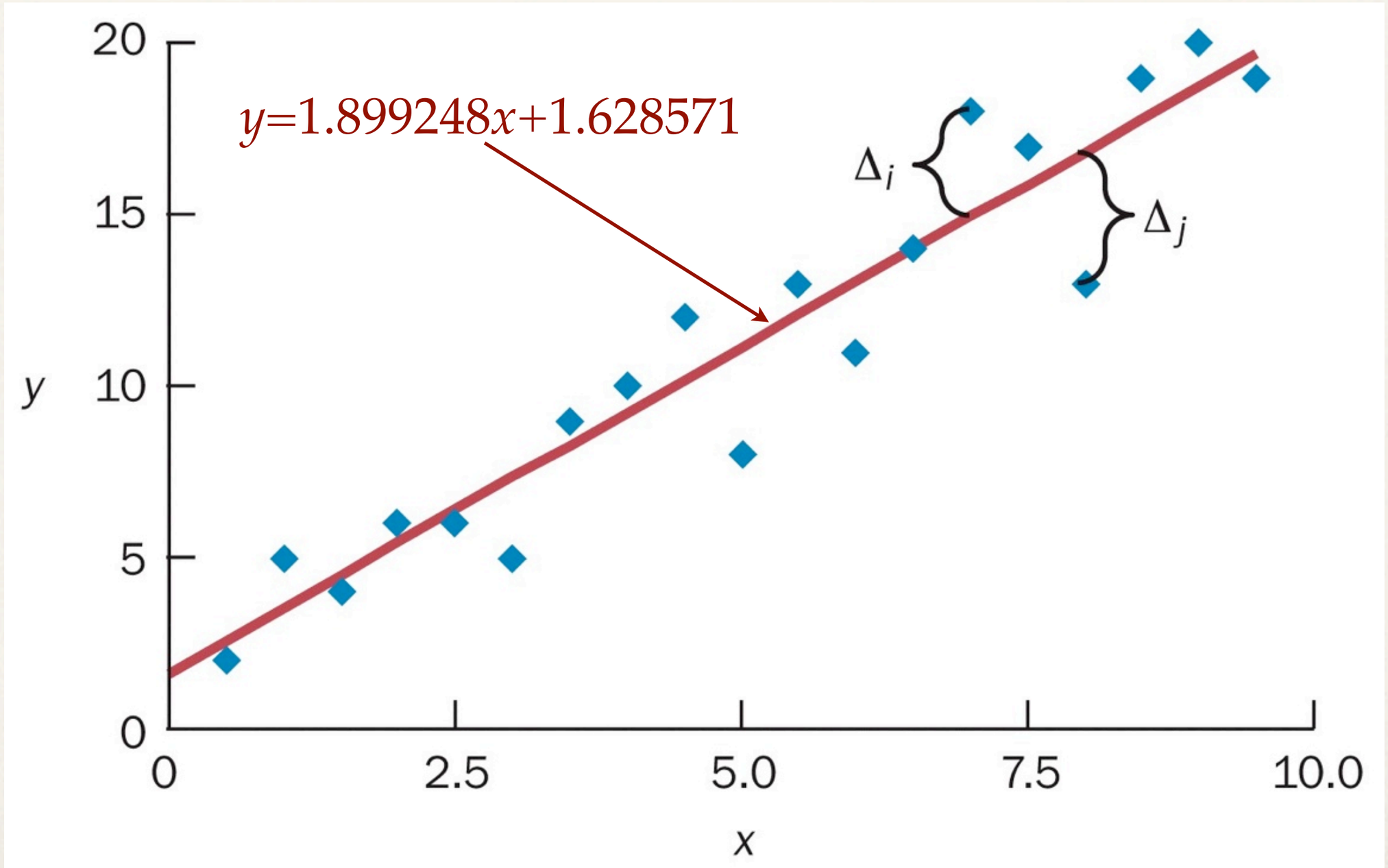


Figure 5.1 A First Course in Systems Biology (© Garland Science 2013)

Linear Systems:
Linear Regression Involving a Single Variable

- Two issues:

- Extrapolations or predictions of y-values beyond the observed range of x are not reliable.

- The algorithm yields linear regression lines whether or not the relationship between x and y is really linear.

Linear Systems:
Linear Regression Involving a Single Variable

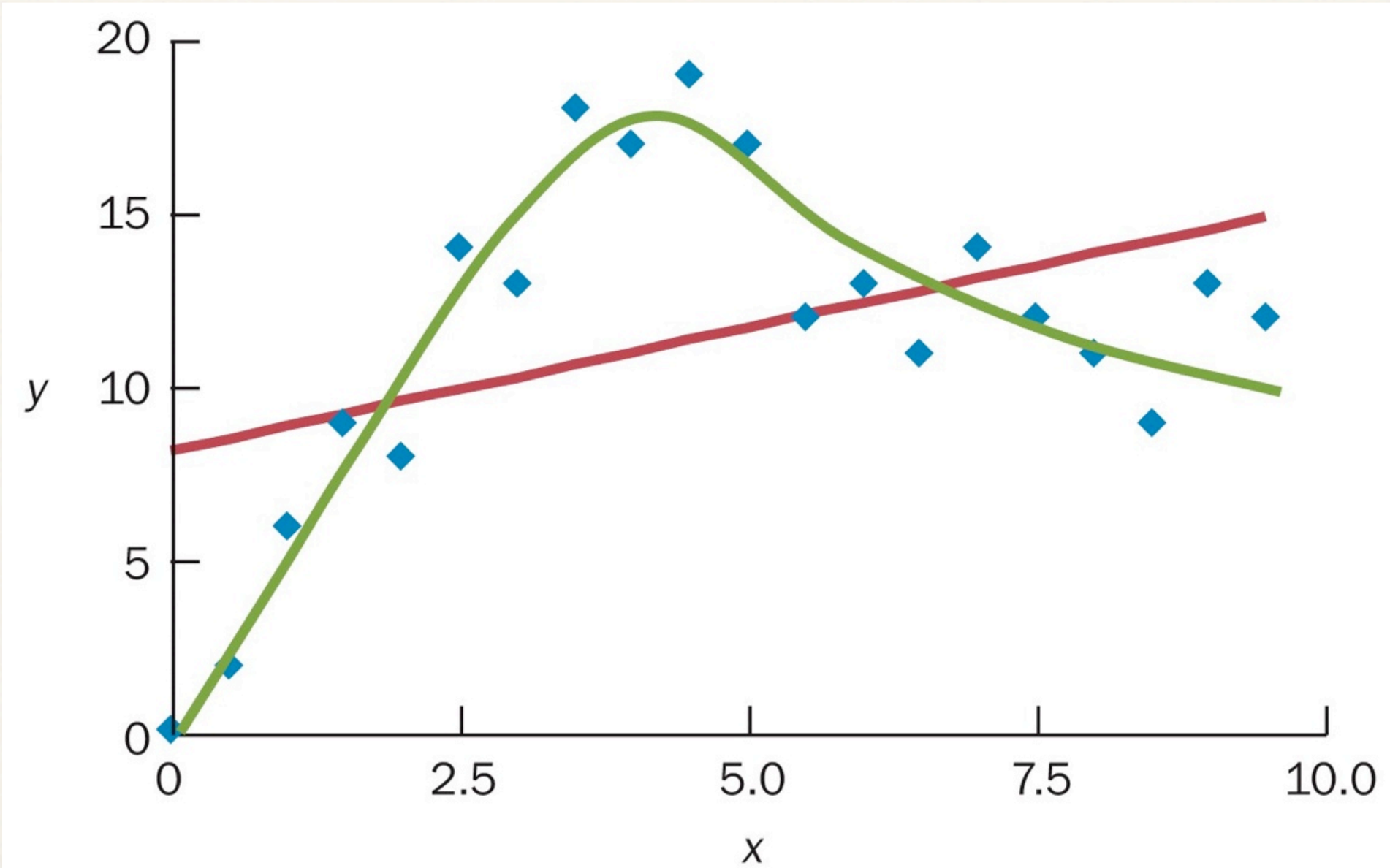


Figure 5.2 A First Course in Systems Biology (© Garland Science 2013)

Linear Systems:
Linear Regression Involving a Single Variable

- Often, simple inspection as in the previous figure is sufficient.

- However, one should consider assessing a linear regression result with some mathematical rigor (e.g., analyze residual errors, ..)

Linear Systems:
Linear Regression Involving Several Variables

- Linear regression can also be executed quite easily in cases of more variables.

- (the function regress is Matlab does the job)

Linear Systems:
Linear Regression Involving Several Variables

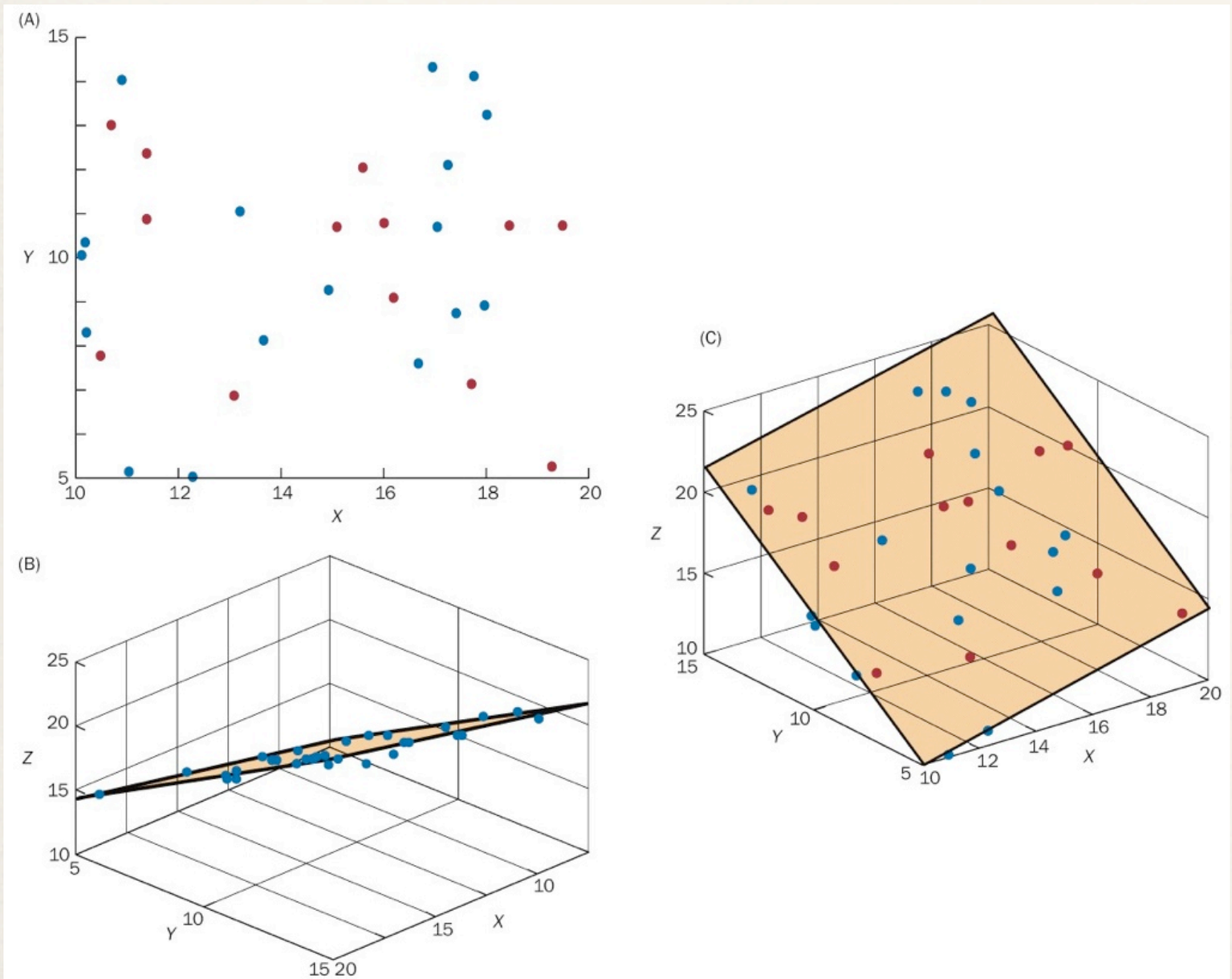
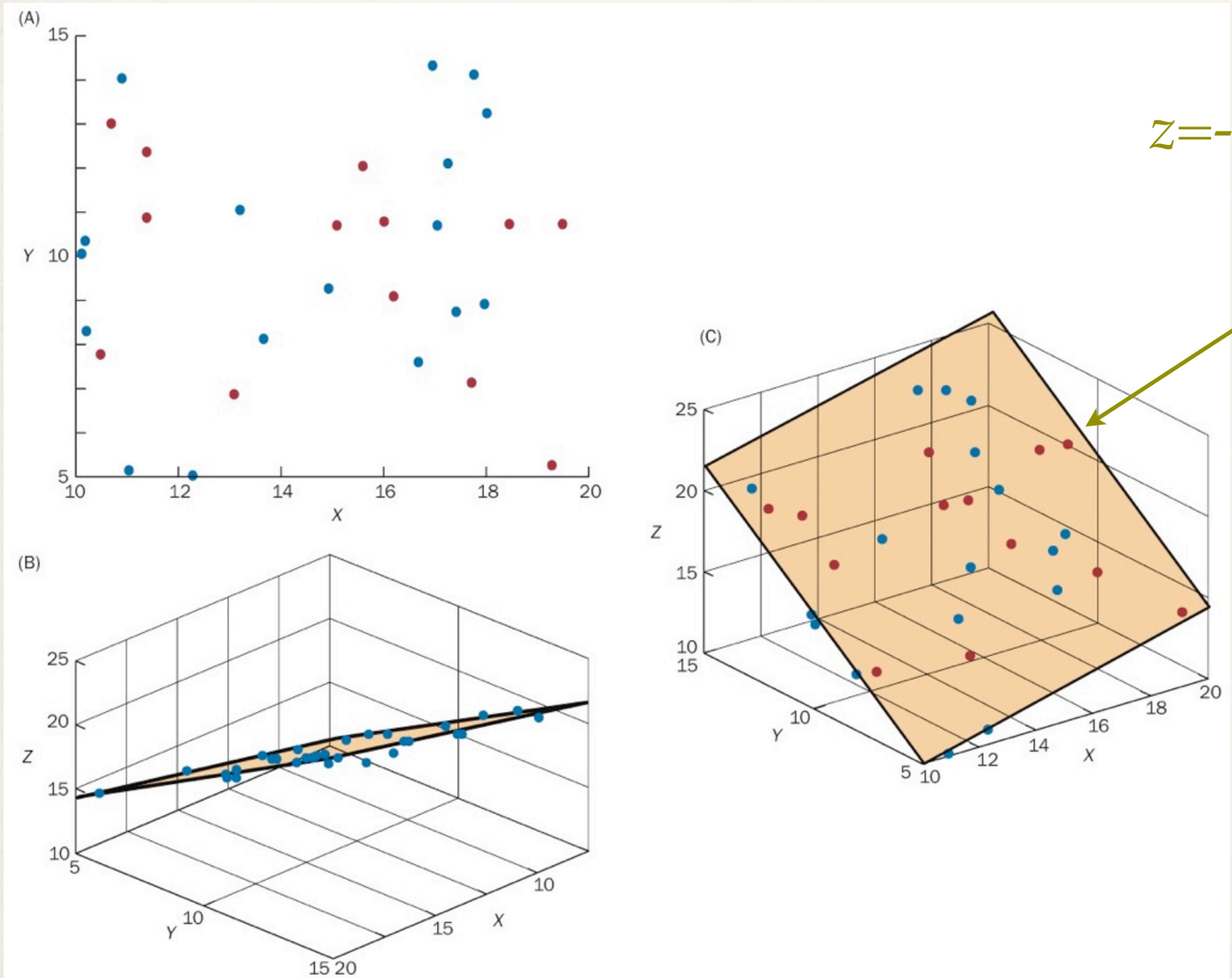


Figure 5.3 A First Course in Systems Biology (© Garland Science 2013)

Linear Systems:
Linear Regression Involving Several Variables



z=-0.0423+0.4344x+1.13y

Figure 5.3 A First Course in Systems Biology (© Garland Science 2013)

Linear Systems:
Linear Regression Involving Several Variables

- Sometimes, a nonlinear function can be turned into a linear one...

- Recall: linearization of Michaelis-Menten rate law.

Linear Systems:
Linear Regression Involving Several Variables

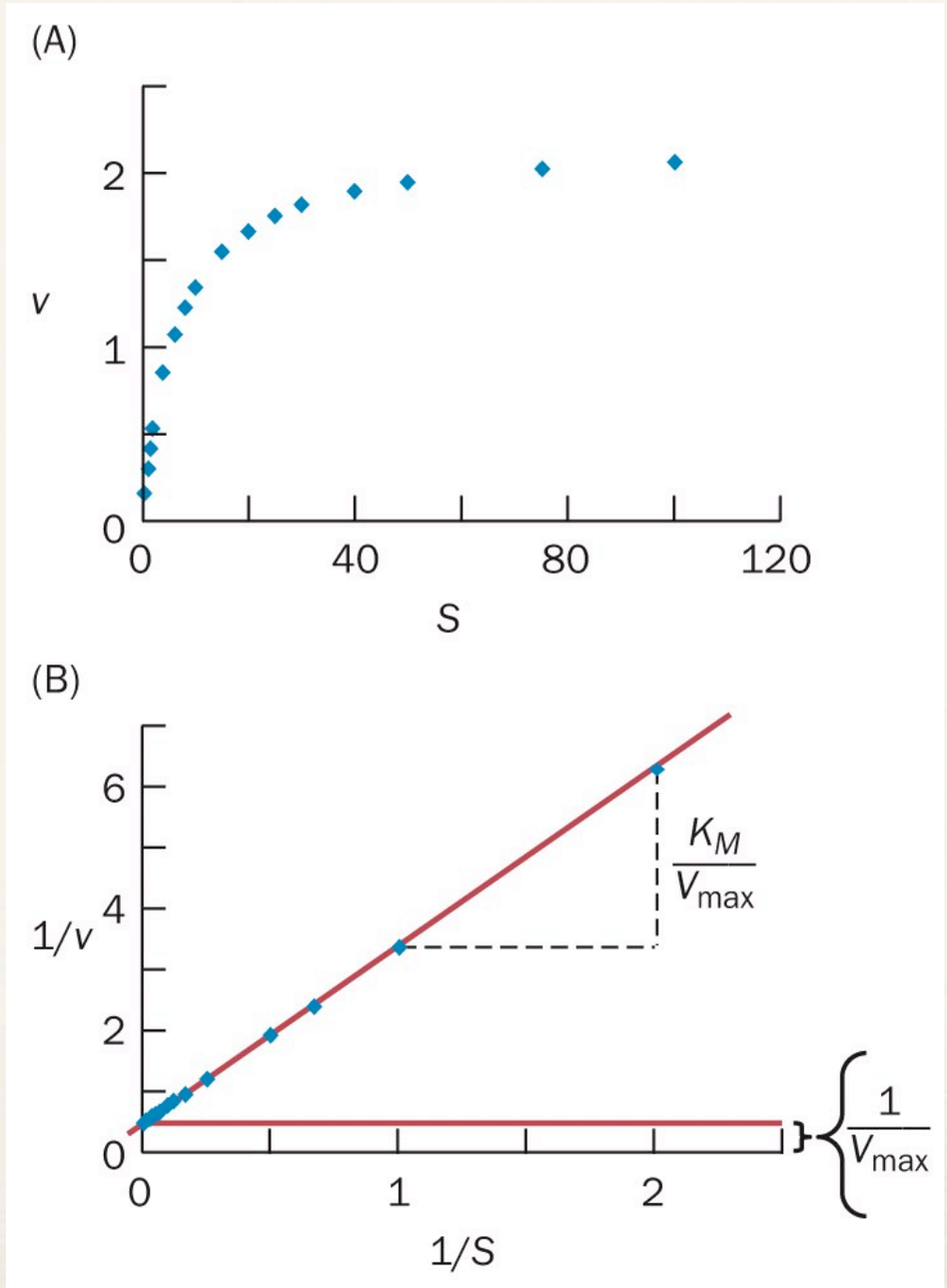


Figure 5.5 A First Course in Systems Biology (© Garland Science 2013)

Linear Systems:
Linear Regression Involving Several Variables

- Sometimes, a nonlinear function can be turned into a linear one...
- E.g., taking the log of an exponential growth function

$$V = \alpha X^g Y^h \qquad \ln V = \ln \alpha + g \ln X + h \ln Y$$

Nonlinear Systems

- Parameter estimation for nonlinear systems is incomparably more complicated than linear regression.

- The main reason is that there are infinitely many different nonlinear functions (there is only one form of a linear function).

Nonlinear Systems

- Even if a specific nonlinear function has been selected, there are no simple methods for computing the optimal parameter values as they exist for linear systems.

- The solution of a nonlinear estimation task may not be unique.

- It is possible that two different parameterizations yield exactly the same residual error or that many solutions are found, but none of them is truly good, let alone optimal.

- Several classes of search algorithms exist for this task (each of which works well in some cases and poorly in others).

Nonlinear Systems

- Class I: exhaustive search (grid search)
  - One has to know admissible ranges for all parameters
  - One has to iterate the search many times with smaller and smaller intervals
  - Clearly infeasible but for very small systems

Nonlinear Systems

- Branch-and-bound methods are significant improvements over grid searches, because they ideally discard large numbers of inferior solution candidates in each step.

- These method make use of two tools:

  - branching: divide the set of candidate solutions into non-overlapping subsets

  - bounding: estimate upper and lower bounds for SSE

Nonlinear Systems

- Class II: hill-climbing (steepest-descent) methods
  - head in the direction of improvement

Nonlinear Systems

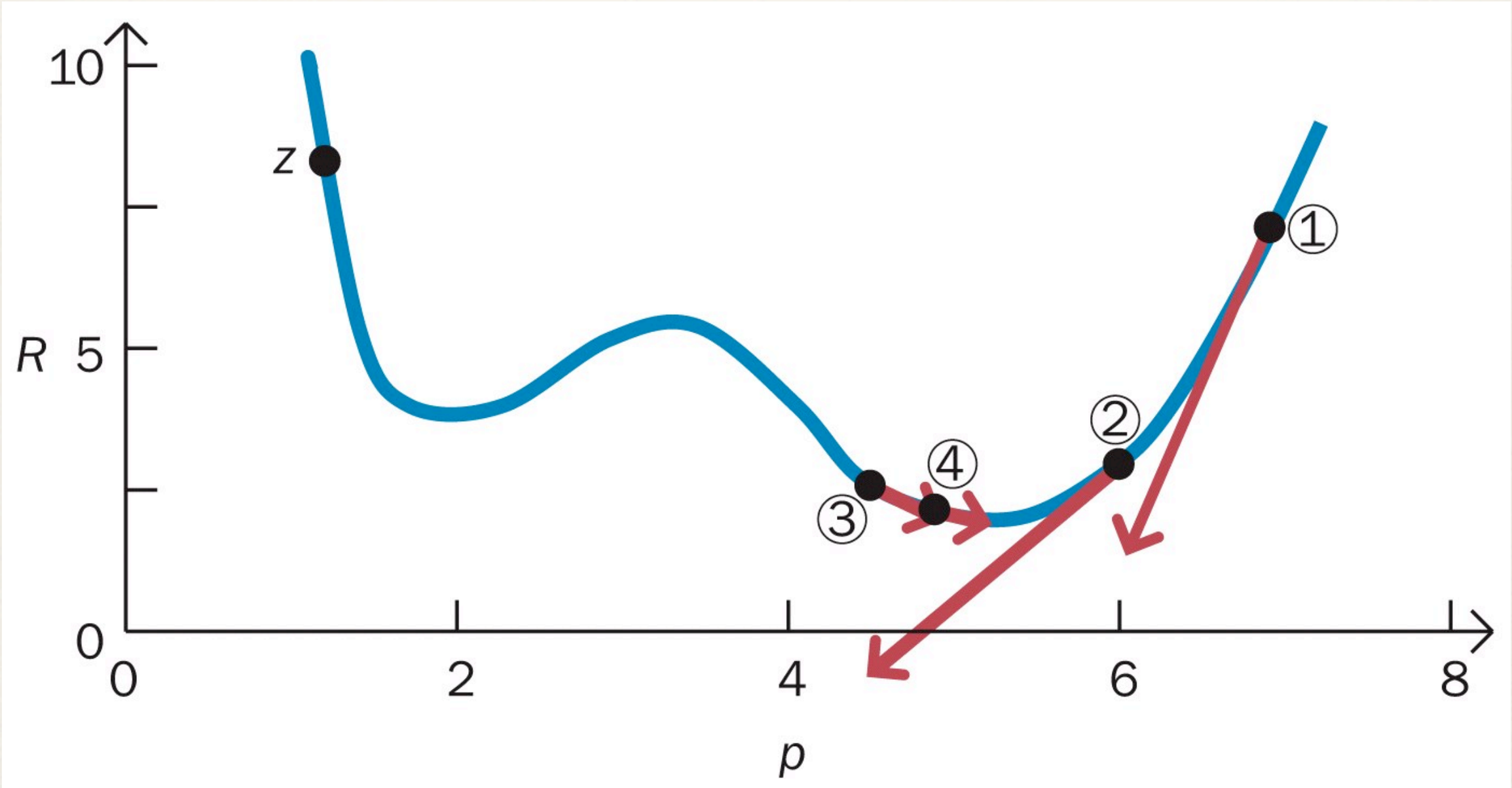


Figure 5.10 A First Course in Systems Biology (© Garland Science 2013)

Nonlinear Systems

- Class II: hill-climbing (steepest-descent) methods
  - Clearly, can get stuck in local optima

Nonlinear Systems

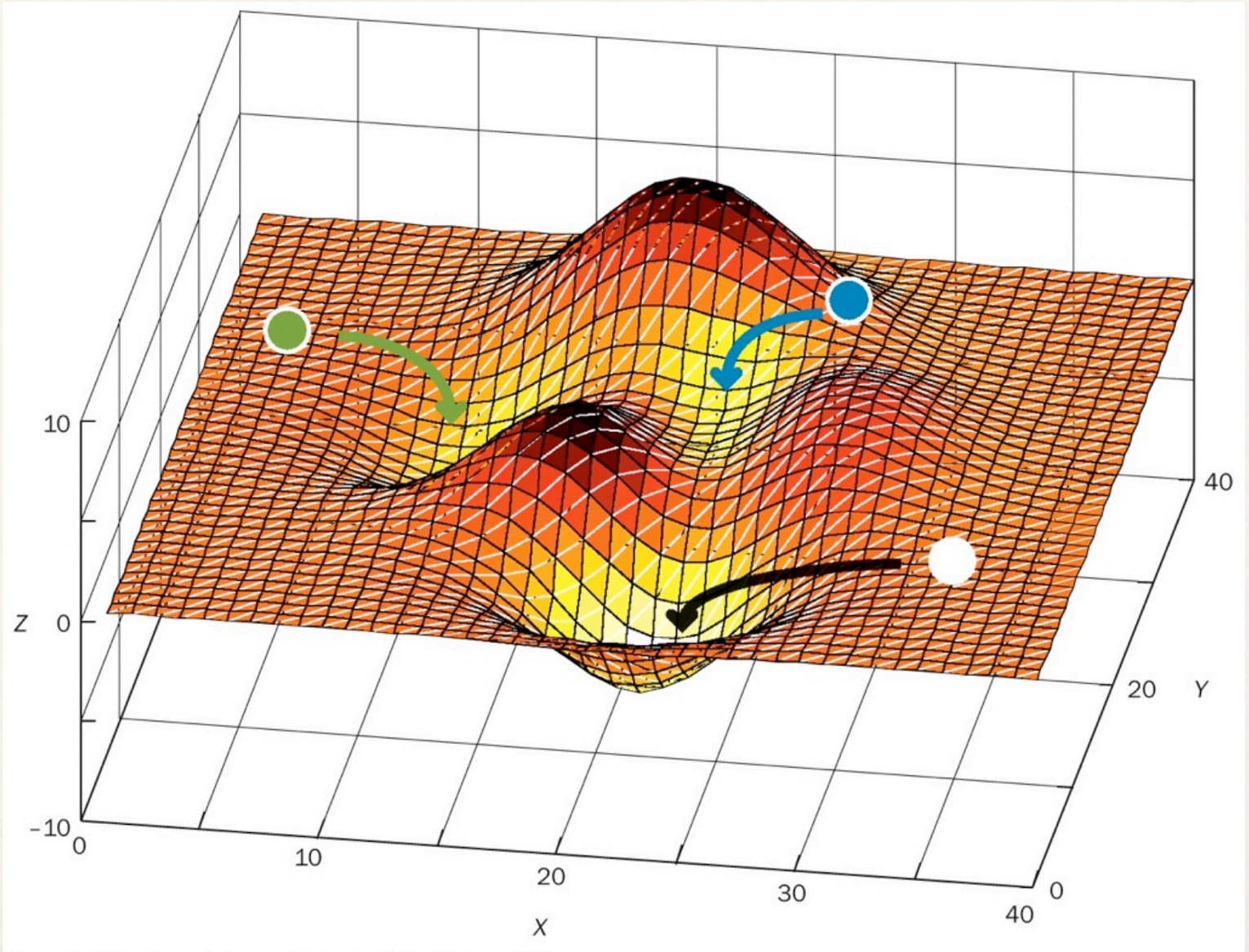


Figure 5.8 A First Course in Systems Biology (© Garland Science 2013)

Nonlinear Systems

- Class III: evolutionary algorithms
  - simulates evolution with fitness-based selection
  - the best-known method in this class is the genetic algorithm (GA)
  - each individual is a parameter vector, and its fitness is computed based on the residual error; mating within the parent population lead to a new generation of individuals (parameter vectors); mutation and recombination can be added as well...

Nonlinear Systems

- A typical individual in the application of a GA to parameter estimation:

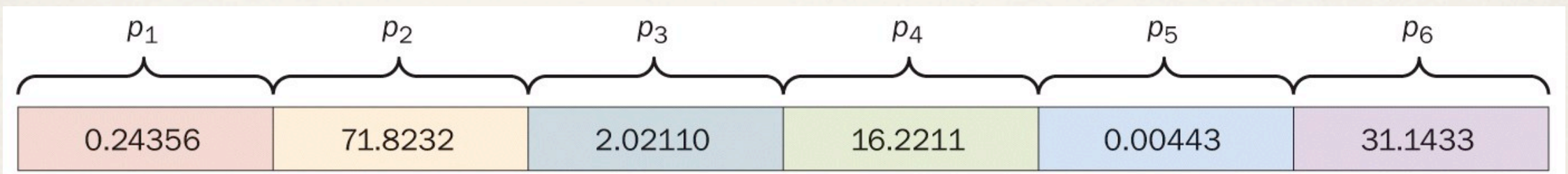


Figure 5.11 A First Course in Systems Biology (© Garland Science 2013)

Nonlinear Systems

- Generation of offspring in a generic GA:

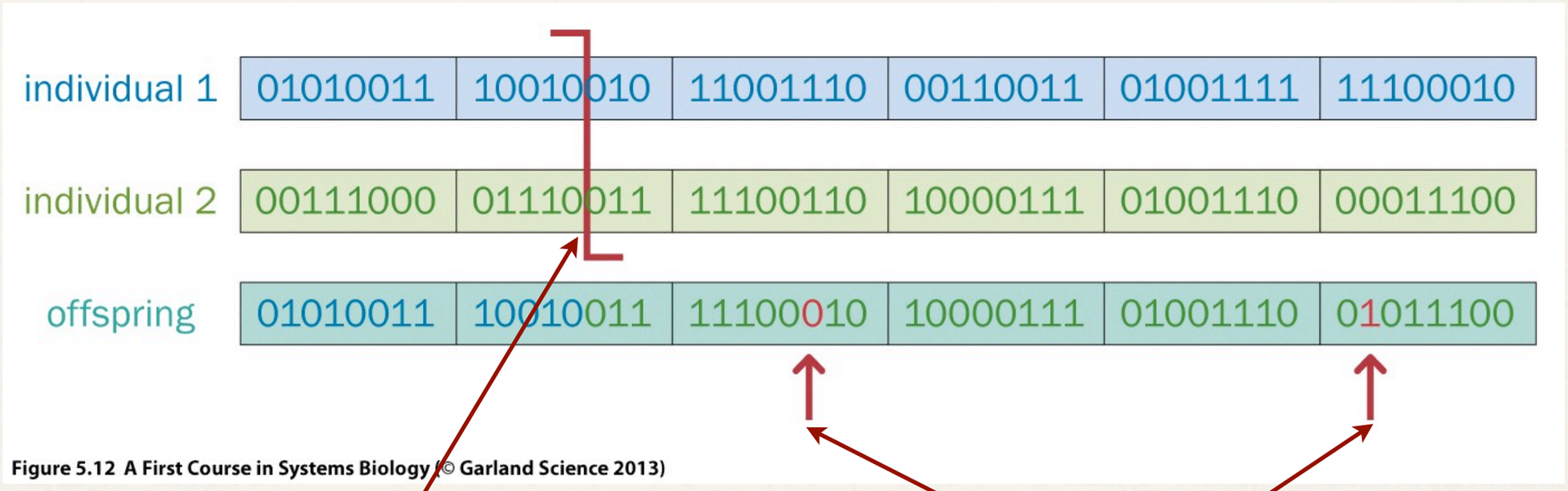


Figure 5.12 A First Course in Systems Biology (© Garland Science 2013)

Nonlinear Systems

- Genetic algorithms work quite well in many cases and are extremely flexible (e.g., in terms of fitness criteria)
- However, in some cases they do not converge to a stable population
- In general, these algorithms are not particularly fast

Nonlinear Systems

- Other classes of stochastic algorithms:
  - ant colony optimization (ACO)
  - particle swarm optimization (PSO)
  - simulated annealing (SA)
  - ....

Typical Challenges

- Noise in the data: this is almost always present (due to inaccuracies of measurements, etc.)

- Noise is very challenging for parameter estimation

Typical Challenges

moderate noise

more noise

much more noise

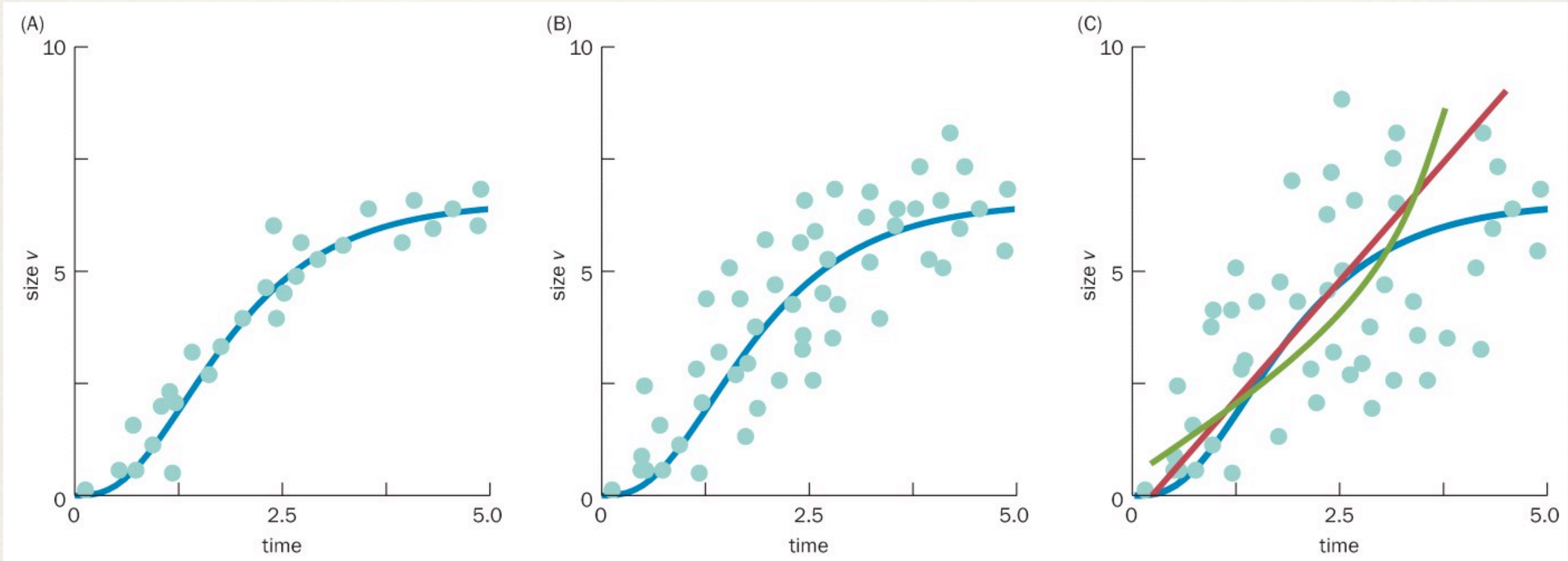


Figure 5.13 A First Course in Systems Biology (© Garland Science 2013)

Bootstrapping

- A noisy data set $\gamma = \mathbf{x}(\theta) + \xi$ will not allow us to determine the true model parameters θ, but only an estimate $\theta'(\gamma)$.

- Each time we repeat the estimation with different data sets, we deal with a different realization of the random error ξ and obtain a different estimator θ'.

- Ideally, the mean value $\langle \theta' \rangle$ of these estimates should be identical to the true parameter value, and their variance should be small.

- In practice, however, only a single data set is available, so we obtain a single data point estimate θ' without knowing its distribution.

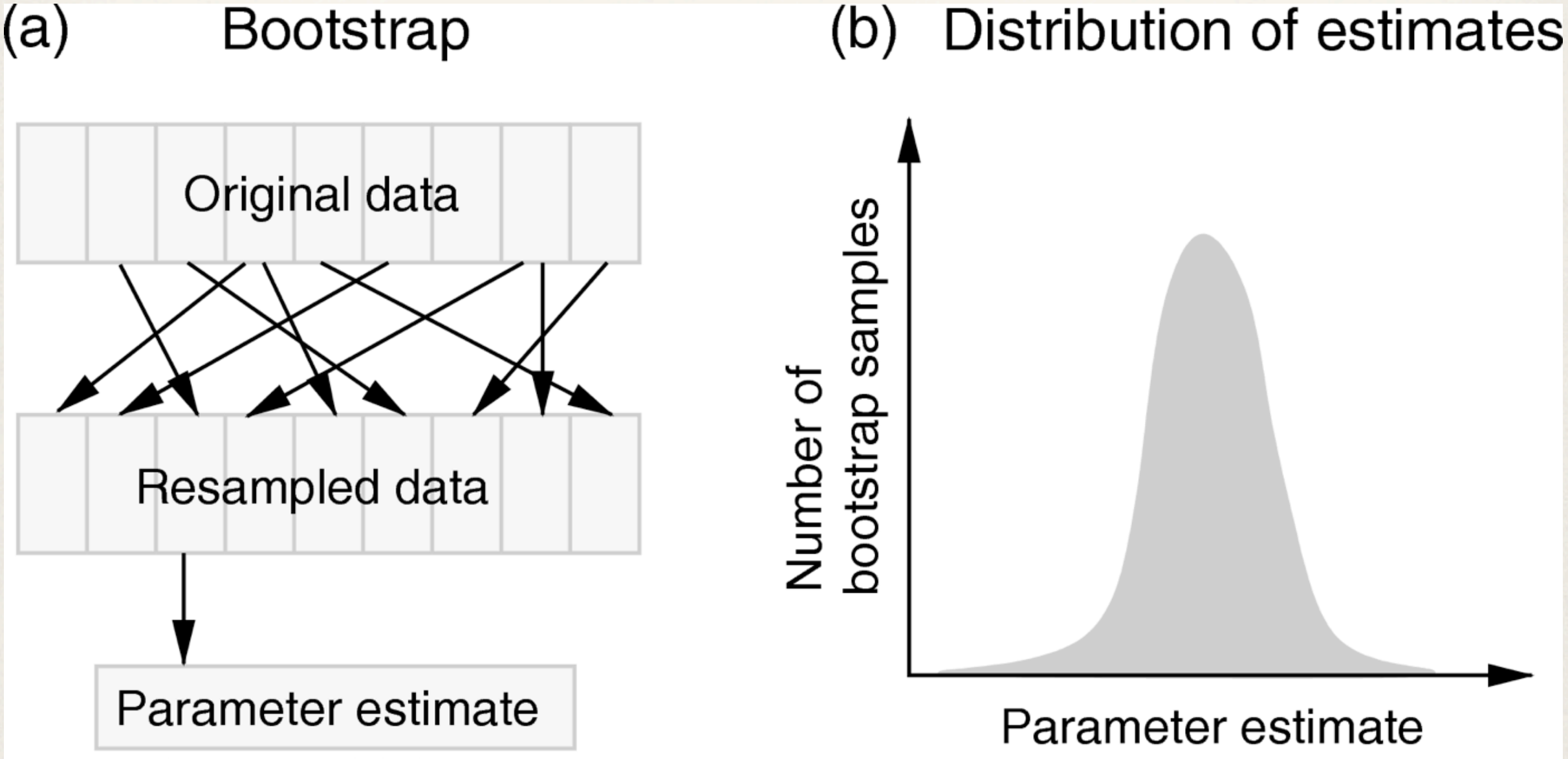
Bootstrapping

- Bootstrapping provides a way to determine, at least approximately, the statistical properties of the estimator θ'.

- First, hypothetical data sets (of the same size as the original data set) are generated from the original data by resampling with replacement, and the estimate θ' is calculated for each of them.

- The empirical distribution of these estimates is then taken as an approximation for the true distribution of θ'.

Bootstrapping



Klipp - Systems Biology
ISBN: 978-3-527-31874-2 fig-04-04

Bootstrapping

- Bootstrapping is asymptotically consistent, that is, the approximation becomes exact as the size of the original data set goes to infinity.

- However, for finite data sets, it does not provide any guarantees.

Typical Challenges

- Non-identifiability: Many solutions might have the same SSE.

- In particular, dependence between variables may lead to redundant parameter estimates.

Typical Challenges

- The task of parameter estimation from given data is often called an inverse problem.

- If the solution on an inverse problem is not unique, the problem is ill-posed and addition assumptions are required to pinpoint a unique solution.

Structure Identification

- Related to the task of estimating parameter values in that of structure identification.

- In this case, it is not known what the structure of the model looks like.

- Two approaches may be pursued:

  - explore a number of candidate models (Michaelis-Menten, sigmoidal Hill functions, etc.)

  - compose a model from a set of canonical models (think: assembling the model from basic building blocks)

Cross-validation

- There exists a fundamental difference between model fitting and prediction.

- If a model has been fitted to a given data set, it will probably show a better agreement with these training data than with new test data that have not been used for model fitting.

- The reason is that in model fitting, we enforce an agreement with the data.

Cross-validation

- In fact, in model fitting, it is often the case that a fitted model will fit the data better than the true model itself!

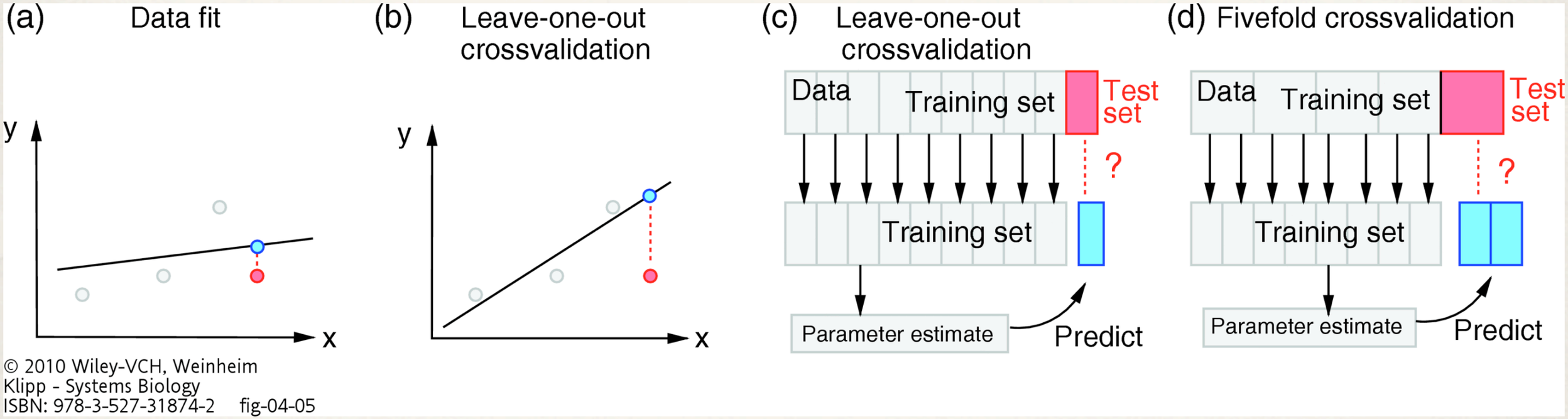
- This phenomenon is called overfitting.

- Strong overfitting should be avoided!

Cross-validation

- How can we check how a model performs in prediction?
- In cross-validation, a given data set (size N) is split into two parts: a training data set of size n, and a test data set consisting of all remaining data.
- The model is fitted to the training data and the prediction error is evaluated for the test data.
- By repeating this procedure for many choices of test sets, we can judge how well the model, after being fitted to n data points, will predict new data.

Cross-validation



Klipp - Systems Biology
ISBN: 978-3-527-31874-2 fig-04-05

Model Selection

- “Essentially, all models are wrong, but some are useful.”

- Useful for what?

Agreement between model and data:	Represent the biological system:	
1: Good data fit	3: Biological details	Complexity ↓ Simplicity
2: Good prediction	4: Reduce to key principles	

Klipp - Systems Biology
ISBN: 978-3-527-31874-2 fig-04-14

- As a rule of thumb, a model with many free parameters may fit given data easily.

- But, as the fit becomes better and better, the average amount of experimental information per parameter decreases, so the parameter estimates and predictions from the model become poorly determined.

- “With four parameters, I can fit an elephant, and with five, I can make him wiggle his trunk.” J. von Neumann

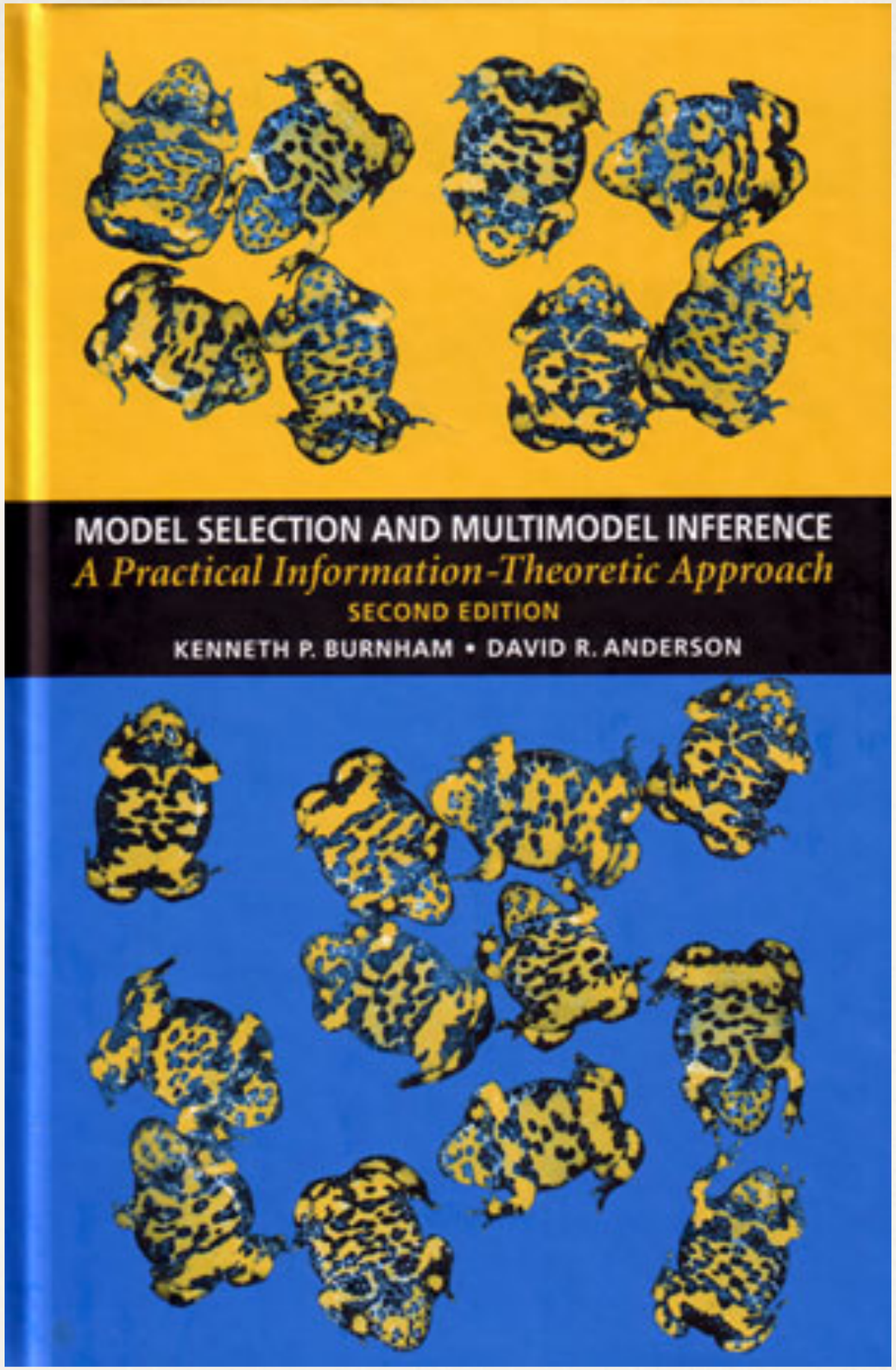
- We can choose between competing models by statistical tests and model selection criteria.

- In statistical tests, we compare a more complex model to a simpler background model. According to the null hypothesis, both models perform well. In the test, we favor the background model unless it statistically contradicts the observed data.

- Selection criteria are mathematical scoring functions that balance agreement with experimental data against model complexity.

Model Selection

- Methods include
  - maximum likelihood and χ^2-test
  - likelihood ratio test
  - information criteria (AIC, AICc, BIC)
  - Bayesian model selection
  - ...


MODEL SELECTION AND MULTIMODEL INFERENCE
A Practical Information-Theoretic Approach
SECOND EDITION
KENNETH P. BURNHAM • DAVID R. ANDERSON

Acknowledgments

- “Systems Biology: A Textbook,” by E. Klipp et al.
- “A First Course in Systems Biology,” by E.O. Voit.