Chapter 2
Reverse Engineering Under Uncertainty

Paul Kirk, Daniel Silk and Michael P.H. Stumpf

Abstract The increased availability of experimental data in systems biology and systems medicine can only lead to better understanding of biological and disease related processes, if we can place them in the context of mechanistic models. Such models can serve as conceptual, but also computational frameworks in which we can reason about, or predict the behaviour of e.g. molecular networks, or cellular processes. Constructing such models, however, remains a formidable challenge: not only are the data noisy and incomplete, but the models that are currently available are hopelessly oversimplified. In this chapter we set out the problems and a list of potential ways of tackling them. The essential premise is always to be aware of the uncertainties inherent in the data and our models.

Keywords Inverse problems · Model selection · Extrinsic versus intrinsic noise · Model misspecification

2.1 Introduction

Reverse engineering the processes that govern the behaviour of biological systems is one of the principal aims of systems biology [46]. From experimental data, we seek to elucidate key aspects of the underlying mechanisms that give rise to observed complex behaviour. We may initially have only very vague, perhaps even wrong, ideas regarding these mechanisms, in which case our first aim may be to use the data in order to generate testable hypotheses. Alternatively, we may have already expressed our existing hypotheses as one or more mathematical models, in which case we may wish to use the data in order to tune their parameters, or to choose between them.

A defining feature of reverse engineering in a biological context is the variety of ways in which we encounter uncertainty [11]. In addition to the usual challenges

P. Kirk · D. Silk · M.P.H. Stumpf (✉)
London, UK
e-mail: m.stumpf@imperial.ac.uk

© Springer International Publishing Switzerland 2016

L. Geris and D. Gomez-Cabrero (eds.), *Uncertainty in Biology*,
Studies in Mechanobiology, Tissue Engineering and Biomaterials 17,
DOI 10.1007/978-3-319-21296-8_2

presented by measurement noise, we must also contend with the inherently stochastic nature of biochemical and biophysical processes. Moreover, given the complexity and interconnectedness of biological processes, we are currently only able to probe incomplete portions of the systems of interest, which has obvious consequences for the analysis [12]. While in vitro studies typically provide us with more control and might even enable us to isolate a particular process, we are still faced with the problem of establishing whether this idealised environment can be representative of the much more complex one that exists in vivo [33]. This combination of measurement noise, incomplete observations, and inherently nonlinear and stochastic underlying processes makes reverse engineering biological systems a particularly difficult task.

In this chapter, we discuss some of the challenges presented by reverse engineering under uncertainty in a biological context. In Sect. 2.2, we provide a broad overview of the *inverse problem* in systems biology, and consider the various ways in which this problem is encountered in practice. We then consider manifestations of uncertainty in Sect. 2.3, and ways in which we can try to cope with them when addressing the inverse problem. In Sect. 2.4, we consider the consequences of uncertainty in the context of modelling, and the potential limitations that uncertainty imposes on what we are able to learn. We offer some final conclusions and advice in Sect. 2.5.

2.2 The Inverse Problem in Systems Biology

An *inverse problem* is one in which we seek to reverse engineer details of a system (or data-generating mechanism) from experimental observations or measurements [49]. Typically, this will involve inferring a model or its parameters from experimental data. In contrast, a *forward problem* is one in which we have a fully specified model and we use it to make predictions or draw conclusions about its behaviour. There is clearly an interplay between inverse and forward problems: a reverse engineered model can subsequently be used for prediction, while a model whose predictions disagree with novel experimental observations might form the basis for a new model. The inverse problem has gained particular prominence in systems biology [22, 55, 58, 61], where we often have access to large quantities of high-throughput data, but may initially lack a deep understanding of how these measurements relate to one another, or what they can tell us about the underlying biological processes [36].

The difficulty of the inverse problem is hard to overstate. Even for simple systems (in terms of the model) it presents formidable challenges and is vastly more complicated than any associated forward problem.

2.2.1 The Different Types of Inverse Problems

We can consider three different, yet closely related, types of inverse problem: (i) we do not have a model and need to reverse engineer one from the data; (ii) we have

a model, the parameters of which need to be estimated/inferred from the data; and (iii) we have a number of distinct candidate models (for which we may or may not know the parameters) and we need to choose between them.

The first type of inverse problem has attracted a lot of attention in systems biology, particularly in the context of *network inference* [22, 38, 39, 50, 57]. Network inference approaches often proceed by first calculating measures of statistical dependence between different biological entities (which form the nodes of the network), and then identifying the pairs of entities between which there is a significant statistical dependence (these define the edges of the network). Some approaches take pains to try to identify direct, causal relationships by eliminating conditional dependencies. Network inference techniques typically have the advantage of being applicable to large-scale problems (e.g. finding dependencies between the expression levels of genes). The resulting network representations tend to be *descriptive* rather then *predictive*, and hence network inference is often seen as a method for hypothesis generation, which may be a first step toward developing more detailed mechanistic models.

The second type of inverse problem describes the problem of estimating the parameters of a known (or assumed) model, which is sometimes known as *model calibration*. In addition to more heuristic methods [6], approaches such as maximum likelihood estimation [56] and Bayesian inference have gained traction in recent years as ways in which to tackle model calibration problems. We consider these methods in more detail in Sect. 2.2.2.

The third type of inverse problem refers to *model selection*. In this case, we wish to choose the 'best' model(s) from a collection (and/or may wish to reject the 'worst'). Usually, our assessment of a model requires us to strike a balance between two criteria: (i) quality of fit; and (ii) complexity. In the interests of parsimony (also known as *Occam's razor*), we ideally wish to maximise the former while minimising the latter, and numerous approaches exist that seek to address this problem. Measures such as the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) do this by combining an assessment of quality of fit with a penalty on the number of parameters (which is taken as a proxy for model complexity). Alternatively, Bayesian approaches usually focus on estimating the *evidence* (or *marginal likelihood*) for different models, and then compare these quantities via the calculation of *Bayes factors*. Marginal likelihood estimation is typically challenging and computationally costly; however, Bayesian approaches have the advantage of naturally embodying the principle of Occam's razor. These and other procedures for model selection are discussed in more detail in [28, 47].

2.2.2 Statistical Inference Approaches

The general problem of fitting a model to data is often approached by considering some function that quantifies the discrepancy (or, alternatively, agreement) between the model's predictions and the observed data, and then tuning the model's parameters

in order to obtain a good fit. Examples of the kind of discrepancy function that might be employed include quantitative distances such as sum of squares and sum of absolute difference errors, or qualitative measures such as the eigenvalues or Lyapunov spectrum of a dynamical system [3, 41]. The choice of discrepancy function is usually based upon heuristic arguments, but is often important, affecting (for example) the degree to which outliers influence the fit. A key problem when adopting such a fitting approach is how to find the minimum of the discrepancy function, and numerous optimisation strategies exist that can be applied for this purpose [7, 53]. Two further important considerations are: (i) the problem of local minima; and (ii) over fitting. The first of these refers to the common problem of the optimisation algorithm getting "stuck" in a local minimum, rather than identifying the parameters that yield the true, global minimum. The second refers to the challenge of how to avoid fitting the experimental noise [40], which will typically result in poor predictive performance.

If our model is probabilistic, we will often be able to define a *likelihood function* [10], $L(\boldsymbol{\theta}) = p(D|\boldsymbol{\theta}, \mathcal{M})$, which scores parameters by assessing how likely the observed data, D, would be under the assumption that those parameters $\boldsymbol{\theta}$ (and our model, $\mathcal{M}$) are correct. In *maximum likelihood* (ML) estimation, we seek the parameters that maximise this likelihood function. In order to improve numerical stability, in practice we often work with the *log* likelihood function. Moreover, due to the way in which optimisation routines are typically implemented, we often think in terms of *minimising* the *negative* log likelihood, which we can consider as a particular kind of discrepancy function that happens to have the advantage of having a formal probabilistic grounding. The challenges of escaping local minima and avoiding over fitting remain.

The Bayesian formalism [16, 45], provides a framework for performing parameter inference, in which assessments of fit (as quantified by the likelihood function) are combined with our prior belief regarding the parameter values. Here, "prior belief" refers to the belief we have before observing the current dataset, and may have been obtained on the basis of previous experiments (e.g. on related biological systems, or in similar conditions). The Bayes rule provides us with a formal mathematical means by which to update our prior belief in light of the observed data, in order to obtain the *posterior distribution*. The posterior quantifies the uncertainty remaining in the values of the parameters after having observed the data, and may be used to derive *credible regions* for the parameter vector. More precisely, we have,

$$p(\boldsymbol{\theta}|D, \mathcal{M}) = \frac{p(\boldsymbol{\theta}|\mathcal{M})p(D|\boldsymbol{\theta}, \mathcal{M})}{\int_{\boldsymbol{\theta}\in\Theta} p(\boldsymbol{\theta}|\mathcal{M})p(D|\boldsymbol{\theta}, \mathcal{M})d\boldsymbol{\theta}}, \tag{2.1}$$

where D is the dataset, $\boldsymbol{\theta}$ is the vector of parameters that is to be inferred, and $\mathcal{M}$ represents the model. In words, we have,

$$\text{Posterior} = \frac{\text{Prior} \times \text{Likelihood}}{\text{Model evidence}}. \tag{2.2}$$

In practice, elucidation of the posterior distribution is rarely possible analytically, and hence we must resort to techniques for obtaining samples from the posterior, such as Markov chain Monte Carlo (MCMC), sequential Monte Carlo (SMC), or nested sampling [25, 42]. For some problems it may not even be possible to write down the likelihood, in which case *approximate likelihood* techniques and approaches such as approximate Bayesian computation might be appropriate [52, 60].

2.2.3 Bypassing the Inverse Problem

It is usually impossible to measure all of the parameters or all of the components of a biological system experimentally, and hence addressing the inverse problem is an unavoidable reality. However, even if we *were* able to measure all of these quantities, they would only be valid for the particular experimental and biological conditions under which the measurements were taken; molecular reaction rates, for example, depend on ambient temperature and pH values among many other things. Given that these conditions are themselves subject to random fluctuations, modelling the variation in these quantities is of vital importance if we wish to understand the sources of uncertainty and variability in the system and in our data.

2.3 Manifestations of Uncertainty

One of the most significant challenges to be overcome when trying to reverse engineer biological processes is the variety of sources of uncertainty that we must take into account. In this section, we describe the various sources of noise that might be important, and discuss strategies for coping with this noise when performing inference.

2.3.1 Sources of Noise

There are many different sources of noise that have an impact on if and how we can reverse engineer a given biological process. On the one hand, we have *experimental noise*, which arises from imprecision and or inaccuracy in the measurement process. On the other, we have the inherently stochastic nature of the underlying biological system [4, 29, 51], which is a component of what we seek to reverse engineer. In the context of cellular noise, this is often investigated in terms of *intrinsic* and *extrinsic* sources.

2.3.1.1 Experimental Noise

In the analysis of experimental error, a distinction is made between the precision and accuracy of an observation [37]. Precision refers to the inherent error distribution associated with a particular type of experiment, and accuracy to the existence of systematic errors in the experimental process. Contributions to the former will vary for repeat observations and include, for example, random fluctuations in the experimental conditions or behaviour of the experimental instruments. In contrast, systematic errors remain unchanged for repeated experiments, and are caused by, for example, imperfect calibration of experimental instruments. If the cause is known, systematic errors should be explicitly modelled in order to avoid bias in any inferred quantities. Otherwise, undetected systematic error can be viewed as a source model misspecification which will be discussed more generally in Sect. 2.4.2.

2.3.1.2 Intrinsic Noise

Cellular behaviour is governed by the biochemical reactions that occur between different molecular species within the cell. The timing of individual reactions is a random quantity, which gives rise to the source of cellular stochasticity known as intrinsic noise. Since each individual reaction only changes the numbers of molecules of the reacting species by one or two, the effects of intrinsic noise are particularly important when there are only low copy numbers of the molecular species of interest.

2.3.1.3 Extrinsic Noise

Extrinsic noise refers to variability in the physical and biological environment within which the intrinsically noisy interactions take place. For example, a collection of cells may vary in cellular volume, be at different stages of the cellular cycle, or have different abundance of RNA polymerase and ribosomes; all of which may contribute to variability in behaviour between cells and subsequent experimental measurements.

2.3.2 Coping with Uncertainty in Inference

Having identified a variety of sources of noise, we now discuss how we should address or capture these when performing parameter inference. The key question is how to model each type of noise, so that we can either derive a likelihood function (and hence adopt a maximum likelihood or Bayesian approach) or else find some other (possibly simulation-based) method for inferring parameters. We consider strategies for coping with each of the three types of noise identified in the previous section.

2.3.2.1 Coping with Experimental Noise

Each source of random experimental uncertainty may be categorised further as applying to either the inputs or outputs of an experiment. In the latter case the measurement error, ϵ, is typically assumed independent of both the parameters and known inputs, (θ, u), and the true state of the system. The likelihood thus factorises into components describing the uncertainty generated by the system and parameters, and by the measurement process,

$$\begin{aligned} L(\theta) &= p(D|\theta, u) \\ &= p(D^*|\theta, u)p(\epsilon) \end{aligned}$$

where D^* is the error-free (i.e. absent of experimental noise) state of the system.

In the less commonly discussed case of uncertain inputs, the true state of the observable is no longer independent of the uncertainty in question, and the likelihood is obtained by integrating over possible values of u,

$$L(\theta) = p(\epsilon) \int p(D^*|\theta, u)du. \tag{2.3}$$

The integral in Eq. 2.3 describes how the error propagates through the system for particular values of θ, and often may only be approximately evaluated. A variety of methods to do so exist, including Monte Carlo approaches [35], Sigma point methods [26], or Gaussian quadrature [43], the appropriateness of each of which is determined by both the complexity of the system model, and the distribution, $p(u)$.

Commonly the total experimental error is summarised as additive and Gaussian. Such an approximation may be justified (as a consequence of the Central Limit theorem) when the errors are the accumulation of large numbers of independent sources of uncertainty. The Gaussian assumption is certainly computationally convenient. For example, if all sources of uncertainty and the data itself are Gaussian distributed, then calculation of the integral in Eq. 2.3 may be undertaken with relative efficiency (e.g. by using the unscented transform [27]). However, it is important to note that the effects of input error (even when assumed Gaussian) on p(y) will almost certainly not be Gaussian in the presence of any non-linearity. Further, care must be taken when measured quantities lie close to limiting boundaries (e.g. abundance or concentration is strictly positive), as this can induce non-Gaussian effects upon the error distribution. In these cases, more sophisticated and computationally expensive Monte-Carlo based approaches are necessary for evaluating the likelihood.

2.3.2.2 Coping with Intrinsic Noise

We assume that the available data comprise intrinsically noisy measurements obtained at discrete time points. While it is possible to derive exact Markov chain Monte Carlo schemes for inference in such situations, their computational cost is usually

prohibitively expensive. However, a number of approaches exist for *simulating* intrinsic fluctuations, and hence several *simulation-based* inference procedures have been proposed. We refer the reader to [19, 59] for examples. At the heart of all of these approaches is simulation using Gillespie's stochastic simulation algorithm (SSA) [18] (see also the top plots in Fig. 2.1 for example realisations). Given a chemical reaction system with known rate constants and initial molecule numbers, the SSA proceeds by using Monte Carlo techniques to simulate both the time until the next reaction, and the next reaction to occur. A number of modifications exist in order accelerate simulation using the SSA, including the Gibson-Bruck algorithm and the τ-leap method [19, 59]. All of these simulation methods have in common that they they provide exact realisations from the underlying (discrete state, continuous time) stochastic kinetic model.

Alternative methods for parameter inference approximate the underlying stochastic kinetic model in order to derive *approximate likelihood* functions. A popular approach is to consider the continuous-state *diffusion approximation* of the true process, which yields a stochastic differential equation (SDE) known as the chemical Langevin equation (CLE). An alternative continuous approximation is given by the linear noise approximation (LNA) [20]. Additionally, several moment expansion and moment closure approaches have been proposed as ways of approximating the underlying model, some of which have also been used in order to allow parameter estimation to be performed.

2.3.2.3 Coping with Extrinsic Noise

Extrinsic noise may be modelled by specifying a probability distribution, $p(\theta, x_0)$, over the parameters and initial conditions [4]. In Fig. 2.1 (right column), we illustrate the effects of extrinsic noise on the oscillations in a model of p53 dynamics, where the extrinsic noise enters the model through fluctuations in just one of the parameters. In this example, we have both intrinsic and extrinsic effects (see also Sect. 2.3.2.4). In the absence of intrinsic stochasticity, extrinsic effects may be simulated in exactly the same way as propagating input uncertainty (discussed in Sect. 2.4.2)—by propagating $p(\theta, x_0)$ through the model. The parameters of the extrinsic noise distribution $p(\theta, x_0)$, may also be the subject of inference given suitable data, such as multiple measurements at single cell resolution.

2.3.2.4 Coping with Mixed Noise Sources

When intrinsic and extrinsic noise are both present, the modelling challenges are more substantial, both conceptually and computationally. The most common approach, originating from [48], is to derive a framework under which each source of noise may be considered separately, whilst other sources are held fixed. The theoretical justification is made via the following decomposition of the stochasticity of cellular products, x, as the direct sum of extrinsic and intrinsic (and experimental) contri-

butions. Defining the extrinsic and intrinsic variables (or parameters) as E and I respectively, the total law of variance gives us,

$$\sigma_x^2 = \underbrace{\sigma^2_{\langle x|E\rangle}}_{\text{Extrinsic}} + \underbrace{\langle\sigma^2_{\langle x|E,I\rangle|E}\rangle}_{\text{Intrinsic}} + \underbrace{\langle\sigma^2_{x|E,I}\rangle}_{\text{Experimental}} \tag{2.4}$$

where the angular brackets represent the expectation. The first term is the variance of the mean values of x with E held fixed, and describes the portion of the total uncertainty arising from extrinsic variability. The second term describes the intrinsic contribution—the mean variance of x when sources of uncertainty other than E and I are averaged out, and E is held fixed. The final term is that part of the total variance

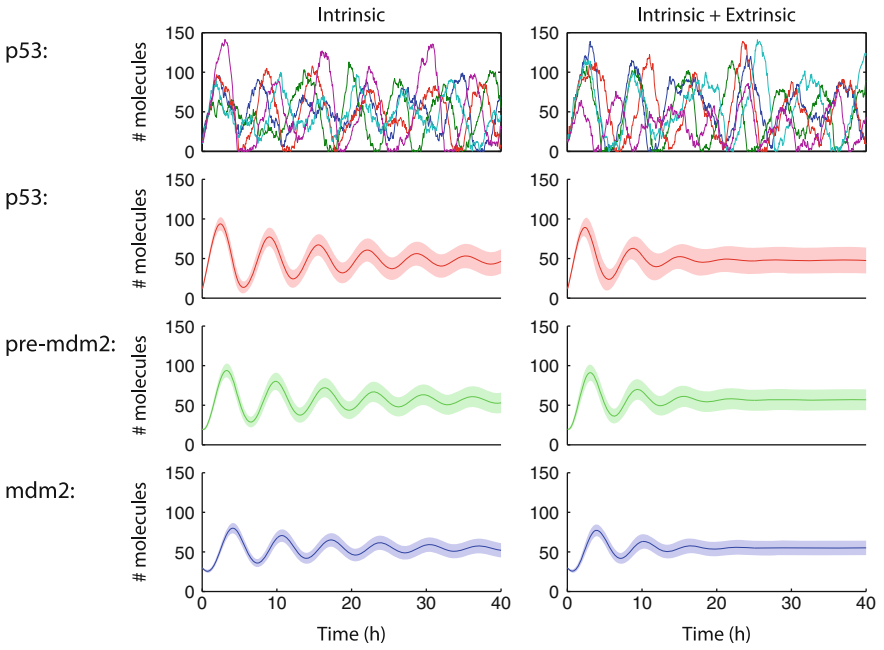

Fig. 2.1 We consider a model of oscillatory p53 dynamics [17]. The model comprises three protein species (p53, precursor of Mdm2 and Mdm2) connected through a nonlinear feedback loop. We take the parameters of the system (see [1] for details) to be $[k_1, k_2, k_3, k_4, k_5, k_6, k_7] = [90, 0.002, 1.7, 1.1, 0.93, 0.96, 0.01]$, with initial conditions fixed at [p53, pre-mdm2, mdm2] = [10, 20, 30] at time $t = 0$. In the *top left*, we show individual realisations of the number of p53 molecules over time, obtained using Gillespie's stochastic simulation algorithm (SSA). Below this, we have 3 plots showing the population mean (*solid line*) and a 1 standard deviation shaded region for the 3 protein species (as indicated), obtained by averaging over many SSA runs. On the *right*, we show the same 4 plots, but this time we illustrate the effects of extrinsic fluctuations by assuming that the k_4 parameter is drawn from a Gamma(12, 0.1) distribution (so that the mode is at $k_4 = 1.1$). While it is difficult to discern any difference from the individual SSA simulations (top plots), it is clear from the plots of the population means that the effect of extrinsic noise in this case is stronger dampening of the oscillations

that is not explained by experimental or intrinsic sources, and which we attribute to uncertainty in the measurement process.

This noise decomposition suggests the innovative dual-reporter experiments—where the products of two genes, regulated by identical promoters are simultaneously measured—in order to quantify intrinsic and extrinsic contributions. Furthermore, it suggests that intrinsic, extrinsic and experimental uncertainty may be modelled jointly by combining their separate strategies in hierarchical fashion. This is demonstrated for intrinsic and extrinsic variability by Toni and Tidor [51], using the linear noise approximation and the unscented transform respectively.

The total law of variance based approach, however, is only accurate when changes in extrinsic variables with time are much slower than fluctuations in intrinsic variables [23]. It turns out that inferring the contributions to total variance from extrinsic and intrinsic sources is reliant upon the history of extrinsic fluctuations and not just their present state. Even if all extrinsic variables can be measured accurately, Eq. 2.4 will introduce errors if the extrinsic variables cannot be assumed constant in time.

2.3.3 Quantifying Information and Knowledge

Given the variety of noise sources that may exist in the underlying processes that generated the data, we may wonder exactly how much information can be extracted from a given dataset. In the context of reverse engineering, our principal concern is the degree to which we will be able to reconstruct the biological process of interest from the available experimental observations. It is therefore useful to be able to quantify the amount of information that our data contain about the parameters that we seek to infer. In the Bayesian formalism, this is conceptually simple. Before we conduct the experiment, the prior distribution describes the knowledge that we have regarding the values of the unknown model parameters. The posterior distribution serves the same role, but *after* observation of the data. The compression from prior to posterior provides an information theoretic measure of the information gain provided by the data. This compression can be quantified by calculating the *Kullback-Leibler divergence* [9] between posterior and prior,

$$d_{KL}\left(p(\boldsymbol{\theta}|D,\mathcal{M}),p(\boldsymbol{\theta}|\mathcal{M})\right)=\int_{\boldsymbol{\theta}\in\Theta}p(\boldsymbol{\theta}|D,\mathcal{M})\log\left(\frac{p(\boldsymbol{\theta}|D,\mathcal{M})}{p(\boldsymbol{\theta}|\mathcal{M})}\right)d\boldsymbol{\theta}. \quad (2.5)$$

Typically, it will not be possible to calculate this divergence analytically; however, there are Monte Carlo methods that permit its estimation.

2.4 Models in Biology and Confidence in Models

2.4.1 Data versus Reality

Despite the increasing range and power of experimental techniques, datasets continue to represent low-dimensional snapshots of the complex cellular environment. It is the task of reverse engineering to interpret the data and fill in the blanks—to explain observed, and allow the prediction of unobserved properties of the real system. It is clear that the quality, quantity, context and subject of experimental observations determines both the inferences that may be drawn and the confidence we associate with them. For example, larger datasets with higher signal to error ratios will in general lead to greater accuracy and precision. However, in many cases the relative utility of different experimental choices can be hard to foresee, e.g., which species should be measured or perturbed (illustrated in Fig. 2.2 and more generally by Liepe et al. [32]), and whether longitudinal datasets or time-point data should be generated in order to reduce the uncertainty in parameter estimates [30].

Here it can be useful to close the loop between experiment and model, by rationally seeking experiments that maximise the expected information available for the inference task at hand. This is known as *experimental design*, of which recent developments in the context of model calibration include the work of Liepe et al. [32] that builds upon existing methods [2, 8, 24, 31, 34, 54], by utilising a sequential approximate Bayesian computation framework to choose the experiment that maximises the expected mutual information between prior and posterior parameter distributions. In so doing, they are able to optimally narrow the resulting posterior parameter or predictive distributions, incorporate preliminary experimental data

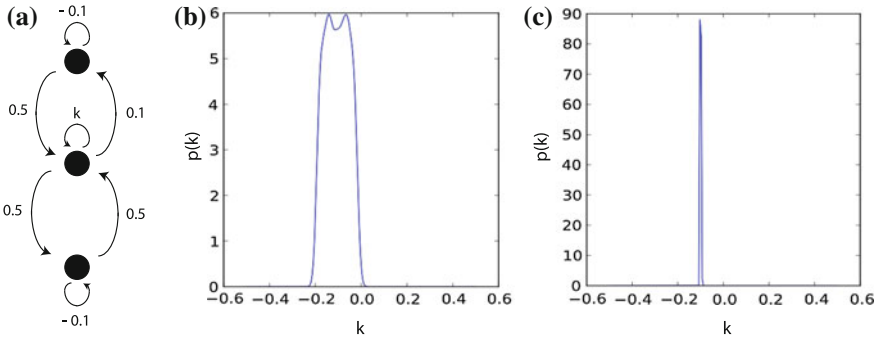

Fig. 2.2 Some experiments are more informative than others. **a** Schematic of a three variable system of ordinary differential equations. *Arrows* represent interactions that are modelled by linear terms with coefficients shown. Inference for k is performed independently for two timeseries datasets that are generated by simulating the model with $k = -0.1$, and measuring the state of **b** the top variable and **c** the middle variable for times $t = 0.5, 1, 1.5, 2$. The broadness of the resulting marginal posterior distributions differ substantially, reflecting the different levels of information contained within the datasets

and provide sensitivity and robustness analyses. Design frameworks also exist for model selection (e.g. [15]), where experiments are sought that maximally distinguish the prior predictive distributions of the competing models.

Although experimental design offers a powerful auxiliary tool to statistical inference, care must be taken in interpreting the confidence associated with inferred models and parameters. For example, it is unsurprising that we assign high confidence to the outcome of a model selection analysis given data from an optimally designed experiment. When each of the models is subject to some level of misspecification, such confidence may be misleading.

2.4.2 Models versus Reality

The complexity of cellular behaviour makes it inevitable that reverse (or forward) engineered systems models will be subject to misspecification errors (when they relate to the observation model, they are called systematic errors). These errors in the model may remain undetected, or they may be introduced knowingly via model reductions aimed at simplifying downstream analyses or at increasing interpretability. In either case, such model uncertainty affects predictions and the outcomes of statistical inferences. For example, inferred values for the physical parameters of a 'wrong' model will also be 'wrong' in order to compensate for misspecification (for example, see Fig. 2.3). Indeed, strictly speaking, Bayesian inference is valid only when a 'true' model is considered.

The effects of parameter and input uncertainty may be quantified by assessing their effect on the likelihood and posterior model predictions. For some classes of model

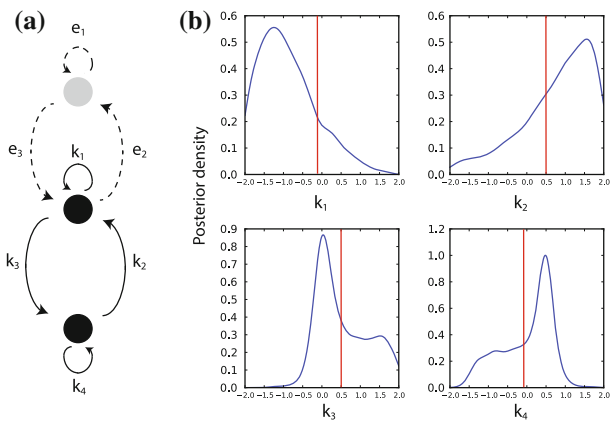

Fig. 2.3 Inference using a 'wrong' model. **a** A misspecified model of a 'true' data generating system are considered. The *grey circle* and dotted arrows represent a true variable and its interactions that are absent from the wrong model. Fixing $(e_1, e_2, e_3) = (-0.1, 0.1, 0.5)$, and $(k_1, k_2, k_3, k_4) = (-0.1, 0.5, 0.5, -0.1)$ a timeseries of 10 data points is simulated from which the k_i are inferred using the incomplete model. **b** Marginal posterior densities for the k_i. Maximum *a posteriori* (MAP) estimates do not coincide with the true parameter values (shown in *red*)

a similar treatment of model uncertainty may be undertaken by capturing the range of possible errors through parametric expansions, and examining the importance of each (e.g. [44]). However more generally, and certainly for mechanistic models, such an approach is undermined by the conceptual and computational difficulties of specifying the complete space of model errors. However, a consideration of the possible sources of model uncertainty may still suggest a collection of possible models that are in reasonable agreement with the data. In this case, the propagation of misspecification may be managed, to an extent, by conditioning upon the whole collection, rather than just on the single best model. This is the basis of the model averaging framework, where the best estimate of the state or parameter of the system, θ, along with confidence intervals may be calculated from the averaged probability under the various models,

$$p(\theta|\mathcal{M}_1, ..., \mathcal{M}_N) = \sum_{i=0}^{N} p(\mathcal{M}_i|D)p(\theta|\mathcal{M}_i)$$

where $p(\mathcal{M}_i|D)$ is the posterior probability of model, $\mathcal{M}_i$, given the data, D, and $p(\theta|\mathcal{M}_i)$ is the posterior distribution for θ under model $\mathcal{M}_i$. While each $\mathcal{M}_i$ is still 'wrong', the averaged prediction of all 'wrong' models at least accounts for a portion of model uncertainty. However the major drawback of averaging, rather than selecting, is to diminish their physical interpretability.

2.4.3 What Can Be Learned from Data?

Frequently we find that parameters, or more often combinations of parameters, can be varied over orders of magnitude without changing the output of a system appreciably [13]. This has major implications for the inverse problem of estimating parameters from data, as large sub-regions in parameter-space may be commensurate with a given dataset. The dependence (or lack thereof) of parameters with respect to data is referred to as *inferability*, which in practice may be quantified as the variance about the Maximum *a posteriori* (MAP) estimate. More formally, the Cramer-Rao inequality [10] gives us a bound on the precision to which a parameter may be estimated in terms of the likelihood,

$$\sigma_\theta^2 \geq \mathcal{I}^{-1}(\theta),$$

with $\mathcal{I}(\theta)$ being the Fisher information matrix (FIM),

$$\mathcal{I}(\theta) = E_\theta \left[\left(\frac{\partial \log(p(D|\theta))}{\partial \theta} \right)^2 \right],$$

and σ_θ^2, the covariance matrix of a vector-valued θ.

The FIM is at the heart of much of statistical inference and can be interpreted as the curvature of the likelihood surface around the maximal value of the likelihood function. It can also be used as a means to consider robustness and sensitivity of dynamical systems. The reason for this is that if a system is sensitive to variation in a parameter, or a combination of parameters, then this means that changing the parameter, e.g. from θ to $\theta + \delta$, will result in a noticeable change to the system output, which in turn means that the likelihood will also be altered appreciably. Notice, however, that inferability is a property of both system and data—it is possible that further observations will render previously 'sloppy' [21] parameters inferable with high certainty (see Fig. 2.2).

Often improved fits to data or better model predictions are interpreted as evidence that more about the true system is being captured. However, it is easy to construct counter-examples where improved data fitting and even predictive power (although desirable in their own right) can be achieved by including more inaccuracies into a misspecified model. It is crucial then not to interpret the physical meaning of any model too assuredly, but instead use them as tools to generate hypotheses for experimental testing (with the result, perhaps, of invalidating the model).

2.5 Conclusion

Reverse engineering is never easy, and probably even harder in biology than in the physical sciences, where sound physical principles can constrain the search space considerably. But once we accept that there is a point to applying quantitative methods and mathematical or computer models in biology, we have to face up to the challenges presented by inverse problems. There have been some arguments, perhaps most notably from Sydney Brenner [5], stating that the inverse problem in molecular and cellular biology is insurmountable and that we should use "the CELLMAP"; how this looks and where it would come from has thus far, sadly or unsurprisingly, been left unspecified.

In order to make progress with the topic of this chapter we have to consider two aspects of reverse engineering. First, problems where models can be tackled by existing methods of reverse engineering. Here we consider only those that make a meaningful and robust attempt at quantifying uncertainty as serious contenders, which restricts us essentially to methods based on statistical and sound probabilistic principles. For such systems it is easy to show that the inverse problem should be tackled in preference of solving sets of forward problems, which rely on experimentally measured parameter values, and which typically are associated with levels of uncertainty that are, it appears, rarely propagated in forward analyses. The best we can make out the elusive "CELLMAP" appears to be a fully parameterized model for the (cellular) system under consideration. Taking the predictions of such a system at face-value ignores uncertainty and does not appear a sound way of making progress.

There are statistical procedures which are provably consistent as the amount of data becomes infinite. This is clearly a situation far from reality but it seems advisable

to use these techniques also in situations where data are rare. The alternative would be to use an approach which is provably sub-optimal as data become perfect and abundant in the hope that it does a good job on poor data.

The second set of problems is more interesting, and probably more widespread: there are numerous systems (and models thereof) for which the inverse problem is indeed insurmountable. Here simple solutions simply do not exist (and a "CELLMAP" is sadly lacking). Two obvious attempts at addressing such problems—each with its own set of caveats—include *partial inference* and *model reduction*. While the details of their respective applicability depend crucially on the specific problem, we can make some general statements.

By *partial* or *composite inference* we mean a pragmatic approach that proceeds by either breaking up the problem into sub-systems for which satisfactory inferential solutions might exist, and then stitching the solutions for such subsystems together. This has the disadvantage that any correlations or interdependencies among subsystems are ignored. Nevertheless, techniques such as composite likelihood approaches [14] can help to make progress in inference problems that are not amenable to a comprehensive or holistic analysis. This will, we believe, continue to be a fruitful area for computational statistics.

Model reduction, on the other hand, requires more domain expertise about the system to be investigated. In the simplest case, it could be an effective model, which, for example, ignores some molecular species, if they exist only briefly and transiently. It could also be a model that looks at lower dimensional spatial problems (although this can be fraught with fundamental problems as mathematical solutions to problems in 1D and 2D can be qualitatively different from solutions in 3D).

Either approach, individually or in combination, may be worthwhile exploring in problems in systems biology (developmental biology seems to be replete with problems that pose challenges to inferential techniques), and is preferable to an analysis of corresponding forward problems for fixed parameters, which would mask uncertainty.

In summary, recent years have shown the fundamental new insights that can result from searching for or determining the origins of uncertainty in biological systems. In some cases, it will turn out that uncertainty is merely a nuisance (e.g. if it enters via the experimental procedure), whereas other types of uncertainty either reveal exciting new biological mechanisms (e.g. extrinsic variability typically points to aspects of a biological system that require further investigation), or are fundamental and inalienable aspects of biomolecular dynamics.

Failure to account for uncertainty in the analysis of biological systems (and in particular in reverse engineering tasks) will likely introduce bias and mask interesting biology. On the other hand, uncertainty becomes easier to deal with once we know where and how it arises.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Ale, A., Kirk, P., Stumpf, M.P.H.: A general moment expansion method for stochastic kinetic models. J. Chem. Phys. **138**(17), 174101 (2013)
2. Apgar, J.F., Witmer, D.K., White, F.M., Tidor, B.: Sloppy models, parameter uncertainty, and the role of experimental design. Mol. Biosyst. **6**(10), 1890–1900 (2010)
3. Barnes, C.P., Silk, D., Stumpf, M.P.H.: Bayesian design strategies for synthetic biology. Interf. Focus **1**(6), 895–908 (2011)
4. Bowsher, G.O., Swain, P.S.: Identifying sources of variation and the flow of information in biochemical networks. Proc. Natl. Acad. Sci. USA **109**(20), E1320–E1328 (2012)
5. Brenner, S.: Sequences and consequences. Philos. Trans. Royal Soc. Lond. Ser. B Biol. Sci. **365**(1537), 207–212 (2010)
6. Camacho, D., Vera Licona, P., Mendes, P., Laubenbacher, R.: Comparison of reverse-engineering methods using an in silico network. Ann. N Y Acad. Sci. 1115:73–89 (2007)
7. Cedersund, G., Sameulsson, O., Ball, G., Tegnér, J., Gomez-Cabrero, D.: Optimization in biology parameter estimation and the associated optimization problem. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Cham (2016, this volume)
8. Chu, Y., Hahn, J.: Integrating parameter selection with experimental design under uncertainty for nonlinear dynamic systems. AIChE J. **54**(9), 2310–2320 (2008)
9. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley-Interscience, New York (2006)
10. Cox, D.R., Hinkley, D.V.: Theoretical Statistics. Chapman&Hall/CRC, New York (1974)
11. Csete, M.E., Doyle, J.C.: Reverse engineering of biological complexity. Science **295**(5560), 1664–1669 (2002)
12. de Silva, E., Thorne, T., Ingram, P., Agrafioti, I., Swire, J., Wiuf, C., Stumpf, M.P.H.: The effects of incomplete protein interaction data on structural and evolutionary inferences. BMC Biol. **4**, 39 (2006)
13. Erguler, K., Stumpf, M.P.H.: Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. Mol. Biosyst. **7**(5), 1593–1602 (2011)
14. Fearnhead, P., Donnelly, P.: Approximate likelihood methods for estimating local recombination rates. J. Royal Stat. Soc. Ser. B Stat. Methodol. **64**(4), 657–680 (2002)
15. Flassig, R.J., Sundmacher, K.: Optimal design of stimulus experiments for robust discrimination of biochemical reaction networks. Bioinformatics (Oxford, England) **28**(23), 3089–3096 (2012)
16. Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B.L.: Bayesian Data Analysis, 2nd edn. Chapman & Hall/CRC , Boca Raton (2003)
17. Geva-Zatorsky, N., Rosenfeld, N., Itzkovitz, S., Milo, R., Sigal, A., Dekel, E., Yarnitzky, T., Liron, Y., Polak, P., Lahav, G., Alon, U.: Oscillations and variability in the p53 system. Mol. Syst. Biol. **2**, 2006.0033 (2006)
18. Gillespie, D.T.: A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. **22**(4), 403–434 (1976)
19. Gillespie, D.T.: Stochastic simulation of chemical kinetics. Annu. Rev. Phys. Chem. **58**(1), 35–55 (2007)
20. Grima, R.: A study of the accuracy of moment-closure approximations for stochastic chemical kinetics. J. Chem. Phys. **136**(15), 154105 (2012)

21. Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., Sethna, J.P.: Universally sloppy parameter sensitivities in systems biology models. PLoS Comput. Biol. **3**(10), e189 (2007)
22. Hartemink, A.J.: Reverse engineering gene regulatory networks. Nat. Biotechnol. **23**(5), 554–555 (2005)
23. Hilfinger, A., Johan, P.: Separating intrinsic from extrinsic fluctuations in dynamic biological systems. Proc. Natl. Acad. Sci. **108**(29), 12167–12172 (2011)
24. Huan, X., Marzouk, Y.M.: Simulation-based optimal Bayesian experimental design for nonlinear systems. J. Comput. Phys. **232**, 288–317 (2013)
25. Johnson, R., Kirk, P., Stumpf, M.P.H.: SYSBIONS: nested sampling for systems biology. Bioinform. **31**, 604–605 (2014)
26. Julier, S., Uhlmann, J.: A general method for approximating nonlinear transformations of probability distributions. Department of Engineering Science (1996)
27. Julier, S., Uhlmann, J., Durrant-Whyte, H.F.: A new method for the nonlinear transformation of means and covariances in filters and estimators. IEEE Trans. Autom. Control **45**(3), 477–482 (2000)
28. Kirk, P., Thorne, T., Stumpf, M.P.: Model selection in systems and synthetic biology. Curr. Opin. Biotechnol. **24**(4), 767–774 (2013)
29. Komorowski, M., Miekisz, J., Stumpf, M.P.H.: Decomposing noise in biochemical signaling systems highlights the role of protein degradation. Biophys. J. **104**(8), 1783–1793 (2013)
30. Komorowski, M., Costa, M.J., Rand, D.A., Stumpf, M.P.H.: Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. Proc. Natl. Acad. Sci. **108**(21), 8645–8650 (2011)
31. Kutalik, Z., Cho, K.-H., Wolkenhauer, O.: Optimal sampling time selection for parameter estimation in dynamic pathway modeling. Biosystems **75**(1–3), 43–55 (2004)
32. Liepe, J., Filippi, S., Komorowski, M., Stumpf, M.P.H.: Maximizing the information content of experiments in systems biology. PLoS Comput. Biol. **9**(1), e1002888 (2013)
33. Liepe, J., Taylor, H., Barnes, C.P., Huvet, M., Bugeon, L., Thorne, T., Lamb, J.R., Dallman, M.J., Stumpf, M.P.H.: Calibrating spatio-temporal models of leukocyte dynamics against in vivo live-imaging data using approximate Bayesian computation. Integr. Biol **4**(3), 335–345 (2012)
34. Lindley, D.V.: On a measure of the information provided by an experiment. Ann. Math. Stat. 986–1005 (1956)
35. Liu, J.S.: Monte Carlo strategies in scientific computing. Springer, Berlin (2008)
36. May, R.M.: Uses and abuses of mathematics in biology. Science **303**(5659), 790–793 (2004)
37. Pugh, E.M., Winslow, G.H.: The Analysis of Physical Measurements. Addison-Wesley series in physics. Addison-Wesley (1966). http://books.google.co.uk/books?id=-vREAAAAIAAJ
38. 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**(5721), 523–529 (2005)
39. Schäfer, J., Strimmer, K.: An empirical Bayes approach to inferring large-scale gene association networks. Bioinformatics **21**(6), 754–764 (2005)
40. Schliemann-Bullinger, M., Fey, D., Bastogne, T., Findeisen, R., Scheurich, P., Bullinger, E.: The experimental side of parameter estimation. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Cham (2016, this volume)
41. Silk, D., Kirk, P.D.W., Barnes, C.P., Toni, T., Rose, A., Moon, S., Dallman, M.J., Stumpf, M.P.H.: Designing attractive models via automated identification of chaotic and oscillatory dynamical regimes. Nat. Commun. **2**, 489 (2011)
42. Skilling, J.: Nested sampling for general Bayesian computation. Bayesian Anal. **1**, 833–860 (2006)
43. Stoer, J., Bulirsch, R., Bartels, R., Gautschi, W., Witzgall, C.: Introduction to Numerical Analysis, vol. 2. Springer, New York (1993)
44. Strong, M., Oakley, J.E., Chilcott, J.: Managing structural uncertainty in health economic decision models: a discrepancy approach. J. Royal Stat. Soc. Ser. C (Applied Statistics) **61**(1), 25–45 (2012)

45. Stuart, A.M.: Inverse problems: a Bayesian perspective. Acta Numerica **19**, 451–559 (2010)
46. Stumpf, M.P.H., Balding, D.J., Girolami, M.: Handbook of Statistical Systems Biology. Wiley, Chichester (2011)
47. Sunnåker, M., Stelling, J.: Model extension and model selection. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Cham (2016, this volume)
48. Swain, P.S., Elowitz, M.B., Siggia, E.D.: Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc. Natl. Acad. Sci. **99**(20), 12795–12800 (2002)
49. Tarantola, A.: Inverse Problem Theory and Methods for Model Selection. SIAM, Philadelphia (2005)
50. Thorne, T., Stumpf, M.P.H.: Inference of temporally varying Bayesian networks. Bioinformatics **28**(24), 3298–3305 (2012)
51. Toni, T., Tidor, B.: Combined model of intrinsic and extrinsic variability for computational network design with application to synthetic biology. PLoS Comput. Biol. **9**(3), e1002960 (2013)
52. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.H.: Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. Royal Soc. Interf. **6**(31), 187–202 (2009)
53. Tucker, W.: Interval methods. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Cham (2016, this volume)
54. Vanlier, J., Tiemann, C.A., Hilbers, P.A.J., van Riel, N.A.W.: A Bayesian approach to targeted experiment design. Bioinformatics (Oxford, England) **28**(8), 1136–1142 (2012)
55. Waldherr, S., Haasdonk, B.: Efficient parametric analysis of the chemical master equation through model order reduction. BMC Syst. Biol. **6**, 81 (2012)
56. Wang, Y., Christley, S., Mjolsness, E., Xie, X.: Parameter inference for discretely observed stochastic kinetic models using stochastic gradient descent. BMC Syst. Biol. **4**, 99 (2010)
57. Werhli, A.V., Grzegorczyk, M., Husmeier, D.: Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical gaussian models and bayesian networks. Bioinformatics **22**(20), 2523–2531 (2006)
58. Wilkinson, D.J.: Stochastic modelling for quantitative description of heterogeneous biological systems. Nat. Rev. Genet. **10**(2), 122–133 (2009)
59. Wilkinson, D.J.: Stochastic Modelling for Systems Biology. CRC PressI Llc, Boca Raton (2011)
60. Wilkinson, R.D.: Approximate Bayesian computation (ABC) gives exact results under the assumption of model error. Stat. Appl. Genet. Mol. Biol. **12**(2), 129–141 (2013)
61. Xu, T.-R., Vyshemirsky, V., Gormand, A., von Kriegsheim, A., Girolami, M., Baillie, Ketley, G.S.D., Dunlop, A.J., Milligan, G., Houslay, M.D., Kolch, W.: Inferring signaling pathway topologies from multiple perturbation measurements of specific biochemical species. Sci. Signal. 3(113), ra20 (2010)

http://www.springer.com/978-3-319-21295-1

Uncertainty in Biology
A Computational Modeling Approach
Geris, L.; Gomez-Cabrero, D. (Eds.)
2016, IX, 478 p., Hardcover
ISBN: 978-3-319-21295-1