Chapter 11
Sloppiness and the Geometry of Parameter Space

Brian K. Mannakee, Aaron P. Ragsdale, Mark K. Transtrum and Ryan N. Gutenkunst

Abstract When modeling complex biological systems, exploring parameter space is critical, because parameter values are typically poorly known a priori. This exploration can be challenging, because parameter space often has high dimension and complex structure. Recent work, however, has revealed universal structure in parameter space of models for nonlinear systems. In particular, models are often *sloppy*, with strong parameter correlations and an exponential range of parameter sensitivities. Here we review the evidence for universal sloppiness and its implications for parameter fitting, model prediction, and experimental design. In principle, one can transform parameters to alleviate sloppiness, but a parameterization-independent information geometry perspective reveals deeper universal structure. We thus also review the recent insights offered by information geometry, particularly in regard to sloppiness and numerical methods.

Keywords Sloppiness · Hessian · Experimental design · Bayesian ensembles · Cost functions · Information geometry

B.K. Mannakee
Graduate Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ, USA
e-mail: mannakee@email.arizona.edu

A.P. Ragsdale
Graduate Interdisciplinary Program in Applied Mathematics, University of Arizona, Tucson, AZ, USA
e-mail: aragsdale@math.arizona.edu

M.K. Transtrum
Department of Physics and Astronomy, Brigham Young University, Provo, UT, USA
e-mail: mktranstrum@byu.edu

R.N. Gutenkunst (✉)
Department of Molecular and Cellular Biology, Tucson, AZ, USA
e-mail: rgutenk@email.arizona.edu

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_11

11.1 Introduction

Mathematical models of cell-signaling, metabolic, and gene networks play a critical role in developing mechanistic understanding of these networks [45]. Building models can be difficult, however, because such networks often have complex nonlinear dynamics and not all components may be known. In fact, important uses of network models are to infer network structure [19] or choose between hypotheses regarding network function [43]. (For more on the challenges in reverse engineering biological networks, see Chap. 2 in this volume [42].) Even when the network is well-known, however, modeling can still be difficult, because mechanistic models typically depend upon a large number of kinetic parameters [29, 52]. Such parameters are often unknown and are difficult to measure experimentally.

In this chapter, we review methods for exploring the spaces of model parameters and data, and we review recent work on *sloppiness*, a general property of complex nonlinear models. Sloppy models have highly anisotropic parameter and data spaces, with complex relationships between parameter values and model output. Sloppiness results in several difficulties for modelers, including that numerical tools used to estimate parameters can be slow, confidence intervals for parameter values and model output can be large, and experiments to improve the model can be difficult to design. We review recent work on how these challenges arise and how they can be overcome.

11.1.1 Parameter Space and Data Space

In a model with N parameters $\theta_1, \ldots, \theta_N$, the parameter space encompasses the set of all possible values for each parameter. Most commonly, parameters are real numbers, in which case the N-dimensional parameter space is a subspace of $\mathbb{R}^N$. A particular realization of parameter values $\theta = [\theta_1, \theta_2, \ldots, \theta_N]$ is a vector representing a single point in parameter space. For biological models, in which parameters may have different units and scales that differ by orders of magnitude, it is convenient to consider logarithmic rather than absolute parameter values, so in this chapter when we speak of parameters and parameter space we are always referring to log-parameters. One can think of the model as a function mapping points or regions in parameter space to output values, or points in data space [18], and a general problem in systems biology is to understand this mapping.

A complete description of this map is a useful mathematical tool. For instance, once the mapping is understood, it is easy to enumerate the possible outcomes a model can generate. Thus, given a model that matches experimental data, one can generate hypotheses about other states the system might enter and perform experiments to look for those states [61]. Additionally, estimating parameter values by fitting them to data [40] employs the reverse map, in that the modeler seeks the point or region in parameter space that maps to the point in data space closest to the measured data. These maps are commonly constructed and analyzed using a cost function.

11.1.2 Cost Functions

In the context of fitting models to data, the cost function measures the difference between the model output for a given set of parameters and the data that is being fit. As such, the cost function acts as a map between parameter space and data space, structuring the parameter space in such a way that moves in parameter space correspond to changes in model fit. This structure is often called the cost landscape, and we will use these concepts, of maps and landscapes, interchangeably here. The most common cost function used in data fitting, and the one we will focus on in this chapter, is the least-squares cost function [62]. (For more on cost functions, see Chap. 7 in this volume [60]). Given a model $y(\theta, t)$ with parameter vector θ we define the least-squares cost function as

$$\begin{aligned} C(\theta) &\equiv \frac{1}{2} \sum_s \sum_c \sum_{T_c} \left[\frac{y_{s,c}(\theta, t) - Y_{s,c}(t)}{\sigma_{s,c,t}} \right]^2 \\ &= \frac{1}{2} \sum_s \sum_c \sum_{T_c} r_{s,c,t}^2(\theta) = \frac{1}{2} \sum_{m=1}^{M} r_m^2(\theta) \,, \end{aligned} \tag{11.1}$$

which is half the squared difference over M data points collected for species s under experimental conditions c at time points T_c, or the sum of squares of the M normalized residuals r between model and data. Measurement uncertainty for each data point is denoted $\sigma_{s,c,t}$. The sum over T_c can be replaced with an integral when fitting continuous data. The best-fit parameter vector θ^* occurs at the global minimum of C.

Most commonly, the model residuals are assumed to be independent and normally distributed. The probability density that the model will produce the data given parameters θ is then

$$\mathrm{P}(\mathrm{Data}|\theta) = \prod_s \prod_c \prod_{T_c} \frac{1}{\sqrt{2\pi}\sigma_{s,c,t}} \exp\left[-\frac{1}{2} \left(\frac{y_{s,c}(\theta, t) - Y_{s,c}(t)}{\sigma_{s,c,t}} \right)^2 \right] . \tag{11.2}$$

In statistics, this probability density is called the likelihood of θ [12]. Taking the negative logarithm of the likelihood function yields the least-squares cost function $C(\theta)$ (Eq. 11.1). Thus, minimizing the cost function to find the best-fit parameters θ^* is equivalent to maximizing the likelihood, so the best-fit parameter vector inherits the statistical properties of the maximum likelihood estimator [26]. This statistical connection, arising from the assumption of normally-distributed residuals, makes the sum-of-squares cost function $C(\theta)$ particularly useful in describing the structure of parameter space. Other assumptions about the distribution of residuals are, however, possible and imply different cost functions. Note that much of what we discuss in this chapter has only been shown for sums-of-squares cost functions.

11.2 Multivariate Sensitivity Analysis

We use sensitivity analysis to explore parameter space, observing how model output changes as model parameters vary. In systems biology, sensitivity analysis is commonly used to quantify the uncertainty associated with both best-fit parameter values, and new model predictions generated using those parameter values. However, methods used in sensitivity analysis also provide additional useful information about parameter space. In this chapter, we are particularly interested in the correlation structure of parameter space, i.e. the relationships among combinations of parameters. Ultimately, we will define sloppiness in terms of these correlations. In this section, we describe one local and three global methods of multivariate sensitivity analysis. We then use an example model to illustrate how sensitivity analysis reveals parameter correlations and how those correlations affect the precision of parameter inference.

11.2.1 Local (The Hessian)

The sensitivity of the fit of the model $y(\theta, t)$ to a given data set is determined by how quickly the cost function $C(\theta)$ increases away from the best-fit parameters θ^*. A useful local perspective on the cost landscape is given by a quadratic expansion of the cost function:

$$C(\theta) \approx C(\theta^*) + \nabla C(\theta^*)(\log\theta^* - \log\theta) + \frac{1}{2}(\log\theta^* - \log\theta)^{\intercal} H (\log\theta^* - \log\theta) \,. \tag{11.3}$$

The gradient $\nabla C(\theta^*)$ of the cost function is, by definition, zero at the best-fit parameters θ^*. The $N \times N$ Hessian matrix H is defined as

$$H_{i,j} \equiv \left. \frac{\partial^2 C}{\partial \log\theta_i \, \partial \log\theta_j} \right|_{\theta=\theta^*} \,. \tag{11.4}$$

Because it involves second derivatives, the Hessian can be difficult to calculate. If the residuals r are small at the best-fit (i.e., the fit to the data is good) then the Hessian

can be well-approximated by $H \approx J^{\mathsf{T}} J$. The Jacobian matrix J is the $M \times N$ matrix of partial derivatives of the residuals with respect to the parameters:

$$J_{m,n} = \frac{\partial r_m}{\partial \log \theta_n} \,. \tag{11.5}$$

The first-derivatives in the Jacobian can be evaluated by finite differences or, for ordinary-differential equation (ODE) models, by integrating sensitivity equations.

The Hessian describes the quadratic behavior of the cost function C near the point θ^*, so analyzing the Hessian corresponds to approximating the level curves of C as N-dimensional ellipsoids in parameter space (Fig. 11.1). The Hessian matrix is positive

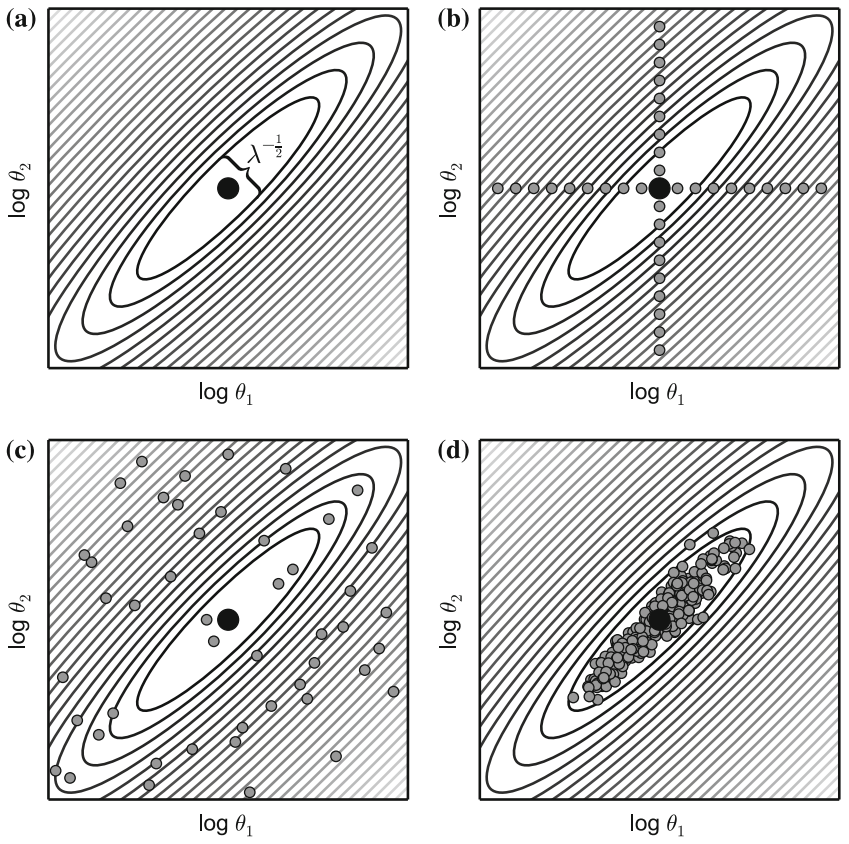

Fig. 11.1 Local and global sensitivity analysis. **a** Local analysis around the best fit point (*black dot*). Ellipses are curves of constant cost calculated from the Hessian matrix. Sensitivity in each direction is proportional to the widths of the curves. **b** Scanning parameters along each axis around the best fit point. **c** Latin Hypersquare scan with uniform priors. Every equal probability bin of each parameter distribution is sampled exactly once. **d** Bayesian parameter vector ensemble for this quadratic model.

definite and symmetric, so it has real eigenvalues λ and eigenvectors v corresponding to the principle axes of those ellipsoids. Additionally, the relationship between the negative log-likelihood $-\log P(\text{Data}|\theta)$ and $C(\theta)$ (Sect. 11.1.2) implies that inverting the Hessian in Eq. 11.4 gives an asymptotic approximation of the covariance matrix of the parameters [22]. Thus, the covariance matrix has eigenvectors v and eigenvalues λ^{-1}, and the widths of the principle axes of the ellipsoids are proportional to $\lambda^{-\frac{1}{2}}$ (Fig. 11.1a).

11.2.2 Global

Local sensitivity analysis accurately measures model sensitivity when parameter space is linear and smooth, so the cost minimum is well defined, but these conditions are not guaranteed to hold in systems biology models. In particular, the relationship between parameters may be nonlinear near the best fit, so the quadratic map imposed by the Hessian may be a poor approximation to the actual cost surface. Figure 11.2b shows such a cost surface, in which strong nonlinearities cause the quadratic approximation to overestimate the variability of the parameters. Moreover, some models may have rough parameter spaces with multiple minima of similar model behavior separated by ridges [25]. In such a landscape, local sensitivity analysis can be misleading, because steep curvatures near a local minimum may obscure the true shape of the parameter space. Global methods of sensitivity analysis address these problems by sampling parameter space in a finite neighborhood around the best fit. Broadly, such methods fall into two categories, scanning methods and Bayesian methods.

Scanning methods sample parameters without regard to the data and look for correlations between locations in parameter space and the model behavior or value of the cost function at those locations. Bayesian methods sample from the posterior distribution of the parameters given the data and use those samples to make inferences about the sensitivity of the model. The challenge in both cases is to generate a sufficiently dense sample of the parameter space that valid inferences can be made about the sensitivity of the model to parameter changes. Here we describe two parameter scanning and two Bayesian methods that are frequently used for multivariate sensitivity analysis in biological modeling.

11.2.2.1 Parameter Scanning

One way to generate parameter sets is to simply scan the parameter space, varying one parameter at a time in small intervals over a range of values (Fig. 11.1b). Only in the unusual case where the principal axes of the model sensitivity line up with the parameter axes will this method provide an accurate measure of the uncertainty in parameter estimates, because it ignores correlations between parameters. Scanning combinations of parameters to capture those correlations is, however, often

prohibitive for large models, because the number of points needed to define a grid in N dimensions grows exponentially with N.

Latin Hypercube Sampling (LHS), a generalization of the Latin Square experimental design, is a method for sampling parameter space in such a way that correlations between parameters can be uncovered. Marino et al. [50] describe an application of LHS in the context of sensitivity analysis, in which each parameter is assigned a probability distribution. These probability distributions incorporate prior information about the range of values a parameter can take, are often normal or uniform, and need not be the same for every parameter. Each of the probability distributions is divided into B equal probability bins, and B parameter vectors are generated by randomly sampling one bin from each parameter distribution without replacement, keeping track of which bin each value came from. The result is a group of parameter vectors such that each value for a given parameter was drawn from a different part of its distribution (Fig. 11.1c). The cost function is evaluated for each of these parameter sets and the correlation between costs and bins describes the sensitivity of the model. Computing partial correlations among parameter combinations reveals the correlation structure of the parameter space. While LHS is computationally efficient due to the Latin Square randomization, its use in analyzing biological models requires special care because nonlinearities in parameter space can render correlation analysis inaccurate [50]. For other approaches to parameter scanning, see Chap. 13 in this volume [77].

11.2.2.2 Bayesian Ensembles

Parameter scanning methods sample parameter vectors without regard to the data. Thus, the resulting sample may contain many vectors that poorly fit the data and add little to our understanding of the relevant distribution of parameters. Bayesian approaches maximize information about the distribution of parameters around the best fit by sampling densely in areas corresponding to good fits and sparsely elsewhere. Bayesian Markov-chain Monte Carlo (MCMC) walks through parameter space have been widely used in systems biology to construct ensembles of parameter sets [9, 13, 14, 24, 27, 79], and flexible approximate Bayesian methods have recently been developed. (For more on Bayesian approaches to sampling parameters and choosing among models, see Chaps. 9 and 10 in this volume [38, 66].)

The goal of Bayesian MCMC is to sample from the posterior distribution $\mathrm{P}(\theta|\mathrm{Data})$ of parameter sets given the observed data. From Bayes' rule:

$$\mathrm{P}(\theta|\mathrm{Data}) = \frac{\mathrm{P}(\mathrm{Data}|\theta)\,\mathrm{P}(\theta)}{\mathrm{P}(\mathrm{Data})}\,, \tag{11.6}$$

where $\mathrm{P}(\mathrm{Data}|\theta)$ is the likelihood defined in Eq. 11.2, $\mathrm{P}(\theta)$ is the prior probability of the parameters, and $\mathrm{P}(\mathrm{Data})$ is the evidence for the data. $\mathrm{P}(\mathrm{Data})$ is often

difficult to calculate but is in many cases an unimportant normalization, leading to the proportionality:

$$\mathrm{P}(\theta|\mathrm{Data}) \propto \mathrm{P}(\theta)\,\mathrm{P}(\mathrm{Data}|\theta)\ . \tag{11.7}$$

This proportionality allows a relative posterior probability to be calculated for any parameter set in terms of the likelihood and the prior. As we saw in Sect. 11.1.2, the likelihood can itself often be calculated in terms of the least squares cost function. The prior distribution reflects pre-existing knowledge of the distribution of parameter values, often from other experiments or analogy with similar molecular parameters. Early work focused on uniform priors [9, 13, 14], while more recent work employs log-normal [24, 27, 32] or gamma [79] priors.

The Markov chain is usually started at the best-fit parameter set and allowed to walk through parameter space sampling the posterior distribution of θ. At each step of the walk the Metropolis-Hastings criterion [17] is applied, such that at the jth step a new random vector θ_{test} is generated and

$$\theta_{j+1} = \begin{cases} \theta_{\mathrm{test}}\ , & \text{with probability } \alpha \\ \theta_j\ , & \text{with probability } 1-\alpha\ , \end{cases} \tag{11.8}$$

where

$$\alpha = \min\left\{1\ , \frac{\mathrm{P}(\theta_{\mathrm{test}}|\mathrm{Data})}{\mathrm{P}(\theta_j|\mathrm{Data})}\right\}\ . \tag{11.9}$$

The walk thus always accepts moves to parameter sets with higher posterior probability and sometimes accepts moves to parameter sets with lower posterior probability. This random walk generates an ensemble of parameter vectors that converges to the posterior distribution [17]. The marginal distributions for each parameter measure the sensitivity of the fit to changes in that parameter, integrating over changes in the other parameters, and provide confidence intervals for the best-fit value. The covariance matrix of the ensemble describes the correlation structure of the cost landscape.

11.2.2.3 Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) allows sampling of approximate posterior parameter distributions when the likelihood function is analytically or computationally intractable. Both Markov-chain [11, 51] and Sequential Monte Carlo (SMC) [64, 67] methods exist. The SMC method uses sequential importance sampling [20] to shorten chain length by preventing the algorithm from getting stuck in areas of low probability [67]. ABC has the advantage that it can be used to sample the parameter space of stochastic models, in addition to deterministic models [67].

The main difference between ABC and likelihood-based MCMC is that at each step, rather than evaluating the likelihood of θ_{test} the algorithm instead generates a new simulated data set $Y'_{s,c}(t)$ from $y_{s,c}(\theta_{\mathrm{test}}, t)$ and computes a distance metric

$\rho(Y_{s,c}(t)\ ,\ Y'_{s,c}(t))$. Possible distance metrics include the euclidian distance, squared distance, and total absolute deviation, among many others. At the jth step

$$\theta_{j+1} = \begin{cases} \theta_{\text{test}}\ , & \text{with probability } \alpha\ , \\ \theta_j\ , & \text{with probability } 1-\alpha\ , \end{cases} \tag{11.10}$$

where

$$\alpha = \begin{cases} 0\ , & \rho(Y_{s,c}(t)\ ,\ Y'_{s,c}(t)) > \varepsilon \\ \min\left\{1\ ,\ \frac{\mathrm{P}(\theta_{\text{test}})}{\mathrm{P}(\theta_j)}\right\}, & \rho(Y_{s,c}(t)\ ,\ Y'_{s,c}(t)) \le \varepsilon\ , \end{cases} \tag{11.11}$$

and ε is some small number chosen to bound the acceptable distance between simulated and real data. This chain generates a collection of parameter vectors drawn from the joint distribution $P(\theta|\rho(Y_{s,c}(t)\ ,\ Y'_{s,c}(t)) \le \varepsilon)$, which can be used in the same way as the joint distribution generated from Bayesian MCMC. In the case of a deterministic ODE model, choosing ρ to be the squared distance and performing an ABC analysis with decreasing ε is equivalent to the maximum-likelihood method used in Bayesian MCMC with a least-squares cost function [67].

11.2.3 Example: Robertson Model

To illustrate the concepts discussed in this section and throughout the chapter, we follow Eydgahi et al. [24] and consider a set of mass-action reactions among chemical species A, B, and C originally formulated by Robertson [58]:

$$\mathrm{A} \xrightarrow{\mathrm{k}_1} \mathrm{B}\ , \qquad 2\mathrm{B} \xrightarrow{\mathrm{k}_2} \mathrm{B}+\mathrm{C}\ , \qquad \mathrm{B}+\mathrm{C} \xrightarrow{\mathrm{k}_3} \mathrm{A}+\mathrm{C}\ . \tag{11.12}$$

These reactions yield the nonlinear system of ODEs:

$$\frac{d\ [\mathrm{A}]}{dt} = \mathrm{k}_3 \cdot [\mathrm{B}] \cdot [\mathrm{C}] - \mathrm{k}_1 \cdot [\mathrm{A}]\ , \tag{11.13}$$

$$\frac{d\ [\mathrm{B}]}{dt} = \mathrm{k}_1 \cdot [\mathrm{A}] - \mathrm{k}_2 \cdot [\mathrm{B}]^2 - \mathrm{k}_3 \cdot [\mathrm{B}] \cdot [\mathrm{C}]\ , \tag{11.14}$$

$$\frac{d\ [\mathrm{C}]}{dt} = \mathrm{k}_2 \cdot [\mathrm{B}]^2\ . \tag{11.15}$$

As initial conditions, we took

$$[\mathrm{A}_0] = 1\ , \qquad [\mathrm{B}_0] = [\mathrm{C}_0] = 0\ . \tag{11.16}$$

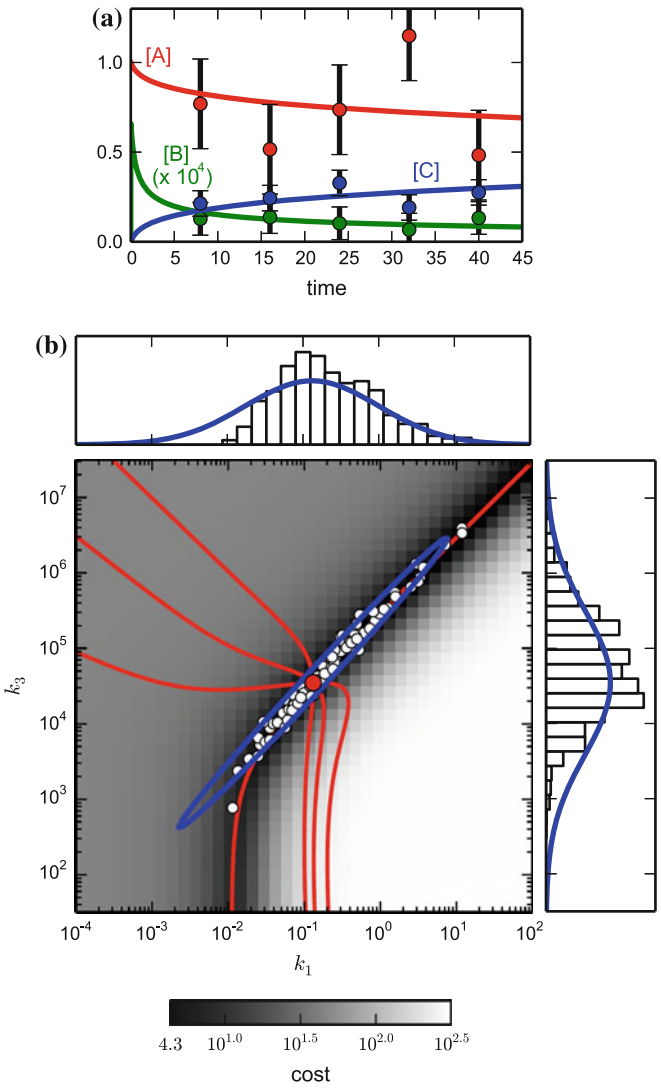

Fig. 11.2 Data fitting for the Robertson model. **a** Simulated data and best-fit trajectories. *Error bars* correspond to one standard deviation. **b** Corresponding cost landscape showing best-fit parameters (*red point*), and the confidence interval from the $J^\intercal J$ approximation to the Hessian (*blue ellipse*). One hundred samples from a Bayesian MCMC ensemble (*white dots*), and geodesic curves starting at the best-fit (*red lines*) are also shown. *Top* and *right panels* show marginal distributions of k_1 and k_3, respectively, inferred from the Hessian approximation (*blue curve*) and the Bayesian ensemble (*white histogram*)

We generated synthetic data points for [A], [B], and [C] by sampling every 8 time units from the model with initial parameters

$$\mathrm{k}_1 = 0.04\ , \qquad \mathrm{k}_2 = 3 \times 10^7\ , \qquad \mathrm{k}_3 = 10{,}000\ , \tag{11.17}$$

and adding normally-distributed noise to each data point with standard deviation equal to 25 % of the maximum value of the corresponding variable. We then fixed k_2 and used least-squares optimization to fit the synthetic data, estimating k_1 and k_3 (Fig. 11.2a).

We conducted sensitivity analysis using both the Hessian matrix (local) and an ensemble of parameter sets sampled by Bayesian MCMC (global), using SloppyCell [55]. In both cases, we added log-normal priors that restricted k_1 and k_3 to remain within three orders of magnitude of the initial values (Eq. 11.17), with 95 % confidence. The quadratic approximation (Fig. 11.2b, blue ellipse) mimics the shape of the cost landscape quite well in the vicinity of the best fit, but it overestimates the variability in these two parameters, due to the strong nonlinearity in their relationship. The ensemble (Fig. 11.2b, white dots), on the other hand, captures the true posterior distribution of parameters.

This example illustrates some of the difficulties encountered when using parameter scanning methods for sensitivity analysis. A simple scan along each parameter axis at the best fit value will dramatically underestimate the variability in the parameter estimates, and the nonlinearities in the landscape will render the correlation analysis used in LHS inaccurate.

11.3 Sloppiness

The topography of the cost landscape plays a critical role in modeling. For example, the cost landscape of the Robertson model (Fig. 11.2b) is highly anisotropic, as indicated by the eigenvalue spectrum in Fig. 11.3a(i). Near the best-fit, the parameter combination k_1/k_3 is tightly constrained (corresponding to the large eigenvalue). By contrast, the parameter combination $k_1 \times k_3$ is loosely constrained (corresponding to the small eigenvalue), so inferred values of k_1 and k_3 have large uncertainty. In 2003, Brown and Sethna noted similar behavior in a much more complex signaling model [13, 14], leading to the discovery of sloppiness.

Brown and Sethna used a system of 15 nonlinear differential equations, involving 48 rate constant parameters, to model the activation of ERK1/2 by epidermal growth factor (EGF) and neuronal growth factor (NGF) in PC12 cells [14]. They fit their model to 68 data points tracking the time-course of activation of several proteins in the system. When they analyzed the corresponding Hessian matrix, they found a surprising regularity in the eigenvalue spectrum (Fig. 11.3a(iv)). The eigenvalues spanned many orders of magnitude roughly evenly, a phenomenon they deemed *sloppiness*. The large eigenvalues and corresponding eigenvectors defined *stiff* combinations of parameters that were tightly constrained, whereas the small eigenvalues

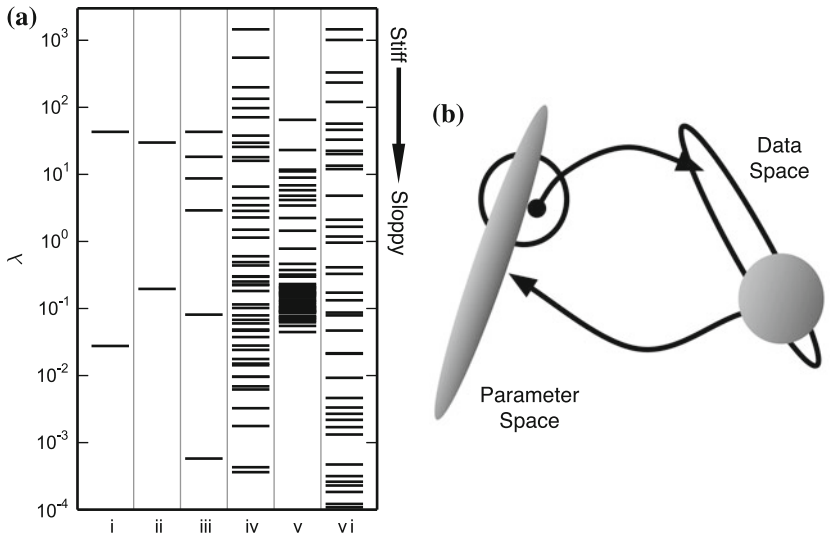

Fig. 11.3 Sloppy parameter spaces, eigenvalues, and data space. **a** Eigenvalue spectra for several sloppy models, illustrating different approaches to parameter and data space. *i* $J^\intercal J$ eigenvalues of the Robertson model and data described in Sect. 11.2.3. *ii* Principle components analysis (PCA) eigenvalues of a Bayesian parameter ensemble for the Robertson model (inverted for comparison with column *i*). *iii* PCA eigenvalues of the model manifold for the Robertson model (inverted and rescaled so the largest eigenvalue matches the largest eigenvalue in column *i*). *iv* $J^\intercal J$ eigenvalues of the Brown and Sethna model for differentiation in PC12 cells [14] that was fit to 68 data points. *v* PCA eigenvalues of a Bayesian parameter ensemble for the PC12 model [32], generated with log-normal priors similar to Robertson model (inverted for comparison with column *iv*). *vi* $J^\intercal J$ eigenvalues for the PC12 model fit to continuous data on all molecular species in the model [32] (rescaled so the largest eigenvalue matches the largest eigenvalue in column *iv*). **b** The sloppy mapping between parameter and data space implies that spherical regions of parameter space map to distorted sloppy regions in model space, and vice-versa [18]

and corresponding eigenvectors defined *sloppy* combinations that were loosely constrained.[1] Moreover, a similar pattern of eigenvalues was found even when considering large amounts of perfectly-fit synthetic data on every species in the model (Fig. 11.3a(vi)), suggesting that sloppiness was a property of the model itself, not the particular data set. As illustrated in Fig. 11.3b, this sloppiness implies that large volumes of parameter space can map to a small volumes in data space, and vice-versa.

The importance of sloppiness to systems biology became more apparent in 2007, when Gutenkunst et al. found sloppiness in a diverse set of sixteen other systems biology models [32]. In a systematic survey of the BioModels database [48], Erguler and Stumpf later found sloppiness in 180 systems biology models [23].

[1] Concurrent with Brown and Sethna's work, Rand et al. independently noted an exponential spacing of eigenvalues for several circadian clock models, although Rand et al. focused their analysis on the stiffest few eigenvalues [57].

Although similar in spirit, sloppiness differs from conventional conceptions of robustness [46]. Typically, when a biological system is deemed *robust* it means that a particular qualitative behavior is insensitive to a particular perturbation. That perturbation may be a change in parameter values [78], temperature [56], or structure of the system. Sloppiness, on the other hand, focuses on the quantitative behavior of the model and its sensitivity to changes in combinations of parameters. A system may be sloppy, but not be robust to changes in individual parameters. For example, at the best-fit set of parameters, the Robertson model is robust to changes in k_1 or k_3 that leave the stiff parameter combination k_1/k_3 unchanged, but it is fragile to changes in either parameter individually.

Brown and Sethna's discovery of sloppiness spawned a large body of literature exploring its theoretical basis. Early work on the origins of sloppiness focused on symmetries between parameter effects [76], but recent connections with information geometry and interpolation theory have revealed a more general origin (Sect. 11.4.1). Although it has been best-studied in the context of systems biology, sloppiness also appears in non-biological models [30, 76], including classic statistical problems such as fitting a sum of exponentials or polynomials to data [76]. In classic physics models for magnetism and diffusion, sloppiness emerges when observations are restricted to large length scales, so microscopic details of the system cease to matter [49]. A similar phenomenon may be occurring in systems biology, where most experiments probe the collective behavior of many interacting reactions. The ubiquity of sloppiness also suggests that it may have implications for biological evolution [18].

In the remainder of this chapter, we focus on the practical implications of sloppiness for modeling biological systems, through building predictive models, designing experiments, and developing numerical methods.

11.3.1 Local and Global Perspectives

We have defined sloppiness in terms of the distribution of the eigenvalues of the Hessian matrix. For nonlinear models, however, the Hessian depends on where in parameter space it is evaluated, as exemplified by the curved basin of the Robertson model (Fig. 11.2b). In the Robertson model and in the Brown and Sethna PC12 model [14], Hessian matrices calculated using multiple parameter sets from the MCMC posterior distribution are all sloppy, with similar eigenvalue spectra but differing eigenvectors [30]. This suggests that the curved basins are everywhere locally sloppy, but a more global perspective can be obtained from Principal Component Analysis (PCA) of the MCMC parameter set ensemble.

PCA is the eigen-decomposition of the covariance matrix of a set of points in space (here we focus on points in parameter space), and it has a broad range of applications in statistics [1, 37]. (For more on PCA and other statistical models in systems biology, see Chap. 6 in this volume [63].) PCA is defined such that the first principal component is the eigenvector with the largest eigenvalue, and it points in the direction that accounts for the largest amount of variance in the positions of the

points. The eigenvector with the second-largest eigenvalue points in the direction that accounts for the second-largest amount of variance in positions of the points, orthogonal to the previous direction, and so on. The Hessian matrix and the covariance matrix share eigenvectors (Sect. 11.2.1), so we can think of performing PCA on an ensemble as the global analog to the local analysis of the Hessian matrix. The eigenvalues of the Hessian are inversely related to those of the principal components, and in Fig. 11.3a(ii, v) we take the inverse of the PCA eigenvalues for comparison.

Because the ensemble captures nonlinearities in the parameter space if they exist, we might wonder whether models with sloppy Hessian eigenspectra also have sloppy PCA eigenspectra. Figure 11.3a(i, ii) shows the eigenvalues for the Robertson example computed using the Hessian and by PCA, respectively. The similar spacing of eigenvalues shows that the aspect ratio of the level curves of the cost manifold are preserved, even as nonlinearities cause them to curve. Figure 11.3a(iv, v) show corresponding Hessian and PCA spectra for the PC12 model [13], where the ensemble in Fig. 11.3a(v) was generated with priors similar to those we used for our Robertson model. The spectrum is truncated from below by the prior,[2] and the largest eigenvalues are reduced due to nonlinearities in the parameter space that are better captured by the ensemble. Although quantitative differences are evident in the eigenspectra generated by the two methods, qualitatively they are both sloppy, spanning several orders of magnitude with eigenvalues that are evenly spaced in the logarithm. In addition to these empirical comparisons of Hessian and PCA eigenspectra, recent work in information geometry (Sect. 11.4.1) also suggests that sloppiness is a global property.

11.3.2 Predictive Modeling from Sloppy Systems

Because sloppiness appears universal in systems biology models, attempting to fit individual parameters in such models is difficult and often uninformative. Even with extensive time-series data, inferred values for individual parameters are often imprecise, because the model is insensitive to changes to most parameter combinations. The common practice of reporting only the means and confidence intervals on individual parameters should thus be avoided. On the other hand, because the model is quite sensitive to changes in a few parameter combinations, with precise measurement of time-series data it is often possible to tightly constrain model predictions, despite large individual parameter uncertainties [32]. Moreover, in most cases precisely modeling and predicting system behavior is more compelling than precisely inferring individual parameters.

[2] A log-normal prior that bounds a parameter θ to be, with $\approx$95 % confidence, between θ_0/F and $\theta_0 \times F$ corresponds to an additional residual in the cost function (Eq. 11.1) of $r = (\log\theta - \log\theta_0)/\log\sqrt{F}$. Such a residual adds $1/(\log\sqrt{F})^2$ to the diagonal elements of the Hessian matrix, bounding the eigenvalues from below. In our case, $F = 10^3$, so the eigenvalues must be greater than $\approx$0.084.

For the Robertson model (Sect. 11.2.3 and Fig. 11.2), the individual parameters k_1 and k_3 are only loosely constrained by time-course data, so our inferred values for these parameters, whether using the Hessian approximation or Bayesian MCMC sampling, span many orders of magnitude, as shown by the top and right panels in Fig. 11.2b. The cost landscape does show a distinct nonlinear canyon of parameter sets that fit the data well (Fig. 11.2b), but this canyon does not align with any individual parameter, so inferring individual parameters with high precision is difficult. On the other hand, combinations of parameters perpendicular to the canyon are tightly constrained.

Because predictions are often more important than individual parameter values, we tested the ability of our synthetic time-course data in the Robertson model to constrain a novel prediction. In particular, we added a new reaction to the model:

$$A + C \xrightarrow{k_4} B\,, \qquad k_4 = 1. \tag{11.18}$$

We then predicted the time course of $[C]$ in this four-reaction model by generating trajectories using the results of our data fit of the original three-reaction model (Eq. 11.12, Fig. 11.2). When we generated a set of predictions assuming that we knew k_1 and k_3 to high precision (95 % confidence interval of ± 50 %), the prediction for $[C]$ was tightly constrained (Fig. 11.4a, b). If we instead knew k_1 precisely, but k_3 imprecisely, the prediction of $[C]$ was uninformative (Fig. 11.4c, d), because the corresponding parameter ensemble includes parameter sets with high cost, rather than exploring only the canyon. When we approximated the stiff and sloppy directions using the Hessian, as in (Fig. 11.4e, f) and generated predictions from this set of parameter combinations, we recovered some constraint on the prediction uncertainties.

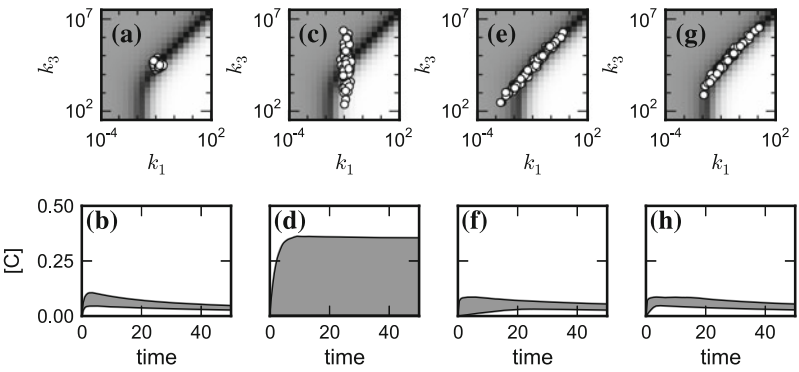

Fig. 11.4 Parameter inferences and prediction 95 % confidence intervals for the Robertson model (Sect. 11.2.3). **a**, **b** Assuming both parameters are measured to $\pm$ 50 % precision. **c**, **d** Assuming k_1 is measured to high precision, but k_3 must be guessed to low precision (95 % confidence interval spanning three orders of magnitude). **e**, **f** Evaluating the prediction using samples from the $J^\intercal J$ approximation to the Hessian matrix. **g**, **h** Evaluating the prediction using samples from the Bayesian MCMC ensemble illustrated in Fig. 11.2b

Taking points from the MCMC-generated ensemble constrained uncertainty even further (Fig. 11.4g, h).

In the Robertson model, both the Hessian and Bayesian-ensemble approaches constrained prediction uncertainty much better than a mixture of well- and poorly-determined rate constants. However, the Hessian did not perform as well as Bayesian MCMC, because the cost manifold is nonlinear; in other words, the canyon of well-fitting parameter sets is curved. The Hessian approximates the stiff and sloppy directions at the best-fit parameters, but away from the best-fit parameters the stiff and sloppy directions inferred from Hessian deviate from the true shape of the cost manifold, so points outside the canyon are sampled. Bayesian MCMC avoids this problem, as there is no assumption of linearity of the cost manifold, so the sampling follows the curve of parameter sets that both fit the data well and yield accurate predictions. The Hessian approximation works well in the Robertson model, but it may fail in more complex models with stronger nonlinearities in the cost landscape [24, 31].

11.3.3 Experimental Design

Motivated by the previous example of precise predictions from a sloppy model without precise parameters, we turn to the design of informative experiments. Experimental design is a large sub-field of systems biology, and many methods have been developed for designing experiments and models to extract optimal information about a quantity of interest [3, 15, 34, 44, 47]. In this section, we discuss several studies that directly address sloppiness. One identifies additional time-series data points that improve system behavior prediction [15], and the others identify experimental conditions that improve parameter inference [3, 34, 68].

11.3.3.1 Optimal Design for Prediction

As illustrated by the Robertson model in Fig. 11.4c, d, precisely measuring individual parameters in a complex model may not improve the predictive power of the model. To overcome this difficulty, Casey et al. developed an approach for designing experiments to improve the prediction of unmeasurable quantities and applied it to a model of epidermal growth factor receptor (EGFR) activation [15].

In the EGFR network, Casey et al. were interested in predicting the dynamics of the triple complex of Cool-1, Cdc42 and Cbl, each of which potentially disrupts receptor down-regulation. The triple complex was not directly measurable, so they relied on a complex systems biology model to predict its dynamics. Casey et al. fit their model to existing experimental data to obtain best-fit parameter values θ^* and an ensemble of parameter sets that fit the data well, but they found that the predicted trajectory for the triple complex had large uncertainty.

Given the large prediction uncertainty from the existing data, Casey et al. set out to design a new experiment to minimize the variance of the prediction. Doing so

required searching over the space of possible experiments and evaluating prediction variance many times. Bayesian sampling (Sect. 11.2.2.2) is the preferred way to estimate prediction variance, but it is computationally very expensive, so Casey et al. used an approximation to the variance of their prediction p:

$$\mathrm{Var}(p) \approx \left.\frac{\partial p}{\partial \theta}\right|_{\theta^*} (J^\intercal J)^{-1} \left.\frac{\partial p}{\partial \theta}\right|_{\theta^*} . \tag{11.19}$$

Here $(J^\intercal J)^{-1}$, the inverse Fisher Information Matrix (FIM), asymptotically approximates the covariance of the parameters, and $\partial p/\partial \theta$ is a linear approximation of the model response to changes in the parameters [15].

Casey et al. employed a sequential experimental design to minimize the prediction variance calculated via Eq. 11.19. They first searched over experimental conditions, measurable molecular species, and timepoints to find the single data point whose addition most greatly reduced the prediction variance. This was computationally feasible because adding a single data point to the collection of measurements is a rank-one update of the Fisher Information Matrix [15]. Assuming that single data point represented the optimal condition and species to measure, Casey et al. then optimized over possible combinations of measured timepoints to design a complete experiment.

Applying their computational analysis, Casey et al. carried out the experiment they had designed. Adding the new data points to their model, they built a new ensemble of parameter sets from which to make predictions. As desired, the new ensemble dramatically reduced uncertainty in the predicted dynamics of the triple complex. Uncertainties on individual parameter values, however, were not substantially smaller. The sloppiness of the EGFR model allowed Casey et al.'s experimental design to improve prediction precision without improving parameter precision, but experimental design can also improve parameter precision.

11.3.3.2 Optimal Design for Parameter Inference

Fitting time-course data typically poorly constrains individual parameter values in sloppy models, but careful experimental design can yield well-constrained parameters. In a recent manuscript, Tönsing et al. argue that sloppiness can be generically caused by autocorrelation and sparseness in the Jacobian matrix (Eq. 11.5) of parameter sensitivities for residuals between model and data [68]. Autocorrelation naturally arises in time-course measurements, and sparseness arises because different predictions may be sensitive to different parameters. Tönsing et al. further show using a model and in silico experiments from the DREAM 6 challenge [52] that careful experimental design can avoid autocorrelation and sparseness, minimizing sloppiness in the resulting parameter inferences.

In a more targeted study, Apgar et al. have shown [3] that carefully designed complementary experiments can in principle tightly constrain all parameter values

in the original Brown and Sethna sloppy model of EGF/NGF signaling in PC12 cells [13, 14]. Apgar et al. sought to design a set of experiments that would together yield uncertainties on all 48 model parameter values of less than 10 %, based on the Hessian approximation. To do so, they considered 164,500 potential experimental conditions, encompassing various levels of EGF and NGF stimulation and protein overexpression or knockdown. To avoid computationally challenging re-optimization of parameter values, they assumed that each experiment would yield data that exactly matched the model prediction. Their design processes employed a greedy algorithm that, at each step, chose the experiment that constrained the most parameters to within 10 % that were not constrained by any earlier experiment. Remarkably, they found that five carefully chosen experiments were enough to tightly constrain all parameters [3]. It was essential that the experiments be chosen in a complementary way; choosing random experiments or even the best individual experiments gave much poorer results. The computational experiments that Apgar et al. considered used continuously sampled species time courses, yielding effectively many more data points than typical experiments, which may account for much of the improvement in parameter constraint [16]. Such dense measurements are, however, becoming increasingly feasible, and even with fewer collected data points, tight constraint on all parameters in the model are still possible [33].

Recently, Hagen et al. have relaxed many of the simplifying assumptions made in Apgar et al.'s work [34]. Most importantly, they considered data at discrete points along the trajectory with some assumed experimental error instead of continuous measurements with zero error. As a result, they had to re-optimize parameters at each stage of the experimental design, so each experiment was chosen on the basis of a model with imperfect parameters. Nevertheless, they found that just six experiments were needed to constrain all parameters to within 10 % as assessed by the Hessian approximation, confirming the previous results.

The experiments designed by Tönsing et al. [68], Apgar et al. [3], and Hagen et al. [34] are complex, and to date they have not been carried out in the lab. This work, however, demonstrates the power of experimental design and offers hope that parameter values can indeed be precisely inferred even for sloppy models.

11.4 Information Geometry Perspective

To this point, we have reviewed work whose focus was analyzing the properties of parameter space. However, we have seen it is often beneficial to focus on the model predictions rather than the parameter values. Recent results focusing on data space rather than parameter space have proven beneficial for understanding the properties of models and for advancing numerical techniques for exploring them. This approach, usually known as information geometry since it combines information theory with differential geometry, is a natural mathematical language for exploring parameterized models. As we have seen, in essence a model is a mathematical mapping from parameters to predictions. This recognition leads to the interpretation of a model as

a manifold embedded in the space of data. The approach is very general, applicable to any parameterized statistical model (although we focus on least squares models in this review) and has many deep connections to statistics [2, 4, 7, 41, 54]. Because differential geometry is foreign to most biologists, much of the technical aspects and insights of information geometry are not immediately accessible to much of the systems biology community. In this section, we give a summary of recent results without assuming a prior understanding of differential geometry, illustrating the types of analyses that can be performed and providing references for further study.

To illustrate the approach for least-squares models, we return to the Robertson model introduced in Sect. 11.2.3. This model has two parameters and was fit to $M = 15$ data points. Any experimental realization of the data can be interpreted as a single point in $\mathbb{R}^M$. Likewise, for any value of the parameters, the model predictions are similarly a point in $\mathbb{R}^M$. As the two model parameters k_1 and k_3 are varied over their allowed ranges, the model sweeps out within the 15-dimensional data space a two-dimensional surface known as the model manifold and denoted by $\mathscr{M}$. In general, for a model of N parameters fit to M data points, the model manifold is the N-dimensional surface embedded in $\mathbb{R}^M$ constructed by varying the model parameters over their physically allowed values.

Figure 11.5d shows a three-dimensional projection of the high-dimensional data space for the Robertson model. The manifold was calculated using a grid of parameter values over the ranges shown in Fig. 11.2b. Model trajectories were sampled at the equally-spaced timepoints for which data was simulated in Fig. 11.2a. The axes of the visualization come from a principal component analysis (Fig. 11.5a–c) that was performed for the resulting set of model trajectories.

The Robertson model manifold illustrates several features of the information geometry perspective that make it a powerful tool for studying models that complements the approach of considering the cost surface in parameter space. First, there is no information loss in the model manifold, i.e. the manifold is mathematically equivalent to the model itself. In contrast the cost surface in parameter space condenses the M numbers making up the prediction and data vectors into a single number.

Second and relatedly, information geometry separates the model, i.e. the manifold embedded in data space, from the data to which it is being fit, i.e. a point in the data space (blue star in Fig. 11.5). This is a useful abstraction which allows one to study the properties of the model itself irrespective of what is experimentally observed. The cost surface in parameter space will vary with the observed data. The best-fit parameters correspond to the point on $\mathscr{M}$ nearest to the data (red point in Fig. 11.5).

Third, the set of points that constitute the model manifold are the same regardless of how the model is parameterized. That is, the geometric properties of the model are the same if a systems biology model is expressed in terms of reaction rates or time constants, in bare or log-parameters. In fact, it will be unchanged if the model is reparameterized in a complicated, highly nonlinear way. Because of this, the geometric perspective places the emphasis on model predictions rather than the parameters. The parameters are not ignored completely, but act as coordinates on the manifold, i.e. labels for specific predictions. The grid lines in Fig. 11.5d–g correspond to a square grid in log-parameter space. In general, differential-geometric objects are

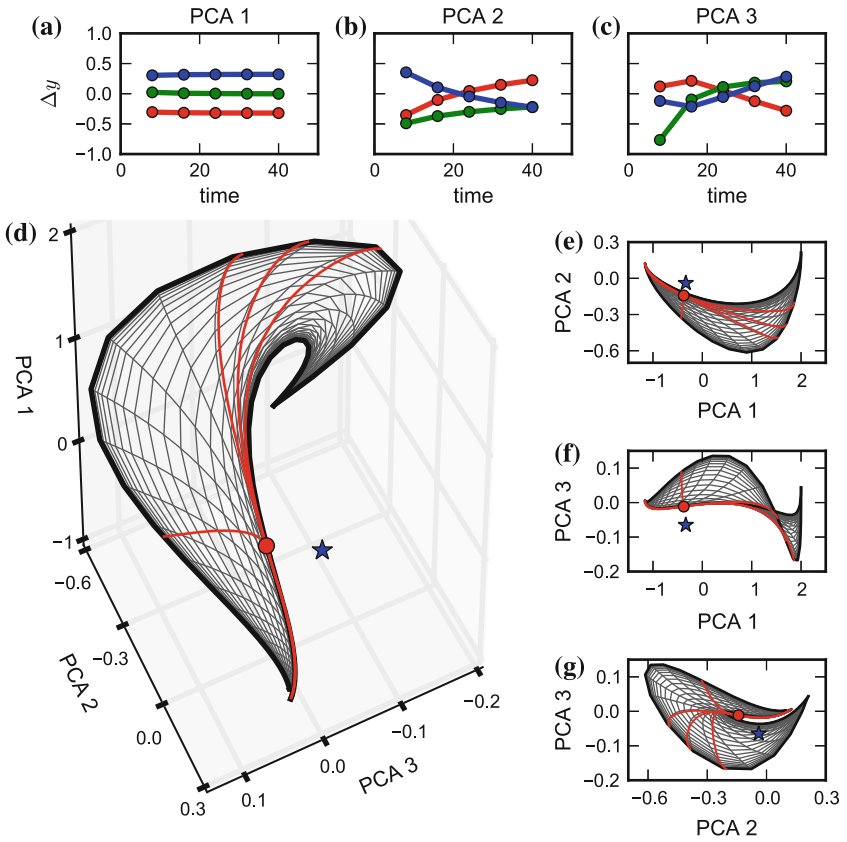

Fig. 11.5 Model manifold evaluated over the parameter space shown in Fig. 11.2b for the Robertson model (Sect. 11.2.3). **a–c** First three principle components of model prediction variation over the manifold. *Colors* identify species as in Fig. 11.2a. The first principle component, for example, represents an increase in [C] and a decrease in [A] that is roughly constant over the sampled timepoints. **d** Projection of the model manifold onto the first three principle components. *Blue star* shows the data, and *red dot* shows the best-fit trajectory. *Red lines* correspond to the geodesics in Fig. 11.2b. **e–g** Projections of the model manifold onto pairs of the first three principle components, as in (**d**)

constructed in terms of derivatives of the predictions with respect to the parameters in such a way that their relevant properties are the same for all possible parameterizations. Indeed, differential geometry is generally concerned with the properties of the manifold that are invariant under such reparameterizations.

A fourth and final point is that the language of differential geometry naturally accommodates the potentially large dimensionality of both the data space and the model manifold. Visualizations of both the cost surface and the model manifold are limited to only a few dimensions, but the geometric properties of high-dimensional spaces can be very different from those of the three-dimensional world in which

our visualizations live. For example, the specific properties of sloppiness are closely tied to the properties of high-dimensional manifolds. The mathematical formalism of differential geometry, however, has no such limitation and provides a framework within which the space can be systematically studied.

A particularly useful differential-geometric tool is a geodesic. A geodesic can be understood qualitatively as the generalization of straight lines to curved surfaces, i.e. the path connecting two points such that its image in data space is as close to a straight line as possible. A perfectly straight line is generally impossible, since the surface is typically curved for nonlinear models. The geodesic is constructed numerically as the solution of a nonlinear differential equation involving first and second derivatives of the model predictions with respect to the parameters. We refer the reader to any introductory text in differential geometry for more details [39, 65]. Several geodesic paths on the model manifold are shown in Fig. 11.5d–g, and the corresponding paths in parameter space are shown in Fig. 11.2b.

11.4.1 Models as Interpolation: Geometric Sloppiness

Geometry helps us understand the phenomenon of sloppiness. The observed universality of sloppiness across a wide range of models is perplexing; its ubiquity suggests some deep connecting principle [32, 76]. However, two other observations suggest otherwise. First, the hierarchy of Hessian eigenvalues can be transformed into any set of positive values by reparameterizing the model. Although such parameterizations might be unnatural from a human perspective, they are mathematically acceptable. Perhaps sloppiness is a reflection of how we humans choose to parameterize models. Is the human-preferred parameterization somehow perverse from a mathematical perspective? Second, sloppiness can be reduced by an appropriate choice of experiments [3, 34, 73]. Perhaps sloppiness is furthermore a reflection on what we choose to measure and not intrinsic to the system itself. Geometric arguments reconcile these apparently contradictory observations.

The key observation is that model manifolds are typically bounded, as is our example manifold in Fig. 11.5d. Considering the cost landscape in Fig. 11.2b, notice that away from the best fit, the cost surface plateaus, apparently approaching a limiting value. In fact, the parameters can be taken to zero and infinity without the cost becoming infinite, implying that the model manifold must be bounded. For any specific model this can be checked numerically using differential geometry. By numerically constructing geodesics, the manifold can be systematically explored to identify boundaries in any direction. Furthermore, by calculating the length of these geodesic paths in data space, one can measure the extent of the manifold in any given direction and calculate its aspect ratio. In this way it was found that typical sloppy model manifolds are not only bounded, but exhibit a hierarchy of widths analogous to the hierarchy of Hessian eigenvalues [71, 72]. For example, PCA analysis of the data space points used to construct the model manifold for our Robertson

model (Fig. 11.5d) reveals an exponential hierarchy of eigenvalues (Fig. 11.3b(iii)).[3] This result is similar to the previously noted observation that sloppiness is generally reflected as a global property as measured by PCA analysis of a Bayesian ensemble (see Fig. 11.3b(ii, v)). However, in the current context the anisotropy is a reflection of an intrinsic property of the entire range of model predictions, rather than the ensemble of parameter values consistent with an instance of data.

This empirical observation of a hierarchy of widths in data space can be explained by applying interpolation theory [71]. Orthogonal geodesic paths identify cross sections of the model manifold. Geometrically, the cross section is formed by fixing the model output along a few axes (the directions orthogonal to the geodesic) and varying the output along others. Now consider a time series of model predictions for which a handful of time points have been fixed and the intermediate time points are allowed to vary. Although not fixed themselves, the values of intermediate time points can often be approximated by interpolating the values of those that are fixed. Therefore, the corresponding cross sections of the model manifold must be bounded by the accuracy of the interpolations.

For one-dimensional time series, the above argument can be made formal using theorems from interpolation theory [71]. In this case, cross sections become more narrow by roughly a constant factor for each additional fixed output, provided that the number of effective degrees of freedom probed by the model predictions is much less than the number of parameters. Qualitatively, this is understood to mean that the "complexity" of the data to be explained is much less than that of the model. In a sense, the model is over parameterized. However, it is often unclear how to remove the unnecessary parameters, because the stiff and sloppy parameter directions are almost always combinations of the bare parameters [32].

This argument suggests that models can be understood as generalized interpolation schemes and explains a number of observations. First, it explains why accurate predictions can be made by sloppy models when parameters are largely unconstrained; the predictions are interpolating from the existing data. It also explains why sloppiness disappears when complementary experiments are chosen; the number of effective degrees of freedom probed by the model becomes comparable to the number of parameters. In this case, the model needs all of the parameters to explain the data, resulting in tight bounds on their estimates.

The connection between manifold widths and the Hessian eigenvalues can be understood by dimensional analysis. The square roots of Hessian eigenvalues have units of data space distance per parameter space distance. If the parameters are expressed in the natural units of the problem, for example by using log-parameters, then we expect the eigenvalues to reflect the natural length scales of the manifold, i.e. the manifold widths as observed. What is interesting about this argument is the implication that the "natural" parameterization preferred by the human modeler is

[3] Note that this PCA was done on points sampled uniformly in parameter space, not data space, and this non-uniformity may bias the resulting eigenvalue summary of the model manifold. We expect, however, that this approach provides a good first approximation to the hierarchy of manifold widths that would be found by geodesics.

actually not perverse after all. Indeed, encoded in this natural parameterization is useful information: the length of the model manifold along several principal axes.

11.4.2 Applications to Numerical Methods

One of the most useful applications of the differential geometry approach involves the development and improvement of numerical methods for exploring parameter space. We now discuss two such improvements: the geodesic acceleration correction to the Levenberg-Marquardt algorithm for least squares data fitting [71, 72] and the Riemannian manifold sampling methods for Markov-Chain Monte Carlo (MCMC) [28].

11.4.2.1 Data Fitting

Fitting multi-parameter models to data via least squares can be notoriously difficult. One reason for this is that as algorithms approach the best fit they become agonizingly slow. This is because the cost surface in the vicinity of the best fit consists of a long, narrow canyon, as illustrated for the Robertson model in Fig. 11.2b. The algorithm must navigate this canyon en route to the best fit. The greater the aspect ratio of the canyon, the smaller the steps the algorithm must take. For many sloppy problems, as we have seen, it is not unusual for the canyon to have aspect ratios of 1000:1 or more, leading to very slow convergence rates.

A second reason that data fitting is difficult is that it is hard for the algorithm to even find the canyon to begin with. Observe in Fig. 11.2b how the cost surface plateaus away from the canyon. Because the cost surface is so flat, it is difficult for the algorithm to know in which direction to move. One typically finds that the results of a fitting algorithm are inconsistent, "converging" to wildly different parameter values depending on the starting point. This is typically attributed to multi-modality, or a rough cost surface with many local minima [25, 53, 59]. Closer inspection and the understanding of bounded model manifolds refines this picture in a very useful way. Specifically, the parameter values that result from failed runs of search algorithms typically contain parameters approaching their physical limits, e.g. zero or infinity. Geometrically, these points correspond to boundaries of the model manifold. The failure of search algorithms is due to them getting stuck in the boundaries of the model manifold en route to the best fit, i.e. being lost on the plateau. It was found that by adding weak, regularizing "prior" terms to the cost function that kept the algorithm away from the limiting parameter value, algorithms were much more successful at finding best fits [72]. These terms should be chosen in a way to force the algorithm to search in the region of parameter space to which the model behavior remains sensitive to changes in the parameters.

A second geometrically-inspired improvement to data fitting is an improvement to the common Levenberg-Marquardt algorithm known as the geodesic acceleration

correction [71, 72, 75]. The motivation for this algorithm is the observation that model manifolds typically have surprisingly small curvatures. (This observation had been noted by statisticians for several decades [5–8, 10, 36] and was finally explained by the same interpolation arguments that explain why the model manifolds are often bounded [71, 72].) Since the manifolds are relatively flat, the ideal path for an algorithm to follow is a geodesic, i.e. a straight line through data space. Notice in Fig. 11.2b how the geodesic path naturally follows the curvature of the canyon in parameter space. In the limit of small curvature, the second-order correction to the Levenberg-Marquardt algorithm reduces to the second-order term in the geodesic equation, which can be easily approximated with little computational cost compared to other aspects of the algorithm. The result is an algorithm that is dramatically faster at finding best fits. An open source FORTRAN implementation of this algorithm is available for download [69].

11.4.2.2 Bayesian Posterior Sampling

As discussed in Sect. 11.2.2.2, MCMC is a powerful technique for exploring parameter space and sampling the Bayesian posterior distribution. One of the challenges to effectively implementing this approach is the need to run the algorithm long enough to gather independent samples of the posterior. For a cost landscape with long narrow canyons, the Markov chain needs to effectively diffuse along the length of the long axes of the canyon for each sample. For the same reason that data fitting algorithms become sluggish in the canyon, the MCMC method also becomes very slow, requiring a very long chain before independent samples can be identified.

In order to alleviate this problem, it was suggested by Girolami and Calderhead that convergence could be improved by taking steps uniform in data space rather than parameter space [28]. Effectively, at each step of the chain, random parameter steps are proposed as a multivatiate normal distribution with covariance chosen so that the corresponding steps in data space have covariance given by the identity. In this way, steps are preferentially aligned with the axis of the canyon, reducing the number of iterations necessary to generate independent samples. For extremely sloppy models with large aspect ratios in the canyons around their best fits, the improvement in convergence rate can be dramatic.

11.4.2.3 Curvature and Beyond

There are many other instances where differential geometry has provided insights and advancements in modeling and numerical methods. One of the most important concepts in differential geometry, and one that is beyond the scope of this review, is curvature. Measures of curvature have been used to quantify nonlinearity in models [5, 10, 36], measure kurtosis [35], and identify the global minimum in least-squares data fitting problems [21].

Information geometry has also led to a new approach to model reduction known as the manifold boundary approximation method [74]. By numerically constructing geodesics to the edge of the model manifold, limiting approximations are identified in the model that can be used to remove sloppy parameter combinations. The net result is a sequence of effective models of decreasing complexity. These reduced models remain expressed in terms of the microscopic parameters, i.e., there are no black boxes, and dramatically highlight the emergent control mechanisms that govern the system's behavior. Differential geometry also provides insights into questions of parameter identifiability, which combined with model reduction techniques, can be powerful tools for constructing appropriate mathematical representations of biological systems [70].

Although a relatively undeveloped approach, information geometry has provided a wealth of insight into modeling and the numerical methods for exploring model behavior. Much of the strength of the approach lies in its generality. Indeed, very little of what is summarized in this section is specific to systems biology. In this respect, applying information geometry to systems biology, with its wide array of models, is a compelling synthesis for the development of new theoretical and computational methods that are likely to not only advance biological understanding, but also find application in other complex systems.

11.5 Conclusion

Mechanistic models in systems biology typically possess a profusion of parameters, and this poses great challenge for modelers. In particular, understanding the multivariate sensitivity of the model to changes in parameter values is critical. Local and global analyses of sensitivity complement each other, and Bayesian methods are particularly powerful for assessing statistical confidence in parameter inferences and model predictions.

Analysis of many models in systems biology and other fields has revealed that nonlinear least-squares models are typically sloppy. Sloppy models have parameter sensitivity eigenvalues that span many decades roughly evenly and thus have highly anisotropic mappings between parameter and data spaces. Consequently, it is difficult to infer precise parameter values from data fits, but some predictions can nevertheless be tightly constrained. Careful experimental design can improve the precision of parameter inferences or model predictions, depending on the goals of the modeler. Information geometry offers a useful parameterization-independent perspective on modeling, and combining it with interpolation theory suggests that sloppiness arises because even complex models are often acting as interpolating functions between available data points. The information geometry perspective also suggests improved algorithms for optimization and Markov-chain Monte Carlo that account for the anisotropic and curved model and parameter spaces common in sloppy models.

Modelers have tackled a huge number of complex nonlinear systems in biology and other fields, and each model is unique. The study of sloppiness has shown,

however, that models of very different systems are nevertheless governed by shared deep statistical properties. Study of sloppy models thus offers insight and tools for not only systems biology, but also many other fields of science.

For readers who want hands-on experience with the methods and ideas discussed here, code implementing our analyses of the Robertson model is bundled with the SloppyCell software [55], available at: http://sloppycell.sourceforge.net.

Acknowledgments B.M. was supported by an ARCS Foundation Fellowship. A.R. was supported by NSF IGERT grant DGE-0654435. R.G. was supported by NSF grant DEB-1146074. We thank Alec Coffman for helpful discussions. R.G. and M.T. particularly thank Jim Sethna for his outstanding support and mentorship.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

1. Abdi, H., Williams, L.J.: Principal component analysis. Wiley Interdiscip Rev: Comput. Statist. **2**(4), 433–459 (2010)
2. Amari, S.I., Nagaoka, H.: Methods of Information Geometry, Translations of Mathematical Monographs, vol. 191. American Mathematical Society, New York (2000)
3. 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)
4. Barndorff-Nielsen, O., Cox, D., Reid, N.: The role of differential geometry in statistical theory. Int. Stat. Rev. **54**(1), 83–96 (1986). doi:10.2307/1403260
5. Bates, D.M., Watts, D.G.: Relative curvature measures of nonlinearity. J. Roy. Stat. Soc. B **42**, 1–25 (1980)
6. Bates, D.M., Watts, D.G.: Parameter transformations for improved approximate confidence regions in nonlinear least squares. Ann. Stat. **9**(6), 1152–1167 (1981)
7. Bates, D.M., Watts, D.G.: Nonlinear Regression Analysis and Its Applications, Wiley Series in Probability and Statistics, vol. 32. Wiley, New York (1988)
8. Bates, D.M., Hamilton, D.C., Watts, D.G.: Calculation of intrinsic and parameter-effects curvatures for nonlinear regression models. Commun. Stat. Simulat. **12**(4), 469–477 (1983). doi:10.1080/03610918308812333
9. Battogtokh, D., Asch, D., Case, M., Arnold, J., Schüttler, H.B.: An ensemble method for identifying regulatory circuits with special reference to the qa gene cluster of Neurospora crassa. Proc. Natl. Acad. Sci. USA **99**(26), 16904–16909 (2002). doi:10.1073/pnas.262658899
10. Beale, E.M.L.: Confidence regions in non-linear estimation. J. Roy. Stat. Soc. B **22**(1), 41–88 (1960)
11. Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate Bayesian computation in population genetics. Genetics **162**(4), 2025–2035 (2002)
12. Birnbaum, A.: On the foundations of statistical inference. J. Am. Stat. Assoc. **57**(298), 269–306 (1962). doi:10.2307/2281641
13. Brown, K., Sethna, J.: Statistical mechanical approaches to models with many poorly known parameters. Phys. Rev. E **68**(2), 021904 (2003). doi:10.1103/PhysRevE.68.021904

14. Brown, K.S., Hill, C.C., Calero, G.A., Myers, C.R., Lee, K.H., Sethna, J.P., Cerione, R.A.: The statistical mechanics of complex signaling networks: nerve growth factor signaling. Phys. Biol. **1**(3–4), 184–195 (2004). doi:10.1088/1478-3967/1/3/006
15. Casey, F.P., Baird, D., Feng, Q., Gutenkunst, R.N., Waterfall, J.J., Myers, C.R., Brown, K.S., Cerione, R.A., Sethna, J.P.: Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model. IET Syst. Biol. **1**(3), 190–202 (2007). doi:10.1049/iet-syb
16. Chachra, R., Transtrum, M.K., Sethna, J.P.: Comment on Sloppy models, parameter uncertainty, and the role of experimental design. Mol. Biosyst. **7**(8), 2522; author reply 2523–4 (2011). doi:10.1039/c1mb05046j
17. Chib, S., Greenberg, E.: Understanding the Metropolis Hastings algorithm. Am. Stat. **49**(4), 327–335 (1995). doi:10.1080/00031305.1995.10476177
18. Daniels, B.C., Chen, Y.J., Sethna, J.P., Gutenkunst, R.N., Myers, C.R.: Sloppiness, robustness, and evolvability in systems biology. Curr. Opin. Biotech. **19**(4), 389–395 (2008). doi:10.1016/j.copbio.2008.06.008
19. De Smet, R., Marchal, K.: Advantages and limitations of current network inference methods. Nat. Rev. Microbiol. **8**(10), 717–729 (2010). doi:10.1038/nrmicro2419
20. Del Moral, P., Doucet, A., Jasra, A.: Sequential Monte Carlo samplers. J. Roy. Stat. Soc. B. Met. **68**(3), 411–436 (2006). doi:10.1111/j.1467-9868.2006.00553.x
21. Demidenko, E.: Criteria for global minimum of sum of squares in nonlinear regression. Comput. Stat. Data An. **51**(3), 1739–1753 (2006). doi:10.1016/j.csda.2006.06.015
22. Efron, B., Hinkley, D.V.: Assessing the accuracy of the maximum likelihood estimator: observed versus expected Fisher information. Biometrika **65**(3), 457–483 (1978). doi:10.1093/biomet/65.3.457
23. 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). doi:10.1039/c0mb00107d
24. Eydgahi, H., Chen, W.W., Muhlich, J.L., Vitkup, D., Tsitsiklis, J.N., Sorger, P.K.: Properties of cell death models calibrated and compared using Bayesian approaches. Mol. Syst. Biol. **9**(644), 644 (2013). doi:10.1038/msb.2012.69
25. Fernández Slezak, D., Suárez, C., Cecchi, G.A., Marshall, G., Stolovitzky, G.: When the optimal is not the best: parameter estimation in complex biological models. PloS One **5**(10), e13,283 (2010). doi:10.1371/journal.pone.0013283
26. Fisher, R.A.: On the mathematical foundations of theoretical statistics. Philos. T. Roy. Soc. Lond. **222**(594–604), 309–368 (1922). doi:10.1098/rsta.1922.0009
27. Flaherty, P., Radhakrishnan, M.L., Dinh, T., Rebres, R.A., Roach, T.I., Jordan, M.I., Arkin, A.P.: A dual receptor crosstalk model of G-protein-coupled signal transduction. PLoS Comput. Biol. **4**(9), e1000185 (2008). doi:10.1371/journal.pcbi.1000185
28. Girolami, M., Calderhead, B.: Riemann manifold Langevin and Hamiltonian Monte Carlo methods. J. Roy. Stat. Soc. B. Met. **73**, 123–214 (2011). doi:10.1111/j.1467-9868.2010.00765.x
29. Gunawardena, J.: Models in sytems biology: the parameter problem and the meaning of robustness. In: Lodhi H.M., Muggleton, S.H. (eds.) Elements of Computational Systems Biology, pp. 19–47. Wiley Hoboken (2010). doi:10.1002/9780470556757.ch2
30. Gutenkunst, R.: Sloppiness, Modeling, and Evolution in Biochemical Networks. Ph.D. thesis, Cornell University (2008). http://www.ecommons.cornell.edu/handle/1813/8206
31. Gutenkunst, R.N., Casey, F.P., Waterfall, J.J., Myers, C.R., Sethna, J.P.: Extracting falsifiable predictions from sloppy models. Ann. NY Acad. Sci. **1115**, 203–211 (2007a). doi:10.1196/annals.1407.003
32. 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 (2007b). doi:10.1371/journal.pcbi.0030189
33. Hagen, D.R., Apgar, J.F., White, F.M., Tidor, B.: Molecular BioSystems reply to comment on Sloppy models, parameter uncertainty, and the role of experimental design. Interface Focus pp. 2523–2524 (2011). doi:10.1039/c1mb05200d

34. Hagen, D.R., White, J.K., Tidor, B.: Convergence in parameters and predictions using computational experimental design. Interface Focus **3**(4), 20130,008–20130,008 (2013). doi:10.1098/rsfs.2013.0008
35. Haines, L.M., O Brien, T.E., Clarke, G.P.Y.: Kurtosis and curvature measures for nonlinear regression models. Stat. Sinica **14**(2), 547–570 (2004)
36. Hamilton, D.C., Watts, D.G., Bates, D.M.: Accounting for intrinsic nonlinearity in nonlinear regression parameter inference regions. Ann. Stat. **10**(38), 393 (1982)
37. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**(6), 417–441 (1933)
38. Hug, S., Schmidl, D., Li, W.B., Greiter, M.B., Theis, F.J.: Bayesian model selection methods and their application to biological ODE systems. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Chem (2016, this volume)
39. Ivancevic, T.T.: Applied Differential Geometry: a Modern introduction. World Scientific, Singapore (2007)
40. Jaqaman, K., Danuser, G.: Linking data to models: data regression. Nat. Rev. Mol. Cell Bio. **7**(11), 813–819 (2006). doi:10.1038/nrm2030
41. Kass, R.E.: The geometry of asymptotic inference. Stat. Sci. **4**(3), 188–219 (1989)
42. Kirk, P., Silk, D., Stumpf, M.P.H.: Reverse engineering under uncertainty. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Chem (2016, this volume)
43. Kirk, P., Thorne, T., Stumpf, M.P.: Model selection in systems and synthetic biology. Curr. Opin. Biotech. **24**(4), 767–774 (2013). doi:10.1016/j.copbio.2013.03.012
44. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by simulated annealing. Science **220**(4598), 671–680 (1983). doi:10.1126/science.220.4598.671
45. Kitano, H.: Systems biology: a brief overview. Science **295**(5560), 1662–1664 (2002). doi:10.1126/science.1069492
46. Kitano, H.: Biological robustness. Nat. Rev. Genet. **5**(11), 826–837 (2004). doi:10.1038/nrg1471
47. Kreutz, C., Timmer, J.: Systems biology: experimental design. FEBS J. **276**(4), 923–942 (2009). doi:10.1111/j.1742-4658.2008.06843.x
48. Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J.L., Hucka, M.: BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res. **34**(Database issue), D689–91 (2006). doi:10.1093/nar/gkj092
49. Machta, B., Chachra, R., Transtrum, M., Sethna, J.: Parameter space compression underlies emergent theories and predictive models. Science **342**(6158), 604–607 (2013). doi:10.1126/science.1238723
50. Marino, S., Hogue, I.B., Ray, C.J., Kirschner, D.E.: A methodology for performing global uncertainty and sensitivity analysis in systems biology. J. Theor. Biol. **254**(1), 178–196 (2008). doi:10.1016/j.jtbi.2008.04.011
51. Marjoram, P., Molitor, J., Plagnol, V., Tavare, S.: Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **100**(26), 15,324–8 (2003). doi:10.1073/pnas.0306899100
52. Meyer, P., Cokelaer, T., Chandran, D., Kim, K.H., Loh, P.R., Tucker, G., Lipson, M., Berger, B., Kreutz, C., Raue, A., Steiert, B., Timmer, J., Bilal, E., Sauro, H.M., Stolovitzky, G., Saez-Rodriguez, J.: Network topology and parameter estimation: from experimental design methods to gene regulatory network kinetics using a community based approach. BMC Syst. Biol. **8**(1), 13 (2014). doi:10.1186/1752-0509-8-13
53. Moles, C.G., Mendes, P., Banga, J.R.: Parameter estimation in biochemical pathways: a comparison of global optimization methods. Genome Res. **13**(11), 2467–2474 (2003). doi:10.1101/gr.1262503
54. Murray, M.K., Rice, J.W.: Differential Geometry and Statistics, Monographs on statistics and applied probability, vol. 48. Chapman & Hall, London (1993)
55. Myers, C.R., Gutenkunst, R.N., Sethna, J.P.: Python unleashed on systems biology. Comput. Sci. Eng. **9**(3), 34–37 (2007). doi:10.1109/MCSE.2007.60

56. Pittendrigh, C.: On temperature independence in the clock system controlling emergence time in Drosophila. Proc. Natl. Acad. Sci. USA **40**(10), 1018–1029 (1954)
57. Rand, D.A., Shulgin, B.V., Salazar, D., Millar, A.J.: Design principles underlying circadian clocks. J. Roy. Soc. Interface **1**(1), 119–130 (2004). doi:10.1098/rsif.2004.0014
58. Robertson, H.: The solution of a set of reaction rate equations. In: Walsh, J. (ed.) Numerical Analysis, an Introduction, pp. 178–182. Academ Press, London (1966)
59. Rodriguez-Fernandez, M., Mendes, P., Banga, J.R.: A hybrid approach for efficient and robust parameter estimation in biochemical pathways. Biosyst. **83**(2–3), 248–265 (2006). doi:10.1016/j.biosystems.2005.06.016
60. Cedersund, G., Samuelsson, 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, Chem (2016, this volume)
61. Savageau, M.A., Coelho, P.M.B.M., Fasani, R.A., Tolla, D.A., Salvador, A.: Phenotypes and tolerances in the design space of biochemical systems. Proc. Natl. Acad. Sci. USA **106**(16), 6435–6440 (2009). doi:10.1073/pnas.0809869106
62. Seber, G.A.F., Wild, C.J.: Nonlinear Regression. Wiley, New York (1988)
63. Shah, M., Chitforoushzadeh, Z., Janes, K.A.: Statistical data analysis and modeling. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Chem (2016, this volume)
64. Sisson, S.A., Fan, Y., Tanaka, M.M.: Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. USA **104**(6), 1760–1765 (2007). doi:10.1073/pnas.0607208104
65. Spivak, M.: A Comprehensive Introduction to Differential Geometry. Publish or Perish (1979)
66. Sunnåker, M., Stelling, J.: Model extension and model selection. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Chem (2016, this volume)
67. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.: Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. J. Roy. Soc. Interface **6**(31), 187–202 (2009). doi:10.1098/rsif.2008.0172
68. Tönsing, C., Timmer, J., Kreutz, C.: Cause and cure of sloppiness in ordinary differential equation models (2014). arXiv:1406.1734
69. Transtrum, M.K.: Geodesic Levenberg-Marquardt source code (2012). http://sourceforge.net/projects/geodesiclm/
70. Transtrum, M.K., Hart, G., Qiu, P.: Information topology identifies emergent model classes. arXiv:1409.6203 (2014)
71. Transtrum, M.K., Machta, B.B., Sethna, J.P.: Why are nonlinear fits to data so challenging? Phys. Rev. Lett. **104**(6), 060,201 (2010). doi:10.1103/PhysRevLett.104.060201
72. Transtrum, M.K., Machta, B.B., Sethna, J.P.: Geometry of nonlinear least squares with applications to sloppy models and optimization. Phys. Rev. E **83**(3), 036,701 (2011). doi:10.1103/PhysRevE.83.036701
73. Transtrum, M.K., Qiu, P.: Optimal experiment selection for parameter estimation in biological differential equation models. BMC Bioinf. **13**, 181 (2012). doi:10.1186/1471-2105-13-181
74. Transtrum, M.K., Qiu, P.: Model reduction by manifold boundaries. Phys. Rev. Lett. **113**(9), 098,701 (2014). doi:10.1103/PhysRevLett.113.098701
75. Transtrum, M.K., Sethna, J.P.: Improvements to the Levenberg-Marquardt algorithm for nonlinear least-squares minimization. arXiv:1201.5885 (2012)
76. Waterfall, J., Casey, F., Gutenkunst, R., Brown, K., Myers, C., Brouwer, P., Elser, V., Sethna, J.: Sloppy-Model Universality Class and the Vandermonde Matrix. Phy. Rev. Lett. **97**(15) (2006). doi:10.1103/PhysRevLett.97.150601
77. Van Schepdael, A., Carlier, A., Geris, L.: Sensitivity analysis in the design of experiments. In: Uncertainty in Biology, A Computational Modeling Approach. Springer, Chem (2016, this volume)
78. von Dassow, G., Meir, E., Munro, E.M., Odell, G.M.: The segment polarity network is a robust developmental module. Nature **406**(6792), 188–92 (2000). doi:10.1038/35018085
79. Xu, T.R., Vyshemirsky, V., Gormand, A., von Kriegsheim, A., Girolami, M., Baillie, G.S., Ketley, 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). doi:10.1126/scisignal.2000517