Chapter 3

Existing approaches for model discrimination and selection

Introductory comments. The forthcoming chapter shall provide just a brief but general survey on model discrimination and selection as well as on the preceding parameter estimation. It is not intended to fully cover all aspects of the state of the art research. For an extended general reading, literature like [14, 15, 234, 238] is recommended. Nevertheless, the review should create awareness of the surfacing problems.

The main focus here will be the strategies or internal characteristics that are applied to either discriminate or select models. The statistically experienced audience, familiar with the existing concepts as well as their application potentialities, might want to skip this chapter. For the others, it is worthwhile reading it before returning to the in-depth motivation of overlap concept and its interpretation in chapter 4.

After these introductory comments, frequently employed methods are presented that can roughly be contributed either to the frequentist (section 3.1 and 3.2) or Bayesian school (section 3.3). It shall be stressed here, that neither of the schools is either favored or declined. The distinction was only chosen to structure the survey. A reflection on both types of approaches follows in section 3.4, trying to focus on the differences between the two concepts and showing some application challanges. Additionally, sensitizing for algorithmic aspects later, the implementation is also reviewed within each section.

3.1 Parameter estimation and model discrimination for regression models

In the typical situation of model discrimination, one has to either select the best or eliminate the worse models from a pool of candidates. They can be represented as a parameterized family, a nested hierarchy of models or structurally non-nested.

The classical process of model discrimination is based on a set of given parameters. Therefore, the entire process starts with a parameter estimation for each model that is under consideration. In a second step, the actual discrimination step, the quality of the goodness–of–fit, resulting from the parameter estimation, is evaluated. Therefore, this section also starts with reviewing parameter estimation.

Linear regression models

In the majority of every day statistical applications, regression models are considered (c.f. [197, 223]). They describe the causal relationship of a known form between random variables. The relationship is given in terms of a model, concatenating the dependent (endogenous) variable **d**, which is explained by the model, and the independent (exogenous, explanatory) variable **X**, which explains or predicts the dependent variables through the model. For the following, a general linear regression model is given

$$\mathbf{d} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}. \tag{3.1}$$

To revert to a notation set in reference literature and to be consistent with the notation introduced in chapter 2, let $\mathbf{X} \in \mathbb{R}^{D \times P}$, $\mathbf{d} \in \mathbb{R}^{D}$, the parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)^T \in \mathbb{R}^P$ and $\boldsymbol{\epsilon}$ the error term. The last mentioned error term indicates that every single measurement will be accompanied by some measurement error. For theoretical reasons and necessities, it is usually assumed to be normally distributed. The distribution's expected value is associated with the measurement data itself; the possible data derivation by the standard deviation of the normal distribution.

Parameter estimation (PE). Conducting a parameter estimation for linear regression models in (3.1) means to solve an *optimization problem* for $\boldsymbol{\theta}$, namely to minimize the distance between the measured data **d** and the model values $\mathbf{X}\boldsymbol{\theta}$

$$\mathcal{F}_{\mathrm{LS}}(\boldsymbol{\theta}) = \min \|\mathbf{d} - \mathbf{X}\boldsymbol{\theta}\|_2. \tag{3.2}$$

The distance between the measured data and model is taken in the $\mathbf{L}^2$ sense. It is also referred to as the *residual.* Therefore, the optimization is also referred to as *least square estimation* or *residual* optimization

$$\begin{aligned} \boldsymbol{\theta}_{\text{LS}} &= \arg\min \|\mathbf{d} - \mathbf{X}\boldsymbol{\theta}\|_2 && (3.3) \\ &= \arg\min \ \mathcal{F}_{\text{LS}}(\boldsymbol{\theta}). && (3.4) \end{aligned}$$

The solution $\boldsymbol{\theta}_{\text{LS}}$ of (3.2) is on the one hand a deterministic result of the optimization problem but on the other hand also a statistical result, namely the result of an estimator $\hat{\boldsymbol{\theta}}$ for the parameters with respect to (3.1).

Analytical solution. Assuming that the variance of the data, representing measurement uncertainty, is known and given by the variance-covariance matrix $\boldsymbol{\Sigma}_{\mathcal{D}}$, in terms of $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathcal{D}})$, then it can be shown (c.f. [25, 84]) that the weighted least squares estimation for parameters $\boldsymbol{\theta}_{\text{LS}}$ (3.4) can be calculated analytically by

$$\hat{\boldsymbol{\theta}}_{\text{LS}} = \left(\mathbf{X}^T \boldsymbol{\Sigma}_{\mathcal{D}}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \boldsymbol{\Sigma}_{\mathcal{D}}^{-1} \mathbf{d}. \tag{3.5}$$

The estimator $\hat{\boldsymbol{\theta}}$ in (3.5) happens to be the one with the smallest variance of all unbiased linear estimators (BLUE-estimator), namely (c.f. [25])

$$\text{Var}(\hat{\boldsymbol{\theta}}_{\text{LS}}) = \left(\mathbf{X}^T \boldsymbol{\Sigma}_{\mathcal{D}}^{-1} \mathbf{X}\right)^{-1}. \tag{3.6}$$

Equation (3.6) can be used to link the variance data $\boldsymbol{\Sigma}_{\mathcal{D}}$ to the quality of the estimated parameters $\boldsymbol{\theta}_{\text{LS}}$, which is expressed by the variance of the estimation of $\hat{\boldsymbol{\theta}}$. Thus, equation (3.6) characterizes the closeness of the estimated parameters to the assumed to be ”true” parameters[1] in the proposed model.

Maximum likelihood estimation (MLE). Another estimation method in parameter estimation is the maximum likelihood principle. It is based on the likelihood function $\mathcal{F}_{\text{ML}}$ which measures the probability of the distance between model and data. In case of the considered linear regression models, it can be expressed by

$$\mathcal{F}_{\text{ML}}(\boldsymbol{\theta}) = \mathbf{P}\left(\|\mathbf{d} - \mathbf{X}\boldsymbol{\theta}\|\right)$$

and

$$\begin{aligned} \boldsymbol{\theta}_{\text{ML}} &= \arg\min \ \mathbf{P}\left(\|\mathbf{d} - \mathbf{X}\boldsymbol{\theta}\|\right) \\ &= \arg\min \ \mathcal{F}_{\text{ML}}(\boldsymbol{\theta}). && (3.7) \end{aligned}$$

[1] Question on the existence of true parameters is revisited in section 3.4.

The statistical setting, namely determining or defining the employed probability measure $\mathbf{P}$, must be done in advance. Similar to (3.4), parameter estimation means to solve an optimization problem, namely to chose the parameters $\boldsymbol{\theta}_{\mathrm{ML}}$ so that they maximize the likelihood function $\mathcal{F}_{\mathrm{ML}}(\boldsymbol{\theta})$. Thus, the parameters $\boldsymbol{\theta}_{\mathrm{ML}}$ are chosen so that the deviation between data and model is most probable. However, the general setup of maximum likelihood estimations allows for a much broader field of applications.

If the model is linear with respect to the parameters $\boldsymbol{\theta}$ like in (3.1) and the data errors are assumed to be normally distributed with known covariance, then the maximum likelihood estimation coincides (c.f. [25]) with the weighted least squares estimation of (3.5).

Significance interpretation. The most common approach to asset the uncertainty or quality of the estimates of $\hat{\boldsymbol{\theta}}$ is the confidence interval or region concept, respectively (c.f. [55]). The underlying concept is usually tailored to suit linear regression models like in (3.1), which results in the definition of the confidence interval or region as

$$\mathcal{I} = \left\{ \boldsymbol{\theta} \in \mathbb{R}^P \mid \frac{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^T \; (\mathbf{X}^T \boldsymbol{\Sigma}_{\mathcal{D}}^{-1} \mathbf{X}) \; (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})}{P} \leq F_{(\alpha)} \right\}. \tag{3.8}$$

For linear models of the form (3.1), the confidence region $\mathcal{I}$ therefore consists of all parameters $\boldsymbol{\theta}$ for which the data–model–deviation can be statistically explained by considering a certain tolerance level (F-distribution) in terms of the rejection probability α. The confidence region or interval translates data variance into parameter variance.

Nonlinear regression models

The class of linear regression models only represents a small portion of real–world problems. In many scientific and engineering applications, nonlinear regression models are used that carry the time t as an independent variable. Since that type of models is used later in chapter 5 and 6, the following notation is introduced

$$\mathbf{d} = \phi(t; \boldsymbol{\theta}) + \boldsymbol{\epsilon}, \tag{3.9}$$

where ϕ is a nonlinear regression model function. A common class of nonlinear regression models are solutions to ODE systems as in (2.1). Their trajectories are prominent examples for nonlinear models

$$\phi(t, \boldsymbol{\theta}) = \Phi^t \, \mathbf{y}_0. \tag{3.10}$$

A good review on nonlinear regression modelling can be found in [234].

Linearization. In many applications, it is necessary to locally approximate the nonlinear models by a linear one, namely by using the first term of its Taylor expansion

$$\mathbf{X} = \mathbf{J}(t, \boldsymbol{\theta}_0)_{(i,j)} = \left. \frac{\partial\, \phi(t, \boldsymbol{\theta})_i}{\partial\, \theta_j} \right|_{=\ 0} \tag{3.11}$$

resulting in the linear approximation

$$\phi(t, \boldsymbol{\theta}) \cong \phi(t, \boldsymbol{\theta}_0) + \mathbf{J}(t, \boldsymbol{\theta}_0) \cdot (\boldsymbol{\theta} - \boldsymbol{\theta}_0)\,. \tag{3.12}$$

To complete the notation and be consistent with chapter 2, the measured data $\mathbf{d}$ is taken at the time points $\mathbf{t} = (t_1, \ldots, t_N)$ and is abbreviated by $\mathbf{d} = (\mathbf{d}(t_1), \ldots, \mathbf{d}(t_N))$. The measurement errors represented by the corresponding variances $\boldsymbol{\sigma}^2$ are abbreviated by $\boldsymbol{\sigma}^2 = (\boldsymbol{\sigma}^2(t_1), \ldots, \boldsymbol{\sigma}^2(t_N))$. With all the notation, the commonly suggested weighted residual for data with non–constant data variance (c.f. [25]) can be formulated in analogy to (3.2):

$$\mathcal{F}_{\mathrm{NLS}}(\boldsymbol{\theta}) = \sum_{j=1}^{D} \sum_{i=1}^{N} \left[\frac{d_j(t_i) - (\phi(t_i; \boldsymbol{\theta}))_j}{\sigma_j(t_i)} \right]^2 . \tag{3.13}$$

All measures $\mathcal{F}_{\mathrm{NLS}}$, $\mathcal{F}_{\mathrm{ML}}$ and $\mathcal{F}_{\mathrm{LS}}$ are from now on referred to as goodness–of–fit measures and are unless specified abbreviated as $\mathcal{F}_{\mathcal{R}}$ as they are used within residual fitting context.

Fit and performance. Writing the least square functionals differently shows an additional challange within the parameter estimation. It can be written in the probabilistic version as the mean squared error (MSE)

$$\mathcal{F}_{\mathrm{MSE}} = \mathbf{E} \left(\frac{1}{n} \| \hat{f} - f \|^2 \right), \tag{3.14}$$

where $\hat{f}$ is the estimated function and f the "true" one. The expectation $\mathbf{E}$ is taken for the preset probability measure, representing the error model of the data. The equation (3.14) can be decomposed into

$$\begin{aligned} \mathcal{F}_{\mathrm{MSE}} &= \mathbf{E} \left[\frac{1}{n} \| (\mathbf{E}\hat{f} - f) + (\hat{f} - \mathbf{E}\hat{f}) \|^2 \right] \\ &= \frac{1}{n} \| \mathbf{E}\hat{f} - f \|^2 + \frac{1}{n} \mathbf{E} \| \hat{f} - \mathbf{E}\hat{f} \|^2 \\ &= \mathrm{Bias}^2 + \mathrm{Variance} \end{aligned} \tag{3.15}$$

The first term, namely the Bias^2, measures how well the model approximates the "true" function. The second term, the variance, describes how well the function can be estimated.

Applying the same arguments to the predictive squared error (PSE), one could assess the predictability based on the estimation

$$\mathcal{F}_{\text{PSE}} = \mathbf{E}\left[\frac{1}{n}\|\mathbf{y}^{+} - \hat{f}\|^{2}\right] \tag{3.16}$$

$$= \sigma^2 + \mathcal{F}_{\text{MSE}}, \tag{3.17}$$

where $\mathbf{y}^{+}$ are new observations of the form $\mathbf{y}^{+} = f + \boldsymbol{\eta}^{+}$ with $\boldsymbol{\eta}^{+}$ being a componentwise independent and identically distributed vector with $\eta_i \sim \mathcal{N}(0, \sigma^2)$, which predicts the performance of the model for new observations.

Problems regression models. Even though the parameter estimation methods are well established, several problems can surface (c.f. [82]) and fallacies can happen (c.f. [179]), which result in statements like: "The whole area of guided regression is fraught with intellectual, statistical, computational and subject matter difficulties." A more historical review on inference problems can be found in the articles by Freedman, where he deals with the uncritical use of regression models and data modelling (c.f. [93, 94]).

The following example is taken from [54]. A bivariate random sample is investigated with one response variable $\mathbf{d}$ and one explanatory variable $\boldsymbol{\theta}$ combined in a linear regression model of the form $\mathbf{E}(\mathbf{d}|\boldsymbol{\theta}) = \mathbf{y} + \mathbf{X}\boldsymbol{\theta}$. A common procedure for regression is to firstly estimate $\boldsymbol{\theta}$, namely by a least-squares estimator for $\hat{\boldsymbol{\theta}}$ and then secondly to fit the regression line, assuming that $\hat{\boldsymbol{\theta}}$ is significantly different from 0. Assuming

$$\mathbf{E}\left[\hat{\boldsymbol{\theta}} \mid \hat{\boldsymbol{\theta}} \text{ is significantly different from } 0\right] \tag{3.18}$$

means that this conditional expectation is *not* equal to $\boldsymbol{\theta}$. The occurring bias of the result, as shown in (3.15), can be neglected when $\boldsymbol{\theta}$ is large. It cannot be neglected when the residual distance is large or when the sample size is small. Other bias sources lies in an underparameterization of the model. Generally, the bias will vanish asymptotically.

Regard the special case $\boldsymbol{\theta} = \mathbf{0}$ then the (unconditional) estimator can be shown to be

$$\hat{\boldsymbol{\theta}}_{\text{PT}} = \begin{cases} \hat{\boldsymbol{\theta}} & \text{is significant} \\ 0 & \text{otherwise.} \end{cases} \tag{3.19}$$

The construction (3.19) is referred to as pretest estimator (c.f. [130]). Two morals can be learnt from it:

(a) A least squares theory does not apply when the same data are used to formulate and fit a model

(b) to the analyst it must always be clear what any inference is conditioned on.

Model discrimination for regression models

As mentioned before, the classical process of model discrimination starts with a parameter estimation for each model. In a second step, the quality of the goodness–of–fit is evaluated. As the error $\boldsymbol{\epsilon}$ in (3.1) is mostly normally distributed, the deviation between model and data within the least squares framework in (3.2) is statistically described by the χ^2- statistics. Therefore, the ratio of the deviation, in terms of the $F-$test can be used to decide on the better model–data–fit for model discrimination (c.f. [26, 37, 159, 197, 240]).

F–Test for non–nested models. For non–nested[2] models the goodness–of–fit in terms of the sum of squares S_1 and S_2 with $S_1 > S_2$ is taken as a ratio

$$F_{\text{non}-\text{nested}} = \frac{S_1}{S_2} \tag{3.20}$$

and tested according to the $F-$statistics with $F_\alpha(N-P, N-P)$, where N denotes the numbers of observations and P the number of parameters (c.f. [40]).

F–Test for nested models. For nested models, that means for models where one model is extended by an additional effect and parameters, also ratio of the sum of squares is taken, however, corrected

$$F_{\text{nested}} = \frac{(S_1 - S_2)}{S_2} \cdot \frac{N - P_1}{P_2 - P_1}. \tag{3.21}$$

The ratio is supposed to be distributed according to $F_\alpha(P_2 - P_1, N - P_2)$, where P_1 and P_2 are the numbers of the parameters for each model, respectively.

For nested model, the general likelihood test is commonly employed (c.f. [31, 178, 117]).

[2] Nested models are models where one is a part of the other. Two models are non–nested models if one model is not the extention of the other one.

Optimal design for regression models

Literature. Optimal or optimum design, both terms are used in literature, is meant to improve the goodness–of–fit and therefore also improve afterwards inference by conducting specific measurements. Optimal design is founded on the works of Atkinson in [17], Box and Lucas in [41], Chernoff in [57], Dette and O'Brien in [73], Kiefer in [135, 136, 137, 138], Kiefer and Wolfowitz in [139] and Smith in [217]. It was first derived for linear models and later extended to generalized linear and nonlinear ones. Good reviews can be found in [18, 19, 20, 87, 187, 192, 216]. For reason of better readability, the citation references mentioned in the paragraph are omitted in the following as long they have already been mentioned here in this introduction.

Idea. The strategy of the optimal (experimental) design for regression models is to choose and weigh the measuring points t_i in such a way that the parameters are estimated best. For example, one could aim at achieving a low uncertainty in the parameters as defined in (3.8) and therefore improve the prediction quality of the model.

Especially in the case of nonlinear regression models, not all measured data points equally influence the result of the parameter estimation. This is due to a inhomogeneous data–parameter–sensitivity–structure within the model's state space. Therefore, according to the idea of optimal design, one would choose the measuring points in such a way that they have the highest possible impact on the parameter estimation's quality if the underlaying model is supposed to be true.

Speaking in mathematical terminology, optimal design involves finding a design ξ consisting out of data points t_i and associated weights ω_i, which are non-negative real numbers summing up to one ($\sum \omega_i = 1$)

$$\boldsymbol{\xi} = \begin{pmatrix} t_1 & t_2 & t_3 & \dots & t_N \\ \omega_1 & \omega_2 & \omega_3 & \dots & \omega_N \end{pmatrix}. \tag{3.22}$$

The *Fisher information matrix* $\mathbf{M}$ describes the parameters' influence on the model's trajectory at the measuring points in terms of their linear sensitivity. On the theoretic level, the matrix is defined in terms of the log-likelihood

$$\mathbf{M}(\boldsymbol{\xi};\boldsymbol{\theta}_0) = \mathbf{E}\left[-\frac{\partial^2 l(\boldsymbol{\xi};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial'\boldsymbol{\theta}}\right]. \tag{3.23}$$

The likelihood function l is the joint probability function of the sample, given the probability distributions that are assumed for the errors. For complex models, the likelihood is often computationally impractical. Therefore, its linear approximation is used

$$\mathbf{M}(\boldsymbol{\xi};\boldsymbol{\theta}_0) = \sum_{i=1}^{N} \omega_i \frac{\partial\,\phi(t_i;\boldsymbol{\theta}_0)}{\partial\,\boldsymbol{\theta}} \frac{\partial\,\phi(t_i;\boldsymbol{\theta}_0)}{\partial\,\boldsymbol{\theta}^T} = \mathbf{J}^T\boldsymbol{\Omega}\mathbf{J}, \tag{3.24}$$

with $\boldsymbol{\Omega} = \mathrm{diag}(\boldsymbol{\omega})$ being the diagonal matrix, composed of the weights and $\mathbf{J}$ the Jacobian with respect to the parameters as defined in (3.11). The matrix $\mathbf{M}^{-1}$ happens to be the first order approximation of the covariance of the maximum likelihood estimator for the parameters $\boldsymbol{\theta}$, as it can be seen for the linear case in (3.5) and (3.8). Therefore, the prediction variance function [20] is defined as

$$\mathbf{DV}(\boldsymbol{\xi};\boldsymbol{\theta}) = \frac{\partial\,\phi(\mathbf{t};\boldsymbol{\theta})}{\partial\,\boldsymbol{\theta}^T}\mathbf{M}^{-1}(\boldsymbol{\xi},\boldsymbol{\theta})\frac{\partial\,\phi(\mathbf{t};\boldsymbol{\theta})}{\partial\,\boldsymbol{\theta}}. \tag{3.25}$$

The very design $\boldsymbol{\xi}_{\mathrm{G}}$ maximizing the prediction variance function of (3.25)

$$\boldsymbol{\xi}_{\mathrm{G}} = \arg\max\, \mathbf{DV}(\boldsymbol{\xi},\boldsymbol{\theta}) \tag{3.26}$$

is called $G-$optimal design. Three other optimizing strategies are based on the Fisher information matrix $\mathbf{M}$. The D–optimal design $\boldsymbol{\xi}_D$, which maximizes the logarithm of the determinant, the A–optimal the trace and the E–optimal the eigenvalue of $\mathbf{M}$. These three designs can be integrated into the so called local $\Phi-$design, which minimizes

$$\boldsymbol{\xi}_{\boldsymbol{\Phi}} = \arg\min\left\{\frac{1}{P}\left(\lambda_1(\boldsymbol{\xi},\boldsymbol{\theta})^k + \ldots + \lambda_P(\boldsymbol{\xi},\boldsymbol{\theta})^k\right)\right\}^{1/P}, \tag{3.27}$$

where $\lambda_1(\boldsymbol{\xi},\boldsymbol{\theta}),\ldots\lambda_P(\boldsymbol{\xi},\boldsymbol{\theta})$ are the P eigenvalues of $\mathbf{M}(\boldsymbol{\xi},\boldsymbol{\theta})^{-1}$ and $k \in (0,\infty)$. The criteria (3.27) corresponds to the $D-$optimal design for $k \rightarrow 0$, to the $A-$optimal for $k = 1$ and $E-$optimal for $k \rightarrow \infty$.

The general equivalence theorem establishes the equivalence of $\boldsymbol{\xi}_D$ and $\boldsymbol{\xi}_G$ for linear and nonlinear models. Another reason for the $D-$optimum design's popularity is that it remains invariant to an even nonlinear reparameterization of the model function and that it is easy to illustrate: The theory of D–optimal experimental design chooses a new design $\boldsymbol{\xi}$ in such a way that the volume of the confidence region in (3.8) is minimized in order to guarantee a high significance of the estimated parameters $\hat{\boldsymbol{\theta}}$.

Other designs. Besides the mentioned design criteria for parameter estimation introduced, other ones do exist and are based for example on the $F-$ or $T-$tests in [26, 52].

Model discrimination. The main work in the field of optimal design focuses on parameter estimation. Nevertheless, there are also some design criteria for model discrimination. The most prominent one, is the $T-$optimum design and was introduced by Atkinson and Federov in [21, 22] for nonlinear models. Later, it was extended to multiresponse dynamical models in [233]. The duality of optimum design for model discrimination problems and parameter estimation problems was proven by Fedorov and Khabarov in [88]. For reasons of completeness, it shall be mentioned that another approach for combining designs for parameter estimation and model discrimination in one step instead of consecutive ones exists in [35, 188]. Further, a review on the differences on sequential verses non–sequential designs for discrimination is documented in [71].

In case of two models ϕ_1 and ϕ_2, one of the models is assumed to be true with known parameters (in our case ϕ_1 and the parameter dependency is omitted). Then the $T-$optimal criteria can be written as

$$\boldsymbol{\xi}_{\mathrm{T}} = \arg\max \left\{ \min_{2} T(\boldsymbol{\xi}, \boldsymbol{\theta}_2) \right\} \tag{3.28}$$

with

$$T(\boldsymbol{\xi}, \boldsymbol{\theta}_2) = \sum_{i=1}^{N} \omega_i \|\phi_1(\xi_i) - \phi_2(\xi_i, \boldsymbol{\theta}_2)\|^2. \tag{3.29}$$

Surfacing problems. For linear regression models, the Fisher information matrix of (3.24) is independent of the true parameter values $\boldsymbol{\theta}_0$ of the unknown parameters $\boldsymbol{\theta}$. For nonlinear models, the Jacobian $\mathbf{J}$ in (3.11) depends on the unknown parameters $\boldsymbol{\theta}$, as the first partial derivatives have to be calculated. Consequently, in order to get a suitable next design $\boldsymbol{\xi}$, one has to adopt a best guess for the parameters $\boldsymbol{\theta}_0$ (see [23]: "... This sensitivity of locally optimum designs to the choice of the parameter value, $\boldsymbol{\theta}_0$, presents problems since on the one hand a poor guess for the true parameter value will lead to an inefficient design and on the other hand a guess close to the true value is overprecise and may result in a non-informative optimum design."). Also due to this linear approximation approach, symbolized by the Fisher matrix $\mathbf{M}$, the designs perform poorly when the model is quadratic (c.f. [244]).

In many applications in chemical and biochemical reactions, experimental optimal design is used (c.f. [46, 72, 226]). However, one should not forget, firstly, that it is sometimes not possible to measure at the points that are suggested by the design and, secondly, in an iterative process of repeated designs, also referred to as sequential design, the cost of real-world experiments sometimes do not allow such experiments.

When the model is unknown, the design criteria must work well regardless of which model is the true one. However, the robustness towards model uncertainty means they do not only work well on average. More, the estimator has to perform well over all models and shall not vary amongst those models with high posterior probability. In [161, 162], Läuter included model uncertainty in the choice of design by averaging the design criteria function over a finite set of possible points.

Implementational aspects

Formulating either the parameter estimation or the optimal design problem results in an optimization problem. For the vast majority of present real–world application, especially in biokinetic or pharmacokinetics applications, the optimization problem itself turns out to be nonlinear. The algorithm has to be tailored to the demands of the problem. A good review on advantages and disadvantages of several nonlinear numerical optimization for biokinetics and metabolic simulations can be found for example in [172].

A prominent problem in optimization refers to the problem of local and global convergence. To put it in a nutshell: Deterministic methods experience a local convergence, whereas there exists stochastic methods that converge globally. However, algorithms allowing for global convergence result in higher computational effort. A good review on present possibilities on computational stochastic optimization can be found in [220]. In the following paragraphs prominent deterministic methods are presented.

Numerical solutions. Unlike in the linear case, an analytical solution to (3.5) is almost always impossible. Therefore, the optimization problem

$$\boldsymbol{\theta}_{\mathrm{NLS}} \quad = \quad \arg\min \ \mathcal{F}_{\mathrm{LSN}}(\boldsymbol{\theta}) \tag{3.30}$$

has to be solved numerically. For nonlinear regression problems, Newton and quasi-Newton methods are the numerical recipe of choice for conducting the parameter estimation. Good and very comprehensive review on these methods can be found in [74]. For all these methods, the target functional $\mathcal{F}$ needs to be differentiable, not all target functionals are suitable for the method.

Newton method [74, 134]. NEWTON method is a gradient based method and calculates the root of a nonlinear functional $\mathcal{F}$

$$\mathcal{F}(w^*) = 0 \tag{3.31}$$

by translating the originally nonlinear problem in (3.31) into a sequence of linear problems and then solving them, where the iteration step w_{inc} is the solution of

$$\mathcal{F}'(w)w_{\text{inc}} = -\mathcal{F}(w), \quad w_{\text{new}} = w + w_{\text{inc}}. \tag{3.32}$$

In order to guarantee the method's convergence, $\mathcal{F}'(w)^{-1}$ must exist within the vicinity of the solution w^*.

Gauss–Newton method [74]. For nonlinear least square problems, the target functionals $\mathcal{F}$, as defined in (3.13), has to be optimized

$$\|\mathcal{F}(w^*)\|_2 = \min_w \|\mathcal{F}(w)\|. \tag{3.33}$$

By considering the necessary condition for extreme points, (3.33) gets reformulated into

$$\frac{1}{2}\,\text{grad}\|\mathcal{F}(w)\|^2 = \mathcal{F}'(w)^T\mathcal{F}(w) = 0. \tag{3.34}$$

Applying the NEWTON–method to (3.34) and neglecting the Hessian $\mathcal{F}''$, results again in a sequence of linear problems as for the NEWTON–method

$$\left(\mathcal{F}'(w)^T\mathcal{F}'(w)\right) w_{\text{inc}} = -\mathcal{F}'(w)^T F(w), \quad w_{\text{new}} = w + w_{\text{inc}}. \tag{3.35}$$

Levenberg–Marquardt method [158, 165]. Depending on the approximation quality of $\mathcal{F}'(w)^T\mathcal{F}'(w)$ and the current iteration values for w, the problem (3.35) can become ill–conditioned. By introducing a limit to the increment vector w_{inc}

$$\|w_{\text{inc}}\|_2 \leq \delta \tag{3.36}$$

and incorporating it as an Lagrange multiplier $p = p(\delta)$, (3.35) is formulated as

$$\left(\mathcal{F}'(w)^T\mathcal{F}'(w) + pI\right) w_{\text{inc}} = -\mathcal{F}'(w)^T F(w), \tag{3.37}$$

where I is the identity matrix.

Due to the restriction in (3.36), the iteration step is contracted or damped. Indeed, the LEVENBERG–MARQUARDT can be reformulated into a *damped* Newton method

$$\left(\mathcal{F}'(w)^T\mathcal{F}'(w)\right) w_{\text{inc}} = -\mathcal{F}'(w)^T F(w), \quad w_{\text{new}} = w + \lambda w_{\text{inc}}, \tag{3.38}$$

for some λ depending on the restriction enforced in (3.36). The user has to choose the appropriate λ. The most successful empirical strategy is the so–called *Armijo strategy* (c.f. [13]).

The LEVENBERG–MARQUARDT shows two interesting limiting cases for p (c.f. [74]). Choosing $p \to 0^+$ with nonsingular $\mathcal{F}'(w)$, one gets the GAUSS-NEWTON-method and for $p \to \infty$ the steepest descent method.

Miscellaneous. Over the past years, a heavy used hybrid method has surfaced, the so called EM–algorithm. It goes back to early works in [70]. A good review can be found in [33, 184]. The EM–algorithm is most widely based on the GAUSS–NEWTON–optimization techniques. It considers unknown parameters as distribution, averages the target functional (**e**xpectation step) and maximizes the functional (**m**aximization step) afterwards. The algorithm converges globally, is sensitive to statistical outliers and has successfully been used for estimating density mixtures structures (c.f. [198]) .

Literature shows that for each application, the optimization methods and algorithms are tailored and usually employ an intrinsic treat of the problem to optimize. Such blueprints could be the penalized likelihood (c.f. [105]), some free energy analogies ([177, 222]), annealing schedules ([45]) via simulated annealing or Markov chains in ([236]).

3.2 Model selection

Occam's razor. In addition to classical model discrimination presented earlier, not only the goodness–of–fit is used to access the suitability of a model. Additional criteria and measures are included in the model selection process. Many of those are subsumed under the group of information theoretic criteria (c.f. [47]) and are commonly based on the *Occam's razor.* It says that the very model should be selected which *fits the observations sufficiently well in the least complex way* (c.f. [238]).

It means that, within the process of model selection, one has to compromise between the two aspects: First, how well does the data fit to the model and second how complex the models needs to be in order to reach that fit.

Goodness–of–fit. As in section 3.1, the measures that assess how well the model fits to the data is referred to as goodness–of–fit measures. Most commonly, derivation of a least squares in (3.2) are used. Measures for other application classes are for example the likelihood for density estimation problems and the classification error for pattern recognition problems.

Model complexity. The measure for model complexity must be tailored for each application individually, since the definition of ”complexity” is rather subjective. For parametric models, the complexity is translated into the *degrees of freedom* and can be expressed by the number of parameters of the model for example. In the case of nonparametric regression models, penalty terms like smoothing terms are used (c.f. [108]).

Trade-off. As mentioned before, goodness–of–fit and model complexity are opposite aspects in model selection. By adding degrees of freedom to a model, for parametric models this means to add parameters. By allowing for more flexibility and adaptability, the approximation quality increases. This procedure might lead to immensely complex models. However, the interpretation possibilities certainly deteriorate. Therefore, both effects have to be considered at the same time, for example by summing up both entities

$$\text{goodness of fit} + \gamma \text{ model complexity}, \tag{3.39}$$

where the constant γ weights both influencing factors. It is a *subjective* factor that has to be chosen by the user beforehand. Looking back on the discrimination strategy listed in the introduction on page 3, γ weights the strategies (D5) and (D6).

Example periodic splines. For periodic splines (c.f. [106, 237]), the model of the following class

$$\begin{aligned} \mathcal{S} &= \{f : f \text{ and } f' \text{ are absolutely continuous}, \\ &\qquad f(0) = f(1), f'(0) = f'(1), \int_0^1 (f''(t))^2 \,\mathrm{d}t < \infty \Big\} \end{aligned}$$

is chosen, that is selected by the following trade–off

$$\mathcal{F}_\gamma = \min_{f \in \mathcal{S}} \left\{ \frac{1}{n} \sum_{i=1}^{D} (y_i - f(t_i))^2 + \gamma \int_0^1 (f''(t))^2 \mathrm{d}t \right\}.$$

The first summand is a least–squares measure, representing the goodness–of–fit. The second one is a penalty term that punishes the ”roughness” or oscillations of the model. In this context, the weighting factor γ is also referred to as *smoothing parameter*. For $\gamma = 0$ the roughness is maintained, for $\gamma = \infty$ the function is forced to stay constant. Therefore, this non–parametric model selection example heavily depends on the choice of γ. Consequently, the result also varies with the selection of γ.

General information criteria. For regression models, a very general goodness–of–fit and model complexity trade–off can be written as

$$\mathcal{F}_{\text{trade off}} = -2l(\boldsymbol{\theta}) + \gamma P, \tag{3.40}$$

where P the number of parameters and

$$l(\boldsymbol{\theta}) = \log \mathbf{P}[\mathbf{d}|\boldsymbol{\theta}] \tag{3.41}$$

is the log–likelihood, which is maximized for the maximum-likelihood estimation of the parameters.

Akaike Information Criteria. The criterion for $\gamma = 2$ is the Akaike information criterion (AIC) (c.f. [4, 5, 6, 7])

$$\mathcal{F}_{\text{AIC}} = -2l(\boldsymbol{\theta}) + 2P. \tag{3.42}$$

The AIC is a large sample approximation of the discrepancy between the assumed true model and the fitted model in terms of the Kullback–Leibler distance (c.f. [146], appendix A). Model selection in terms of AIC means to choose the very model among the candidates that is closest in the Kullback–Leibler sense. However, the AIC tends to accept the more complex model (c.f. [193]). The following criteria does not.

Schwarz or Bayesian Information Criterion. Substituting the logarithm of the sample size $\gamma = \log N$ in (3.40), one gets the Schwarz's (SIC, [214]) or Bayesian information criteria (BIC)

$$\mathcal{F}_{\text{BIC}} = -2l(\boldsymbol{\theta}) + P \log N. \tag{3.43}$$

Minimum description length. A last model selection criteria shall mentioned here. It is the minimum description length (MDL) originated in field of data compression (c.f. [107, 201, 202])

$$\mathcal{F}_{\text{MDL}} = -l(\boldsymbol{\theta}) + \frac{k}{2} \log \left(\frac{n}{2\pi} \right) + \log \int \sqrt{|\mathbf{M}(\boldsymbol{\theta})|} \, \mathrm{d}\boldsymbol{\theta}, \tag{3.44}$$

where $\mathbf{M}$ is the Fisher matrix also surfacing in experimental design (3.24). Model selection by means of MDL is more accurate than AIC and BIC, however, demands higher computational effort (c.f. [180]).

Before continuing, two remarks shall be stated. First, for constant data variance likelihood function in (3.40) can be exchanged by a residuum measure like (3.30). Second, further information on similarities and differences between the AIC and BIC criteria concerning their reasoning and interpretation can be found in [49, 144].

Model selection problems. In many applications, models selected based on criteria like (3.39), are object to further statistical inference (c.f. [61]). Since this is then a concatenation of two statistical decisions, both incorporate statistical uncertainties. Therefore, the consequence of possible error made in the first decision, on the second one is of interest. However, only a few papers can be found on this interaction (c.f. [191]) showing results like

- model parameter estimates are asymptotically consistent (corresponding to a asymptotically vanishing bias) when model selection criteria are consistent like the AIC and BIC criteria (c.f. [59]) or
- the asymptotic distribution of parameter estimators is unaffected by model selection if the selection procedure is consistent but in some cases, for exmaple AIC or Mallow's C_p, the asymptotic distribution will be different from the 'usual' distribution.

As intuitively expected, the transported uncertainty from the model selection process results in higher variance of the succeeding estimation. This can also be seen as an analogy of the $\mathcal{F}_{\mathrm{PSE}}$ in (3.16). But even more, the asymptotic results, usually considered to cope with bias (see (3.14)), do not match with the actual one. This has also been observed by Zhang for linear regression models with final prediction error criterion (c.f. [67, 243]). In other words, the model selection bias is therefore not merely a result of small sample size but also an inherent one, as the property of the estimators may depend not only on the selected model, but also on the selection process (c.f. [113, 114]). Several approaches have been proposed to ease the problem, like a partitioning the sample into two disjunct subsets (c.f. [112, 114]).

Within the statistics community, this fact is also referred to as the "optimism principle" (c.f. [189]). It states that the model prediction is too optimistic on a new set of data.

Underestimation and cross-validation. This section is continued with a comment on cross-validation, a method that is also used within the model selection process. The goodness–of–fit–measures generally tend to underestimate the error of the model (c.f. [81, 109]). This can be seen in comparing $\mathcal{F}_{\text{MSE}}$ (3.14) and $\mathcal{F}_{\text{PSE}}$ of (3.16), where the first underestimates the second, when it comes to model prediction.

The mentioned effect results from using the same set of data for calibrating and validating the model. In order to avoid such types of underestimation, one could split the data set into one for calibration and one for validation. Therefore, the idea of cross-validation recycles the data by switching the role of training and test samples (c.f. [224]): One repeatedly omits some data points, calibrates the model and tests the prediction with respect to the left out ones. However, unless the available data sample is sufficiently large, this approach is not applicable.

The effect of under- and even overfitting also surfaces in the previously mentioned selection criteria, resulting in advanced criteria like unbiased risk method (c.f. [237]) or bias correcting AIC criteria like the AIC_C in [124].

Remark. Reviewing the two previous sections, one realizes that one has to be very careful when conducting statistical inference in the field of model selection and discrimination. One has to check, whether it is reasonable to apply the concept in question and whether there are unexpected pitfalls resulting from the intrinsic properties of the approaches.

3.3 Bayesian analysis

Literature. A good review on parametric Bayesian analysis can be found in [42, 100, 157]; for nonparametric Bayesian data analysis [176] is recommended.

Philosophy. The statistical concepts and methods for model discrimination and selection methods (including the preceding parameter estimation), that have been introduced in the previous sections, do all belong to frequentist methods. There, all statistical inferences are solely done on the basis of the measured data, the model and some significance level.

On the contrary, the Bayesian methods (c.f. [15, 65, 66]) of this sections, additionally incorporate gained knowledge from other sources, namely apriori information for the parameters or the models, respectively. This extra information is given in terms of a distribution and is built into the target functional. By means of the Bayesian theorem, one can then calculate an aposteriori distribution, the result of the estimation, for the entity of interest:

$$\text{aposteriori density} = \text{standardized likelihood} \times \text{apriori density}. \tag{3.45}$$

The Bayesian paradigm distinguishes itself from other statistical approaches by demanding that prior to obtaining the data, the statisticians considers his degrees of belief for the circumstance and represents it in the form of probabilities (c.f. [157]). The fundamental tenet of the Bayesian approach: The data does not create beliefs; they rather modify existing beliefs.

Critics. The challenge within the Bayesian framework lies in the provision of this very prior information. This necessity is a main source for the critics for the Bayesian approach (c.f. [8]).

Bayesian parameter estimation.

Calculation. For Bayesian parameter estimation, the aposteriori distribution for the parameters $\boldsymbol{\theta}$ is calculated

$$\mathbf{P}[\boldsymbol{\theta} \,|\, \mathbf{d}] = \frac{\mathbf{P}[\mathbf{d} \,|\, \boldsymbol{\theta}]}{\int\limits_{\Theta} \mathbf{P}[\mathbf{d} \,|\, \boldsymbol{\theta}]\mathbf{P}[\boldsymbol{\theta}]\mathrm{d}\boldsymbol{\theta}} \cdot \mathbf{P}[\boldsymbol{\theta}] = \frac{\mathbf{P}[\mathbf{d} \,|\, \boldsymbol{\theta}]}{\mathbf{P}[\mathbf{d}]} \cdot \mathbf{P}[\boldsymbol{\theta}], \tag{3.46}$$

where the first factor of (3.46) is the *standardized likelihood*, $\mathbf{P}[\boldsymbol{\theta}]$ is the prior distribution for the parameters and $\mathbf{P}[\mathbf{d} \,|\, \boldsymbol{\theta}]$ the likelihood.

Influence of the prior. The choice of the prior distribution depends on individual judgement in the light of the information and experience available at the time. This somehow involves the trust in the parameters or the model.

For the following illustrations, a one-dimensional example is used, where one is measuring the identity. This means, the parameters and data should "coincide". In an experiment, an entity for $\boldsymbol{\theta}$ is measured. Let its value be 11. The experimenter assumes that the measurement's realization was drawn from a normal distribution. Two cases have been prepared, one case with a large likelihood variance, representing a case where the measurement is *uncertain* and one case with a small deviation, where the measurement is *certain*. Let therefore the likelihood function $\mathbf{P}[\mathbf{d} \,|\, \boldsymbol{\theta}]$ be normally distributed with

$$\mathbf{P}_{\text{uncertain}}[\mathbf{d} \,|\, \boldsymbol{\theta}] \sim \mathcal{N}(11, 1) \tag{3.47}$$

$$\mathbf{P}_{\text{certain}}[\mathbf{d} \,|\, \boldsymbol{\theta}] \sim \mathcal{N}(11, 0.1). \tag{3.48}$$

Additionally, two different parameter priors $\mathbf{P}_A[\boldsymbol{\theta}]$ and $\mathbf{P}_B[\boldsymbol{\theta}]$ are investigated. Similar to the beliefs in the variance interpretation, A represents a strong belief, B a weak one in the parameter range

$$\mathbf{P}_A[\boldsymbol{\theta}] \sim \mathcal{N}(12, 0.4) \quad \text{and} \quad \mathbf{P}_B[\boldsymbol{\theta}] \sim \mathcal{N}(10, 3). \tag{3.49}$$

The result of the aposterior estimation can be graphically displayed: The "uncertain" likelihood scenario, in figure 3.1, the "certain" one in figure 3.2.

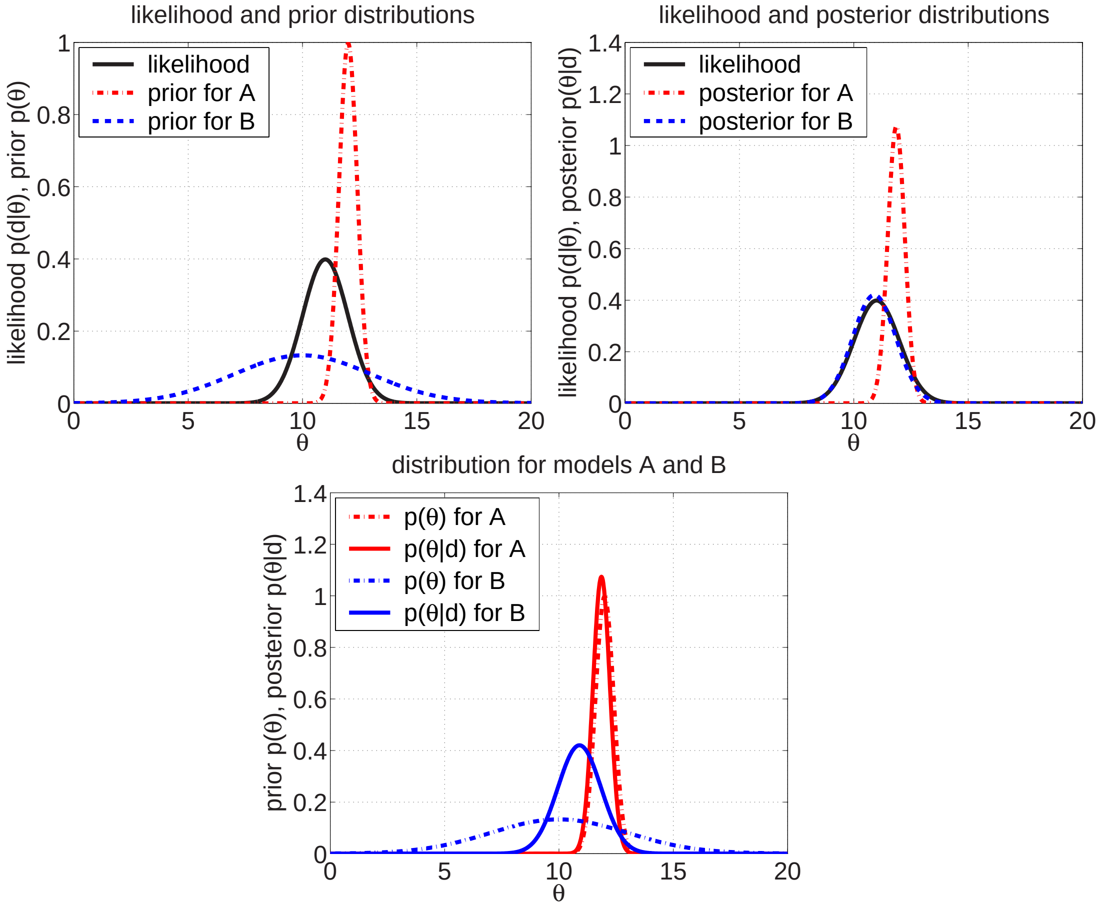

Figure 3.1: Prior, posterior and likelihood distribution for a Bayesian estimation with a large likelihood variance

The illustrations truly support that the data do not create believes, they just modify existing ones: In figure 3.1 the variance of the parameter prior $\mathbf{P}_A$ is larger than the one of the likelihood. Consequently, the posterior belief is close to the "more likely" likelihood. On the contrary, the strong prior "absorbs" the likelihood. Generally, this can also be turned into an advantage, as a single, rather inaccurate, observation cannot have much impact on relatively strongly held aprior beliefs.

For a small variance of the likelihood function in figure 3.2, the aposterior distributions is close to the likelihood belief. If the variance of prior and likelihood are of the same magnitude, then the belief ends up in the "middle".

As a result of the little illustration, one has to admit, that the result of the estimation strongly depends on the choice of the prior. Reversely speaking, this estimation method can be interpreted as a sensitivity analysis.

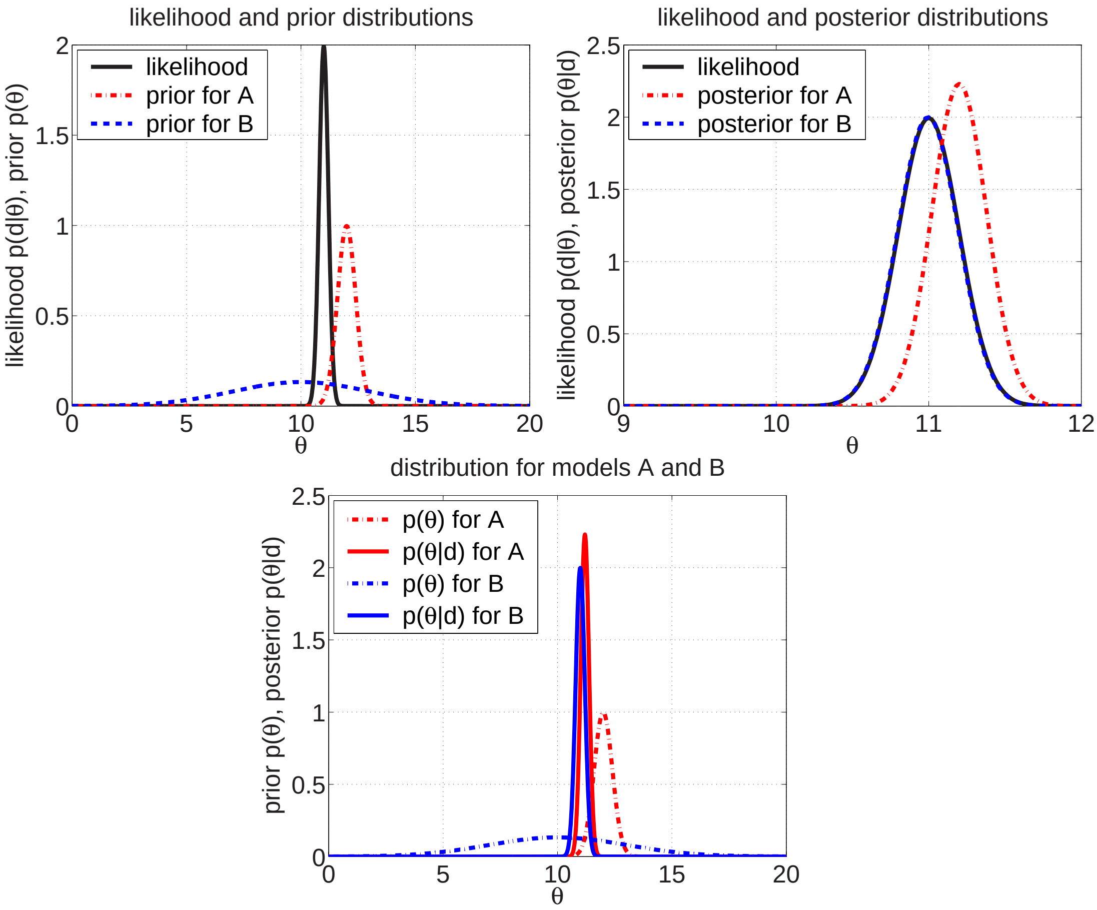

Figure 3.2: Prior, posterior and likelihood distribution for a Bayesian estimation with a small variance

Weak prior. With a more drastic assumptions of weak prior information, the Bayesian approach is transferred into *maximum likelihood estimation.* With no prior information available, one is tempted to assume an uniform distribution on some interval. The situation of a ”flat” prior distribution but ”peaking” likelihood also occurs when a moderate size of experiments is paired with diffuse prior beliefs, or even when a very large experiment is paired with moderately strong prior beliefs. In either ways, inserting a constant prior in (3.46), the aposterior distribution simplifies to

$$\mathbf{P}[\boldsymbol{\theta}\,|\,\mathbf{d}] \approx \frac{\mathbf{P}[\mathbf{d}\,|\,\boldsymbol{\theta}]}{\int\limits_{\Theta} \mathbf{P}[\mathbf{d}\,|\,\boldsymbol{\theta}]\,\mathrm{d}\boldsymbol{\theta}} \sim \mathbf{P}[\mathbf{d}\,|\,\boldsymbol{\theta}], \tag{3.50}$$

being the standardized likelihood function.

As a consequence, Bayesian analysis is only helpful, when the prior is known and sensible. Otherwise, the frequentist methods have to be favored. Parameter estimation strongly depends on the prior. In applications when dealing with physiological data, which are based on inhomogeneous populations, the Bayesian approach can be applied, as the prior then is interpreted as an ensemble of individuals among the population.

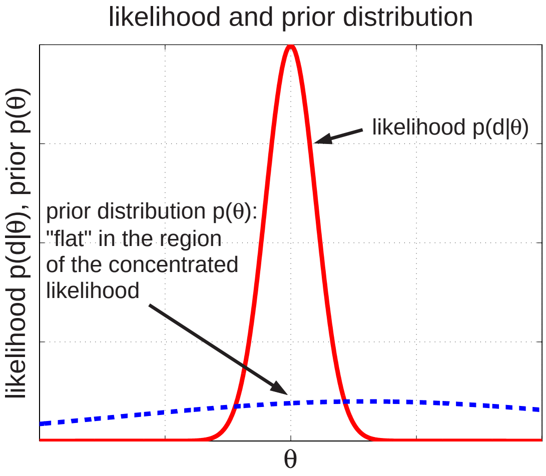

Figure 3.3: Weak prior and ”peaking” likelihood

Bayesian averaging

One elegant way to consider model uncertainty for predicting responses is the *Bayesian averaging* (c.f. [116]). After for example a model elimination process, a pool of candidate models

$$\mathcal{M} = (\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k)$$

remains. If one can no further eliminate, but wants to predict an observable Γ, one could calculate a prediction for every model $\mathcal{M}_i$ and average on them. Assigning a model uncertainty in terms of a model priors $\mathbf{P}[\mathcal{M}_i]$ (c.f. [60]), one can consider the entire pool $\mathcal{M}$ of the models and average in the "Bayesian way".

The very parameter vector $\boldsymbol{\theta}$, associated with the model $\mathcal{M}_k$, is abbreviated $\boldsymbol{\theta}_{(k)}$. As in the parameter estimation setup, one again assumes that the observed data $\mathbf{d}$ are generated, however, now by each model in question separately. Analogously, the *model likelihood* for model $\mathcal{M}_i$ is defined as

$$\mathbf{P}[\mathbf{d} \,|\, \mathcal{M}_k] \;=\; \int \mathbf{P}[\mathbf{d} \,|\, \boldsymbol{\theta}_{(k)}, \mathcal{M}_k]\, \mathbf{P}[\boldsymbol{\theta}_k \,|\, \mathcal{M}_k]\, \mathrm{d}\boldsymbol{\theta}_k, \tag{3.51}$$

where $\boldsymbol{\theta}_{(k)}$ denotes the parameters corresponding to the model $\mathcal{M}_k$, $\mathbf{P}[d \,|\, \boldsymbol{\theta}_{(k)}, \mathcal{M}_k]$ the *likelihood* and $\mathbf{P}[\boldsymbol{\theta}_{(k)} \,|\, \mathcal{M}_k]$ the parameter prior under model $\mathcal{M}_k$. Then the posterior probability for model $\mathcal{M}_k$ is calculated by

$$\mathbf{P}[\mathcal{M}_k \,|\, \mathbf{d}] \;=\; \frac{\mathbf{P}[\mathbf{d} \,|\, \mathcal{M}_k]}{\sum\limits_{k=1}^{K} \mathbf{P}[\mathbf{d} \,|\, \mathcal{M}_k]\; \mathbf{P}[\mathcal{M}_k]}, \tag{3.52}$$

where the model priors $\mathbf{P}[\mathcal{M}_k]$ assumes that $\mathcal{M}_k$ is the true model (given that one of the considered models is true).

Let Γ be a future observable. Then the posterior distribution as well as the posterior mean and variance are (c.f. [78, 194])

$$\begin{aligned}
\mathbf{P}[\Gamma \,|\, \mathbf{d}] &= \sum_{k=1}^{K} \mathbf{P}[\Delta \,|\, \mathcal{M}_k, \mathbf{d}]\; \mathbf{P}[\mathcal{M}_k \,|\, \mathbf{d}], \\
\mathbf{E}[\Gamma \,|\, \mathbf{d}] &= \sum_{k=1}^{K} \mathbf{E}[\Delta \,|\, \mathbf{d}, \mathcal{M}_k]\; \mathbf{P}[\mathcal{M}_k \,|\, \mathbf{d}], \\
\mathbf{Var}[\Gamma \,|\, \mathbf{d}] &= \sum_{k=1}^{K} (\mathbf{Var}[\Delta \,|\, \mathbf{d}, \mathcal{M}_k] + \mathbf{E}[\Delta \,|\, \mathbf{d}, \mathcal{M}_k])\, \mathbf{P}[\mathcal{M}_k \,|\, \mathbf{d}] - \mathbf{E}[\Delta \,|\, \mathbf{d}]^2.
\end{aligned}$$

In contrast to $\mathcal{F}_{\text{PSE}}$, Madigan and Raftery showed in [163] that averaging over all models provides a better predictive ability as it decreases the prediction variance. However, the prediction for Γ depends on the choice of the candidate model pool.

Model discrimination

The first approaches to Bayesian model discrimination were done by Box and Hill in [39]. Today, the most commonly employed method is the *Bayesian factor* introduced by Kaas and Raftery in [131] and is the analogue of likelihood ratio tests within the frequentist framework.

A *Bayesian factor* $B(\mathcal{M}_i, \mathcal{M}_j)$ for two models $\mathcal{M}_i$ and $\mathcal{M}_j$ is defined as the ration of the posterior odds and the model priors resulting in the ratio of the model likelihoods for two models

$$B(\mathcal{M}_i, \mathcal{M}_j) = \frac{\mathbf{P}[\mathcal{M}_i \,|\, \mathbf{d}] \; \mathbf{P}[\mathcal{M}_j]}{\mathbf{P}[\mathcal{M}_j \,|\, \mathbf{d}] \; \mathbf{P}[\mathcal{M}_i]} = \frac{\mathbf{P}[\mathbf{d} \,|\, \mathcal{M}_i]}{\mathbf{P}[\mathbf{d} \,|\, \mathcal{M}_j]}, \tag{3.53}$$

where $\mathbf{P}[\mathcal{M}_i]$ is the model prior for $\mathcal{M}_i$, $\mathbf{P}[\mathbf{d} \,|\, \mathcal{M}_i]$ the model likelihood for model $\mathcal{M}_i$. The marginal distribution of the data $\mathbf{d}$ under model $\mathcal{M}$, the Bayesian factors, therefore, choose the very model for which the marginal likelihood of the data is maximum. The value of a factor gives evidence of the preference between two models (see table 3.1).

Bayes factor	Interpretation
$B(\mathcal{M}_i, \mathcal{M}_j) < 1/10$	Strong evidence for $\mathcal{M}_j$
$1/10 < B(\mathcal{M}_i, \mathcal{M}_j) < 1/3$	Moderate evidence for $\mathcal{M}_j$
$1/3 < B(\mathcal{M}_i, \mathcal{M}_j) < 1$	Weak evidence for $\mathcal{M}_j$
$1 < B(\mathcal{M}_i, \mathcal{M}_j) < 3$	Weak evidence for $\mathcal{M}_i$
$3 < B(\mathcal{M}_i, \mathcal{M}_j) < 10$	Moderate evidence for $\mathcal{M}_i$
$B(\mathcal{M}_i, \mathcal{M}_j) > 10$	Strong evidence for $\mathcal{M}_i$

Table 3.1: Bayes factors scale by Jeffrey [128], adapted by Wassermann [239]

As in the case of parameter estimation, the Bayesian factors are sensitive to the choice of the model priors.

Approximation. As the model likelihood $\mathbf{P}[\mathcal{M}_i \,|\, \mathbf{d}]$ is very hard to compute, often an approximation is used (c.f. [180]). For example, the difference of the BIC (3.43) for two models is

$$\mathcal{F}_{\mathrm{BIC}_{\mathcal{M}_\mathrm{i}}} - \mathcal{F}_{\mathrm{BIC}_{\mathcal{M}_\mathrm{j}}} = -2\log\left(B(\mathcal{M}_i, \mathcal{M}_j)\right). \tag{3.54}$$

Bayesian experimental design for model discrimination. Over the past years, the classical optimal design framework has been enriched by a Bayesian perspective (c.f. [215]). In comparison to the experimental design approach introduced in section 3.1, the prior knowledge is incorporated in terms of a probabilistic model $\mathbf{P}(Z|\boldsymbol{\theta}, \mathbf{d})$, where ϱ denotes the experiment or model, respectively, and Z the observed data under experiment $\boldsymbol{\xi}$. Similar to the classical approach, a target function, in this context referred to as *utility function*, is maximized with respect to new measurement candidates $\boldsymbol{\xi}$. Let U be such a utility function. Most commonly the shannon information and or the quadratic loss criteria (c.f. [53]) is used. Than the optimal design or decision rule problem means to maximize the *posterior* expected utility. Assuming a collection of possible experiments, abbreviated by $\mathbf{d}$, the Bayesian solution to the experimental design problem, is the very experiment $\boldsymbol{\xi}^*$ maximizing the gained decision

$$\mathcal{F}(\boldsymbol{\xi}^*) = \max \int\limits_{\Omega} \max_{d} \int\limits_{\Theta} U(\mathbf{d}, \boldsymbol{\theta}, \varrho, Z)\, p(\boldsymbol{\theta}|Z, \mathbf{d})\, p(Z|\mathbf{d})\, \mathrm{d}\boldsymbol{\theta}\, \mathrm{d}Z. \tag{3.55}$$

Miscellaneous. Beyond Bayesian factors, there are other methods. As in section 3.2 concepts respecting the trade–off between model fit and model complexity do also exist within the Bayesian framework (c.f. [221]). They are referred to as the deviance information criterion (DIC). Another less sophisticated approach is using the Bayesian approach as a sensitivity analysis in [186]. General for non–nested model it is advisable to apply the Bayesian concept.

Computational Advances. The computation of the model likelihood is very demanding. Due to the general increasing computation performance and algorithmic techniques, the Bayesian approaches got revitalized. Good review articles are [12, 58, 203].

Bayesian analysis heavily depend on Monte Carlo experiments (c.f. [99, 110, 141]). For estimating of the likelihood, more recent methods like the harmonic mean estimator (c.f. [183]), important sampling (c.f. [95]), reciprocal importance estimator (c.f. [98]), bridge sampling (c.f. [96, 173]) or a reverse jump MCMC strategy (c.f. [104]). A comparison of these algorithmic alternatives can be found in [174].

Bayesian vs. frequentists. This section shall be closed with some remarks on the comparison of Bayesian factors and frequentist model testing framework.

For frequentist testings, one of the models has to be the null hypothesis. A common practise to get around it, is to switch and test both. Then both models may or may fail to be rejected. Therefore, a clear result is not guaranteed. Frequentists tests tend to reject the null hypotheses almost systematically in very large samples, whereas Bayesian factors do not (c.f. [131]). Further, frequentist tests are hard to implement for non–nested models. Bayesian factors are easy to apply for nested as well as non-nested models and allow for a broader field of applications.

A last remark refers to the type of statement one generates. Both approaches involve a pool of models. Multiple frequentist tests guide a search for the best model that can give very misleading results (c.f. [92]). Since Bayesian factors take model uncertainty into account, this problem can be avoided (c.f. [195]).

In [199], the authors Ren, Sun and Dey showed that even in the simple case of estimating and predicting normal populations, it strongly depends upon the setup whether a Bayesian or Frequentists approach is more suitable. Bayesian model selection results in better decisions in favor of the true model than maximum likelihood (c.f. [174]).

In classical or frequentists approach, it is somehow assumed that the correct structure of the model is known and the ”true” parameter value of the model parameters were to be estimated. Within the Bayes framework, there is no ”true” parameter value. The posterior density is a quantitative, probabilistic description of the knowledge about the parameters in the model (c.f. [28]).

A very interesting comparison between AIC and BIC can be found in [49]. It is shown there, that actually the differentiation between the frequentist and Bayesian interpretation does not apply to both of the criteria.

3.4 Miscellaneous

In the previous sections, several model discrimination and selection methods have been presented. However, now it is time to take a closer look at the method’s assumptions. Therefore, they are now reviewed to see the problems of the existing ones and to understand later in the next chapter how the model–data–overlap is going to cope with the challanges of the existing approaches.

Is there a true model? In the presented concept of the previous sections, the model in question is somehow assumed to be existing and to some extent correct. Within the statistical community, the discussion whether there is such an object like a *true model* is controversial.

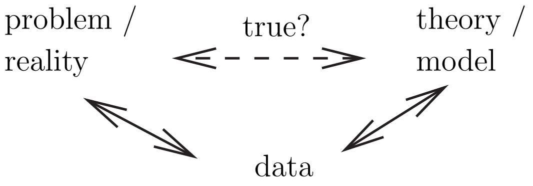

Figure 3.4: Constellation between reality, model and data. The communication between reality and model takes place only indirectly through the data.

The reality and the model do not communicate directly with each other, but instead only indirectly through the measured data. Therefore, one *cannot* assume that the data represent all aspects of the reality and that the measured data are without any measurement error.

Looking from an application point of view, the dilemma can be put into the following nutshell: On the on hand, one has to rely on something, especially a model, when performing some calculations, simulations and prediction. On the other hand, however, one can actually never be sure whether the reality is either described, explained or predicted by the model correctly. This thesis paper is not the suitable place to enter that discussion, nevertheless, some statements on the topic should be itemized here as well as the consequences for the statistical inference (c.f. [54]):

- Fildes and Howell in [91]: "It is a truism of forecasting that the model chosen is misspecified"
- Tiao and Tsay in [231]: "If one accepts the premise that any model is, at best, an approximation, then parameter estimation should be treated more in the context if the use for which the model is to be put rather than as an end in itself."
- Tsay in [230]: "Since all statistical models are wrong, the maximum likelihood principle does not apply"
- Tukey in [232]: When dealing with statistics we are "assuming that we always know what in fact we never know"

As a consequence many statisticians suggest to rather search for a *parsimonious* model that gives an adequate approximation for the investigates circumstance (c.f. [38, 152, 182]). One should concentrate on determining the model's accuracy and usefulness rather than testing it (c.f. [151]). The best summary for that attitude is the well-known statement by Box: "All models are wrong, but some are useful" (c.f. [38, 68]), which takes the problem of model uncertainty to the point.

Sources model uncertainty. As model uncertainty is surfacing frequently, the reseaons for it shall be reviewed systematically. According to [54, 79, 115], there are three main sources for model uncertainty

(U1) Uncertainty about the structure of the model;

(U2) Uncertainty about estimates of the model parameters, assuming that one knows the structure of the model;

(U3) Unexplained random variation in the observed variables even when one knows the structure of the model and the values of the model parameters.

According to [54], model uncertainty based on nescience in model structure of (U1) can be broken down further, namely

(S1) Model misspecification (e.g. omitting a variable by mistake),

(S2) Specifying a general class of models of which the true model is a special, but unknown case or

(S3) Choosing two or more models of quite different structures.

A broad selection of statistical methods for dealing with the aspects (U2), (U3) as well as (S2) is available. However, for cases like (U1), (S1) or (S3), the existing concepts do not allow strong inference possibilities as in the previously mentioned cases. When occurring, the errors arising from sources like (U1), are hard to handle and are worse than the errors from the sources that can be handled easily (c.f. [54]).

For example in the case of multiple-regression models this means: The theoretical statements document the very errors that result from having estimates of regression coefficients, rather than their true values. In comparison to the errors resulting from misspecification, like omitting a variable or neglecting non–linear terms, these errors are much smaller in many application contexts.

Model building process. The general aim of modelling is to gain information from the measured data, more precisely: Firstly, to extract information on the interdependence between response and input variables, and secondly, to predict the responses of the investigated circumstance in the future or of other input variables. The model building process consists out of the consecutive steps:

(M1) model formulation (or model specification),

(M2) model fitting (or model estimation),

(M3) model checking (or model validation) and

(M4) the combination of data from multiple sources (e.g. meta-analysis).

By iterating through them, the modeler tries to reduce model uncertainty and to gain as it is has been previously mentioned, a parsimonious model. He can choose between three main strategies to reduce model uncertainty (c.f. [43]):

(R1) Data investigation as well as isolating and incorporating additional effects

(R2) Considering a pool of plausible models

(R3) Algorithmic modelling

From the modelling aspect, the first two approaches belong to the concept of *data modelling* in contrast to *algorithmic modelling*. From the point of model discrimination at early stage modelling as well as enhancing intermediate models, only (R1) is of interest. Therefore, it is the question of how to incorporate the sources of model uncertainty into the discrimination process. The alternative (R2) is taken care of by Bayesian model averaging (as shown on page 34).

Algorithmic modelling. In case of algorithmic modelling, only the measured data but no specific model is given at the beginning. By means of an iterative algorithm, one wants to find a suitable model function that predicts the data response. The models are validated by measuring the predictive accuracy. It is a very direct and intuitive implementation of the inductive nature of statistics (data → model) in comparison to the deductive nature of probability theory (model → behavior).

Algorithmic modelling is not very commonly applied within the statistical community, but very widely used in engineering, for example in neural networks, learning machines, decision trees or splines. It is sometimes also referred to as automated model generation (c.f. [164]). Meanwhile many applications like speech recognition, image recognition or handwriting recognition are based on this modelling concept. More generally, it is very fruitful when the problem under consideration is very complex and cannot be modelled by summing up single effects. Further, one needs many observations to calibrate (train) the model in order to make decisions, whether the model are more or less reliable within the calibrated area.

The result of algorithmic modelling is a model showing data–model interdependencies that *cannot* directly be associated with certain scientific documented effects and mechanisms. The algorithm abandons interpretability and favors goodness–of–fit to rule out model uncertainty. Therefore, the reliability on long term forecasts is controversial. As the model's goodness–of–fit is guaranteed by the algorithm itself, additional criteria, as suggested in section 3.2, have to be taken into consideration for model discrimination and selection.

Data modelling. In contrast to algorithmic modelling, data modelling assumes that the data is generated by some stochastic data model. More specifically, the data is generated by independent draws from the response variable, which is a function of the input variables, parameters and random noise. The parameters are estimated for the measured data as well as the given model by means of some estimator and are then used for information and prediction. This data modelling approach dominates within the statistical community for analyzating effect and was therefore described in detailed in the previous sections.

In data modelling, the appropriate goodness–of–fit is usually checked by some statistical test. Unfortunately, those tests often have little power unless the direction of the alternative is precisely specified. Therefore, defects are hard to detect and models will not be rejected unless the lack of fit is seizable (c.f. [32, 43]). Another problems is the statistical indistinguishability of two models with different structure. In [170] it is explicitly stated: "Data will often point with almost equal emphasis on several possible models, and it is important that the statistician recognizes and accepts this." For small dimensional models, it is often observed that the tests result in a large number of models whose fit is acceptable. A third point that is often being raised when criticizing data modelling is that selecting and calibrating the model as well as making prediction is done on the same set of data. To overcome this source of bias, cross validation, as introduced in [179], is a "natural route" to reduce the bias. This, however, requires a large sample size (see page 18).

In comparison to algorithmic modelling, the question of true models surfaces here.

Data modelling does not guarantee the ”exact” model in terms of goodness–of–fit, but checks on the already known effects that are given in algebraic terms. Therefore, model–data–deviations are expected and model uncertainty must be taken into consideration.

The strategy can be summarized: In data modelling, one has a model and associates an effect a known and interpretable one to it. In this context, testing means to decide on an effect. By isolating the effects, one hopes to make trustworthy predictions. Knowing that not all effects can be detected, unmeasured variables are subsumed as noise and the prediction is most certainly not perfect.

Pool of plausible models. As an alternative and a comparison between the two positions (R1) and (R3) shown so far, one can consider a pool of models that has been identified as ”useful” in the sense that they represent a sufficiently close approximation of the data (c.f. [190]). This idea is the motivation for Bayesian model averaging (see page 34), an approach frequently employed in time series analysis, where different models describe different sequences. This shows the belief that a single model cannot describe all data alone.

Optimism vs. pessimism. Looking at both opposite modelling philosophies (R1) and (R3), one could call the algorithmic to be the optimistic and the data modelling to be the pessimistic approach. It means that the algorithmic modelling trusts the data as they are, whereas for data modelling, one only accepts the things that can be statically proven and interpreted. In both cases, however, one is never certain whether the model is true or not.

Sensitivity analysis. Reflecting all the approaches in this section with all the advantages and disadvantages, it shows that no general master approach exists. One needs a mixture of tailored solutions, which depends on the stage of modelling, on the type of model uncertainty one expects and on the strategy one chooses to cope with it.

Within the last years, it was suggested, that sensitivity analysis should be included in all parts of modelling (see (M1) - (M4)), especially in model validation (c.f. [205, 206, 208, 207]): ”I propose a form of organized sensitivity analysis in which a neighborhood of alternative assumptions is selected and the corresponding interval of inference is identified.” This idea is based on works by Leamer in [153, 154, 155]: ”Conclusions are judged to be sturdy only if the neighborhood of assumptions is wide enough to be credible and the corresponding interval of inference is narrow enough to be useful." It is therefore worthwhile to follow up this suggestion and treat sensitivity analysis as uncertainty analysis.

In the next chapter, it is going to be shown how the model–data–overlap incoporates model sensitivity into the model discrimination process and that it is a suitable tool for analyzing and coping with model uncertainty settings like (S3) and (U3).