Chapter 4

The overlap – illustration and motivation

Introductory comments. In chapter 2, the model–data–overlap concept was introduced. The previous chapter reviewed existing concepts for model discrimination and selection as well as showing challenges with different concepts and approaches. It is now time to return to the overlap concept and to show how it differs from the approaches of chapter 3. Within the comparison, a central question will be how model uncertainty and model–data–deviation are assessed, handled and interpreted.

The general aim now is to show that the model–data–overlap, as introduced in chapter 2, is a suitable concept for parameter estimation, model selection and discrimination when constructing intermediate models (see (M1) on page 40) and lacking structural knowledge (see (S1) and (U3)). After an extended motivation, illustration and conceptional classification, the short remainder of this chapter introduces the notation and the mathematical implementation of the model–data–overlap. In section 4.3, an illustrative model with non–linear parameters is used to demonstrate the difference between the maximum likelihood and model–data–overlap interpretation.

At this points, it shall be explicitly mentioned and stressed that the model–data–overlap is an alternative to the existing methods and concepts for model discrimination and parameter estimation. It is a tool within the modelling process that shows some beneficial properties existing approaches do not. Since tailored mixtures of tools are necessary within the application context, the model–data–concept adds flavor to the already colorful bouquet of statistical concepts.

A final remark: In the following many arguments refer to statements and results

made and documented in chapter 3. For reasons of better readability, the citation references are omitted in the following chapter if they have already been specified in the previous one.

4.1 Motivation and reasoning

Intermediate models. In numerous applications in chemical or biokinetic engineering, one wants to model experimentally observed effects. Numerous experiments consist out of many reactions with more than two reactants and catalytic components. The lineup of reactants, parameters or catalytic components is usually motivated by previously conducted experiments, though, in a different context, and by theoretical considerations (c.f. [62, 89]). As a result, a huge pool of complex model candidates arises.

This is due to the common procedure to suggest models by ordinary differential equations (ODE), where the right hand side is a function of reactants and parameters, which determine the dynamical behavior as well as the ease to assemble new models as it was described in chapter 1.

Nevertheless, due to the nescience of the actual underlaying processes, one has to anticipate a systematic deviation between the data and the model, especially when eliminating candidate models on the way to the *parsimonious* one (see section 3.4).

The situation described above is typical for modelling. For the mentioned intermediate candidate models, one lacks prior knowledge on their trustworthiness. Furthermore, one has to anticipate some model uncertainty and cannot assume that a model is *true* or *parsimonious*, as it has been classified on page 39. As uncertainty at this modelling stage is unavoidable, it has to be included either in the model itself for the discrimination process or in the result's interpretation.

Incorporating model uncertainty in terms of parameter and model prior beliefs, as demonstrated in the Bayesian approaches in section 3.3, and exploiting them, is conceptional not advisable, as they are unknown and not reliable in the context shown above. Uninformative priors result in likelihood estimations, as seen in (3.50), and suggest frequentist methods as introduced in section 3.1. Abandoning the Bayesian idea for a moment and returning to the frequentist framework, one has to acknowledge that model uncertainty is not accounted for in the parameter estimation procedure there. To illustrate it, the following paragraphs will show the interpretation of the classical parameter estimation.

Interpretation classical parameter estimation. Any parameter estimation method, relying on goodness–of–fit criteria alone, assumes the existence of a *true* or *parsimonious* model. This assumption is integrated in the data model itself, where data is expressed by the model as it is done for example in regression models (3.1). These models are characterized by a model–data–closeness. Model uncertainty is not anticipated. Parameter estimation in this context is seen as *model calibration*: It is a fine tuning that allows to trust the model's prediction. However, the available data is not always perfectly reliable. If new experimental data is added, the result of the parameter estimation with respect to the extended set of data might be (slightly) different. One typically interprets this effect as an uncertainty of the present estimation's result, caused by the incompleteness of the available data, and describes this uncertainty by means of some confidence intervals for the present estimation result as defined in (3.8). The construction of confidence intervals, therefore, deeply and necessarily depends on the model itself and translates *data uncertainty* into *parameter uncertainty*. No systematic model–data–deviation and consequently no model uncertainty is expected.

Optimal experimental design, as a sophisticated calibration method for parameter estimation, suggests new data measurement points in order to optimally improve the trust in the model's prediction. They are calculated in such a way that the parameter uncertainty is reduced, for example by minimizing the confidence intervals (see (3.8)). As data from future experiments influence the parameter estimation's result but is not available, its effect can only be captured by predicting its influence by means of the model itself (for example in (3.26). Conclusively, estimation of parameters, their uncertainties and enhancements by experimental design is in particular crucially based on the trust in the validity of the model under consideration.

There is another point that one should be aware of when it comes to interpreting the parameter estimation result: It concerns the question of bias. Among others, the reason lies in the asymptotic character of many estimation methods and their interpretations as well as in the small sample size available (see also page 18). However, the limited access to data is typical when constructing models.

Intermediate result. Reflecting the last paragraphs, one can draw the following conclusions. If one cannot establish a trust in a model than strictly speaking neither the parameter estimation methods nor their methods enhancing them can be applied without caution. These difficulties carry on for the model discrimination process by testing, as the entities used for deciding are calculated on the

basis on the previously calibrated model. In this context, a model is rejected when the model–data–deviation becomes too large. However, rejecting and therefore revealing a structural deficit is often not possible (see page 41). Several models can algebraically be very different, but however, are virtually indistinguishable in terms of their fit to a set of data and give very different predictions outside the range of the data. Additionally, testing depends on a user given significance level, which is hard to interpret when a systematic model–data–deviation has to be anticipated.

If one takes the model–data–deficiency seriously, one especially has to challenge the concept of a model describing or explaining the data. A model describes the measured data if and only if one can calibrate the model parameters so that the model will reproduce the data except for some acceptable error. It has already been shown that the plain fact of describing the data in terms of some goodness–of–fit criterium is not sufficient to accept or select the model. Some authors claim that a suitable model has to predict further measurements correctly, others content themselves with some explanation of the data. In most cases explanation means coherence of the calibrated model with some of a hitherto successful theories. To explain a measurement in this sense, no prediction of data and no statement about the causes of the measured effects is necessary. Nevertheless, here it will not be discussed whether the so far described alternatives of this common picture are right or misleading.

As a consequence one needs an approach to discriminate between competing model *without any explanation or prediction* based criteria. The model–data–overlap does it in such a way that it does provide knowledge about the model without any reference to other established theories, since it just compares model and data variability. However with this consequence one does not imply, that the data in itself favors a unique model or that a justification of models is possible without reference to prediction. Instead one can argue that the overlap concept ensures a kind of *model ranking*.

Additional criteria. Returning to the situation of several candidate models, where a systematic model–data–deviation is assumed: One promising strategy to deal with the experienced uncertainty and to cope with the interpretation problems mentioned above is the consideration of a second criteria, accompanying the goodness–of–fit criteria. Very commonly the model selection criteria of section 3.2 are employed.

However, when proposing such a second criterium or more thoroughgoing sug-

gesting a different criterium, one is confronted with the following questions: What should be measured or expressed by the second criterium? How is model uncertainty interpreted there? How is it incorporated?

A meanwhile well–established approach to a second criterium, is to add the aspect of model complexity as proposed by the Occam's razor, also referred to (D6) discrimination strategy on page 3. Within this framework, model complexity is interpreted as some model uncertainty: The model with the least complexity is more plausible and more suitable, therefore it is more likely.

A second criterium can be added to the goodness–of–fit criteria as seen in (3.39), which is a concept reverting within its interpretation to a trustworthy model. There, the second criterium is regarded as a penalty term and reflects a user given assumption or belief. The model complexity term in functionals based on (3.40) are subjective and application driven. Even more, in the AIC of (3.42) or in the BIC criterium of (3.43) criterium, only general model information like number of parameters or observations are considered. However, it actually does not incorporate and reveal anything about the algebraic structure of the model, especially not the influence of the parameter on the model in terms of sensitivity.

Sensitivity as a new criterium. However, a lack in structural knowledge as a source of model uncertainty, also referred to (S3) on page 39, is most common in chemical and biokinetic model engineering. Therefore, for discriminating models, one would have to choose a criterium to access the structural model uncertainty; a criterium that does evaluate not only the number of parameters for example but rather their impact. As a consequence, one could think of some sort of sensitivity analysis as it is has already been proposed by Leamer or Saltelli in the publications mentioned (see page 42), especially in [207]. However, no exact advise was given on how to incorporate it, just the hint to do it is stated and supported.

If structural deficits are likely to occur, then other properties like parameter sensitivity give evidence, whether the model is suitable or not. It is advisable to do so, as sensitivity can check whether a model can take on certain values (including their vicinity see (S3) and (U3)). It therefore assesses the adequacy of the chosen model–parameter–structure. Incorporating sensitivity as a criterium for model discrimination seems to be an alternative and possibility: *the model–data–overlap* approach. It can be seen as a discrimination strategy combining (D1), (D3) and (D5) in contrast to the scheme combination (D5) and (D6) (see page 3). Furthermore, sensitivity can also be used to cope with (S3) and (U3).

Model variability. The considerations up to now are embraced by the model–data–overlap through incorporating the mentioned sensitivity in terms of *model variability*. For dynamical systems for example, this new entity *model variability* can be defined as in (2.2). It shall reflect the general ability of the model's trajectory to change by means of the parameter. Theoretically, also perturbations of initial values can be considered. However, experimental experiences have shown that the parameter effects dominate and are focused on henceforth. Nevertheless, by considering initial values as parameters themselves, one can revert the case to the parameter one. This allows for a new line of reasoning: Model–data deviations are not only caused by noise, but also by distributed parameters. Consequently, the systematical difference between measurement errors as a data property and data deviations by parameter distributions vanishes. The model–data–deviation and structural model uncertainty is now handled by a different interpretation of the parameters, namely by assuming that they are distributed themselves.

Distributed parameters. Considering parameter–model–sensitivity as above, parameters are now conceived as distributions themselves. It is an interpretation, that for example is also supported by chemical interpretation, where the parameters can be modelled by stochastic processes depending on time. This means, that they may change their values randomly during time evolution, for example representing a kind of noisy behavior.

Beside the sensitivity motivation to consider distributed parameters, there are also application driven arguments as well as pragmatic considerations. First, in statistical mechanics, it is common to interpret all values of observables to be the expectation of a state variable with respect to some state space density. The same applies for macroscopic parameters: Even if taking the physical homogeneity of the density in state space for granted, the idea of different and changing realizations of a system state implies fluctuations of the coupled parameters of the system for each measurement. Second motivation: For biokinetics physiological models, realization of the same physiological processes stem from different values of the same set of parameters. To take this fact into account, distributed parameters are used. Another reason is less obvious: Suppose one investigates a time series from a completely new experiment. Then one could face coupled parameters determining local and non-periodic effects in our measurement data. Hence, there is no warrant to have invariant parameters during one measurement. And further, many measurements can hardly realize under the same initial conditions. Clearly distributed parameters form a possible account to deal with this uncertainty, because filtering periodic rates of the measured signal will not

help. The last argument is the existence of non–additive noise, coupling into the degrees of freedom of every single reaction in a non–deterministic way. In order to represent this effect in the model concerning different measurements, one has to consider distributed parameters for each measurement.

Intermediate resume. After showing numerous new aspects of the model–data–overlap, it is expedient to perform an intermediate resume. The overlap focuses on the model and data variability. The model variability reveals the model–parameter–sensitivity structure and the data variability shows how the data is distributed. By comparing both entities, one can assess, whether it is possible for a model to somehow reproduce the measured data including its variability. It is analyzes, whether the structural model–data–deviations are too large in order to discard the model under consideration.

Model–data–separation. The section shall be continued with some remarks on methodical and algorithmic questions. By matching model and data variability through the overlap, one explicitly ”separates” model and data information. In contrast to for example classical regression models as in (3.9), one does not explicitly assume that the data is reproduced or generated by the model within some acceptable error. Therefore, terms like description and explanation of the data by the model do not apply here. The same applies for Bayesian approaches, where the data is described by the likelihood function.

Due to the model–data–separation, one is reminded of algorithmic modelling. In the presented case, however, the model under consideration is not ”arbitrarily” generated by some algorithm to fit the data, but is still associated to some effect interpretation that is based on experience and trust in the application. In case of the overlap, the model–data–separation is mainly due to some kind of sensitivity considerations.

Strictly separating data and model properties has also another advantage. When constructing intermediate models, it is not possible to find sensible test statistics as well as common model–data–statistics. Neither it is possible to guarantee that all the assumptions discrimination and estimation methods need are fulfilled as one lacks model knowledge.

By abandoning the idea of comparing data by an adapted or extended residual setting, but comparing the shapes of the distribution instead in the overlap, one might be reminded of algorithmic modelling and the information theoretic framework applied for model selection in section 3.2. But as said before, the necessity

to consider distribution reverts to the involvement of sensitivity. Furthermore, the interpretation is different.

The idea of using distribution comparing measures for model selection and discrimination did also surface a few years ago for discrete models, where an optimal design for discriminating competing models was derived on the basis of the HELLINGER distance (c.f. [34]). There, however, the models themselves were compared pair and group-wise after being calibrated. As the huge pool increases, the comparison combination grows. The models were weighted by introducing a prior distribution on the model space in a Bayesian setting, which again involves non–existing prior knowledge.

A brief survey on the distribution comparing measures can be found in appendix A. In many applications in algorithmic modelling, especially in the context of information theoretic methods, the KULLBACK–LEIBLER–measure is used, as defined in (A.1). When calculating the $\mathcal{D}_{KL}(f_1||f_2)$, the roles of f_1 and f_2 cannot be interchanged as f_1 is favored in the interpretation. The KULLBACK–LEIBLER–measure denotes the information lost when the model presented by f_2 is used to approximate the full reality of the model that is represented by f_1. This asymmetric interpretation is passed on to the AIC criterium of (3.42): In [5] AKAIKE showed how the KULLBACK–LEIBLER information can be estimated, based on the maximized log–likelihood function as defined in (3.41).

Despite all the difference between the distribution comparing methods and the overlap, one similarity can be found: Neither information theoretic criteria nor the model–data–overlap do constitute a statistical "test" in any sense. There is no null hypothesis, no α-level, no asymptotic test statistic, no P-values, and no "reject" or "accept" decision. Instead, there is a concept of statistical evidence and the level of support to each model, representing alternative hypotheses (c.f. [48]). It seems that the inferences that can be made by the overlap, namely the model ranking and parameter estimation, are "less strict" than in case of the numerous approaches shown in chapter 3. However, when arguing about it, one should not forget about the applicability problems of the other methods as many assumption cannot be met for intermediate models.

Different data interpretation possible. After showing the conceptional and modelling advantages, there are also interpretational benefits from the data perspective. One advantage is to consider model parameters to be distributed is due to a lack in data. If the data basis is poor and only rough ideas of their values exist, it might be beneficial to consider entities, not necessarily normally

distributed, data distributions as an entity. A better specification of the data in terms of purely numeric values by reducing the measurement error is sometimes also not possible since some points are not possible to measure or the experiment's costs do not justify a measurement run.

From the modeler point of view, one could use an arbitrary data distribution, in contrast to the asymptotic statistical theories, where inference calculus dictates the characteristics of the distribution, usually normal ones. It is further possible to "punish" or "penalize" possible data values, for example by enforcing strictly positive data distributions when for example modelling concentrations in human bodies. To this end, one can regard the data deviations as measured data under different combinations of parameter values.

Error models. Due to the arbitrariness concerning the data distribution used, the overlap shows more flexibility when it comes to considering error models. Since many statistical inference methods rely one some asymptotic considerations, normally distributed error terms are assumed. By separating looking at model and data variability distribution, ones is free to easily introduce different data variables (c.f. [102]).

4.2 Overlap notation

After only verbally describing the model–data–overlap, it is now time to analytically describe it for dynamical systems as introduced in (2.1). The overlap at time t is defined as the scalar product of the model variability $\mathcal{M}_t$ and data $\mathcal{D}_t$ distribution. One way to guarantee a ranking is to introduce a upper bound

$$\mathcal{F}_{\mathcal{O}}(t) = \langle \mathcal{M}_t, \mathcal{D}_t \rangle_2 \leq \|\mathcal{D}_t\|_2 \, \|\mathcal{M}_t\|_2 \leq 1, \tag{4.1}$$

by enforcing $\|\mathcal{D}_t\| = 1$ and $\|\mathcal{M}_t\| = 1$.

Model variability. For dynamical system, as they have already been introduced in chapter 2, the model variability distribution $\mathcal{M}_t$ at time t is defined as

$$\mathcal{M}_t(A) = \frac{1}{C_t} \int_{\Theta} \mathbf{1}_A(\Phi^t \, \mathbf{y}_0) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}, \tag{4.2}$$

the constant C_t is chosen such that $\|\mathcal{M}_t\|_2 = 1$ for each t.

Data variability. Unless the experimental setting dictates something else, the data distribution is assumed to be normal with $\mathcal{N}(\mathbf{d}(t), \mathbf{\Sigma}(t)^2)$ with $\mathbf{\Sigma}(t) = \mathrm{diag}(\sigma_1(t), \ldots, \sigma_D(t))$. Then for each dimension i^{th} of data variability $\mathcal{D}_t$ is defined as

$$\mathcal{D}_t(x) \sim \frac{1}{\sqrt[4]{\pi}\,\sqrt{\sigma_i(t)}}\, e^{-\frac{\left(x-d_i(t)\right)^2}{2\sigma_i(t)^2}}. \tag{4.3}$$

Model parameter distribution. To calculate the model variability $\mathcal{M}_t$ (4.2), one needs to model the parameter distribution π. As already mentioned, it is a convenient assumption to use a normal distribution also such that $\pi \sim \mathcal{N}(\boldsymbol{\theta},\ \mathbf{\Delta\theta}^2)$, with

$$\mathbf{\Delta\theta} = \begin{pmatrix} \Delta\theta_1 & 0 & \ldots\ldots\ldots & 0 \\ 0 & \Delta\theta_2 & 0 \quad \cdots & 0 \\ & & \ddots & \\ 0 & \cdots & 0 \quad \Delta\theta_{p-1} & 0 \\ 0 & \ldots\ldots\ldots & 0 & \Delta\theta_p \end{pmatrix}. \tag{4.4}$$

However, later in chapter 6, also other families for the parameter distribution, like the log–normal distribution can be considered in the calculations.

The definition of the data variability $\mathcal{D}$ in (4.3) makes it evident again: In contrast to classical deviation functionals like $\mathcal{F}_{\mathrm{NLS}}$ or $\mathcal{F}_{\mathrm{ML}}$ in (3.13) or (3.7), respectively, the overlap functional $\mathcal{F}_{\mathcal{O}}$ does not merely depend on the data $\mathbf{d}(t)$ and model trajectory values $\Phi^t\,\mathbf{y}_0$, but also directly on the measurement standard deviations $\boldsymbol{\sigma}(t)$ as well as on the parameter variabilities $\mathbf{\Delta\theta}$:

$$\mathcal{F}_{\mathcal{O}} = \mathcal{F}_{\mathcal{O}}(\theta, \mathbf{\Delta\theta}, \mathbf{d}(t), \mathbf{\Sigma}(t)). \tag{4.5}$$

In this notation, one sees again: Compared to approaches presented in chapter 3, $\mathbf{\Sigma}(t)$ becomes a quantitative input data.

In the context of the normally distributed model parameters, overlap optimization means to choose the hyperparameters $\boldsymbol{\theta}$ as well as $\mathbf{\Delta\theta}$ so that $\mathcal{F}_{\mathcal{O}}$ is maximal. The second hyperparameter of the normal distribution is the variance itself, one could also refer to the variance as parameter variability. For other distribution families, the equality of hyperparameters and statistical momenta does not always hold.

4.3 Illustrative model and comparison

This section gives a simplified example for the model–data–overlap that demonstrates the change in interpretation when nonlinear parameters are included in

the model.

The illustration is done by the main interaction term of the so–called causal structure equation (c.f. [44]) surfacing in economy, more specifically in impact factor analysis. A causal model describes the interaction between impact factors and the success of a company. By formulating such an equilibrium condition as in

$$\eta = B\,\eta + \Gamma\,\xi + \zeta. \tag{4.6}$$

It is very easy to add, to eliminate or to substitute single factors. For real–world surveys one does not know the the equilibrium condition holds. Even worse, one is aware that a systematic model–data–deviation is expected. The surfacing structural model uncertainty suggests the use of the model–data–overlap!

However, another motivation can be given. The causal structure model are used when modelling the impact factors for performance of companies within an industry sector. As a result one wants to identify the most significant ones for a company's success. However, since the companies within a sector are not identically (c.f. revenue, number of employees, etc.), the parameters cannot be represented as single numeric values but are distributed. Consequently, the overlap is applicable. At this point, the application background shall be left alone and it is worthwhile to focus on the equation (4.6).

The equation (4.6) can be interpreted as a state of equilibrium, as it describes the processes arriving to or departing from the state η, respectively. The parameter B represents the portion remaining in B, whereas Γ characterizes the transfer process from $\xi \to \eta$. In the formulation above, the state space variables η, ξ and ζ as well as the parameters B and Γ are linear.

However, looking at the state η only, one has to reformulate (4.6) to

$$\eta = \frac{\Gamma\,\xi + \zeta}{1 - B}. \tag{4.7}$$

In this formulation (4.7), the parameter B has a nonlinear impact on η, the other quantities involved, remain linear. For reasons of simplicity, the term $\Gamma\,\xi + \zeta$ is set to be equal 3 during the further course of the calculations. Therefore, one only has to deal with the very simply, but nonlinear model:

$$\eta = \frac{3}{1 - B}. \tag{4.8}$$

Simulation. For the following overlap optimization, the data distribution was assumed to be normal with $\mathcal{N}(6, 1.25)$. A Monte–Carlo–simulation is used to

calculate the model variability that is needed to numerically calculate the model–data–overlap.

The simulation pictures can be seen in figure 4.1. The overlap optimization results in a parameter distribution for $B \sim \pi(B) \sim \mathcal{N}(0.52, 0.103)$ and a model–data–overlap number of 98.7 %.

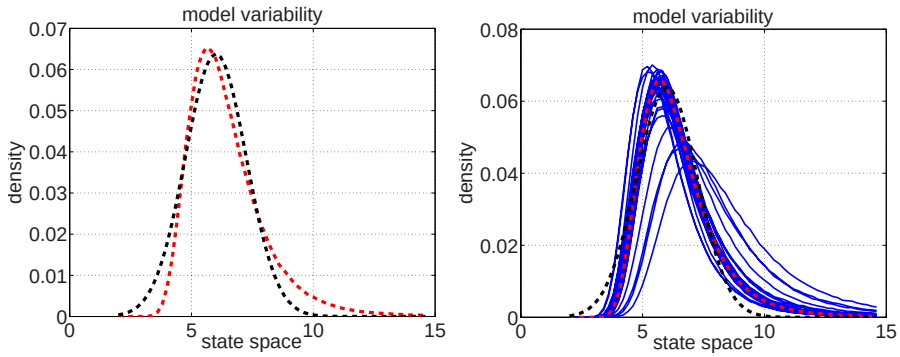

Figure 4.1: The light line symbolizes the optimal model variability and the black the data distribution. The dark distributions on the right show intermediate results of the model variability distribution within the iterative optimization process.

Interpretation. A classical frequentist regression for B would result in $B = 1/2$. With that value, the model would exactly match the data value 6, which is the mean value for the data distribution used.

In the overlap setting, the best way for the model to match the data is to interpret the parameter B to be normally distributed with $B \sim \pi(B) \sim \mathcal{N}(0.52, 0.103)$. That means, the model variability and data peaks are apart. This shift contradicts the usual intuition in terms of maximum-likelihood: There the peaks, symbolizing the highest probabilities of data realization, would match.

Looking at the shapes of the distributions, one is not surprised about the overlap result. In a very sloppy language: Due to the skewness of the model variability distribution, its mean has to be shifted to the ”left” in order for the bigger ”right tail” of the distribution to cover more data distribution. Of course, the results of the classical and the overlap optimization do not differ much. Due to the simplicity of the problem, this is not surprising. However, it does show that the model–data–overlap interpretation differs to the one of purely calibration methods of section 3.1.

Difference to classical concepts. At this points, it is worthwhile to generally see how different the classical residual and the overlap concept work for models. In the following one has to distinguish between the residual and the overlap case. As a target function for the residual concept take for example a goodness–of–fit criterium, similar to (3.13). Let $\boldsymbol{\theta}_{\mathcal{R}}$ be the optimal parameters in the sense of $\mathcal{F}_{\mathcal{R}}$; associated with them is a single trajectory, $\mathbf{y}_{\mathcal{R}}(t) = \Phi^t_{\theta_{\mathcal{R}}}\mathbf{y}_0$. In contrast, one has to consider the optimal distribution $\mathcal{M}_t(\pi_{\mathcal{O}})$ resulting from the optimal parameter distribution $\pi_{\mathcal{O}}$. If one wants to select a single trajectory representing this distribution, one should take the average trajectory

$$\mathbf{y}_{\mathcal{O}}(t) \;=\; \mathbf{E}\mathcal{M}_t(\pi_{\mathcal{O}}), \tag{4.9}$$

where the expectation $\mathbf{E}$ is taken for each instance t separately.

To graphically illustrate the conceptual difference between the overlap $\mathcal{F}_{\mathcal{O}}$ and the classical residual approach $\mathcal{F}_{\mathcal{R}}$, consider two different candidate models M_1 and M_2 like in figure 4.2. For simplicity, one assumes that the $\mathcal{F}_{\mathcal{R}}$-optimal trajectory $\mathbf{y}_{\mathcal{R}}$, and the mean trajectory $\mathbf{y}_{\mathcal{O}}$ of the $\mathcal{F}_{\mathcal{O}}$-optimal distribution $\mathcal{M}(\pi_{\mathcal{O}})$ coincide.

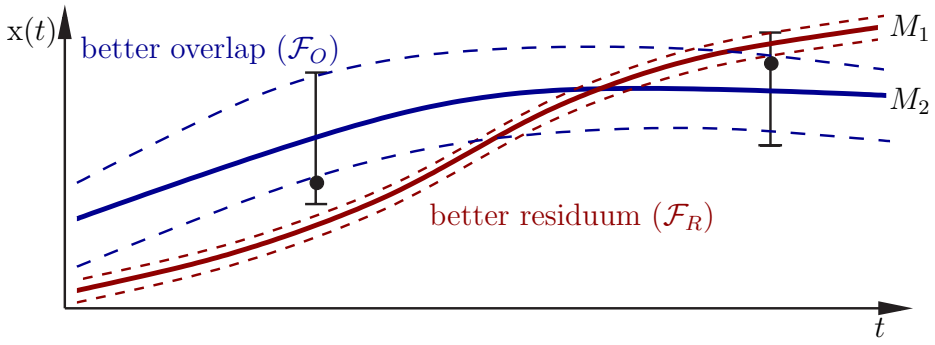

Figure 4.2: Different residual and overlap interpretation: The 95 %-confidence intervals of the data distribution $\mathcal{D}_t$ are symbolized by error bars, the inner 95 %-quantils of the model variability distribution $\mathcal{M}_t(\pi_{\mathcal{O}})$ by dashed strips around each fitted trajectory. A smaller residual (lower model) does not imply a large model–data–overlap (upper model) and vice versa.

The lower candidate model M_1, shows the smaller residual, but due to its low model variability, it does not reproduce the data distributions as well as the other one. In the residual framework, the candidate model M_1 is preferred. In the overlap interpretation, however, one would prefer the other one with the higher model variability, namely model M_2. Due to its higher model variability it matches the data distribution better and has therefore the higher capability of

reproducing the data distributions.

Compared to figure 4.2, a more extreme setting is shown in figure 4.3. Due to non-existing model variability for the left model at the measuring point $t = 1$, there is no probability to reproduce any data given by $\mathcal{D}_1$. On the other hand, the right model is capable of taking values that can be justified by $\mathcal{D}_1$.

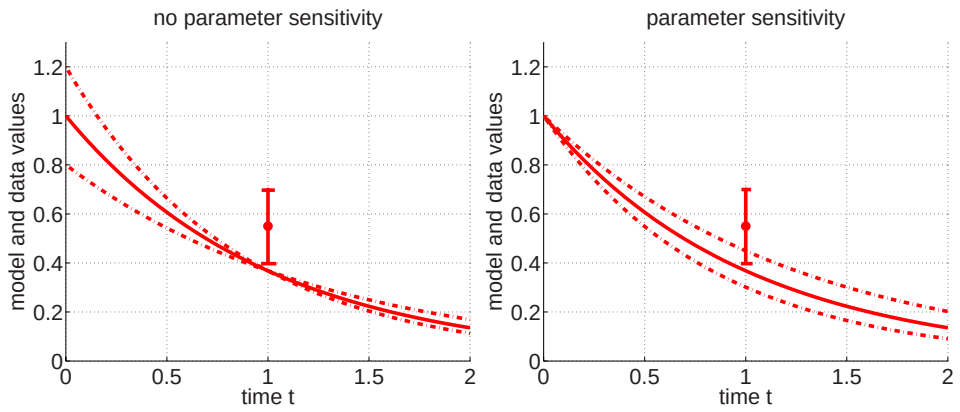

Figure 4.3: In both examples, the residual is the same. A non vanishing probability to reproduce the data is not given in the left, but in the right model.

Both illustrations show that the distance between data and model alone does not give sufficient information about the quality of the model–data–fit in the sense of reproducibility.

Interpretation overlap and variability. The overlap optimization delivers two quantities: the overlap number itself and the hyperparameters of the parameter distribution π. Besides the qualitative ranking, the hyperparameters can be used for further interpretation, namely for calculating statistical momenta of the parameter distribution. The expected value can be associated and interpreted as it has been done in (4.9). The variance can be used as a sensitivity information. It shows how data variability effects the parameter estimation result. It can therefore be used when validating and assessing the models suitability.

Comparison with the Bayesian approach. After looking at the comparison to the residual framework, it is now time to do so for Bayesian methods. In the

frequentist's world, parameter estimation and model discrimination are consecutive steps. By optimizing the overlap functional $\mathcal{F}_{\mathcal{O}}$ of (2.4), one simultaneously estimates the parameters as well as the overlap number that allows for ranking and for conducting model discrimination. A similar direct one–step–evaluation can also be found in the Bayesian factors concept of (3.53). However as mentioned before, the Bayesian approach cannot be applied for model uncertainty that is caused by a systematic model–data–deviation. In the Bayesian approach, uncertainty is attributed to the model in total rather than dealing with the uncertainty within the model. In contrast, for the overlap approach, uncertainty is interpreted through matching model and data variability.

Another difference between the two concepts concerns the different interpretation of the distributed parameters within the parameter estimation. In the Bayesian setting, the apriori parameter distribution is a given belief that does not apply when constructing intermediate models. The aposteriori distribution for the parameter reveals the certainty given by the parameter prior and the likelihood of the data instead of some structural model–data–uncertainty. For the model–data–overlap, the distribution shows the uncertainty from the model and data.

Optimization scheme. As mentioned before, the calculation of the overlap functional is algorithmically speaking an optimization. The entities to optimize are the hyperparameters of the parameter distribution π. To avoid confusion, the term "parameter uncertainty" is subsequently used to name the influence of perturbed data on a calibrated model. On the contrary, "parameter variability" combines both effects: perturbed data and model–data–deviation.

By matching the shapes of data and model variability distribution, the parameter variability itself becomes a result of the optimization. In the frequentist setting, the parameter uncertainty is a result of a downstream calculation as seen in figure 4.4. It shows again that the overlap combines parameter estimation and the calculation of a model ranking entity.

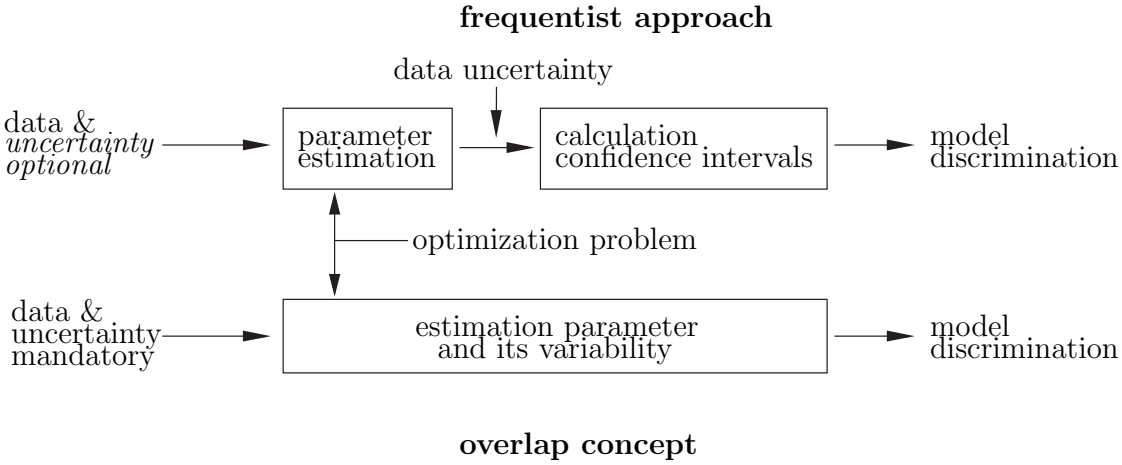

Figure 4.4: Parameter variability as optimized quantity

There is a fundamental difference between the existing concepts when it comes to the input entities. In order to conduct the overlap optimization, one explicitly has to include the quantitiative information on the data variability, for example in terms of a distribution. This means, the data deviation becomes input data itself.

Hybrid methods. There is also a possibility of considering hybrid models. When hardly no uncertainty is expected in some dimensions of the models, than one could use classical residual for those dimensions, whereas for other dimensions the overlap could be used. This aspect is, however, not further intensified within this thesis paper.

4.4 Linear case

If the data and model variability are normally distributed with $\mathcal{N}(\mu_D, \sigma_D)$ and $\mathcal{N}(\mu_M, \sigma_M)$, respectively, then the overlap can be calculated analytically. A detailed derivation for the one–dimensional case is shown in appendix A

$$\mathcal{F}_L = \sqrt{\frac{2\,\sigma_M\,\sigma_D}{\sigma_M^2 + \sigma_D^2}}\; e^{-\frac{(\mu_M-\mu_D)^2}{2(\sigma_M^2+\sigma_D^2)}}. \tag{4.10}$$

Discussion. This simplified scenario of (4.10) allows discussing overlap characteristics. In figure 4.5, the one–dimensional case is documented. The data

variability is distributed with $\mathcal{N}(0, 1)$.

For constant model variability, the overlap is also Gaussian. For a constant mean deviation between model and data peak, the overlap shows a different behavior. The larger the deviation, the larger the optimal variance becomes. It shows that the model deviation, due to uncertainty has to be compensated in terms of a higher variability. Further, for each non–optimal overlap number, there are two

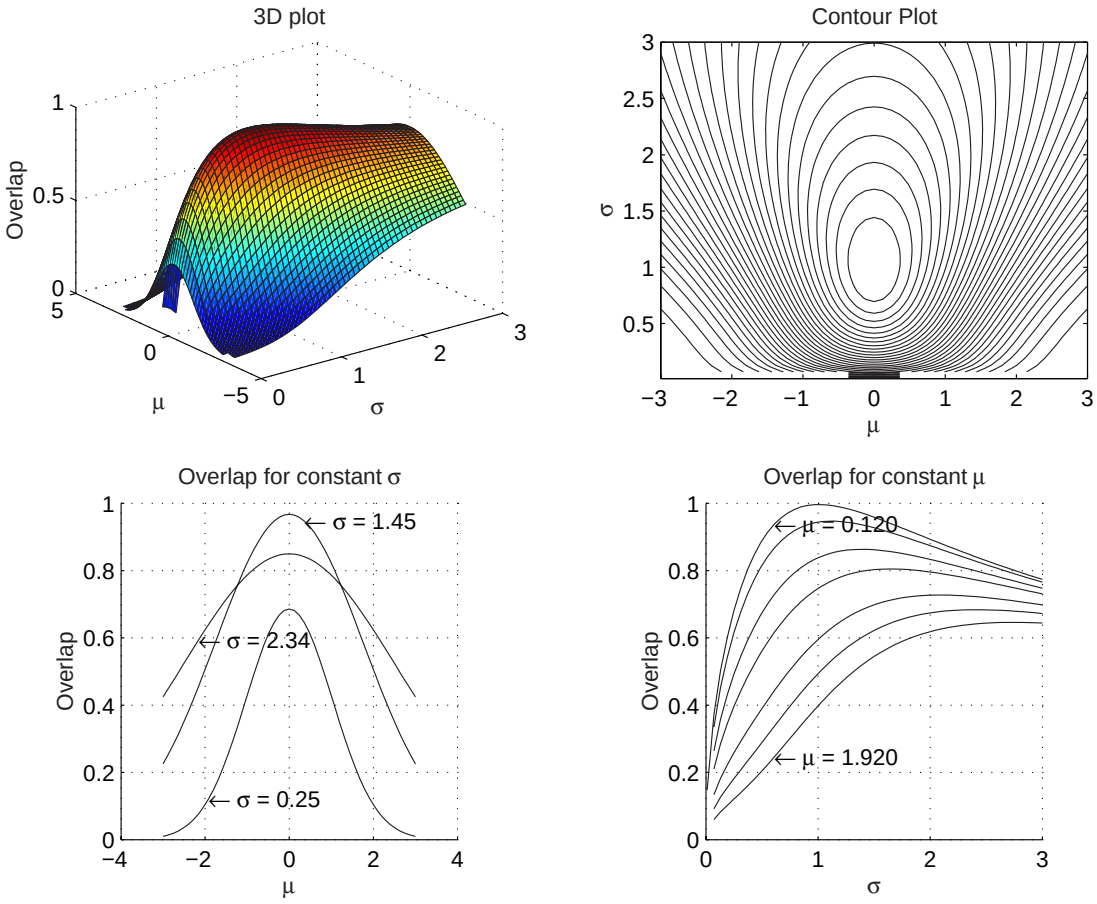

Figure 4.5: Overlap normally distributed variabilities

realizations for the variability. So, one cannot uniquely associate one variability for each overlap number and mean deviation (see figure 4.6).

Looking at the contour plot in figure 4.5 and the double scenario shown in figure 4.6, one wants to enforce a small mean deviation in the sense of (4.10), in order to achieve a low variability as well. When algorithmically performing the

overlap optimization, it is advisable to introduce a punishment term to prevent large mean deviation.

The lower subplots in figures 4.5, again explicitly show the possibility of having the same overlap number for two different σ_M. As it can be seen in the upper right subplot of figure 4.5, for some given model mean μ, the maximum overlap is not unique.

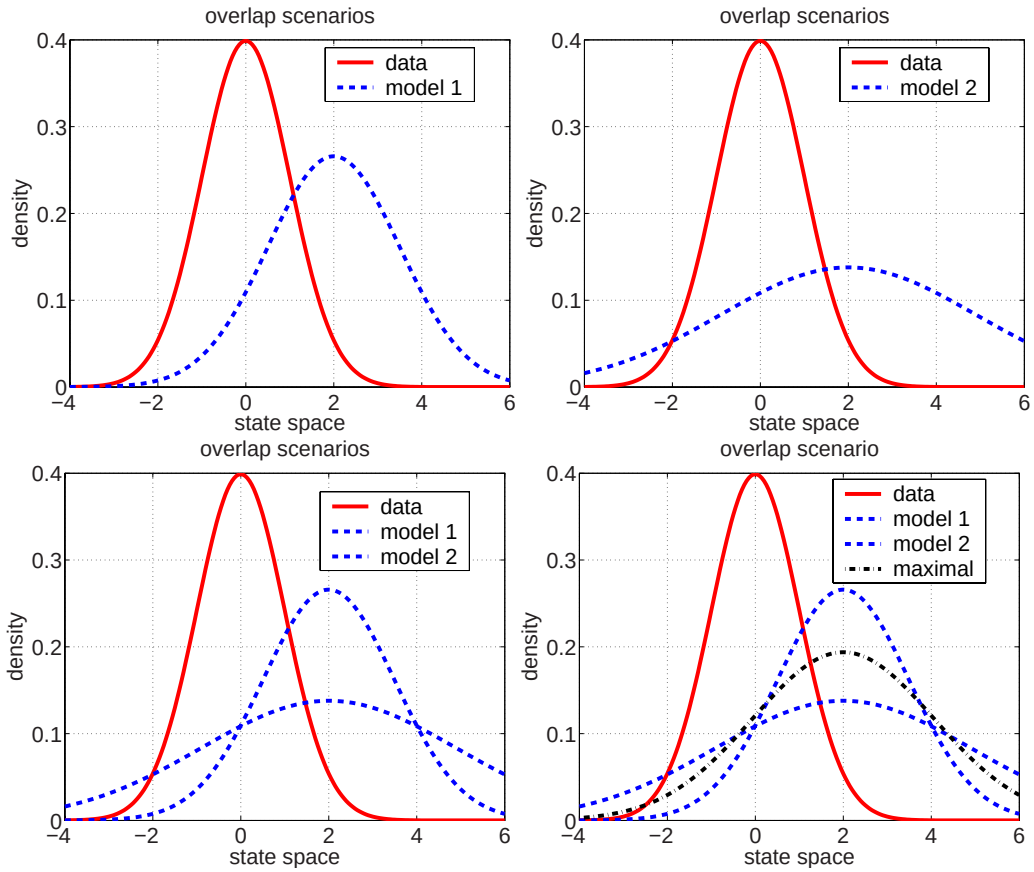

Figure 4.6: The upper two plots show have the same overlap number, the optimal overlap is shown in the lower right figure.

Maximum. For a constant mean deviation, one can also calculate the optimal variance of the model variability analytically.

$$\sigma_M = \sqrt{(\mu_M - \mu_D)^2 + \sqrt{(\mu_M - \mu_D)^4 + \sigma_D^4}} \;\geq\; \sigma_D \tag{4.11}$$

A detailed derivation can be found in appendix A. The equation (4.11) is illustrated in figure 4.7. For maximal overlap, the variances σ_M and data σ_D do not coincide. Nonsurprisingly, a systematic model–data–gap is accompanied by an increased uncertainty expressed in a higher model variability.

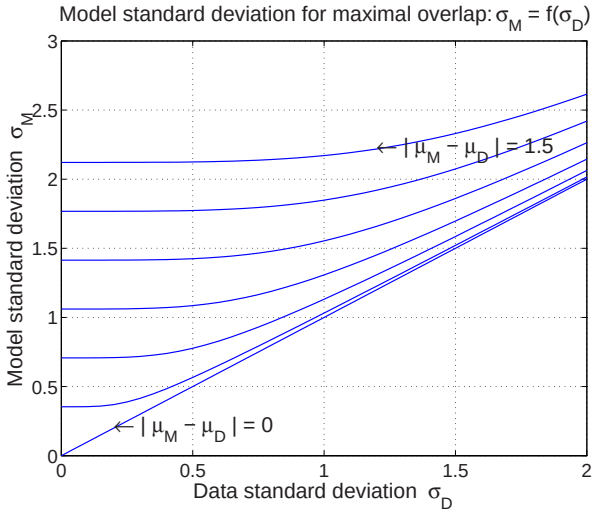

Figure 4.7: Model σ_M and data σ_D standard deviation for maximal overlap

Linear Approximation. The section shall be closed by mentioned that for locally constant variances in (4.10), the traditionally least square criterium is the linear approximation for the overlap criterium for small model–data–deviations.