REVIEWS

Linking data to models: data regression

Khuloud Jaqaman and Gaudenz Danuser

Abstract | Mathematical models are an essential tool in systems biology, linking the behaviour of a system to the interactions between its components. Parameters in empirical mathematical models must be determined using experimental data, a process called regression. Because experimental data are noisy and incomplete, diagnostics that test the structural identifiability and validity of models and the significance and determinability of their parameters are needed to ensure that the proposed models are supported by the available data.

Structural identifiability
A model is structurally identifiable if its parameters can be uniquely estimated by fitting the model to experimental data. Structural identifiability is related to the sensitivity of process output to parameter variations.

Variance
A measure of the dispersion of a variable around its average. Its square root is the standard deviation.

Covariance
A measure of how two variables vary relative to each other.

Department of Cell Biology, the Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California, 92037, USA. Correspondence to G.D. e-mail: gdanuser@scripps.edu
doi:10.1038/nrm2030
Published online
27 September 2006

The objective of systems biology is to elucidate the behaviour of a multicomponent system, taking into consideration the network of interactions between the components of the system. Because of the complexity of biological systems, this goal requires the use of mathematical models that provide a framework for determining the outcome of numerous and simultaneous time-dependent and space-dependent processes[1,2].

A model consists of a set of rules (for example, $A + B \underset{k_b}{\overset{k_f}{\rightleftharpoons}} C$ in a system with three components A, B and C) and the corresponding parameters (in this case, the rate constants k_f and k_b). Some mathematical modelling approaches involve extracting rules from experimental data[3–9], whereas others define empirical rules based on *a priori* hypotheses about the mechanism of interest[10–15] (see the accompanying Reviews by Aldridge, Burke, Lauffenburger and Sorger in *Nature Cell Biology* and by Janes and Yaffe in this issue). In approaches that define empirical rules based on *a priori* hypotheses, model parameters are determined using experimental data, a process called regression.

However, experimental data are not perfect and models contain many unknown parameters. Therefore, the structural identifiability of models must be investigated before regression can be used. Also, parameter variances and covariances, which are estimated by regression, must be used to verify the overall validity of a model in representing the data and to ensure the significance and determinability of its parameters. Without these diagnostics, regression results can be inaccurate and subsequent conclusions can be dubious.

This review is intended as a basic User's guide to data regression in the context of data-driven mechanistic modelling of biological systems (FIG. 1). We present common regression schemes and essential pre-regression and post-regression diagnostic tests for the evaluation of overall model validity. We also discuss issues that are related to the regression of stochastic data. Throughout the review, the simple example presented in BOX 1 will be used to illustrate regression concepts and techniques.

Pre-regression diagnostics

Prior to the use of regression, the structural identifiability of a model must be assessed. Given the structure of a model, is it possible to uniquely estimate its unknown parameters? What experimental data are required to achieve unique parameter identification[16–18]? If some of the necessary data are not available, then structural identifiability analysis reveals which parameters must be eliminated *a priori*. Otherwise, unidentifiable parameters might lead to regression instability. Importantly, the structural identifiability of a model is independent of the regression scheme that is employed.

The structural identifiability of a model can be assessed by testing the sensitivity of measured output to changes in model parameters[17,18]. Let $r = \{r_1, r_2, \ldots, r_n\}$ be a set of measurable system output and $\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ be the set of unknown model parameters. The sensitivity coefficients $s_{r,\alpha}$ of the output with respect to the parameters are defined as:

$$s_{r_i,\alpha_j} = \frac{\partial r_i}{\partial \alpha_j},\ i = 1,\ldots,n,\ j = 1,\ldots,m \qquad (1)$$

The symbol ∂ means 'infinitesimally small change'. Equation 1 defines an $n \times m$ matrix, in which the i,jth entry evaluates the change in the output r_i in response to a small change in the parameter α_j.

A model is structurally identifiable if its sensitivity coefficient matrix satisfies the following two conditions: each column has at least one large entry (that is, each parameter has a strong influence on at least one measurable output); and the columns of the coefficient matrix

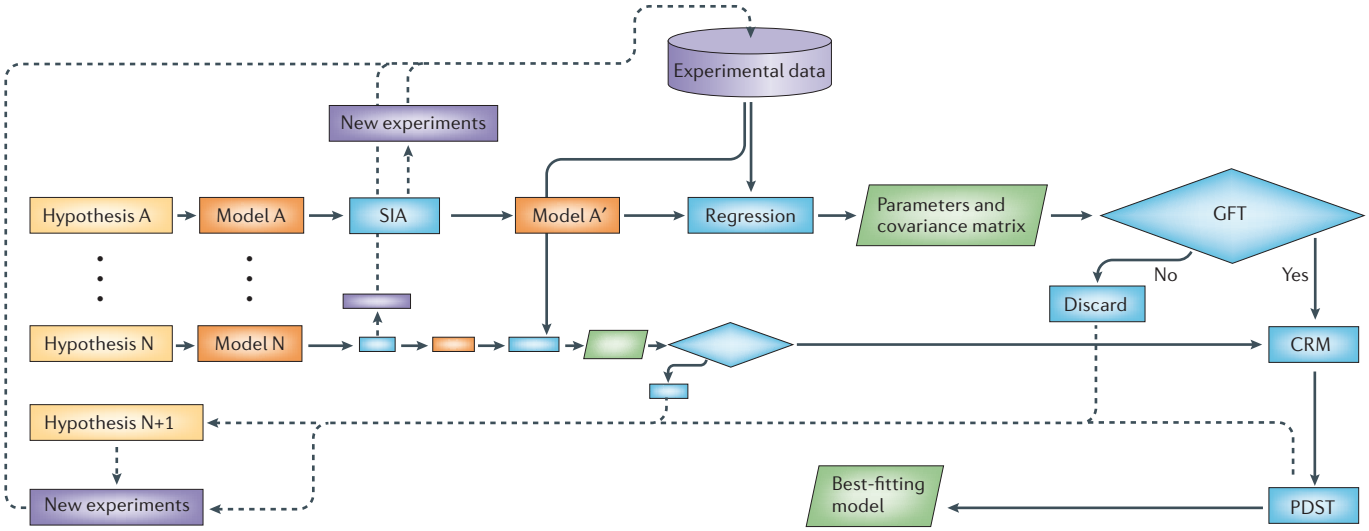

Figure 1 | **Workflow of data-driven mechanistic modelling that employs regression as well as pre-regression and post-regression diagnostics.** A number of hypotheses about the mechanism under study are used to devise a number of models. Structural identifiability analysis (SIA) is performed on each model, possibly leading to a modification of the model or to the design of new experiments that yield necessary data. Then, model parameters and their variance–covariance matrices are determined by fitting each model to the available experimental data using a regression scheme, such as maximum likelihood and least squares. The goodness-of-fit (GFT) of each model is tested, and models that do not pass the GFT are discarded. Models that pass the GFT are compared and ranked (CRM), for example using the F-test or Bayesian information criterion, to determine the best-fitting model. Then the parameters of the models that pass the GFT are tested for significance and determinability (parameter significance and determinability test (PSDT)) to evaluate the best-fitting model and to reveal any shortcomings in the data and models. New hypotheses can be generated, and new experiments that yield data that shed light on new aspects of the process can be identified based on models that do not pass the GFT and PSDT.

Significance
A parameter is statistically significant if, given the uncertainty in its estimate due to noise in the input data, the probability that the parameter magnitude is different from zero not just by chance exceeds the confidence required by the investigator.

Determinability
A measure for the capability to infer the value of a model parameter from the available input data, independent of the values of other parameters.

Regression instability
A measure for the variation of regression results in the presence of data noise. A regression is unstable if the estimates of model parameters significantly differ when one additional data point is added to the set of input data.

Linear independence
A set of parameters is linearly independent if none of its parameters can be written as a linear combination of the other parameters.

Normal distribution
A bell-shaped distribution that is fully characterized by its mean μ and variance σ^2. It is usually written as $N(\mu,\sigma^2)$.

Residual
The difference between an observation and the corresponding model prediction.

are linearly independent[17] (that is, the effects of parameters on the output must be uncorrelated from each other). When parameter effects are correlated, fluctuations in the estimate of one parameter will be compensated for by fluctuations in the estimate of other parameters. In the worst case, one parameter can have arbitrary values that are always perfectly counteracted by one or several other parameters in the model.

Although the concept is straightforward, sensitivity analysis becomes complicated for large models that contain many unknown parameters. Computational techniques have been developed to assess the structural identifiability of models and to design experiments accordingly[17,18]. Also, the sensitivity coefficients depend on the initial guess of the values of parameters. Therefore, identifiability analysis must be done iteratively and sensitivity coefficients must be updated using parameter estimates that have been obtained from regression.

In the example presented in BOX 1, all of the models are identifiable. For instance, in model C, $s_{r,\alpha_t^C} = t$, $s_{r,\alpha_l^C} = l$ and $s_{r,\alpha_a^C} = 1/a$ (for which r is plant growth rate, t is temperature, l is sunlight exposure and a is altitude). The sensitivity coefficients have non-zero magnitude and are independent of each other, and, therefore, model parameters can be determined by regression.

Regression schemes

The two most common regression schemes that are used for parameter estimation are maximum likelihood (ML) and least squares (LS) (see the regression box in FIG. 1).

In ML estimation, the likelihood of a parameter set is equated to the probability of obtaining the available experimental data from a process that is represented by the model tested. Therefore, in ML estimation, the most likely parameter values are determined as the parameters that maximize the probability of observing the experimental data[19–21]. The probability is defined as a function of the differences between the model-predicted data and the experimentally observed data, and it increases as these differences decrease.

Differences between experimental and model-generated data arise from model inadequacy and measurement noise. In most regression schemes, measurement errors are considered as the sole source of differences between model-generated and experimental data, whereas errors that are due to model inadequacy are ignored and dealt with by post-regression diagnostics.

Assuming that errors in different measurements are not correlated and that they follow a normal distribution, the likelihood L of the parameter set α is given by the probability of observing the available data set r:

$$L(\alpha|r) = \prod_{i=1}^{n} P(\Delta r_i), \qquad (2)$$

$$P(\Delta r_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\Delta r_i^2 / 2\sigma_i^2)$$

Π means product, $P(\Delta r_i)$ is the probability of obtaining a residual Δr_i, representing the measurement error in data point i, and σ_i^2 is the variance of the distribution of Δr_i.

Box 1 | Illustrative data-regression problem

Suppose a researcher is interested in testing whether plant growth rate (r) depends on temperature (t), sunlight exposure (l) and/or altitude (a). In this system, there is one dependent variable, r, and three independent variables, t, l and a. Data sets of corresponding plant growth rate, temperature, sunlight exposure and altitude are fitted with four models: in model A, $r = \alpha_t^A t$ (hypothesis: growth rate is proportional to temperature — α denotes an unknown proportionality constant); in model B, $r = \alpha_t^B t + \alpha_l^B l$ (hypothesis: growth rate is proportional to temperature and light); in model C, $r = \alpha_t^C t + \alpha_l^C l + \alpha_a^C / a$ (hypothesis: growth rate is proportional to temperature and light and is inversely proportional to altitude); and in model D, $r = \alpha_t^D t^2 + \alpha_l^D l^2 + \alpha_a^D / a^2$ (hypothesis: growth rate is proportional to the square of temperature and light and is inversely proportional to the square of altitude).

Data sets are obtained through the following approach: an initial set of 100 measurements (data set 1) was generated using $r_i = 2.5t_i + 0.2l_i + 45/\alpha_i + \varepsilon_i$, $i = 1 \ldots 100$ (where ε_i denotes measurement error); the sets $\{t_i\}$ and $\{l_i\}$ were randomly chosen between 1.5 and 30 and between 100 and 500, respectively, whereas $\{\alpha_i\}$ was coupled to $\{t_i\}$ through $\alpha_i \approx 30/t_i$. Each measurement error ε_i was drawn from a normal distribution with a mean of zero and a standard deviation of σ_i. A second set of 100 measurements (data set 2) was generated with the relationship between $\{\alpha_i\}$ and $\{t_i\}$ taken as $\alpha_i \approx 30/t_i$, $\alpha_i \approx 40/t_i$, $\alpha_i \approx 15/t_i$, $\alpha_i \approx 35/t_i$ and $\alpha_i \approx 20/t_i$ for each one-fifth of the data points, effectively decoupling $\{\alpha_i\}$ and $\{t_i\}$ and expanding their range of values.

Results of regression and post-regression diagnostics are shown in Supplementary information S1 (table).

Robust
An estimation technique is said to be robust if it is insensitive to deviations in the model and the input data from the ideal assumptions about them that were used in formulating the estimation process.

Outlier
A data point with an error that does not belong to the assumed distribution of measurement errors.

Lorentzian distribution
A distribution that resembles the normal distribution, but with lower probability for values that are close to the mean and higher probability for values that are farther from the mean.

Linear
The models $y = \alpha_1 x + \alpha_2 x^2$ and $y = \alpha \exp(-x)$ are linear functions of the parameters α.

Nonlinear
The models $y = (\alpha_1 x + \alpha_2 x)^2$ and $y = \alpha_1 \exp(-\alpha_2 x)$ are nonlinear functions of the parameters α.

Closed-form solution
A solution that can be expressed analytically in terms of a finite number of operations (for example, addition, multiplication, square root, and so on).

Global optimization
The search for the lowest minimum or highest maximum of an objective function that has multiple minima or maxima. Such a function is called non-convex.

Central limit theorem
The central limit theorem states that any variable that is calculated as the sum of a large number of variables, even if they are not normally distributed, will be normally distributed.

Nonparametric methods
Statistical methods that do not assume an underlying distribution for the data being analysed.

Maximizing L under the assumption of normally distributed measurement noise (equation 2) is equivalent to minimizing the weighted sum of squared residuals S:

$$S(\alpha|r) = \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \Delta r_i^2 \qquad (3)$$

The weight of each squared residual is the inverse of the variance of its distribution. Equation 3 is the familiar objective function for LS estimation of the model parameters α.

The strategy above for finding the most likely parameters assumes that only the dependent variable measurements are subject to observational error. When this assumption is not a valid approximation, total LS regression schemes that account for error in both the dependent and independent variables must be used[22,23]. These regression procedures require advanced numerical treatments that are beyond the scope of this review.

Furthermore, the assumption of normally distributed residuals makes the above regression schemes not robust with respect to outliers[21,24]. Using other noise distributions ($P(\Delta r_i)$) to construct L, such as the Lorentzian distribution, leads to more robust regression[21]. The method of least median of squares, which assumes that residuals are normally distributed but minimizes the median of squared residuals instead of their sum (equation 3), is theoretically the most robust regression method[24].

Linear versus nonlinear models. The ML and LS formulations make no assumptions about the model itself. However, the procedures of maximizing L or minimizing S depend on whether the model being fitted is a linear or a nonlinear function of the unknown parameters.

When a model is linear, LS regression has a closed-form solution[25]. For nonlinear models, regression is more complicated and requires global optimization of L or S. Innumerable strategies have been developed for global optimization[26–28], and further efforts are required to ensure the development of new optimization strategies that can fit nonlinear models to large sets of noisy data — a common problem in systems biology.

Regardless of whether a model is linear or nonlinear, ML and LS regression generate estimates of parameter values $\hat{\alpha}$ and their variance–covariance matrix $\hat{V}$[25,29,30]. Elements on the diagonal of the square matrix $\hat{V}$ represent parameter uncertainties as variances, whereas elements off the diagonal define the interdependencies between parameters as covariances. Parameter variances and covariances decrease as the size of the fitted data set increases ($\hat{V} \sim 1/n$), although the interdependencies between parameters stay unaltered unless new types of experimental data are employed in the regression.

Assuming that the parameter vector is normally distributed, $\hat{\alpha}$ and $\hat{V}$ are sufficient to fully characterize the parameter vector distribution. The assumption of normality holds when the measurement errors are normally distributed and the model is linear or close to linear. It also holds asymptotically for regressions of a large number of observations. In this case, the central limit theorem[20] predicts that the estimated parameters are normally distributed, regardless of the distribution of measurement errors. When neither one of these conditions is satisfied, for instance if a model is highly nonlinear and the number of available data points is small, then the estimated parameters might deviate from the normal distribution. In this case, nonparametric methods, such as the bootstrap and the jackknife[31,32], can be applied to infer an empirical representation of the distribution of the parameter vector. From this, any function of the parameters, such as their expectation value, variance and higher moments, can be calculated.

Bayesian inference. The likelihood in equation 2 can also be used in a third parameter estimation approach called Bayesian inference[33–35]. In this approach, a prior distribution $P(\alpha)$ of the unknown parameters is specified based on *a priori* knowledge, which is then multiplied by the likelihood $L(\alpha|r)$ to obtain a posterior estimate of the parameter distribution $P(\alpha|r)$:

$$P(\alpha|r) = \frac{L(\alpha|r)\, P(\alpha)}{\int L(\alpha|r)\, P(\alpha) d\alpha} \qquad (4)$$

The denominator in equation 4 is a normalization constant. From the posterior $P(\alpha|r)$, parameter averages and variance–covariance matrices, as well as any other function of the parameters, can be calculated. Bayesian inference does not make any assumptions about the

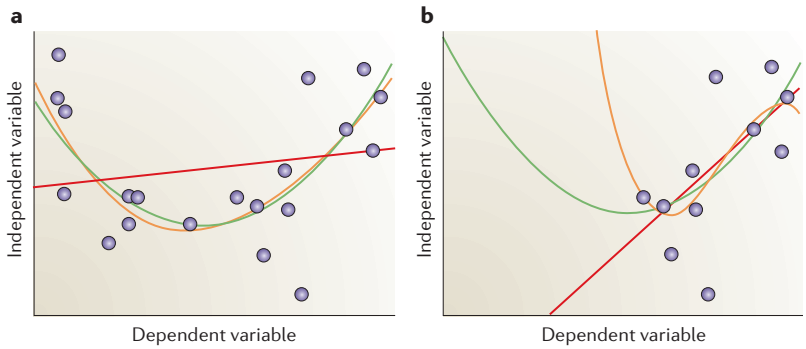

Figure 2 | **Diagrams of post-regression diagnostics. a** | The model goodness-of-fit test (GFT) identifies models that fit the data well, such as the second-order polynomial (green) and third-order polynomial (orange) curves, and models that are far off from the data, such as the first-order polynomial (red line). The F-test and Bayesian information criterion can then be used to decide which model is the most suitable among those that pass the GFT. The green and orange curves fit the data equally well; therefore the model with a smaller number of parameters (the green curve) is the most suitable model. **b** | When the available data spans only a limited sub-region of the space that is needed to uniquely determine model parameters, multiple models can fit the data equally well, but then they behave completely differently outside the region that is used for fitting (see all three lines). This lack of model uniqueness can be detected by testing parameter determinability and significance.

posterior $P(\alpha|r)$. It is therefore especially suitable for parameter estimation in highly nonlinear models. On the other hand, it requires the specification of the prior $P(\alpha)$. Because equation 4 does not have a closed form solution, $P(\alpha|r)$ is inferred by stochastic simulations, such as the Gibbs sampler[35].

Post-regression diagnostics

Post-regression diagnostics are necessary for the evaluation of model validity. First, they are used to determine the best-fitting model and, therefore, the most likely hypothesis when several models exist. Second, they are used to test the significance and determinability of model parameters to reveal shortcomings in the models and/or the data. These are the minimum tests that must be done to gain trust in a model and to surmise that it is a reasonable representation of the process that has generated the available experimental data.

Model goodness-of-fit. When a model fits the data well, model inadequacy does not contribute to the residuals Δr_i in equation 2 and equation 3, and the Δr_i values are solely due to the normally distributed measurement noise. Under these assumptions, the minimized sum S_{min} in equation 3 has an expected value that is equal to the number of degrees of freedom ν[20,25].

Model goodness-of-fit (GFT in FIG. 1) can be assessed by testing the null hypothesis ($H_0 : S_{min} = \nu$) against the alternative hypothesis ($H_A : S_{min} \neq \nu$). Under H_0, the test-statistic, $T = S_{min}$ is chi-square distributed with ν degrees of freedom[20,25]. If the p-value of T is smaller than a certain significance value, then H_0 is rejected.

Rejecting H_0 with $S_{min} > \nu$ implies that the model is not suitable to describe the data, and the residuals are not only due to measurement noise but also reflect the inadequacy of the model (FIG. 2a). It can also mean that some *a priori* variances of the measurement error have been underestimated. Rejecting H_0 with $S_{min} < \nu$ indicates that the measurement-error variances have been overestimated. The underestimation and overestimation of measurement errors might skew model parameters, as it might change the relative weights of observations, and must be mended[25].

In the example presented in BOX 1, only models B and C are suitable to describe data set 1 (p-value = 0.42 and 0.45, respectively). Model A is insufficient (p-value = 0), as it assumes dependence on temperature only, whereas model D has an incorrect form (p-value = 0.0004) — instead of a linear relationship, model D postulates a quadratic relationship between growth rate, temperature, light exposure and altitude.

The most suitable model. As models are approximations of reality, it is likely that more than one model fits the data to an acceptable degree. Generally, models with a larger number of parameters are more flexible and fit the data better than models with a smaller number of parameters. However, the degree of interdependency between parameters increases as the number of parameters in a model increases. Therefore, the guiding principle in choosing the most suitable model is that a simpler and more parsimonious model is preferred over a complicated one if they both fit the data to the same degree (FIG. 2a).

The F-test (compare and rank models (CRM) in FIG. 1) can be used to check whether the fit of a model with extra parameters is significantly better than the fit of another model with a smaller number of parameters. It tests H_0 against H_A:

$$H_0: \frac{S^{(2)}_{min}}{\nu^{(2)}} = \frac{S^{(1)}_{min}}{\nu^{(1)}} \quad (5)$$

$$H_A: \frac{S^{(2)}_{min}}{\nu^{(2)}} < \frac{S^{(1)}_{min}}{\nu^{(1)}} \quad (6)$$

The superscripts $^{(1)}$ and $^{(2)}$ indicate the model with a smaller and a larger number of parameters, respectively. The test-statistic is:

$$T = \frac{S^{(2)}_{min}}{\nu^{(2)}} \Big/ \frac{S^{(1)}_{min}}{\nu^{(1)}} \quad (7)$$

T follows an F-distribution with $\nu^{(1)}$ and $\nu^{(2)}$ degrees of freedom[25]. Rejecting H_0 indicates that model 2 fits the data significantly better than model 1, which justifies the introduction of extra parameters to fit the data. The sensitivity of the F-test decreases with increasing number of degrees of freedom.

A method that is more appropriate for comparing and ranking models that are fitted to large data sets (CRM in FIG. 1) is the Bayesian information criterion (BIC). This method assigns a 'score' (also called BIC) to each model based on its likelihood L, the number m of estimated parameters in it and the number n of fitted data points[36]. The BIC is given by:

$$BIC(\alpha|r) = -2\ln L(\alpha|r) + m\ln n \quad (8)$$

In the case of normally distributed measurements, $-2\ln L = S_{min}$. The BIC decreases as the model likelihood increases, and increases as the number of parameters

Number of degrees of freedom
The number of degrees of freedom in a regression is the number of data points that were used in the regression minus the number of estimated parameters.

Null hypothesis
A statement that is tested for possible rejection under the assumption that it is true.

Alternative hypothesis
A statement that is placed in opposition to the null hypothesis.

Test-statistic
The variable calculated from the available data in order to test whether the null hypothesis can be rejected. Its distribution under the null hypothesis is usually known.

Chi-square distribution
A variable that is calculated as the sum of the squares of ν variables that are $N(0,1)$-distributed has a Chi square (χ^2)-distribution with ν degrees of freedom.

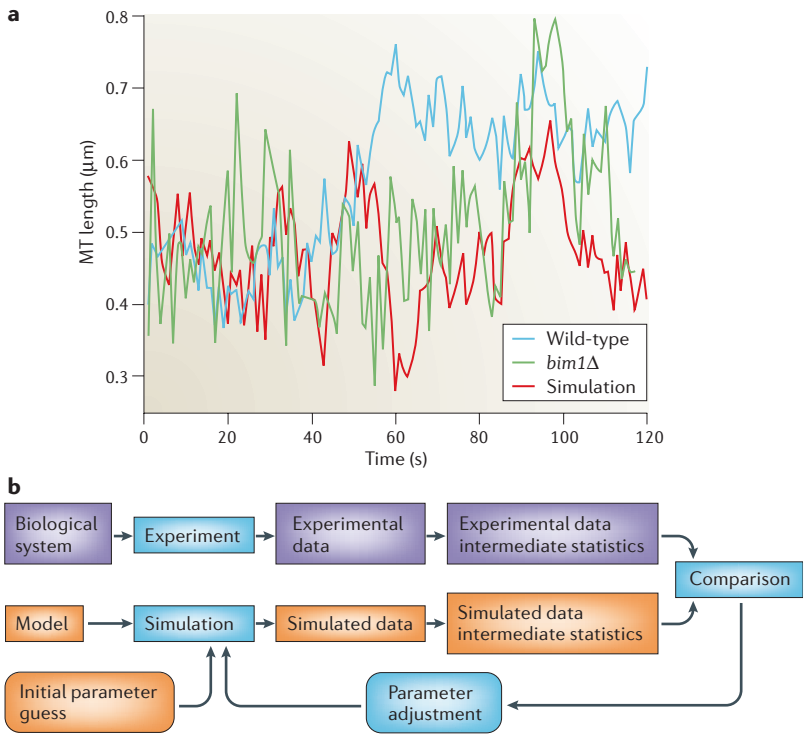

Figure 3 | **Modelling of probabilistic processes. a** | Data from probabilistic processes, such as microtubule (MT) length trajectories, are stochastic and cannot be compared directly point-by-point. The data must be analysed to extract its characteristics that can then be compared to decide whether MT dynamics in wild-type are different from those in a mutant (*bim1*Δ), and whether the model-generated dynamics are equivalent to those measured experimentally. **b** | Flowchart for the estimation of parameters in probabilistic models. Experimental and model-generated data are analysed to obtain intermediate statistics that characterize them. These statistics are then compared, and model parameters are updated until the statistics of model-generated data match those of experimental data. Identifying the appropriate statistics and the optimal strategy for matching the descriptors of model-generated and experimental data are major challenges within this framework.

P-value
The probability of obtaining a test-statistic at least as extreme as the one observed, assuming that the null hypothesis is true. It is effectively the probability of wrongly rejecting the null hypothesis when it is actually true.

Significance value
The value below which a p-value supports rejecting the null hypothesis.

F-distribution
A variable that is calculated as the ratio of two Chi-square-distributed variables divided by their degrees of freedom ν_1 and ν_2, has an F-distribution with ν_1 and ν_2 degrees of freedom.

in a model increases. Among competing models, the model that minimizes the BIC is the most suitable to describe the available data. Because the first term in the BIC grows linearly with n, whereas the second term is proportional to $\ln(n)$, the penalty for having too many parameters is diminished as the data set gets larger.

Comparing the fits of models B and C to data set 1 (BOX 1) using the F-test shows that the use of an extra parameter in model C (dependence on altitude) is not justified (p-value = 0.52). In agreement with this finding, model B has the minimum BIC (Supplementary information S1 (table)). Therefore, in our example, both model-selection criteria indicate that the rate of growth depends only on temperature and light exposure.

Parameter determinability. After identifying the most suitable model, it must be confirmed that the available data uniquely determine the value of each parameter (parameter significance and determinability test (PSDT) in FIG. 1). If a model is structurally identifiable yet some of its parameters turn out to be undeterminable, this indicates that there are hidden dependencies within the available data. If structural identifiability analysis has not been performed on a model, its parameters might be undeterminable due to the structure of the model.

The simplest quantity that can be used to test parameter determinability is the cross-correlation between parameters. For two parameters i and j, cross-correlation is defined as:

$$\kappa_{ij} = \frac{\hat{V}_{ij}}{\sqrt{\hat{V}_{ii}}\sqrt{\hat{V}_{jj}}} \quad (9)$$

$\hat{V}$ is the variance–covariance matrix of the parameter estimates. The cross-correlation is normalized between −1 and +1; 0 indicates no dependency between parameters, whereas ±1 indicates complete dependency. A large cross-correlation between two parameter estimates (for example, $|\kappa| > 0.95$) indicates that the two parameters are weakly determinable because of their strong influence on each other. Like the lack of structural identifiability, weak parameter determinability can cause regression instability.

Another approach for assessing the determinability of model parameters is by measuring the contribution of each parameter to the trace of the variance–covariance matrix[37]. Parameters with a high contribution are weakly determinable.

Determinability analysis can also be performed on less optimal models to elucidate hidden dependencies between variables. The conclusions from this analysis can suggest the design of new experiments that might yield more complete data and new models that might better capture the system behaviour.

In our illustrative example (BOX 1), the parameters that correspond to temperature and altitude in the fit of model C to data set 1 exhibit large cross-correlation ($\kappa = -0.9998$) and therefore are weakly determinable. Because these two parameters are identifiable according to the pre-regression diagnostics, this weak determinability must follow from the strong, yet hidden, coupling between temperature and altitude in the available data. In data set 1, the relationship temperature = constant/altitude (BOX 1) was assumed. By contrast, there is less coupling between temperature and altitude in data set 2, in which the range of measurements has been expanded (BOX 1). This weaker coupling allows the extraction of the individual contributions of temperature and altitude to the growth rate. The best-fitting model for data set 2 is model C, without any inappropriately strong correlations between its parameters ($\kappa = -0.4$ between temperature and altitude). The strong dependency between temperature and altitude in data set 1 is the reason that the fit of model C to that data set is not significantly better than the fit of model B.

Parameter significance. Another important issue to consider during the evaluation of a model is whether the estimated parameters are significantly different from zero. A difference from zero indicates a significant relationship between the dependent and independent variables (PSDT in FIG. 1). Parameters that are in principle identifiable can turn out to be insignificant due to the large uncertainty in the available data compared to the sensitivity coefficient in equation 1.

Box 2 | **Microtubule dynamics: an example of probabilistic modelling**

The history of modelling microtubule (MT) dynamic instability over the past two decades nicely shows the importance of choosing appropriate intermediate statistics (descriptive parameters) for the characterization of a probabilistic process. Generally, MT length trajectories have been characterized by their average growth and shrinkage speeds and their average times spent in growth and in shrinkage[45,55,56]. However, this set of statistics cannot distinguish between MT growth with memory (the probability of switching from growth to shrinkage depends on the time spent in growth) and without memory (the probability of switching is independent of the time spent in growth), which are characterized by different growth-time distributions. To distinguish between these two modes of dynamic instability, an extra statistic, the variance of growth times, has been used[57]. Furthermore, growth and shrinkage speeds have inherent variability[58]. Therefore, if a mutation changes the growth-speed distribution without affecting the average growth speed, for example, the effect of that mutation will go unnoticed unless the set of statistics is expanded from only averages to overall distributions[59]. Even speed and time distributions can fail to detect changes in MT dynamics, because these statistics do not fully capture the coupling between MT states over time[60]. To capture these important characteristics, we have used stochastic time-series analysis tools that explicitly account for time coupling[60]. With each extra statistic, more detailed information about MT-length trajectories is extracted, and so the models that are devised better reflect the mechanisms that underlie our observations.

Due to the interdependence between parameters, reflected by non-zero cross-correlation coefficients, significance cannot be assessed for each parameter individually, but only for groups of interdependent parameters. However, the grouping of interdependent parameters in large models is ambiguous, rendering the direct testing of parameter significance a non-trivial task.

The simplest approach to test the significance of interdependent parameters is to transform them into a new parameter set (orthogonal set) $\tilde{\alpha}_1 \ldots \tilde{\alpha}_m$ with zero covariances between the components. This orthogonalization transformation uses a technique called eigenvalue decomposition[38]. The components of this set can be checked for significance, independently of each other, by testing the null hypothesis (H_0:$\tilde{\alpha}_i = 0$) against the alternative hypothesis (H_A:$\tilde{\alpha}_i \neq 0$), for $i = 1 \ldots m$. The test-statistic $T = \tilde{\alpha}_i / \tilde{\sigma}_{\alpha_i}$, where $\tilde{\sigma}_{\alpha_i}$ is the standard deviation of $\tilde{\alpha}_i$, follows a Student's t-distribution with ν degrees of freedom. Rejecting H_0 implies that $\tilde{\alpha}_i$ is significantly different from zero. Components that are found to be insignificant are set to zero, and then the original parameter set is recovered by inverting the orthogonalization transformation, now with all insignificant values eliminated.

Testing the significance of the parameters in models B and C that are estimated by fitting data set 1 shows that the parameters of model C have insignificant values that get eliminated (before test: $\alpha_t = 6.55$, $\alpha_l = 0.21$, $\alpha_a = -68$; after test: $\alpha_t = 4.33$, $\alpha_l = 0.21$, $\alpha_a = 0.14$). Both models fitted to data set 2, on the other hand, do not have insignificant values.

Modelling of probabilistic processes

Many biological processes are probabilistic, such as gene expression[39,40], synaptic transmission[41,42], chemical reactions with low copy number of molecules[43,44] and microtubule dynamic instability[45] (FIG. 3a). In the case of probabilistic models, the formulation that was discussed above for estimating parameters by minimizing the differences between individual experimental and model-predicted data points is no longer applicable. Instead, parameters in probabilistic models can be estimated through indirect inference (FIG. 3b), a method that was developed mostly in econometrics for the estimation of parameters in stochastic dynamic models[46–49].

In the method of indirect inference, first an intermediate model is fitted to the simulated and experimental data to yield intermediate statistics that characterize the data. Then the unknown model parameters in the stochastic simulation are varied in order to minimize the differences between the intermediate statistics of simulated and experimental data. Both steps can be achieved through ML or LS regression, and the diagnostics discussed above can be used to evaluate regression quality and model goodness-of-fit.

The match between model-generated and experimental data is limited to information captured by the intermediate statistics. Characteristics of the experimental data that are not captured by the intermediate statistics will not be reflected in the model, whereas model-generated data might have characteristics that bear no resemblance to reality. Therefore, the reliable estimation of model parameters, and consequently the proper choice of models, is heavily dependent on the use of unique and complete intermediate statistics that fully characterize the experimental and model-generated data (see BOX 2 for an example of stochastic data characterization).

To circumvent the use of intermediate statistics, alternative methods based on Bayesian inference have been developed[50–54]. Within the Bayesian approach, the process underlying the observed data is assumed to be a Markov chain, from which the probability of obtaining the observed data is directly constructed. This likelihood is then multiplied by the prior parameter distribution to obtain the posterior parameter distribution (equation 4). The posterior distribution can then be used to calculate the expectation value and higher moments of model parameters, as in the deterministic case.

Both of the approaches that were introduced above are at the forefront of research on how to link probabilistic models to stochastic data. As the sensitivity of measurement increases and data on single-molecule behaviour become more available, devising reliable methods for estimating parameters in probabilistic models will become increasingly important for achieving the ultimate goal of systems biology — understanding cellular behaviour as a result of the underlying stochastic molecular interactions.

Conclusions

Modelling is an essential tool in systems biology. However, given the complexity of biological systems and the scarcity and incomplete nature of data, modelling can be misleading. In this review, we have presented a minimal set of strategies for regression as well as pre-regression and post-regression tools that should be

Trace
Sum of the diagonal elements of a matrix.

Student's t-distribution
A distribution that is similar to $N(0, 1)$, except that it has heavier tails. It is a function of the number of degrees of freedom ν, and converges to $N(0, 1)$ as ν gets larger.

Probabilistic process
A process in which the current state of a system does not uniquely determine its next state, but defines a set of possible states with their transition probabilities.

Markov chain
A chain of events in which what happens at time point $t + 1$ only depends on what has happened at time point t, and not on any previous time points.

employed to evaluate and diagnose postulated models from a statistical perspective. The ultimate validation of a model that passes these tests, however, will stem from its power to predict system behaviour in response to perturbations and conditions that are not included in the data that are used for parameter estimation.

Unfortunately, in systems biology, we are still far from having these basic quality controls established as the minimum requirements for the publication of a model. As a community, we should jointly set out to elevating our rigour in explaining experimental observations with numerical approaches.

1. Arkin, A. P. Synthetic cell biology. *Curr. Opin. Biotechnol.* **12**, 638–644 (2001).
2. Kitano, H. Systems biology: a brief overview. *Science* **295**, 1662–1664 (2002).
3. Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A. & Nolan, G. P. Causal protein-signaling networks derived from multiparameter single-cell data. *Science* **308**, 523–529 (2005).
4. Woolf, P. J., Prudhomme, W., Daheron, L., Daley, G. Q. & Lauffenburger, D. A. Bayesian analysis of signaling networks governing embryonic stem cell fate decisions. *Bioinformatics* **21**, 741–753 (2005).
5. Bulashevska, S. & Eils, R. Inferring genetic regulatory logic from expression data. *Bioinformatics* **21**, 2706–2713 (2005).
6. Segal, E., Friedman, N., Kaminski, N., Regev, A. & Koller, D. From signatures to models: understanding cancer using microarrays. *Nature Genet.* **37**, S38–S45 (2005).
7. Janes, K. A. *et al.* Systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science* **310**, 1646–1653 (2005).
8. Janes, K. A. *et al.* Cue-signal-response analysis of TNF-induced apoptosis by partial least squares regression of dynamic multivariate data. *J. Comp. Biol.* **11**, 544–561 (2004).
9. Heard, N. A., Holmes, C. C., Stephens, D. A., Hand, D. J. & Dimopoulos, G. Bayesian coclustering of *Anopheles* gene expression time series: study of immune defense response to multiple experimental challenges. *Proc. Natl Acad. Sci. USA* **102**, 16939–16944 (2005).
10. Sprague, B. L. *et al.* Mechanisms of microtubule-based kinetochore positioning in the yeast metaphase spindle. *Biophys. J.* **84**, 3529–3546 (2003).
11. Bentele, M. *et al.* Mathematical modeling reveals threshold mechanism in CD95-induced apoptosis. *J. Cell Biol.* **166**, 839–851 (2004).
12. Gardner, M. K. *et al.* Tension-dependent regulation of microtubule dynamics at kinetochores can explain metaphase congression in yeast. *Mol. Biol. Cell* **16**, 3764–3775 (2005).
13. Rodriguez-Fernandez, M., Mendes, P. & Banga, J. R. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems* **83**, 248–265 (2006).
14. Mendes, P. & Kell, D. B. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics* **14**, 869–883 (1998).
15. Schoeberl, B., Eichler-Jonsson, C., Gilles, E. D. & Muller, G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nature Biotechnol.* **20**, 370–375 (2002).
Ordinary differential equation-based model of the epidermal-growth-factor-signalling network with parameters that were estimated using sensitivity analysis and least-squares regression of concentration time-courses.
16. Bellman, R. & Astrom, K. J. On structural identifiability. *Math. Biosci.* **7**, 329–339 (1970).
17. Yao, K. Z., Shaw, B. M., Kou, B., McAuley, K. B. & Bacon, D. W. Modeling ethylene/butene copolymerization with multi-site catalysts: parameter estimability and experimental design. *Polymer Reaction Eng.* **11**, 563–588 (2003).
18. Gadkar, K. G., Gunawan, R. & Doyle, III, F. J. Iterative approach to model identification of biological networks. *BMC Bioinformatics* **6**, 155 (2005).
Presents details of structural identifiability analysis and its application to parameter estimation and optimal experimental design.
19. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning — Data Mining, Inference and Prediction* (Springer, New York, 2001).
20. Papoulis, A. in *Probability, Random Variables, and Stochastic Processes* (ed. Editions, M.-H. I.) (McGraw-Hill, New York, 1991).
21. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge Univ. Press, New York, 1992).
22. Golub, G. H. & Van Loan, C. F. An analysis of the total least squares problem. *SIAM J. Numer. Anal.* **17**, 883–893 (1980).
23. Danuser, G. & Strickler, M. Parametric model fitting: from inlier characterization to outlier detection. *IEEE Trans. Patt. Anal. Mach. Intell.* **20**, 263–280 (1998).
24. Rousseeuw, P. J. Least median of squares regression. *J. Am. Stat. Ass.* **79**, 871–880 (1984).
25. Koch, K.-R. *Parameter Estimation and Hypothesis Testing in Linear Models* (Springer, Berlin, 1988).
26. Pardalos, P. M. & Romeijn, H. E. *Handbook of Global Optimization Volume 2* (Kluwer Academic, Dordrecht, 2002).
27. Horst, R. & Pardalos, P. M. *Handbook of Global Optimization* (Kluwer Academic, Dodrecht, 1995).
28. Moles, C. G., Mendes, P. & Banga, J. R. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.* **13**, 2467–2474 (2003).
29. Kleinbaum, D. G., Kupper, L. L., Muller, K. E. & Nizam, A. *Applied Regression Analysis and Multivariable Methods* (Duxbury, 1997).
30. Seber, G. A. & Wild, C. J. *Nonlinear Regression* (Wiley-Interscience, Hoboken, 2004).
References 29 and 30 are comprehensive textbooks on linear and nonlinear regression and important related diagnostics.
31. Efron, B. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* **68**, 589–599 (1981).
32. Potvin, C. & Roff, D. A. Distribution-free and robust statistical methods: viable alternatives to parametric statistics. *Ecology* **74**, 1617–1628 (1993).
33. Coleman, M. C. & Block, D. E. Bayesian parameter estimation with informative priors for nonlinear systems. *AIChE J.* **52**, 651–667 (2005).
34. Barenco, M. *et al.* Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biol.* **7**, R25 (2006).
35. Chen, M., Shao, Q. & Ibrahim, J. G. *Monte Carlo Methods in Bayesian Computation* (Springer, New York, 2000).
Presents many computational techniques for carrying out Bayesian inference.
36. Schwarz, G. Estimating dimension of a model. *Ann. Stat.* **6**, 461–464 (1978).
37. Gruen, A. W. Data-processing methods for amateur photographs. *Photogramm. Rec.* **11**, 567–579 (1985).
38. Golub, G. H. & Van Loan, C. F. *Matrix Computations* (Johns Hopkins Univ. Press, Baltimore, 1983).
39. Pedraza, J. M. & van Oudenaarden, A. Noise propagation in gene networks. *Science* **307**, 1965–1969 (2005).
40. Rosenfeld, N., Young, J. W., Alon, U., Swain, P. S. & Elowitz, M. B. Gene regulation at the single-cell level. *Science* **307**, 1962–1965 (2005).
41. Bennett, M. R. & Kearns, J. L. Statistics of transmitter release at nerve terminals. *Prog. Neurobiol.* **60**, 545–606 (2000).
42. Redman, S. Quantal analysis of synaptic potentials in neurons of the central nervous system. *Physiol. Rev.* **70**, 165–198 (1990).
43. Morton-Firth, C. J. & Bray, D. Predicting temporal fluctuations in an intracellular signalling pathway. *J. Theor. Biol.* **192**, 117–128 (1998).
44. Spudich, J. L. & Koshland, D. E. Non-genetic individuality — chance in single cell. *Nature* **262**, 467–471 (1976).
45. Mitchison, T. & Kirschner, M. Dynamic instability of microtubule growth. *Nature* **312**, 237–242 (1984).
46. Smith, Jr A. A. Estimating nonlinear time-series models using simulated vector autoregression. *J. Appl. Econometrics* **8**, S63–S84 (1993).
Introduces methods of indirect inference for the estimation of parameters in probabilistic models.
47. Gourieroux, C., Monfort, A. & Renault, E. Indirect inference. *J. Appl. Econometrics* **8**, S85–S118 (1993).
48. Jiang, W. & Turnbull, B. The indirect method: inference based on intermediate statistics — a synthesis and examples. *Stat. Sci.* **19**, 239–263 (2004).
49. Gallant, A. R. & Tauchen, G. Which moments to match? *Econometric Theory* **12**, 657–681 (1996).
50. Golightly, A. & Wilkinson, D. J. Bayesian sequential inference for stochastic kinetic biochemical network models. *J. Comp. Biol.* **13**, 838–851 (2006).
51. O'Neill, P. D. & Roberts, G. O. Bayesian inference for partially observed stochastic epidemics. *J. Royal Stat. Soc. A* **162**, 121–129 (1999).
52. Gibson, G. J., Kleczkowski, A. & Gilligan, C. A. Bayesian analysis of botanical epidemics using stochastic compartmental models. *Proc. Natl Acad. Sci. USA* **101**, 12120–12124 (2004).
53. Smith, A. F. M. & Roberts, G. O. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo Methods. *J. Royal Stat. Soc. B* **55**, 3–23 (1993).
54. Wilkinson, D. J. *Stochastic Modelling for Systems Biology* (CRC Press, Boca Raton, 2006).
Discusses issues that are related to probabilistic modelling and the estimation of parameters in stochastic models using Bayesian inference.
55. Walker, R. A. *et al.* Dynamic instability of individual microtubules analyzed by video light-microscopy — rate constants and transition frequencies. *J. Cell Biol.* **107**, 1437–1448 (1988).
56. Shaw, S. L., Yeh, E., Maddox, P., Salmon, E. D. & Bloom, K. Astral microtubule dynamics in yeast: a microtubule-based searching mechanism for spindle orientation and nuclear migration into the bud. *J. Cell Biol.* **139**, 985–994 (1997).
57. Odde, D. J., Cassimeris, L. & Buettner, H. M. Kinetics of microtubule catastrophe assessed by probabilistic analysis. *Biophys. J.* **69**, 796–802 (1995).
58. Gildersleeve, R. F., Cross, A. R., Cullen, K. E., Fagen, A. P. & Williams, R. C. Microtubules grow and shorten at intrinsically variable rates. *J. Biol. Chem.* **267**, 7995–8006 (1992).
59. Dorn, J. F. *et al.* Interphase kinetochore microtubule dynamics in yeast analyzed by high-resolution microscopy. *Biophys. J.* **89**, 2835–2854 (2005).
60. Jaqaman, K. *et al.* Comparative autoregressive moving average analysis of kinetochore microtubule dynamics in yeast. *Biophys. J.***91**, 2312–2325 (2006).

Acknowledgements
This work was supported in part by a National Institutes of Health grant. K.J. is a Paul Sigler/Agouron fellow of the Helen Hay Whitney Foundation.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Gaudenz Danuser's homepage: http://lccb.scripps.edu

SUPPLEMENTARY INFORMATION
See online article: S1 (table)
Access to this links box is available online.