

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/138823> holds various files of this Leiden University dissertation.

Author: Chen, X. (G.)

Title: Prediction sets via parametric and nonparametric Bayes: With applications in pharmaceutical industry

Issue date: 2021-01-05

**Prediction Sets via Parametric and Nonparametric Bayes:
with Applications in Pharmaceutical Industry**

Proefschrift
ter verkrijging van
de graad van Doctor aan de Universiteit Leiden
op gezag van Rector Magnificus Prof. mr. C. J. J. M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op dinsdag 5 januari 2021
klokke 15:00 uur

door

Xiangyi (Gregory) Chen

geboren te Xiamen, China
in 1987

Promotor: Prof.dr. A.W. van der Vaart

Samenstelling van de promotiecommissie:

Prof.dr. E.R. Eliel

Prof.dr. J.J. Meulman

Prof.dr. E.R. van den Heuvel (Technische Universiteit Eindhoven)

Prof.dr. J.J. Goeman

Dr. T. W. Nagler

To my dear mother, Ms. Luo Yan-Peng

Table of contents

Introduction	1
1 Parametric Prediction Region based on LMM	5
1.1 Introduction	5
1.2 Definitions and setup	7
1.2.1 Frequentist definitions	7
1.2.2 Bayesian definitions	8
1.2.3 Comparison	9
1.2.4 One-sided and two-sided tolerance intervals	10
1.3 Normally distributed future observation	10
1.3.1 Computational strategy	13
1.4 Data from a linear mixed model	17
1.4.1 Approximations to the conditional posterior mean	18
1.4.2 Numerical examples	20
1.5 Frequentist justification of the Bayesian procedure	22
1.5.1 Normal predictions	25
1.6 Annex : Extra Insights on KMW solution in Section 1.3.1	29
1.7 Annex: Gibbs Sampler for one-way Random Effect model	29
1.7.1 Scenario 1: Formulation (a) and Prior (a)	30
1.7.2 Scenario 2: Formulation (a) and Prior (b)	32
1.7.3 Scenario 3: Formulation (b) and Prior (a)	33
1.7.4 Scenario 4: Formulation (b) and Prior (b)	34

1.7.5	Additional Note	35
2	Another Parametric Application:	
	Setting Control Limit at Release for Stability Assurance	37
2.1	Introduction	37
2.2	Problem formulation	38
2.3	Benchmark approach in the industry	39
2.4	Our Population Model	40
2.5	CoT formulation	41
2.5.1	CoT in population model	41
2.5.2	Sample Inference for CoT	42
2.6	CoI formulation	42
2.6.1	CoI in population model	42
2.6.2	Practicality of CoI	44
2.6.3	Sample Inference for CoI	46
2.7	Calculation with Example datasets	49
2.8	Conclusion	50
2.9	Proofs	51
2.10	Annex: Asymptotics of the Bayesian Estimators for CoI Formulation	54
3	Nonparametric Prediction Regions	59
3.1	Introduction	59
3.2	Univariate Nonparametric Prediction Intervals	60
3.3	Data Depth and DD-based Prediction Sets	61
3.3.1	Data Depth	61
3.3.2	DD-based Prediction Sets	63
3.4	Conformity Scores and Conformal Prediction Sets	64
3.5	Bayesian Nonparametric Prediction Sets	65
3.6	Bayesian Nonparametric Prediction Sets via Dirichlet Mixtures	66
3.6.1	Posterior Sampling and SUGS	67

3.6.2	SUGS Drawbacks and Remedies	71
3.6.3	Derive a Prediction Set	72
3.7	Numerical Evaluation	75
3.7.1	Univariate Case	76
3.7.2	Multivariate case	78
3.8	Annex: Stick-Breaking Process	83
3.9	Annex: Generalized Pólya Urn Scheme	83
3.10	Annex: Pitman-Yor and Dirichlet Process Mixture Model	84
3.10.1	Block Gibbs Sampler in Stick-Breaking Scheme	85
3.10.2	Gibbs Sampler in Pólya Urn Scheme	88
4	Nonparametric Tests and Prediction Sets for Functional Data	91
4.1	Introduction	91
4.2	Functional Data Depth (FDD)	92
4.2.1	Definition and Overview	92
4.2.2	ID and WID	93
4.2.3	PBD, RPD and MPD	95
4.2.4	BD and MBD	96
4.2.5	HRD and MHRD	97
4.3	Topic I - Visualization of Outliers	98
4.4	Topic II - Prediction Sets	102
4.5	Topic III - Two Functional Sample Testing	105
4.5.1	Rank Test	105
4.5.2	Permutation Test	106
4.5.3	Bootstrap Test	106
4.5.4	Example Analysis	106
5	An Application: Image-Based Filter for Silicon Oil Particles	109
5.1	Introduction	109
5.2	Industrial Background	110

5.3	Three Benchmark Approaches	111
5.4	Data and Image Pre-processing	112
5.5	Construction of Image-Based (IB) Filter	115
5.5.1	IB-Test 1: General SO Pattern Recognition	115
5.5.2	IB-Test 2: Reference books of SO stereotypes	120
5.6	Performance Evaluation	121
5.7	Conclusion	125
5.8	Discussion	125
5.9	Appendix 1: Gallery on variations of SO/NSO images	128
5.10	Appendix 2: Gallery of the Reference book in Test 2 of IB-Filter	132
5.11	Appendix 3: Gallery of example images from the testing dataset.	134
References		137
Summary		143
Samenvatting		147
Acknowledgment		151
Curriculum Vitae		153

Introduction

This thesis consists of five chapters on how to construct prediction sets for different types of data and models. These chapters can be treated as individual articles. The motivation of the models comes from applications in pharmaceutical manufacturing and quality control.

Statistics in Pharmaceutical Manufacturing (PM)

Manufacturing of a drug product starts on a small scale for clinical trials, and then is elevated to a much larger, commercial scale once regulators have approved its efficacy and safety based on trial outcomes. Scaling up a manufacturing process is usually not straightforward. The key aim is to ensure a stable process that can robustly and continuously deliver high quality drug products. While the best definition of drug quality might be its efficacy and safety, it is infeasible (and unethical) to measure them for every manufactured batch before it is released to the patients. Hence, prior to commercial launch, regulators and the manufacturer determine various attributes and their corresponding release criteria (the release specification) in the filing of the drug ¹. These attributes describe chemical, biological and/or physical aspects of the drug product that are relevant to its clinical efficacy and safety. Laboratory methods to measure these attributes typically also need refinement or to be newly developed in order to achieve satisfactory accuracy and precision.

Statistical applications in pharmaceutical manufacturing can emerge anywhere that data are generated. The majority of them concern the following three subjects: the manufacturing process, laboratory methods, and quality control and assessment. Applying a suitable experimental design and statistical analysis helps to create high-performance manufacturing processes and laboratory methods, to provide rigorous evidence to the regulators during the validation phase, to monitor performance across different production sites throughout the commercial phase, and to shed light on root causes of emerging trouble. In addition, statistics are the key resort to determine a sound quality control strategy (e.g. setting alert limits or system suitability tests) or to perform assessments using probabilistic requirements (e.g. assessing stability risk of a batch given data of empirical batches).

¹ As a side remark to avoid an overly-simplified perception of the quality control effort on drug batch release: verifying that sample measures of all attributes meet their corresponding specifications is only a part of the release decision and only a small piece in the full chain of quality assurance on the continuous manufacturing process.

With the increasing awareness (and enthusiasm) for "data science" and the advancement of laboratory techniques, manufacturers are increasingly interested in exploring how to exploit metadata (e.g. auxiliary manufacturing system information) and new types of data (e.g. images or videos) to enhance the understanding of the true quality of a whole batch of drugs, and to improve efficiency and reduce the cost of a process or a laboratory. For example, a chromatography profile (e.g. a response factor² as function of time) is traditionally reduced to an area under curve in a pre-specified time window after baseline adjustment, but now the whole profiles may be analyzed as functional data. Another example is the application of emerging imaging techniques to detect physical defects earlier in the manufacturing process.

Prediction Set and Its Usage in PM

A prediction set, also known as a tolerance set, is a region that will capture a prescribed proportion (say $1 - \delta$, where δ is typically in $[0.01, 0.1]$) of future observations. This prescribed proportion is also known as the *desired coverage*. The distribution of a future sample is typically unknown but can be inferred from past data. This brings in estimation uncertainty next to the randomness of future observations. If the past data is sufficiently informative, estimates of unknown parameters can simply be plugged in, and estimation uncertainty ignored. More realistically both sources of randomness should be taken into account, and it is desirable to make precise how the desired coverage $1 - \delta$ should be achieved. The two prevailing definitions are to require the region to have $1 - \delta$ coverage on average, or to have coverage at least $1 - \delta$ with a certain level of confidence. The "average" and the "confidence" here can refer to the sample distribution of the data in a frequentist sense, or refer to a posterior distribution in the Bayesian framework. These two formulations give different types of tolerance regions. In addition, special consideration might be needed to form a bounded prediction set. For example, there are multiple ways to derive a two-sided 95% prediction interval for a known univariate gamma distribution. One may choose to use the density level set that covers 95% of mass, or build the interval symmetrically around the mean, or take the lower and upper ends to be the 0.025 and 0.975 quantile respectively. In the multivariate case, the shape of a prediction set is even more adaptable to the purpose of the application.

In the realm of quality control deriving a prediction set is equally, if not more, common as making a point estimate or point prediction. Related tasks include: determining the empirical performance range of process attributes to detect a future abnormal operation, determining the representative range of quality attributes for continuous monitoring or for supporting a change of release specification, determining the regular range in a reference group and use it as criterion to check individuals from a test group in a comparability study.

²Typically, the ratio between a signal produced by an analyte of interest and the quantity of the analyte.

Outline of the Thesis

This thesis starts with reviewing and contrasting frequentist and Bayesian definitions of tolerance regions in Chapter 1. We give conditions under which for large samples a Bayesian tolerance region also has frequentist validity, and study this for smaller samples in a simulation study. A computational strategy is provided and compared with some benchmarks for computing a Bayesian two-sided tolerance interval for a Gaussian future variable, and this is applied in details to the case of possibly unbalanced linear mixed models. The motivation to consider a mixed-effect model comes from the context of quality control in pharmaceutical industry, where data usually include several replicates per batch and hence the sample are not i.i.d.

Chapter 2 is motivated by a common task in stability risk control, and involves deriving a conditional one-sided prediction limit in a linear random-coefficients model. Given empirical stability data, the shelf life (SL) and the stability specification, the objective is to determine a limit for a measurement at release (i.e. time zero) that incorporates stability risk. This is particularly important when equal release and stability specifications are registered for a product. We discuss the two practically most relevant formulations of "incorporate stability risk". The first one is to ensure that the calculated limit at release and the stability specification will define the riskiness of a future batch consistently, provided it has similar stability behavior as the empirical batches. The second one is to set a control limit at release that can screen out future batches that are at high risk of failing the stability specification at the end of their shelf life. This second formulation is natural for quality assurance and popular in practice, but turns out to be stringent. We provide insight when this control strategy is too ambitious and impractical to be implemented. We also introduce a heat-map visualization that allows practitioners to determine when the limits calculated in the second formulation are feasible, and sheds light on directions to improve in the other case. In addition, the current industrial benchmark is reviewed and contrasted to our two formulations. Computational algorithms for both formulations are given and illustrated on an example dataset.

Chapter 3 discusses how to construct the two types of Bayesian tolerance regions mentioned in Chapter 1 in a nonparametric setting. Past and future observations are assumed to be sampled from a unknown distribution F . Instead of assuming a parametric family for F and then setting a prior on the population parameters, we use a Dirichlet Process mixture (DPM) of Gaussian distributions as prior for F . We start with a brief introduction on the classical nonparametric prediction sets for univariate observations, and then proceed with an introduction to data depth. Data depths ranks data points in a center-outwards manner, with a higher depth value indicating that a point is more central relative to the sample or a distribution. This depth paradigm provides a way to build nonparametric prediction sets, that is particularly useful in the multivariate case, since ranking of higher-dimensional data points is not straightforward otherwise. After an introduction to data depth, detailed procedures are provided for constructing the two types of Bayesian tolerance regions with the aforementioned nonparametric prior for F . The validity of these regions is investigated by simulation under various univariate and

multivariate setups, where performance is also compared with several benchmark nonparametric approaches.

Chapter 4 discusses data depth methodology in functional data analysis for (i) outlier detection, (ii) constructing prediction sets, and (iii) performing two-sample testing. These topics are motivated by examples from bio-pharmaceutical manufacturing. We start with an overview of functional data depth (FDD), and propose two new depth measures. The new types of FDD are particularly useful in outlier detection when the population is multi-modal (e.g. a mixture of Gaussian processes with distinct mean functions), which commonly occurs in industrial practice. We illustrate the sensitivity of different FDDs to different types of outliers via a simulated dataset, and show the potential advantage of centering the curves before calculating FDDs. For topic (ii), we review two existing approaches to build a prediction set for functional data, and propose an alternative. The calculation procedure of our proposed approach is detailed via an example dataset. Next a simulation study is deployed to examine its validity for modest sample size. For topic (iii), we discuss three different approaches to two-sample testing with functional data. The performance of the three approaches is investigated based on a widely-known dataset.

Chapter 5 presents an application of multivariate data depth to build prediction sets for image recognition. The objective is to classify images as showing either a silicon oil (SO) or a non-silicon oil (NSO) particle. SO particles have a clear profile, but NSO particles may have many, also unknown forms. Therefore, it would be difficult to form a training set that contains sufficiently many examples of both types. Since a classifier based on supervised learning would not be able to generalize its success against NSO stereotypes that are not seen in the training data, our chosen method is based on agreement of a future image with the reference set of SO particle images. In machine learning terms our approach is unsupervised, rather than supervised. We start with an orientation on the industrial background and problem statement, and next provide a pre-processing procedure, where the main contribution is to standardize the input images of varying resolutions to the same dimension. The construction of our SO filter (unsupervised classifier) based on 798 raw training SO images is then elaborated. The performance of the filter is compared with three industrial benchmarks on a testing set with 1843 SO images and 1624 NSO images.

Chapter 1

Parametric Prediction Region based on LMM

1.1 Introduction

The concept of *tolerance region* is of central importance in quality control. A tolerance region is a prediction set for an unobserved or future observation, which takes account of both the random nature of this observation and the uncertainty about its distribution. Parameters of the latter distribution are estimated using past data, where it is desired to account for the statistical error resulting from simply plugging in the estimated values.

Conventional tolerance regions take the uncertainty of estimated parameters into account in one of two ways. Either the region captures the future observation a fraction of $1 - \alpha$ times on average over both future and past observations (the $(1 - \alpha)$ -*expectation tolerance region*), or it captures the future observation with probability at least $1 - \delta$ with $1 - \alpha$ confidence over past observations (the (δ, α) -*tolerance region*). (See Section 1.2 for precise definitions.) The second way appears to be preferred in the pharmaceutical industry. The “on average” and “confidence” can refer to the sampling distribution of the data in a frequentist sense, but can also refer to a posterior distribution in the Bayesian statistical framework. The main focus on the present paper is the second, but we do relate it to the frequentist setup.

Frequentist tolerance regions have been well studied in the literature. A general reference is the book [28], especially for the situation that the data are i.i.d. For the linear mixed model (LMM), the paper [51] provides an elegant solution to build one- and two-sided (δ, α) -tolerance intervals, and includes a comprehensive review of the literature. One purpose of the present paper is to provide a Bayesian approach for the general LMM.

The Bayesian formulation of tolerance regions dates back to at least 1964, but the subsequent literature is relatively small. In his paper [1] Aitchison derived Bayesian (δ, α) -tolerance regions from a decision-theoretic framework, and contrasted them to the frequentist counterparts. In [2] he extended his discussion to $(1 - \alpha)$ -tolerance regions, which are a natural way to build Bayesian prediction intervals.

A one-sided (δ, α) -tolerance interval for a univariate future observation is usually easy to compute, but two-sided tolerance intervals pose challenges, both conceptually and computationally. There are at least two common approaches: intersecting two one-sided tolerance intervals, or fixing one degree of freedom of the interval (e.g. the midpoint of the interval). The former approach is identical to specifying probability masses in the two tails of the distribution of the future variable separately and gives a valid construction, in view of Bonferroni's inequality, but it yields longer intervals than necessary. (They are called "anti-conservative" in the pharmaceutical industry in reference to the customers, whereas statisticians might use the term "conservative".) See [21] for an example application. The second approach, fixing one degree of freedom, is the conventional choice, especially in the frequentist framework, but requires untangling the dependence of the interval on the unknown true parameter. Solutions are often not available in analytical form and computationally more challenging. Wolfinger in [67] proposed an algorithm to derive a two-sided Bayesian interval for a future normal variate, which was refined by Krishnamoorthy and Mathew [28]. Their algorithms have been widely adopted in practice, and also in other literature (e.g. [27], [57], [56]).

Among other contributions to the literature we mention the following: [35] used the empirical Bayes method to construct a one-sided tolerance interval given an i.i.d. sample from a normal distribution; [22] derived a Bayesian tolerance interval that contains a proportion of observations with a specified confidence; [14] and [68] focused on the sample size needed to attain a certain accuracy; [57] and [56] allowed data from the unbalanced one-way random effects model and the balanced two-factor nested random effects model; [36] discussed probability matching priors (PMP) in the one-sided case to ensure second-order frequentist validity; [41] extended this to the two-sided case; [42] incorporated it to a balanced one-way random effects model, and evaluated its performance against the frequentist method MLS in [28].

Although the PMP approach has merit when the sample size is small, it is analytically demanding even when data are i.i.d., and it seems difficult to extend to the general LMM setting. The algorithms of Wolfinger [67] and Krishnamoorthy and Mathew [28] can be extended to LMM, but they overly simplify the target function during optimization and may result in less satisfactory performance.

In this paper we propose a computationally efficient solution for the general case that the future observation possesses a normal distribution. We show that this is easy to implement given any data model for which a sample from the posterior distribution is available. We investigate when the shortest interval is centered at the posterior mean of the parameter. We discuss the interval in particular for the linear mixed model, and within this context show its good performance by simulation. We illustrate

the method on an example that is representative for pharmaceutical applications. Finally we also prove that the Bayesian interval has frequentist validity in the case of large samples.

1.2 Definitions and setup

Given are observed data X , with a distribution P_θ depending on a parameter θ , and future unobserved “data” Z , with a distribution Q_θ depending on the same parameter θ . In both cases the sample space is arbitrary. A tolerance region is a set $\mathcal{R}(X)$ in the sample space of Z that captures Z with a “prescribed probability”. It will typically be constructed using the observation X to overcome the problem that θ , and hence the law of Z , is unknown. There are various ways to make the “prescribed probability” precise, and these can be divided into frequentist and Bayesian definitions. The probability statement will refer to both X and Z , and is fixed by one or two parameters α and δ , which are typically chosen small, e.g. 5%.

The parameter θ will typically be chosen to identify the distribution of Z . The distribution of X may also depend on unknown “nuisance” parameters. For simplicity of notation we do not make this explicit in the following. We shall use the notation P or P_θ for general probability statements, which may be reduced to P_θ or Q_θ if the event involves only X or Z .

1.2.1 Frequentist definitions

The most common frequentist definition is the (δ, α) -tolerance region. For a set R abbreviate $Q_\theta(R) = P_\theta(Z \in R)$. Then $\mathcal{R}(X)$ is an (δ, α) -tolerance region if

$$P_\theta \left(x: Q_\theta(\mathcal{R}(x)) \geq 1 - \delta \right) \geq 1 - \alpha, \quad \forall \theta. \quad (1.1)$$

If we let $Q_\theta(\mathcal{R}(X))$ denote the probability or coverage of $\mathcal{R}(X)$ under Q_θ , for X held fixed, then we can also write the display in the shorter form $P_\theta \left(Q_\theta(\mathcal{R}(X)) \geq 1 - \delta \right) \geq 1 - \alpha$, where the outer probability P_θ refers to X , and the inequality must hold for all possible values of the parameter θ . The latter reminds us of the definition of confidence sets, and indeed it can be seen that $\mathcal{R}(X)$ is a frequentist (δ, α) -tolerance region if and only if the set $\mathcal{C}(X) = \{\theta: Q_\theta(\mathcal{R}(X)) \geq 1 - \delta\}$ is a confidence set for θ of confidence level $1 - \alpha$.

An alternative is the α -expectation tolerance region, which requires that

$$\int Q_\theta(\mathcal{R}(x)) dP_\theta(x) \geq 1 - \alpha, \quad \forall \theta. \quad (1.2)$$

With the notational convention as before, the display can be written in the shorter form $E_\theta Q_\theta(\mathcal{R}(X)) \geq 1 - \alpha$, which is again required for all possible parameter values.

Both definitions have the form of requiring that $E_\theta \ell \left[Q_\theta \left(\mathcal{R}(X) \right) \right] \geq 1 - \alpha$, for all θ , and some given loss function ℓ . In the two cases this loss function is given by $\ell(q) = 1\{q \geq 1 - \delta\}$ for (1.1), and $\ell(q) = q$ for (1.2), respectively, where $1\{\cdot\}$ is the indicator function.

1.2.2 Bayesian definitions

In the Bayesian setup the parameter θ is generated from a prior distribution Π , and the densities p_θ and q_θ are the conditional densities of X and Z given θ , respectively. To proceed, it is necessary to make further assumptions that fix the joint law of (θ, X, Z) . The typical assumption is that X and Z are independent given θ .

A natural Bayesian approach is to refer to the predictive distribution of Z , and define a tolerance region $\mathcal{R}(X)$ to be a set such that $P(Z \in \mathcal{R}(X) | X) \geq 1 - \alpha$, i.e. a credible set in the posterior law of Z given X . The inequality can be written in terms of the posterior distribution $\Pi(\cdot | X)$ of θ given X as

$$\int P(Z \in \mathcal{R}(X) | X, \theta) d\Pi(\theta | X) \geq 1 - \alpha.$$

Under the conditional independence assumption this becomes

$$\int Q_\theta(\mathcal{R}(X)) d\Pi(\theta | X) \geq 1 - \alpha. \quad (1.3)$$

This is like a frequentist α -expectation tolerance region (1.2), but with the expectation with respect to X under P_θ replaced by the expectation with respect to θ under the posterior distribution.

An alternative, derived from a utility analysis by Aitchison [1], is the *Bayesian (δ, α) -tolerance region*, which is a set $\mathcal{R}(X)$ such that

$$\Pi\left(\theta: Q_\theta(\mathcal{R}(X)) \geq 1 - \delta | X\right) \geq 1 - \alpha. \quad (1.4)$$

This may be compared to (1.1). We can also say that $\mathcal{R}(X)$ is a *Bayesian (δ, α) -tolerance region* if and only if the set $\mathcal{C}(X) = \{\theta: Q_\theta(\mathcal{R}(X)) \geq 1 - \delta\}$ is a *credible set* at level $1 - \alpha$.

Both types of Bayesian regions satisfy $\int \ell \left[Q_\theta \left(\mathcal{R}(X) \right) \right] d\Pi(\theta | X) \geq 1 - \alpha$, for the appropriate loss function ℓ . Solving the region $\mathcal{R}(X)$ from such an equation may seem daunting, but good approximations may be easy to obtain using stochastic simulation. This is true even for complicated data models, as long as one is able to implement an MCMC procedure for generating a sample from the posterior distribution given X . We make this concrete in Section 1.3.1 for a Gaussian variable Z , and illustrate this in Section 1.4 for an unbalanced linear mixed model (LMM).

1.2.3 Comparison

The frequentist and Bayesian definitions differ in the usual way in that the frequentist probability in (1.2) and (1.1) refers to the possible values of x in the sample space, whereas the Bayesian probability in (1.3) and (1.4) conditions on the observed value of X and refers to the distribution of the parameter.

As is the case for credible sets versus confidence sets, the Bayesian approach may feel more natural.

An advantage of Bayesian tolerance set is that while their form (1.4) is determined by the future variable Z , through the prediction problem Q_θ , the model for the data X enters only through the posterior distribution $\Pi(\theta \in \cdot | X)$. If in the former the dependence on the parameter θ is not too complicated, then the problem is solvable for even complicated data models. In contrast, the frequentist problem permits explicit solutions only in very special cases, although approximations and asymptotic expansions may extend their use (see [51]).

Neither the frequentist nor the Bayesian formulation restrains the shape of the region $\mathcal{R}(x)$. One may prescribe a fixed form and/or seek to optimize the shape with respect to an additional criterion, such as the volume of the region. The Bayesian formulation is again easier to apply, as the optimization will be given the data X . In the case of frequentist region it may be necessary to optimize an expected quantity instead.

In general the two approaches give different tolerance regions, but the difference may disappear in the large sample limit. The requirements of the frequentist and Bayesian *tolerance regions* $\mathcal{R}(X)$ for loss function ℓ and level α , can be given symmetric formulations, as:

$$\mathbb{E} \left(\ell \left[Q_\theta \left(\mathcal{R}(X) \right) \right] | \theta \right) \geq 1 - \alpha, \quad \forall \theta, \quad (1.5)$$

$$\mathbb{E} \left(\ell \left[Q_\theta \left(\mathcal{R}(X) \right) \right] | X \right) \geq 1 - \alpha. \quad (1.6)$$

In the first the expectation is taken with respect to X , which gives an integral over x with respect to the density p_θ . In the second the expectation is relative to θ , which leads to an integral relative to the posterior distribution given X . The integral of the first relative to the prior is identical to the integral of the second relative to the Bayesian marginal distribution of X , but there is no reason that a Bayesian tolerance region would also be a frequentist tolerance region.

However, Bayesian and frequentist inference typically merge if the informativeness in the data tends to a limit. For instance, this is true for regular parametric models in the sense that Bayesian credible sets are frequentist confidence sets, in the limit, with corresponding levels. The prior is then washed out and the Bayesian credible sets are equivalent to confidence sets based on the maximum likelihood estimator. This equivalence extends to tolerance regions, under some conditions. We defer a discussion to Section 1.5.

1.2.4 One-sided and two-sided tolerance intervals

For a one-dimensional future variable Z it is natural to choose $\mathcal{R}(x)$ an interval in the real line. The endpoints of such an interval are referred to as *tolerance limits*.

The single finite tolerance limit of a Bayesian *one-sided interval* is determined by meeting the (δ, α) or α tolerance criterion. The pair of tolerance limits of a Bayesian *two-sided interval* might be optimized to give an interval of minimal length, next to requiring that the tolerance criterion is met.

One-sided tolerance limits possess a straightforward interpretation and implementation. In particular, the (δ, α) -type has a simple description in terms of confidence intervals and posterior quantiles:

- $(-\infty, U(X)]$ is a frequentist (δ, α) -tolerance interval if and only if it is a $(1 - \alpha)$ -confidence interval for the induced parameter $Q_\theta^{-1}(1 - \delta)$; it is a Bayesian (δ, α) -tolerance interval if $U(X)$ is the $(1 - \alpha)$ -quantile of the posterior distribution of $Q_\theta^{-1}(1 - \delta)$ given X .
- $[L(X), \infty)$ is a frequentist (δ, α) -tolerance interval if and only if it is a $(1 - \alpha)$ -confidence interval for $Q_{\theta+}^{-1}(\delta)$; it is a Bayesian (δ, α) -tolerance interval if $L(X)$ is the $(1 - \alpha)$ -quantile of the posterior distribution of $Q_{\theta+}^{-1}(\delta)$.

Here $Q_\theta^{-1}(u) = \inf\{z: Q_\theta(-\infty, z] \geq u\}$ is the usual quantile function of Z , and $Q_{\theta+}^{-1}(u)$ is the right-continuous version of this quantile function. (The distinction between the two is usually irrelevant, and linked to the arbitrary convention of including the boundary point in the tolerance intervals.) The assertions follow by inverting the inequality $Q_\theta(\mathcal{R}(X)) \geq 1 - \delta$, using the fact that for a cumulative distribution function Q and its quantile functions: $Q^{-1}(u) \leq x$ if and only if $u \leq Q(x)$, and $Q_+^{-1}(u) < x$ if and only if $u < Q(x-)$, for $u \in [0, 1]$ and $x \in \mathbb{R}$.

A valid two-sided interval might be constructed as the intersection of two one-sided intervals, each at half of the error rate, but this will be conservative and lead to needlessly wide intervals. It makes good sense to try and minimize the length of the interval. We consider this in the next section for the case that the future observation Z is univariate Gaussian.

1.3 Normally distributed future observation

Consider the case that the future observation Z is univariate Gaussian with mean ν and variance τ^2 . Thus the parameter is $\theta = (\nu, \tau)$, and $Z \sim Q_\theta = N(\nu, \tau^2)$. The probability that the future observation is captured within a candidate tolerance interval $[L, U]$ is

$$Q_\theta[L, U] = \Phi\left(\frac{U - \nu}{\tau}\right) - \Phi\left(\frac{L - \nu}{\tau}\right).$$

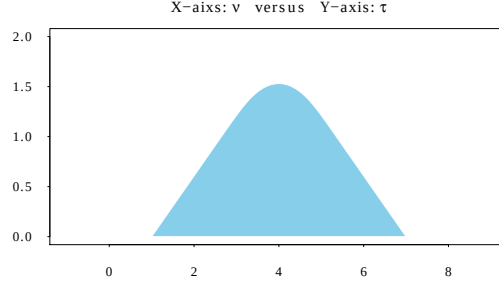


Fig. 1.1 The set $G_{A,B,\delta}$ of pairs (v, τ) such that $\Phi((U-v)/\tau) - \Phi((L-v)/\tau) \geq 1 - \delta$, for $A = 4, B = 3, \delta = 0.1$. The number A is its horizontal point of symmetry and B is the half-length of its base. The base of the set (the line segment at height $\tau = 0$) corresponds to the tolerance interval $[L, U]$.

It is convenient to parametrize the interval $[L, U]$ by its midpoint $A = (L + U)/2$ and half length $B = (U - L)/2$. For given $[L, U]$, or $\{A, B\}$, and $\delta \in (0, 1)$, define the set

$$G_{A,B,\delta} = \{\theta = (v, \tau) : Q_\theta[L, U] \geq 1 - \delta\}.$$

For given $[L, U]$ the set $G_{A,B,\delta}$ is shaped as in Figure 1.1. It is symmetric about the vertical line $v = (L + U)/2$ and intersects the horizontal axis $\tau = 0$ in the interval $[L, U]$. Changing A moves the set $G_{A,B,\delta}$ horizontally, while changing B changes its shape, with bigger B making the set both wider and taller. Although we use the normal distribution as our example, similarly shaped sets and conclusions would be obtainable for other unimodal symmetric distributions.

It follows that $\mathcal{R}(X) = [L(X), U(X)]$ satisfies inequality (1.1) and hence is a frequentist (δ, α) -tolerance interval for Z if and only if

$$P_\theta(x : \theta \in G_{A(x), B(x), \delta}) \geq 1 - \alpha, \quad \forall \theta.$$

In other words $[L(X), U(X)]$ is an (δ, α) -tolerance interval for Z if and only if $G_{A(X), B(X), \delta}$ is an $(1 - \alpha)$ -confidence region for $\theta = (v, \tau)$. Setting a joint confidence interval for location and dispersion is a familiar problem, but here the shape is restrained to the form $G_{A,B,\delta}$ and the focus will be on minimizing $B = B(X)$ (in some average sense). Solutions will depend on the type of data X . Standard solutions are available in closed form for the simplest models, and more generally as approximations.

Similarly $\mathcal{R}(X) = [L(X), U(X)]$ satisfies inequality (1.4) and hence is a (δ, α) -Bayesian tolerance interval for Z if

$$\Pi(\theta : \theta \in G_{A(X), B(X), \delta} | X) \geq 1 - \alpha. \quad (1.7)$$

It is natural to choose $A(X)$ and $B(X)$ to satisfy this inequality in such a way that $B(X)$ is minimal. In the resulting optimization problem the posterior distribution $\Pi(\theta \in \cdot | X)$ is a fixed probability distribution on the upper half plane and optimization entails shifting and scaling the shape shown in Figure 1.1 in a position such that it captures posterior mass at least $1 - \alpha$, meanwhile minimizing

its width. In Section 1.3.1 we show how to achieve this given a large sample from the posterior distribution.

The data X determines the posterior distribution, but does not enter the optimization problem. The parameter θ may not be the full parameter characterising the distribution of X , but our computational strategy will work as long as θ is a function of this full parameter. For instance, if X follows a linear regression model with predictor “time” and Z is an observation at time 0, then v will be a function the regression intercept; if X follows a random-effects model, then typically v will depend on the fixed effects and τ^2 will be a specific linear combination of the variance components, depending on practical interests.

Frequentist methods typically choose $A(X)$ equal to a standard estimator of v . One might guess that the Bayesian solution will be to take $A(X)$ equal to the posterior mean $E(v|X)$ of v . This would be convenient as it would reduce the optimization of (A, B) to the problem of only optimizing B . However, the posterior mean does not necessarily give the minimal length interval. The following lemma gives a sufficient condition.

Lemma 1.3.1. *Suppose that the conditional distribution of v given (X, τ) is unimodal and symmetric with decreasing density to the right of its mode and has mean $E(v|X, \tau)$ that is free of τ . Then the shortest (δ, α) -Bayesian tolerance interval $[L, U]$ for a future variable $Z \sim N(v, \tau^2)$ is centered at the posterior mean $E(v|X)$.*

Proof. We can decompose the probability on the left side of (1.7) as

$$\Pi(G_{A,B,\delta}|X) = \int \Pi(v \in (G_{A,B,\delta})_\tau | X, \tau) d\Pi(\tau|X),$$

where $(G_{A,B,\delta})_\tau$ is the section of $G_{A,B,\delta}$ at height τ . By the unimodality and monotonicity the integrand is maximized over A for every τ and a given B by choosing $A = E(v|X, \tau)$. If this does not depend on τ , then this common maximizer A will maximize the whole expression. Since we need to determine A and B so that the expression is at least $1 - \alpha$, maximizing it over A will give the minimal B . By assumption $A = E(v|X, \tau) = E(v|X)$. ■

The condition of the preceding lemma is not unreasonable, but depends on the prior, as illustrated in the following simple example. In Section 1.4 we show that for a LMM the condition is approximately satisfied. Then choosing A equal to the posterior mean is a fast computational shortcut that may perform almost as well as the optimal solution.

Example 1.3.1. The simplest possible data model is to let $X = (X_1, \dots, X_n)$ be a random sample from the $N(v, \tau^2)$ -distribution. This example was already discussed by Aitchison [1].

For the standard priors $\tau^2 \sim IG(\alpha_0, \beta_0)$ and $v|\tau \sim N(a, \tau^2/b)$, the conditional posterior distribution of v given τ is normal with mean

$$E(v|\tau, X) = \frac{ba + n\bar{X}}{b + n}.$$

Since this is independent of τ , the preceding discussion shows that the shortest (δ, α) -tolerance interval is centred at this posterior mean.

Choosing the prior variance $\text{var}(v|\tau)$ proportional to τ^2 , which is customary, is crucial for this finding. For instance, if we set the prior v to be independent of τ , say $v|\tau \sim N(a, b^{-1})$, then the conditional posterior mean changes to

$$E(v|\tau, X) = \frac{ba\tau^2 + n\bar{X}}{b\tau^2 + n}.$$

This is $\bar{X}$ if $\tau = 0$ and shrinks to the prior mean a as $\tau \rightarrow \infty$. For illustration, let $a = 0$, $b = 0.1$, $\alpha_0 = \beta_0 = 0.01$. Given data with $n = 3$, $\bar{x} = 10$, $s^2 = 1$, we approximated the posterior distribution of (v, τ) given X by a Gibbs sampler, using the full conditional posteriors, where $s^2 = \sum (x_i - \bar{x})^2 / (n - 1)$:

$$\begin{aligned} v|\tau, X &\sim N\left(\frac{ba\tau^2 + n\bar{X}}{b\tau^2 + n}, \frac{\tau^2}{b\tau^2 + n}\right), \\ \tau^2|v, X &\sim IG\left(\alpha_0 + \frac{n}{2}, \beta_0 + \frac{(n-1)s^2 + n\bar{X}^2 + nv^2 - 2nv\bar{X}}{2}\right). \end{aligned}$$

The contour plots of the posterior distribution in the left panel of Figure 1.2 show dependence between v and τ given X , and ensuing functional dependence of $E(v|\tau, X)$ on τ . Using Algorithm 2, as explained in Section 1.3.1, we computed the shortest tolerance interval (i.e. the smallest $\hat{B}$) for every possible location A of the interval, for A in a neighbourhood of the posterior mean $E(v|X)$. The right panel of Figure 1.2 shows $\hat{B}$ as a function of $\hat{A}$. The minimum value is *not* taken at the location of the posterior mean $E(v|X)$, which is indicated by a dashed line.

Admittedly the data in this example has been tweaked to illustrate the principle. Inspection of the vertical scale for $\hat{B}$ shows that the global minimal length of a tolerance interval is only slightly smaller than the length of the interval centred at the posterior mean of v . In Section 1.4 we show, by theoretical derivation, a similar phenomenon for linear mixed models, and in Section 1.5 we study this approximation in a large sample context.

1.3.1 Computational strategy

In this section we elaborate on the computation of the two-sided (δ, α) Bayesian tolerance interval for a normally distributed univariate future variable Z , as discussed in Section 1.3. We also compare our approach to the one taken in [67] and [28]. We assume given a large sample of values $\theta_j = (v_j, \tau_j)$

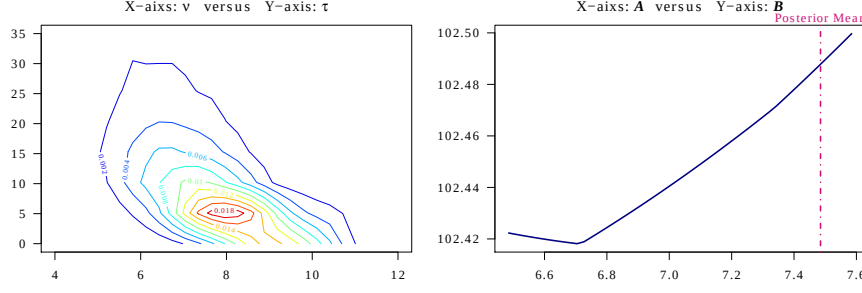


Fig. 1.2 Left panel: density-level contourplots of MCMC approximation to realization of posterior distribution $v, \tau | X$ in Example 1.3.1 (prior parameters: $a = 0, b = 0.1, \alpha_0 = \beta_0 = 0.01$; data $\bar{x} = 10, s^2 = 1, n = 3$). Right panel: corresponding half-lengths B (vertical axis) of the $(\delta = 0.05, \alpha = 0.1)$ -tolerance interval centered at A (horizontal axis); the minimal length is not taken at the posterior mean $E(v | X)$, whose location is indicated by the abscissa of the dotted line.

from the posterior distribution of $\theta = (v, \tau)$ given X . This could be the result of an MCMC run of a sampler for the posterior distribution, or, depending on the data model, of using an analytic formula for the posterior distribution. We shall use the sample values (v_j, τ_j) to approximate expectations under the posterior distribution, whence they need not be independent, and values from a (burnt-in) MCMC run will indeed qualify. Possible dependence together with sample size will determine the error due to simulation.

The idea is to replace the posterior distribution in (1.4) by the empirical distribution of the values $\{\theta_j\}_{j=1}^J$. For a given interval $\mathcal{R}(X) = [L, U]$ we can (in theory) compute the values $Q_{\theta_j}[L, U]$ and next search for the interval $[L, U]$ of minimal length $U - L$ such that

$$\frac{1}{J} \# \left(Q_{\theta_j}[L, U] \geq 1 - \delta \right) \doteq 1 - \alpha,$$

where $\doteq$ means approximately equal, yielding a slightly conservative or anti-conservative solution in case exact equality is not attainable due to discretization.

Algorithm 1 was proposed by Wolfinger [67] and later refined (or corrected) by Krishnamoorthy and Mathew [28], Chapter 11. This algorithm has a convenient graphical representation and has been widely adopted in practice. The idea is to compute for every θ_j the quantiles $L_j = Q_{\theta_j}^{-1}(\delta/2)$ and $U_j = Q_{\theta_j}^{-1}(1 - \delta/2)$, yielding intervals $[L_j, U_j]$ with $Q_{\theta_j}[L_j, U_j] \geq 1 - \delta$, and next setting the tolerance interval $[L, U]$ equal to an interval that is symmetric about the posterior mean and contains a fraction $1 - \alpha$ of the intervals $[L_j, U_j]$ (Krishnamoorthy and Mathew, referred to as KM), or is contained in a fraction α of these intervals (Wolfinger, W). The graphical interpretation is to plot the points (U_j, L_j) in the x - y -plane and search for a point (U, L) on the line $y + x = 2\hat{v}$, for $\hat{v}$ the posterior mean or some other useful estimator, such that a fraction $1 - \alpha$ of the points are in the left-upper quadrant relative to the point $[L, U]$ (see Figure 11.1 in [28] for an example). The KM method results in an interval that is more confident (KM) than the prescribed level $1 - \alpha$, and appears not to optimize the length of the interval.

Algorithm 1: WKM solution for two-sided tolerance interval**Data:** Given $\alpha, \delta, \{(v_j, \tau_j)\}_{j=1}^J$

- 1 Let $\hat{A} = \sum_j v_j / J$;
- 2 Calculate two quantiles sequences: $\{L_j \equiv Q_{v_j, \tau_j}^{-1}(\frac{\delta}{2})\}_{j=1}^J$ and $\{U_j \equiv Q_{v_j, \tau_j}^{-1}(1 - \frac{\delta}{2})\}_{j=1}^J$;
- 3 Find a point $(\hat{L}, \hat{U})$ such that $\hat{L} + \hat{U} = 2\hat{A}$ satisfying one of the following ;
- 4 (W) $\arg \min_{\hat{L}, \hat{U}} \left| \frac{\#S}{J} - \alpha \right|$, where $S = \{(L_j, U_j) : L_j \leq \hat{L}, U_j \geq \hat{U}\}$;
- 5 (KM) $\arg \min_{\hat{L}, \hat{U}} \left| \frac{\#S}{J} - 1 + \alpha \right|$, where $S = \{(L_j, U_j) : L_j \geq \hat{L}, U_j \leq \hat{U}\}$;

Result: two-sided tolerance interval $[\hat{L}, \hat{U}]$

Here we propose another algorithm that directly utilizes (1.4). We seek to minimize B under the constraint that the interval $[L, U] = [A - B, A + B]$ satisfies (1.4). This takes two steps: for fixed A we optimize over B ; next we perform a grid search over A . Because given A , the optimizer over B will yield equality in (1.4), $\hat{B}$ will be the solution to

$$\Pi \left[\Phi \left(\frac{A + B - v}{\tau} \right) - \Phi \left(\frac{A - B - v}{\tau} \right) \geq 1 - \delta | X \right] = 1 - \alpha. \quad (1.8)$$

The posterior mean $E(v|X)$ will typically be close to the optimal solution for A , and is a good starting point for this parameter. As a fast approximation we may also set A to this value and omit the grid search. The posterior mean can be approximated by the average of the sample values v_j .

In practice we replace the posterior distribution in equation (1.8) by an average over the sample values (v_j, τ_j) . Given $\hat{A}$ we approximate $\hat{B}$ by the $(1 - \alpha)$ -quantile of the points g_j computed as the solutions to

$$Q_{v_j, \tau_j}[\hat{A} - g_j, \hat{A} + g_j] \equiv \Phi \left(\frac{\hat{A} + g_j - v_j}{\tau_j} \right) - \Phi \left(\frac{\hat{A} - g_j - v_j}{\tau_j} \right) = 1 - \delta. \quad (1.9)$$

The motivation for this procedure is that $(v_j, \tau_j) \in G_{A,B,\delta}$ if and only if $Q_{v_j, \tau_j}[A - B, A + B] \geq 1 - \delta$, whence precisely the points (v_j, τ_j) with $g_j \leq \hat{B}$ satisfy $Q_{v_j, \tau_j}[\hat{A} - \hat{B}, \hat{A} + \hat{B}] \geq 1 - \delta$ and hence are inside the set $G_{\hat{A}, \hat{B}, \delta}$, whereas the other points are outside this set. This makes the posterior mass of the set equal to $1 - \alpha$ up to simulation error.

Given $\hat{A}$ and (v_j, τ_j) , the function $g \mapsto Q_{v_j, \tau_j}[\hat{A} - g, \hat{A} + g]$ in (1.9) is increasing, from the value 0 when $g = 0$ to 1 as $g \rightarrow \infty$ (see Figure 1.3). The solutions g_j to each equation (1.9) can be found fast by a Newton-Raphson algorithm, with some caution on choosing the initial value for g_j (the algorithm will diverge if the initial value is chosen in the domain where $Q_{v_j, \tau_j}[\hat{A} - g_j, \hat{A} + g_j]$ is very close to 1). An appropriate algorithm is listed in Algorithm 2. Note that the (middle) expression in (1.9) does not change if $\varepsilon := \hat{A} - v_j$ is replaced by $-\varepsilon$.

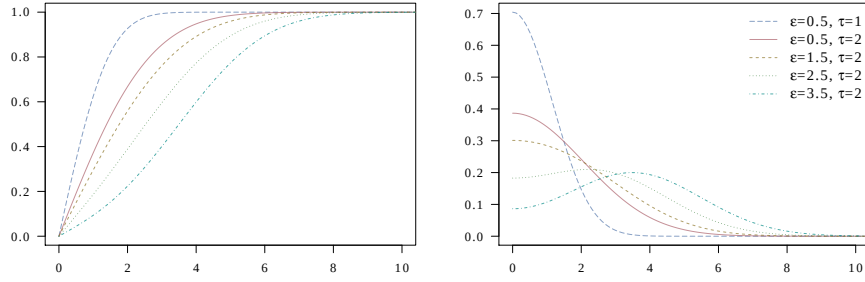


Fig. 1.3 Left panel: plot of the curves $g \mapsto \Phi(\frac{\varepsilon+g}{\tau}) - \Phi(\frac{\varepsilon-g}{\tau})$ for various settings of ε and τ . Right panel: derivatives of these curves.

Algorithm 2: Proposed solution for two-sided tolerance interval

Data: Given $\alpha, \delta, \{(v_j, \tau_j)\}_{j=1}^J$

- 1 Let $\hat{A} = \sum_j v_j / J$;
- 2 **for** $j = 1, 2, \dots, J$ **do**
- 3 Solve equation (1.9) by a Newton-Raphson algorithm as follows;
- 4 **if** $|\hat{A} - v_j| < \tau_j$ **then**
- 5 $g_0 = |\hat{A} - v_j| + \tau_j$
- 6 **else**
- 7 $g_0 = |\hat{A} - v_j|$
- 8 set initial value for g_j at g_0 ;
- 9 **while** $\omega > 0.0001$ **do**
- 10 let $q_{v_j, \tau_j}[\hat{A} - g_j, \hat{A} + g_j]$ be the first-order derivative of $Q_{v_j, \tau_j}[\hat{A} - g_j, \hat{A} + g_j]$;
- 11 $g_j = g_j - \frac{Q_{v_j, \tau_j}[\hat{A} - g_j, \hat{A} + g_j] - 1 + \delta}{q_{v_j, \tau_j}[\hat{A} - g_j, \hat{A} + g_j]}$;
- 12 $\omega = Q_{v_j, \tau_j}[\hat{A} - g_j, \hat{A} + g_j] - 1 + \delta$;
- 13 The above loop results in $\{g_j\}_{j=1}^J$, and let $\hat{B}$ be its $(1 - \alpha)$ th sample quantile;

Result: two-sided tolerance interval $[\hat{A} - \hat{B}, \hat{A} + \hat{B}]$

1.4 Data from a linear mixed model

In this section we apply the preceding to a model that is representative for practice in pharmaceutical quality control: the *linear mixed model* (LMM). We assume that the data X are acquired in an LMM design, and that the future variable Z is defined in terms of the same LMM. We concentrate attention to the two-sided (δ, α) -tolerance interval.

In the LMM we observe a vector $X = U\beta + V\gamma + e$, for known (deterministic) matrices U and V of covariates, a vector of fixed effects parameters β , an unobserved random effect vector γ , and an error vector e . Assume that γ and e are independent, with

$$\gamma \sim N(0, D), \quad e \sim N(0, \sigma^2 I). \quad (1.10)$$

Then the data X follows a $N(U\beta, C)$ -distribution, for $C = VD V^T + \sigma^2 I$, and the full parameter is (β, D, σ^2) .

Consider predicting a new observation $Z = u^T \beta + v^T \gamma' + e'$ with given fixed and random effects coefficients u and v and *newly generated* random effect vector γ' and error e' , with $\gamma' \sim \gamma$ and $e' \sim N(0, \sigma^2)$. Thus γ' is assumed equal in distribution to, but independent of γ , and similarly for e' . This target for prediction is reasonable in many contexts, but sometimes another choice, in particular for γ' , may be more relevant. Typically γ will carry a group structure matched by a block structure in V . The vector v will then have nonzero coordinates corresponding to a single group. The target corresponds to setting the distribution Q_θ of Z equal to $N(v, \tau^2)$, with

$$v = u^T \beta, \quad \tau^2 = v^T D v + \sigma^2.$$

The “prediction” parameter $\theta = (v, \tau)$ is of smaller dimension than the full parameter” (β, D, σ^2) governing the distribution of the data X , whence part of the latter full parameter should be considered a nuisance parameter. To set a Bayesian tolerance interval we need a posterior distribution of θ given the data X . This will typically be inferred from a posterior distribution of the full parameter, resulting from a prior distribution on (β, D, σ^2) .

The (conditional) posterior distribution for a conditional prior $\beta | D, \sigma^2 \sim N(0, \Lambda)$, where Λ may depend on (D, σ^2) , satisfies

$$\beta | D, \sigma^2, X \sim N \left((U^T C^{-1} U + \Lambda^{-1})^{-1} U^T C^{-1} X, (U^T C^{-1} U + \Lambda^{-1})^{-1} \right).$$

In general $E(v | D, \sigma^2, X) = u^T E(\beta | D, \sigma^2, X)$ will depend on D and σ^2 (hidden in C) and hence typically also on τ^2 . Therefore Lemma 1.3.1 does not apply, and there appears to be no reason that a shortest tolerance interval would be centered at the posterior mean of v . To obtain the shortest interval Algorithm 2 should be augmented with a search on possible centerings A .

As in the standard *i.i.d.* model in Example 1.3.1, the dependence on σ^2 can be removed by choosing the variances Λ and D proportional to σ^2 . If $D = \sigma^2 D_0$ and $\Lambda = \sigma^2 \Lambda_0$, then C will be $\sigma^2 (V D_0 V^T + I) =: \sigma^2 C_0$ and the conditional posterior mean of β will be $(U^T C_0^{-1} U + \Lambda_0^{-1})^{-1} U^T C_0^{-1} X$. However, the dependence on D_0 (through C_0) remains, in general.

Letting the prior covariance matrix Λ tend to infinity corresponds to the noninformative prior. If all other quantities are fixed and $\Lambda \rightarrow \infty$, then

$$E(\beta | D, \sigma^2, X) \rightarrow (U^T C^{-1} U)^{-1} U^T C^{-1} X.$$

The limit is the maximum likelihood estimator of β in the model where C is known. Since this is still dependent on C (and hence D and σ^2), it seems that for both the Bayesian and frequentist tolerance intervals the two parameters ν and τ cannot be separated in general. The choice $\Lambda = \lambda (U^T C^{-1} U)^{-1}$ leads to $\lambda / (1 + \lambda)$ times the maximum likelihood estimator.

1.4.1 Approximations to the conditional posterior mean

For special designs the dependence of $E(\beta | D, \sigma^2, X)$ on (D, σ^2) is only mild and can be quantified. We discuss some examples.

Example 1.4.1 (One-way random effects). Suppose X is a vector with coordinates $X_{ik} = \beta + \gamma_i + e_{ik}$, for $i = 1, \dots, m$ and $k = 1, \dots, n$, ordered as $(X_{11}, \dots, X_{1n}, X_{21}, \dots, X_{mn})$, where $\beta \in \mathbb{R}$ and $\gamma = (\gamma_1, \dots, \gamma_m)^T$ with i.i.d. $\gamma_i \sim N(0, d^2)$, so that $D = d^2 I_m$, for I_m the $(m \times m)$ -identity matrix. As prior on β we choose a one-dimensional normal distribution $N(0, \lambda^2)$.

The matrix U is the mn -vector 1_{mn} with all coordinates equal to 1, while V is the $(mn \times m)$ -matrix with the column having 1s in rows $(i-1)n+1$ to in and 0s in the other rows. Then $V^T V = n I_m$, and $U^T V = n 1_m$, and it can be verified that $C 1_{mn} = (nd^2 + \sigma^2) 1_{mn}$ and hence $C^{-1} U = (nd^2 + \sigma^2)^{-1} 1_{mn}$. The coefficient vector of $E(\beta | D, \sigma^2, X)$ is

$$(U^T C^{-1} U + \lambda^{-2})^{-1} U^T C = \left(mn + \frac{nd^2 + \sigma^2}{\lambda^2} \right)^{-1} 1_{mn}^T.$$

For $\lambda = \infty$, this is free of d^2 and σ^2 , while for finite, fixed λ and $m, n \rightarrow \infty$, the coefficient vector is $(mn)^{-1} \left(1 + O(d^2 / (m\lambda^2)) + O(\sigma^2 / (mn\lambda^2)) \right)$.

The dependence on d and σ can be removed by choosing $d = d_0 \sigma$ and $\lambda = \lambda_0 \sqrt{nd_0^2 + 1} \sigma$.

Example 1.4.2. Suppose $X_{ik} = u_{ik}^T \beta + v_{ik}^T \gamma_i + e_{ik}$, ordered as in the preceding example, but now with observed covariates $u_{ik} \in \mathbb{R}^p$ and $v_{ik} \in \mathbb{R}^q$, fixed effects parameter $\beta \in \mathbb{R}^p$ and i.i.d. random effects

$\gamma_i \sim N_q(0, D_q)$. The corresponding matrices U and V are

$$U = \begin{pmatrix} U_1 \\ \vdots \\ U_m \end{pmatrix}, \quad V = \begin{pmatrix} V_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & V_m \end{pmatrix}, \quad U_i = \begin{pmatrix} u_{i1}^T \\ \vdots \\ u_{in}^T \end{pmatrix}, \quad V_i = \begin{pmatrix} v_{i1}^T \\ \vdots \\ v_{in}^T \end{pmatrix}.$$

Then $C = VDV^T + \sigma^2 I$ is an $(mn \times mn)$ -block-diagonal matrix with blocks $V_i D_q V_i^T + \sigma^2 I_n$, and

$$U^T C^{-1} U = \sum_{i=1}^m U_i^T (V_i D_q V_i^T + \sigma^2 I_n)^{-1} U_i, \\ U^T C^{-1} = \left(U_1^T (V_1 D_q V_1^T + \sigma^2 I_n)^{-1}, \dots, U_m^T (V_m D_q V_m^T + \sigma^2 I_n)^{-1} \right).$$

The matrices $U_i^T V_i$ and $V_i^T V_i$ are of dimensions $p \times q$ and $q \times q$, and are sums over the n observations (for $k = 1, \dots, n$) per group i , as defined by the random effect γ_i . In convential asymptotics we could view them as n times a matrix of fixed order. Then, for $D_0 = \sigma^{-2} D_q$,

$$\begin{aligned} (V_i D_0 V_i^T + I)^{-1} &= I - V_i (D_0^{-1} + V_i^T V_i)^{-1} V_i^T \\ &= I - V_i (V_i^T V_i)^{-1} \left[D_0^{-1} (V_i^T V_i)^{-1} + I \right]^{-1} V_i^T \\ &= I - V_i (V_i^T V_i)^{-1} \left[I - D_0^{-1} (V_i^T V_i)^{-1} + (D_0^{-1} (V_i^T V_i)^{-1})^2 + \dots \right] V_i^T \\ &= P_{V_i^\perp} + V_i (V_i^T V_i)^{-1} D_0^{-1} (V_i^T V_i)^{-1} V_i^T \\ &\quad - \sum_{k=2}^{\infty} (-1)^k V_i (V_i^T V_i)^{-1} (D_0^{-1} (V_i^T V_i)^{-1})^k V_i^T, \end{aligned} \quad (1.11)$$

where $P_{V_i^\perp}$ is the projection on the orthocomplement of the linear span of the columns of V . If $V_i^T V_i$ is large, then the series on the right can be neglected, as its terms contain multiple terms $(V_i^T V_i)^{-1}$. There is then still dependence on D and σ^2 in the second term, which may dominate.

In a full random effects model we shall have $u_{ik} = v_{ik}$, for every (i, k) , and then $U_i = V_i$. Then $U_i^T P_{V_i^\perp} = 0$ and $U_i^T V_i (V_i^T V_i)^{-1} = I_p$. If $n^{-1} V_i^T V_i$ stabilizes as $n \rightarrow \infty$,

$$\begin{aligned} \sigma^2 U^T C^{-1} U &= \sum_{i=1}^m \sum_{k=1}^{\infty} (-1)^{k-1} D_0^{-1} ((V_i^T V_i)^{-1} D_0^{-1})^{k-1} = m \left(D_0^{-1} + O\left(\frac{1}{n}\right) \right), \\ \sigma^2 U^T C^{-1} &= \sum_{k=1}^{\infty} (-1)^{k-1} \left(D_0^{-1} (V_i^T V_i)^{-1} \right)^k V_i^T \Big|_{i=1}^m \\ &= \left(D_0^{-1} (V_i^T V_i)^{-1} \left(I + O\left(\frac{1}{n}\right) \right) V_i^T \right) \Big|_{i=1}^m. \end{aligned}$$

Thus we find that

$$\begin{aligned} & (U^T C^{-1} U + \Lambda^{-1})^{-1} U^T C^{-1} \\ &= \left(\frac{1}{m} \left(I + O\left(\frac{D_0}{n}\right) + \frac{D_0 \Lambda^{-1}}{m} \right) \right)^{-1} \left((V_i^T V_i)^{-1} \left(I + O\left(\frac{1}{n}\right) \right) V_i^T \right)_{i=1}^m. \end{aligned}$$

To first order this is free of D and σ^2 with relative remainders of the order D_0/n and $D_0 \Lambda^{-1}/m$.

If every random effect is matched by a fixed effect with the same covariate vector (supplying a common mean value to the random effects), but there are more fixed than random effects, then $u_{ik}^T = (v_{ik}^T, \bar{u}_{ik}^T)$, which implies $U_i = (V_i, \bar{U}_i)$ for an $(n \times (p - q))$ -matrix $\bar{U}_i$. The preceding formulas must then be adapted to, where the approximations refer to ignoring the series in (1.11), for $W_i = V_i(V_i^T V_i)^{-1}$,

$$\begin{aligned} \sigma^2 U^T C^{-1} U &\doteq \sum_{i=1}^m \left[\begin{pmatrix} 0 & 0 \\ 0 & \bar{U}_i^T P_{V_i^\perp} \bar{U}_i \end{pmatrix} + \begin{pmatrix} D_0^{-1} & D_0^{-1} W_i^T \bar{U}_i \\ \bar{U}_i^T W_i D_0^{-1} & \bar{U}_i^T W_i D_0^{-1} W_i^T \bar{U}_i \end{pmatrix} \right], \\ \sigma^2 U^T C^{-1} &\doteq \left(\begin{pmatrix} 0 \\ \bar{U}_i^T P_{V_i^\perp} \end{pmatrix} + \begin{pmatrix} D_0^{-1} W_i^T \\ \bar{U}_i^T W_i D_0^{-1} W_i^T \end{pmatrix} \right)_{i=1}^m. \end{aligned}$$

It is reasonable to expect that the matrices $\bar{U}_i^T W_i = \bar{U}_i^T V_i(V_i^T V_i)^{-1}$ will settle down as, as will the matrices $n^{-1} \bar{U}_i^T P_{V_i^\perp} \bar{U}_i$. Then up to lower order terms

$$(U^T C^{-1} U)^{-1} U^T C^{-1} \doteq \sum_{i=1}^m \left(\begin{pmatrix} D_0^{-1} & D_0^{-1} W_i^T \bar{U}_i \\ \bar{U}_i^T W_i D_0^{-1} & \bar{U}_i^T P_{V_i^\perp} \bar{U}_i \end{pmatrix} \right)^{-1} \left(\begin{pmatrix} D_0^{-1} W_i^T \\ \bar{U}_i^T P_{V_i^\perp} \end{pmatrix} \right)_{i=1}^m.$$

Here the three appearances of D_0^{-1} in the top row cancel each other (as follows by factorizing out the block diagonal matrix with blocks D_0^{-1} and I_{p-q} from the inverse matrix), while the factor $\bar{U}_i^T W_i D_0^{-1}$ in the bottom row is a factor $1/n$ smaller in order than the matrix $\bar{U}_i^T P_{V_i^\perp} \bar{U}_i$ and hence can be set to 0 up to order $1/n$. It follows that again the matrix is free of D and σ^2 up to order $1/n$.

1.4.2 Numerical examples

We evaluate the small sample performance of our proposed algorithm via a simulation study, with data generated from a one-way random effects model. We observe $(X_{ik}: i = 1, 2, \dots, m; k = 1, 2, \dots, n_i)$, where X_{ik} is the k^{th} value of the i^{th} group and satisfies $X_{ik} = \nu + \gamma_i + e_{ik}$ for i.i.d. $\gamma_i \sim N(0, d^2)$ independent of the i.i.d. error $e_{ik} \sim N(0, \sigma^2)$. We are interested in the two-sided (δ, α) -tolerance interval related to a future observation $Z = \nu + \gamma + e \sim Q_\theta = N(\nu, \tau^2)$, where γ and e are independent copies of the γ_i and e_{ik} . The parameter is $\theta = (\nu, \tau^2)$, for $\tau^2 = d^2 + \sigma^2$.

We used 6 different parameter settings. In every setting the overall mean was set equal to $v = 0$, and the group variance to $d^2 = 1$. The intra-correlation $\sigma^2/(d^2 + \sigma^2)$ was chosen equal to the numbers 0.1, 0.3, 0.5, 0.7, 0.9. In every setting the number of groups was $m = 6$ and the group sizes were $(n_1, \dots, n_6) = (2, 3, 4, 2, 3, 4)$. This simulation setup is the same as in [51], and facilitates a cross-comparison against the performance of a standard frequentist solution.

We computed the $(\delta = 0.1, \alpha = 0.05)$ -tolerance interval $[L, U]$ by Algorithm 2, for every of $K = 1000$ replicates of the data in each parameter setting, and computed the true coverage $Q_\theta[L, U]$ of these intervals using the true parameter θ of the simulation. We dub an interval with true coverage no less than $1 - \delta$ as “qualified”, and compared the empirical fraction of qualified intervals out of the K replicates to the nominal value $1 - \alpha$. The procedure is considered to perform well in the frequentist sense if this empirical fraction is close to this nominal value. Here we must allow for the simulation error, which has a standard error of $\sqrt{p(1-p)/K}$, for p the true coverage, which is unknown, but hopefully close to $1 - \alpha$.

For each simulated dataset the posterior distribution of (v, d, σ^2) was approximated by a standard Gibbs sampler (with the vector of random effects γ added in as a fourth parameter), before utilizing Algorithm 2. Two setups of priors were deployed, both with independent priors on the three parameters v, d, σ . The first is the vanilla setup with vague marginal prior distributions $v \sim N(0, 1000)$, $d^2 \sim IG(0.001, 0.001)$ and $\sigma^2 \sim IG(0.001, 0.001)$. The second uses the same prior on σ^2 , but uses a t -distribution for v given by the hierarchy $v | \sigma_0 \sim N(0, \sigma_0^2)$ and $\sigma_0^2 \sim IG(0.001, 0.001)$, and the prior on d given by the structural equation $d = |\xi|\omega$ for independent $\xi \sim N(0, 1)$ and $\omega^2 \sim IG(0.001, 0.001)$. The latter specification can also be understood as over-parameterizing the distribution of the random effect two parameters instead of one, as $\gamma | \xi, \omega \sim N(0, \xi^2 \omega^2)$. This “data augmentation” or “parameter expansion” is meant to enhance the mixing rate of the Gibbs sampler, in particular when the number of groups m is small. See [6] for a comparison of methods (including non-Bayesian methods) to fit the LMM. A more accurate approximation to the posterior should have positive impact on the subsequent tolerance interval. Finer amendments should be possible for concrete cases.

In all simulation settings the tolerance intervals were constructed both by fixing the center point A at the posterior mean $E(v|X)$, and by seeking an optimal value of A to minimize the half length B of the interval. Thus four tolerance intervals were calculated based on each simulated dataset. To save on computation time the optimization over A was carried out only approximately. Still for 85% of the simulation cases a shorter interval was obtained than the interval at the posterior mean, in a few cases as much as 20% shorter, but in 75% of the cases no more than a few percentage points. Table 1.1 reports the quotients of the lengths.

The proportions of qualified intervals (the ones which attain true coverage $1 - \delta = 0.9$) are listed in Table 1.2. They are reasonably close to the nominal value $1 - \alpha = 0.95$, with deviations in both directions up to several percentage points. The performance seems to depend on the true intra-correlation. This dependence follows a similar pattern as for the high-order asymptotic solution

in [51] without correction for imbalance, and is close to their solution that includes correction when the intra-correlation is small or very big. The tolerance intervals centered at the (approximately) optimal value have shorter length and attain lower confidence, but their performance seems to surpass slightly the intervals centered at the posterior mean $E(v|X)$.

The difference in performance of Algorithm 2 between the two prior setups is within the order of the simulation error.

Table 1.1 Interval length at optimal value of A relative to fixing A at $E(v|X)$

	<i>under Vanilla setup</i>				<i>under Parameter Expansion setup</i>			
	<i>Min</i>	<i>0.25th qu.</i>	<i>Median</i>	<i>0.75th qu.</i>	<i>Min</i>	<i>0.25th qu.</i>	<i>Median</i>	<i>0.75th qu.</i>
<i>intra – correlation</i>								
0.1	0.8420	0.9958	0.9984	0.9993	0.8625	0.9969	0.9991	0.9997
0.3	0.8244	0.9943	0.9984	0.9994	0.8569	0.9969	0.9989	0.9997
0.5	0.8192	0.9905	0.9981	0.9992	0.8374	0.9954	0.9985	0.9995
0.7	0.8193	0.9897	0.9979	0.9992	0.8020	0.9947	0.9984	0.9994
0.9	0.8102	0.9889	0.9980	0.9993	0.8058	0.9925	0.9980	0.9993

Table 1.2 Approximated Confidence for $(\delta = 0.1, \alpha = 0.05)$ -Bayesian tolerance interval

	<i>under Vanilla setup</i>		<i>under Parameter Expansion setup</i>	
	$A = E(v X)$	$A = \text{Optimal}$	$A = E(v X)$	$A = \text{Optimal}$
<i>intra – correlation</i>				
0.1	0.972	0.969	0.968	0.963
0.3	0.964	0.955	0.955	0.949
0.5	0.936	0.921	0.925	0.917
0.7	0.925	0.907	0.915	0.911
0.9	0.952	0.941	0.940	0.935

1.5 Frequentist justification of the Bayesian procedure

In this section we show that Bayesian tolerance regions are often also approximate frequentist tolerance regions, of corresponding levels. We consider an asymptotic setup, with data $X = X_n$ indexed by a parameter $n \rightarrow \infty$, in which the *Bernstein-von Mises theorem* holds. The latter theorem (see [59], Chapter 10) entails that the posterior distribution $\Pi_n(\cdot|X_n)$ of θ can be approximated by a normal distribution with deterministic covariance matrix, centered at an estimator $\hat{\theta}_n = \hat{\theta}_n(X_n)$,

$$\Pi_n(\cdot|X_n) - N\left(\hat{\theta}_n, \frac{1}{n}\Sigma_\theta\right) \xrightarrow{P} 0, \quad (1.12)$$

(in total variation norm), where the estimators $\hat{\theta}_n = \hat{\theta}_n(X_n)$ satisfy

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{\theta}{\rightsquigarrow} N(0, \Sigma_\theta). \quad (1.13)$$

For instance, under regularity conditions this is valid for X_n a vector of n i.i.d. observations from a smooth parametric model, with $\hat{\theta}_n$ the maximum likelihood estimator and Σ_θ the inverse Fisher information matrix. The Bernstein-von Mises theorem can be used to show that Bayesian and frequentist inference (testing and confidence sets) *merge* for large sample sizes. In this section we investigate this for tolerance intervals.

We shall show that Bayesian tolerance regions $\mathcal{R}_n(X_n)$ such that the functions

$$h \mapsto Q_{\hat{\theta}_n(X_n) + h/\sqrt{n}}(\mathcal{R}_n(X_n)), \quad n = 1, 2, \dots, \quad (1.14)$$

stabilize asymptotically to a *deterministic* function are asymptotically frequentist tolerance regions, for any given loss function ℓ and level α . The crux of this *stability condition* is that the randomness which enters the functions (1.14) through X_n in $\hat{\theta}_n(X_n)$ asymptotically cancels the randomness which enters through X_n in $\mathcal{R}_n(X_n)$: the Bayesian tolerance regions $\mathcal{R}_n(X_n)$ should be “asymptotically pivotal” with respect to the estimators $\hat{\theta}_n$. Some type of stability condition appears to be necessary, because the shape of a Bayesian tolerance region is left free by its definition.

An informal proof of the frequentist validity of Bayesian tolerance regions is as follows. Replacing the posterior distribution in (1.6) by its normal approximation (1.12) from the Bernstein-von Mises theorem, we find that

$$\int \ell \left[Q_{\vartheta}(\mathcal{R}_n(X_n)) \right] dN \left(\hat{\theta}_n, \frac{1}{n} \Sigma_\theta \right) (\vartheta) \doteq 1 - \alpha. \quad (1.15)$$

By the substitution $\vartheta = \hat{\theta}_n + h/\sqrt{n}$ this can be rewritten in the form

$$\int \ell \left[Q_{\hat{\theta}_n + h/\sqrt{n}}(\mathcal{R}_n(X_n)) \right] dN(0, \Sigma_\theta)(h) \doteq 1 - \alpha. \quad (1.16)$$

By the stability assumption the integrand

$$h \mapsto g_n(h; X_n) := \ell \left[Q_{\hat{\theta}_n + h/\sqrt{n}}(\mathcal{R}_n(X_n)) \right] \quad (1.17)$$

in this expression is asymptotically the same as a deterministic function $h \mapsto g_\infty(h)$. In view of (1.13) the integral in (1.16) is then approximately equal to $E_\theta g_\infty(\sqrt{n}(\theta - \hat{\theta}_n))$, which in turn, again by stability, is asymptotically the same as $E_\theta g_n(\sqrt{n}(\theta - \hat{\theta}_n); X_n)$, or

$$E_\theta \ell \left[Q_{\hat{\theta}_n + \sqrt{n}(\theta - \hat{\theta}_n)/\sqrt{n}}(\mathcal{R}_n(X_n)) \right] = E_\theta \ell \left[Q_\theta(\mathcal{R}_n(X_n)) \right].$$

Thus the final expression, which is the frequentist level of the tolerance region $\mathcal{R}_n(X_n)$, is asymptotically equal to $1 - \alpha$.

For an (δ, α) -tolerance region $\ell(Q_\theta(\mathcal{R}_n(X_n)))$ is the indicator of the set $\hat{G}_n = \{\theta: Q_\theta(\mathcal{R}_n(X_n)) \geq 1 - \delta\}$ and the function (1.17) is the indicator of the set

$$\hat{H}_n = \sqrt{n}(\hat{G}_n - \hat{\theta}_n).$$

Thus the stability condition is that the latter sets approximate to a deterministic set, as $n \rightarrow \infty$. Condition (1.16) becomes

$$N(0, \Sigma_\theta)(\hat{H}_n) \doteq 1 - \alpha. \quad (1.18)$$

This equality allows to “solve” one aspect of the sets $\hat{H}_n$; in general additional constraints will be imposed to define their shape. As the normal distribution in this display is fixed, it is not unnatural that these constraints would render the sets $\hat{H}_n$ also to become fixed, in the limit: stability is natural.

Theorem 1.5.1. *Suppose that (1.12)–(1.13) hold, the loss function ℓ is bounded, and suppose that there exist (deterministic) functions $f_{n,1}, f_{n,2}: \mathbb{R}^d \rightarrow \mathbb{R}$ with the property that $f_{n,i}(h_n) \rightarrow f_\infty(h)$ for some function f_∞ and any sequence $h_n \rightarrow h$ with limit h in a set of probability one under the normal distribution in (1.13) and such that*

$$f_{n,1}(h) \leq \ell\left[Q_{\hat{\theta}_n + h/\sqrt{n}}(\mathcal{R}_n(X_n))\right] \leq f_{n,2}(h), \quad h \in \mathbb{R}^d. \quad (1.19)$$

Then $\int \ell(Q_\theta(\mathcal{R}_n(X_n))) d\Pi(\theta|X_n) \rightarrow 1 - \alpha \in (0, 1)$ in probability implies that $E_\theta \ell(Q_\theta(\mathcal{R}_n(X_n))) \rightarrow 1 - \alpha$, as $n \rightarrow \infty$, for every θ .

Proof. We may assume without loss of generality that the functions $f_{n,i}$ are uniformly bounded. Then the condition $f_{n,i}(h_n) \rightarrow f_\infty(h)$ for every sequence $h_n \rightarrow h$ implies that $E f_{n,i}(Y_n) \rightarrow E f_\infty(Y)$, whenever the sequence of random vectors Y_n tends in distribution to the random vector Y , in view of the extended continuous mapping theorem (see [58], Theorem 1.11.1). Thus by (1.13), for $i = 1, 2$,

$$E_\theta f_{n,i}(\sqrt{n}(\hat{\theta}_n - \theta)) \rightarrow \int f_\infty dN(0, \Sigma_\theta).$$

By assumption the function g_n given in (1.17) is sandwiched between $f_{n,1}$ and $f_{n,2}$. Therefore $E_\theta \ell(Q_\theta(\mathcal{R}_n(X_n))) = E_\theta g_n(\sqrt{n}(\hat{\theta}_n - \theta); X_n)$ tends to the same limit.

By (1.12) we have

$$\int \ell(Q_\theta(\mathcal{R}_n(X_n))) d\Pi(\theta|X_n) = \int g_n(h; X_n) dN(0, \Sigma_\theta)(h) + o(1).$$

Again by sandwiching of $g_n(h; X_n)$ this is asymptotic to $\int f_{n,i} dN(0, \Sigma_\theta)$, and hence tends to $\int f_\infty dN(0, \Sigma_\theta)$.

■

1.5.1 Normal predictions

An (α, δ) tolerance interval $\mathcal{R}_n(X_n) = [A_n - B_n, A_n + B_n]$ for a one-dimensional Gaussian variable $Z \sim N(\nu, \tau^2)$ is the base (the section at $\tau = 0$) of a set of the form

$$G_{A,B,\delta} = \left\{ \theta = (\nu, \tau) : \Phi\left(\frac{A+B-\nu}{\tau}\right) - \Phi\left(\frac{A-B-\nu}{\tau}\right) \geq 1 - \delta \right\}.$$

The values A_n and B_n are determined so that $\Pi(G_{A_n,B_n,\delta} | X_n) \geq 1 - \alpha$, for $\Pi(\cdot | X_n)$ the posterior distribution of $\theta = (\nu, \tau)$, and so that the length $2B_n$ of the interval is minimal.

Under (1.12) the posterior distribution contracts (at rate $1/\sqrt{n}$) to the Dirac measure at $\hat{\theta}$, which tends to the true parameter under (1.13). Hence the equation $\Pi(G_{A_n,B_n,\delta} | X_n) \geq 1 - \alpha$ forces $\hat{\theta} = (\hat{\nu}, \hat{\tau})$ to be contained in $G_{A_n,B_n,\delta}$ with probability tending to one. This implies that $A_n - \hat{\nu} \rightarrow 0$ in probability and hence $\Phi(B_n/\hat{\tau}) - \Phi(-B_n/\hat{\tau}) \rightarrow 1 - \delta$, whence $B_n/\hat{\tau} \rightarrow \xi_{\delta/2}$, for ξ_{δ} the upper δ -quantile of the standard normal distribution.

The function $\theta \mapsto \ell\left(Q_{\theta}([A - B, A + B])\right)$ corresponding to the (α, δ) tolerance interval is the indicator of the set $G_{A,B,\delta}$, and the stability condition (1.19) is that the (indicator functions) of the sets $\hat{H}_n = \sqrt{n}(G_{A_n,B_n,\delta} - \hat{\theta})$ are asymptotically deterministic. These sets can be written $\hat{H}_n = \{(g, h) : K_n(g/\sqrt{n}, h/\sqrt{n}) \geq 0\}$, for the stochastic processes

$$K_n(g, h) = \Phi\left(\frac{A_n + B_n - \hat{\nu} - g}{\hat{\tau} + h}\right) - \Phi\left(\frac{A_n - B_n - \hat{\nu} - g}{\hat{\tau} + h}\right) - (1 - \delta). \quad (1.20)$$

By a second-order Taylor expansion we see that these processes satisfy (1.21) with

$$\begin{aligned} \hat{a}_n &= K_n(0, 0) = \Phi\left(\frac{A_n + B_n - \hat{\nu}}{\hat{\tau}}\right) - \Phi\left(\frac{A_n - B_n - \hat{\nu}}{\hat{\tau}}\right) - (1 - \delta), \\ \hat{b}_n &= \frac{\partial}{\partial h} K_n(0, 0) = \psi\left(\frac{A_n + B_n - \hat{\nu}}{\hat{\tau}}\right) \frac{1}{\hat{\tau}} - \psi\left(\frac{A_n - B_n - \hat{\nu}}{\hat{\tau}}\right) \frac{1}{\hat{\tau}}, \\ V(g, h) &= h. \end{aligned}$$

Here $\psi(x) = \phi(x)x = -\phi'(x)$. (Note that the partial derivative of K_n relative to its first argument g vanishes at $(0, 0)$.) Since A_n , B_n , $\hat{\nu}$ and $\hat{\tau}$ tend in probability to nontrivial limits, the conditions of Lemma 1.5.1 are satisfied and hence the sets $\hat{H}_n$ are asymptotically sandwiched between pairs of deterministic sets. Functions $f_{n,i}$ as in (1.19) can be constructed from these sets by letting $\varepsilon \rightarrow 0$ and $M \rightarrow \infty$ slowly with n .

It follows that the conditions of Theorem 1.19 are satisfied and hence the Bayesian (α, δ) tolerance sets are asymptotic frequentist (α, δ) tolerance sets.

The convergence $A_n - \hat{\nu} \rightarrow 0$ in probability means that the tolerance intervals are asymptotically centered at the (asymptotic) posterior mean. If the posterior distribution of θ is exactly normal

$N(\hat{v}, \hat{\tau}, \Sigma_\theta)$ with a diagonal covariance matrix, then $E(v|X_n, \tau)$ is free of τ and Lemma 1.3.1 shows that the tolerance interval is centered exactly at the posterior mean. This is more generally true if the posterior distribution of v given (τ, X_n) is normal with mean free of τ .

For a nondiagonal matrix Σ_θ this is not necessarily true, and in general the normal distribution will be an approximation only. The approximation $A_n - \hat{v} \rightarrow 0$ can then be improved to order $n^{-1/4}$, and an asymptotic expression for the half length B_n of the interval is as follows.

Theorem 1.5.2. *If the Bernstein-von Mises theorem (1.12)–(1.13) holds, then the Bayesian (α, δ) tolerance interval $[A_n - B_n, A_n + B_n]$ of minimal length is a frequentist (α, δ) tolerance interval. Its center and half length satisfy $A_n = \hat{v} + o_P(n^{-1/4})$ and $B_n = \hat{\tau}\xi_{\delta/2} + (\xi_\alpha/\xi_{\delta/2})n^{-1/2} + o_P(n^{-1/2})$.*

Proof. Since $A_n - \hat{v} \rightarrow 0$ and $B_n \rightarrow \xi_{\delta/2}$, there exist intervals I_n and J_n around v_0 and $\xi_{\delta/2}$ that shrink to these points that contain A_n and B_n with probability tending to one. Define functions $F_n, G_n: I_n \times J_n \rightarrow \mathbb{R}$ by

$$F_n(A, B) = \Pi(G_{A,B,\delta}|X_n),$$

$$G_n(A, B) = N(0, \Sigma_\theta) \left\{ (g, h): h \leq \frac{\sqrt{n}K_n(0, 0; A, B)}{-2\psi(\xi_{\delta/2})/\hat{\tau}} \right\},$$

Here $K_n(h, g; A, B)$ is the expression on the right side of (1.20), but with (A_n, B_n) replaced by a generic (A, B) . We shall show that these functions satisfy the conditions of Lemma 1.5.2 with $c_n = \hat{\tau}\xi_{\delta/2}$ and $\xi(\alpha) = (\xi_\alpha/\xi_{\delta/2})n^{-1/2}$, whence the theorem follows from the lemma.

In view of (1.12) we have $\sup_{A,B} |F_n(A, B) - N(0, \Sigma_\theta)(\hat{H}_{A,B,\delta})| \rightarrow 0$, for

$$\hat{H}_{A,B,\delta} = \left\{ (g, h): K_n(g/\sqrt{n}, h/\sqrt{n}; A, B) \geq 0 \right\}.$$

The supremum here and in the following is taken over $(A, B) \in I_n \times J_n$. By a second-order Taylor expansion, for functions $\eta_{n,1}$ and $\eta_{n,2}$ with $\sup_{A,B} \eta_{n,i}(A, B) \rightarrow 0$ and a constant b independent of (A, B) ,

$$\begin{aligned} & \sqrt{n} \left| K_n(g/\sqrt{n}, h/\sqrt{n}; A, B) - K_n(0, 0; A, B) \right. \\ & \quad \left. - \left(-2\psi(\xi_{\delta/2})/\hat{\tau} + \eta_{n,1}(A, B) \right) \frac{g}{\sqrt{n}} + \eta_{n,2}(A, B) \frac{h}{\sqrt{n}} \right| \leq b \frac{g^2 + h^2}{\sqrt{n}}. \end{aligned}$$

As in the proof of Lemma 1.5.1, this shows that, with $C_n = \{(g, h): g^2 + h^2 \leq L_n\}$ and $L_n \rightarrow \infty$ sufficiently slowly

$$\sup_{A,B} \left| N(0, \Sigma_\theta)(\hat{H}_{A,B,\delta} \cap C_n) - N(0, \Sigma_\theta) \left\{ (g, h) \in C_n: h \leq \frac{\sqrt{n}K_n(0, 0; A, B)}{-2\psi(\xi_{\delta/2})/\hat{\tau}} \right\} \right| \rightarrow 0.$$

Because $N(0, \Sigma_\theta)(C_n^c) \rightarrow 0$, this is then also true without intersecting the sets by C_n . This finishes the proof that $\sup_{A,B} |F_n(A, B) - G_n(A, B)| \rightarrow 0$.

From the unimodality of the normal distribution it is clear that $A \mapsto G_n(A, B)$ is maximal at $A = \hat{v}$, for every B . It is elementary to verify by Taylor expansion that $G_n(\hat{v}, \hat{B}) \rightarrow 1 - \alpha$, for $\hat{B} = \hat{\tau} \xi_{\delta/2} + (\xi_\alpha / \xi_{\delta/2}) n^{-1/2}$.

Finally, again by Taylor expansion there exist functions $\eta_{n,3}$ with $\sup_{A,B} \eta_{n,3}(A, B) \rightarrow 0$ such that

$$K_n(0, 0, A, \hat{B}) \leq K_n(0, 0, \hat{v}, \hat{B}) + (A - \hat{v})^2 \left(\phi'(\hat{B}/\hat{\tau})/\hat{\tau} + \eta_{n,3}(A, B) \right).$$

As $\phi'(x) < 0$, for $x > 0$, we obtain that $\sqrt{n}K_n(0, 0, A, \hat{B})$ is strictly smaller than $\sqrt{n}K_n(0, 0, \hat{v}, \hat{B})$ if $\sqrt{n}(A - \hat{v})^2$ is bounded away from zero, but then

$$N(0, \Sigma_\theta) \left\{ (g, h) : h \leq \frac{\sqrt{n}K_n(0, 0, A, \hat{B})}{-2\psi(\xi_{\delta/2})/\hat{\tau}} \right\}$$

is strictly smaller than its asymptotic value $1 - \alpha$ at $A = \hat{v}$ unless $\sqrt{n}(A - \hat{v})^2 \rightarrow 0$. This verifies the last displayed condition of Lemma 1.5.2. ■

Lemma 1.5.1. *Suppose that for every $M > 0$ the stochastic processes $(K_n(h) : h \in \mathbb{R}^d)$ satisfy*

$$\sup_{\|h\| \leq M/\sqrt{n}} \sqrt{n} |K_n(h) - \hat{a}_n + \hat{b}_n Vh| \xrightarrow{P} 0, \quad (1.21)$$

for random variables $\hat{a}_n$ and $\hat{b}_n > 0$ such that $\hat{b}_n^{-1}$ is bounded in probability, and a linear map $V : \mathbb{R}^d \rightarrow \mathbb{R}$. If the sets $\hat{H}_n = \{h \in \mathbb{R}^d : K_n(h/\sqrt{n}) \geq 0\}$ satisfy $N(0, \Sigma)(\hat{H}_n) \rightarrow 1 - \alpha \in (0, 1)$, then $\sqrt{n}\hat{a}_n/\hat{b}_n \rightarrow \xi_\alpha \sqrt{V\Sigma V^T}$ and for every $\varepsilon, M > 0$, with probability tending to 1,

$$\begin{aligned} & \left\{ h \in \mathbb{R}^d : \|h\| \leq M, Vh \leq \xi_\alpha \sqrt{V\Sigma V^T} - \varepsilon \right\} \subset \hat{H}_n \subset \\ & \subset \left\{ h \in \mathbb{R}^d : Vh \leq \xi_\alpha \sqrt{V\Sigma V^T} + \varepsilon \text{ or } \|h\| > M \right\}. \end{aligned}$$

Proof. Define $\varepsilon_n(h) = \sqrt{n}(K_n(h) - \hat{a}_n + \hat{b}_n Vh)$ and set $\hat{\varepsilon}_n = \sup_{\|h\| \leq M/\sqrt{n}} |\varepsilon_n(h)|$, for given M . Then by assumption $\hat{\varepsilon}_n \rightarrow 0$ in probability, and $|K_n(h/\sqrt{n}) - \hat{a}_n + \hat{b}_n Vh/\sqrt{n}| \leq \hat{\varepsilon}_n$, for every h with $\|h\| \leq M$. From the latter inequality we find that

$$\begin{aligned} \|h\| \leq M, K_n(h/\sqrt{n}) \geq 0 & \Rightarrow \hat{b}_n Vh \leq \sqrt{n}\hat{a}_n + \hat{\varepsilon}_n, \\ \|h\| \leq M, \hat{b}_n Vh \leq \sqrt{n}\hat{a}_n - \hat{\varepsilon}_n & \Rightarrow K_n(h/\sqrt{n}) \geq 0. \end{aligned}$$

This implies that

$$\left\{ \|h\| \leq M, Vh \leq (\sqrt{n}\hat{a}_n - \hat{\varepsilon}_n)/\hat{b}_n \right\} \subset \hat{H}_n \subset \left\{ Vh \leq (\sqrt{n}\hat{a}_n + \hat{\varepsilon}_n)/\hat{b}_n \text{ or } \|h\| > M \right\}.$$

Combining this with the fact that $N(0, \Sigma)(\hat{H}_n) \rightarrow 1 - \alpha \in (0, 1)$ and the fact that $Vh \sim N(0, V\Sigma V^T)$ if $h \sim N(0, \Sigma)$, we conclude that there exists $\delta_M > 0$ such that $\delta_M \rightarrow 0$ as $M \rightarrow \infty$ such that $(\sqrt{n}\hat{a}_n - \hat{\epsilon}_n)/\hat{b}_n \leq \xi_{\alpha-\delta_M}\sqrt{V\Sigma V^T} + o_P(1)$ and $(\sqrt{n}\hat{a}_n + \hat{\epsilon}_n)/\hat{b}_n \geq \xi_{\alpha+\delta_M}\sqrt{V\Sigma V^T} + o_P(1)$, for every M . This implies that $\sqrt{n}\hat{a}_n/\hat{b}_n = \xi_\alpha\sqrt{V\Sigma V^T} + o_P(1)$. We substitute this in the last display to obtain the result of the lemma. ■

Lemma 1.5.2. *Let $F_n, G_n: I_n \times J_n \rightarrow \mathbb{R}$ be functions on rectangles $I_n \times J_n \subset \mathbb{R}^2$ that are nondecreasing in their second argument, such that*

$$\sup_{A,B} |F_n(A,B) - G_n(A,B)| \rightarrow 0,$$

and such that for numbers $c_n \in J_n$ and continuous functions $\xi: (0, 1) \rightarrow \mathbb{R}$,

$$\sup_A G_n(A, c_n + \xi(\alpha)/\sqrt{n}) = G_n(0, c_n + \xi(\alpha)/\sqrt{n}) \rightarrow 1 - \alpha, \quad \text{every } \alpha \in (0, 1),$$

$$\limsup_n \sup_{A: |A| > \delta_n n^{-1/4}} G_n(A, c_n + \xi(\alpha)/\sqrt{n}) < 1 - \alpha, \quad \text{some } \delta_n \rightarrow 0.$$

Then $B_n := \inf(B: \sup_A F_n(A, B) \geq 1 - \alpha)$ satisfies $B_n = c_n + \xi(\alpha)/\sqrt{n} + o(n^{-1/2})$, and $\hat{A} := \operatorname{argmax}_A F_n(A, B_n)$ satisfies $\hat{A} = o(n^{-1/4})$.

Proof. The functions $\bar{F}_n$ and $\bar{G}_n$ defined by $\bar{F}_n(B) = \sup_A F_n(A, B)$ and similarly for G_n satisfy $\sup_B |\bar{F}_n(B) - \bar{G}_n(B)| \rightarrow 0$. Combined with the assumptions on G_n , this gives that $\bar{F}_n(c_n + \xi(\alpha)/\sqrt{n}) \rightarrow 1 - \alpha$, for every α . The definition of B_n and monotonicity of $\bar{F}_n$ now readily give that $c_n + \xi(\alpha_2)/\sqrt{n} \leq B_n \leq c_n + \xi(\alpha_1)/\sqrt{n}$ for every $\alpha_1 < \alpha < \alpha_2$, eventually, or equivalently $\xi(\alpha_1) \leq \sqrt{n}(B_n - c_n) \leq \xi(\alpha_2)$, eventually. By the continuity of ξ it follows that $\sqrt{n}(B_n - c_n) \rightarrow \xi(\alpha)$.

By the uniform approximation of F_n by G_n , we have that

$$\sup_{|A| > \delta_n n^{-1/4}} F_n(A, B_n) = \sup_{|A| > \delta_n n^{-1/4}} G_n(A, B_n) + o(1),$$

which is strictly smaller than $1 - \alpha$, eventually, by assumption. Similarly $F_n(0, B_n) = G_n(0, B_n) + o(1) \rightarrow 1 - \alpha$. It follows that the maximum of $A \mapsto F_n(A, B_n)$ is taken on the interval $(-\delta_n n^{-1/4}, \delta_n n^{-1/4})$.

■

1.6 Annex : Extra Insights on KMW solution in Section 1.3.1

Let us rewrite (1.8) as

$$\Pi \left[\overbrace{\left(Q_\theta(-\infty, \hat{A} + B] - 1 + \frac{\delta}{2} \right)}^{E_1} - \overbrace{\left(Q_\theta(-\infty, \hat{A} - B] - \frac{\delta}{2} \right)}^{E_2} \geq 0 \mid X \right] = 1 - \alpha.$$

The KM criterion is to seek a value of B (denoted as $\hat{B}^*$) satisfying $\Pi(E_1 \geq 0, E_2 \leq 0 \mid X) = 1 - \alpha$. With provided $\hat{A}$ and δ , it should be clear that $\{\theta: E_1 \geq 0, E_2 \leq 0\} \subset \{\theta: E_1 - E_2 \geq 0\}$ for $\forall B > 0$, and both sets expand as B grows. Therefore, the B satisfying $\Pi(\{\theta: E_1 \geq 0, E_2 \leq 0\} \mid X) = 1 - \alpha$ must be larger than the B satisfying $\Pi(\{\theta: E_1 - E_2 \geq 0\} \mid X) = 1 - \alpha$. This is identical to say, if we plug $\hat{B}^*$ into $\Pi(E_1 - E_2 \geq 0 \mid X)$, it results in a total posterior mass strictly larger than $1 - \alpha$.

Now, rewrite (1.8) as

$$\Pi \left[\left(1 - Q_\theta(-\infty, \hat{A} + B] \right) + Q_\theta(-\infty, \hat{A} - B] \leq \delta \mid X \right] = 1 - \alpha.$$

The W criterion tries to optimize

$$\Pi \left[\overbrace{\left(1 - \frac{\delta}{2} - Q_\theta(-\infty, \hat{A} + B] \right)}^{E_1} + \overbrace{\left(Q_\theta(-\infty, \hat{A} - B] - \frac{\delta}{2} \right)}^{E_2} > 0 \mid X \right] = \alpha,$$

by requiring $\Pi(E_1 \geq 0, E_2 \geq 0 \mid X) = \alpha$, which is more stringent than it should be. Follow the similar reasoning, the estimate $\hat{B}^*$ fulfilling the W criterion will result in a total posterior mass strictly larger than α , and in turn less than $1 - \alpha$ on its dual side.

1.7 Annex: Gibbs Sampler for one-way Random Effect model

In this annex, we work out the Gibbs sampling scheme for deriving Bayesian estimates for a one-way random effect model. We consider two common formulations of one-way the random effect model together with two choices of priors, namely:

Formulation (a) $y_{ij} = \mu_i + \varepsilon_{ij}, \quad \mu_i \sim N(\mu, \tau^2)$

Formulation (b) $y_{ij} = \mu + a_i + \varepsilon_{ij}, \quad a_i \sim N(0, \tau^2)$

Prior (a) $\mu \sim N(\mu_0, \sigma_0^2)$ where μ_0 and σ_0^2 are given

Prior (b) $\mu \sim N(\mu_0, \sigma_0^2)$ where μ_0 is given and $\sigma_0^2 \sim IG(a_0, b_0)$

This gives four combinations, which are treated in Sections 1.7.1-1.7.4, with more precise definitions in the start of each section. To be clear, the density function of a inverse-gamma (denoted by IG) distribution is parameterized in the following way:

$$x \sim IG(\alpha, \beta), \quad f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp\left[\frac{-\beta}{x}\right], \quad (x > 0).$$

1.7.1 Scenario 1: Formulation (a) and Prior (a)

Given data points y_{ij} , being the j^{th} measure from the i^{th} group, we assume that

$$y_{ij} = \mu_i + \varepsilon_{ij}, \quad \mu_i \sim N(\mu, \tau^2), \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

where $i \in \{1, 2, \dots, m\}$, $j \in \{1, 2, \dots, n_i\}$, and any pair from $\{\mu_i\} \cup \{\varepsilon_{ij}\}$ are assumed mutually independent. To further simplify the expressions, let θ denote the vector of parameters $(\mu, \tau, \sigma)^t$ that govern the distribution of y_{ij} , and let y_i be the vector $(y_{i1}, y_{i2}, \dots, y_{in_i})^t$. Then y_i is independent to $y_{i'}$ for all $i \neq i'$ given θ . The prior of θ is set as the following:

$$\begin{aligned} p(\theta) &= p(\mu)p(\tau^2)p(\sigma^2), \quad \text{where} \\ \mu &\sim N(\mu_0, \sigma_0^2), \quad \tau^2 \sim IG(a_1, b_1), \quad \sigma^2 \sim IG(a_2, b_2) \\ \mu_0, \sigma_0^2, a_1, b_1, a_2, b_2 &\text{ are given.} \end{aligned}$$

The posterior can then be derived as below.

$$\begin{aligned} p(\theta | \text{Data}) &\propto p(\theta)p(\text{Data} | \theta) \\ &= p(\theta)p(y_1, \dots, y_m | \mu_1, \dots, \mu_m, \mu, \tau^2, \sigma^2) \\ &= \frac{p(\theta)p(y_1, \dots, y_m, \mu_1, \dots, \mu_m | \mu, \tau^2, \sigma^2)}{p(\mu_1, \dots, \mu_m | \mu, \tau^2)} \\ &= \frac{p(\theta)\mathcal{Q}}{p(\mu_1, \dots, \mu_m | \mu, \tau^2)} \propto p(\theta)\mathcal{Q} \end{aligned}$$

First, we work out the details of $\mathcal{Q}$:

$$\begin{aligned}
\mathcal{Q} &= p(y_1, \dots, y_m \mid \mu_1, \dots, \mu_m, \mu, \tau^2, \sigma^2) p(\mu_1, \dots, \mu_m \mid \mu, \tau^2) \\
&= \prod_{i=1}^m \left((2\pi)^{-\frac{1}{2}} (\tau^2)^{-\frac{1}{2}} \exp \left[-\frac{(\mu_i - \mu)^2}{2\tau^2} \right] \prod_{j=1}^{n_i} (2\pi)^{-\frac{1}{2}} (\sigma^2)^{-\frac{1}{2}} \exp \left[-\frac{(y_{ij} - \mu_i)^2}{2\sigma^2} \right] \right) \\
&= \prod_{i=1}^m \left((2\pi)^{-\frac{1}{2}} (\tau^2)^{-\frac{1}{2}} \exp \left[-\frac{(\mu_i - \mu)^2}{2\tau^2} \right] (2\pi)^{-\frac{n_i}{2}} (\sigma^2)^{-\frac{n_i}{2}} \exp \left[-\frac{\sum_j (y_{ij} - \mu_i)^2}{2\sigma^2} \right] \right) \\
&= (2\pi)^{-\frac{m - \sum n_i}{2}} (\tau^2)^{-\frac{m}{2}} (\sigma^2)^{-\frac{\sum n_i}{2}} \exp \left[-\frac{\sum_i (\mu_i - \mu)^2}{2\tau^2} \right] \exp \left[-\frac{\sum_i \sum_j (y_{ij} - \mu_i)^2}{2\sigma^2} \right] \\
&= (2\pi)^{-\frac{m - \sum n_i}{2}} (\tau^2)^{-\frac{m}{2}} (\sigma^2)^{-\frac{\sum n_i}{2}} \exp \left[-\frac{\sum_i \mu_i^2 - 2\mu \sum_i \mu_i + m\mu^2}{2\tau^2} \right] \\
&\quad \exp \left[-\frac{\sum_i \sum_j y_{ij}^2 - 2\sum_i \mu_i n_i \bar{y}_i + \sum_i n_i \mu_i^2}{2\sigma^2} \right] \tag{1.22}
\end{aligned}$$

Below we derive the conditional distributions of each element of θ given the other coordinates and the observations, the so-called full conditionals, used in a Gibbs sampling scheme. For brevity we denote the full set of conditioning variables by $\dots$, even though this set is different in different instances. It is helpful to remember, that a quadratic function $Ax^2 + Bx + c$ can be re-written as $A(x + \frac{B}{2A})^2 - \frac{B^2}{2A} + C$ if $A \neq 0$. This allows to recognize the values of the mean and the variance if the conditional posterior is Gaussian.

Posterior for $\mu_i, \forall i \in \{1, 2, \dots, m\}$

$$\begin{aligned}
p(\mu_i \mid \dots) &\propto \exp \left[-\frac{\mu_i^2 - 2\mu \mu_i}{2\tau^2} - \frac{-2\mu_i n_i \bar{y}_i + n_i \mu_i^2}{2\sigma^2} \right] \\
&\propto \exp \left[-\frac{(\sigma^2 + n_i \tau^2) \mu_i^2 - 2(n_i \bar{y}_i \tau^2 + \mu \sigma^2) \mu_i}{2\tau^2 \sigma^2} \right] \tag{1.23}
\end{aligned}$$

The above formula implies that $\mu_i \mid \dots \sim N(\frac{n_i \bar{y}_i \tau^2 + \mu \sigma^2}{\sigma^2 + n_i \tau^2}, \frac{\tau^2 \sigma^2}{\sigma^2 + n_i \tau^2})$.

Posterior for μ

$$\begin{aligned}
p(\mu \mid \dots) &\propto \exp \left[-\frac{\mu^2 - 2\mu_0 \mu + \mu_0^2}{2\sigma_0^2} - \frac{m\mu^2 - 2\mu \sum_i \mu_i}{2\tau^2} \right] \\
&\propto \exp \left[-\frac{(\tau^2 + m\sigma_0^2) \mu^2 - 2(\mu_0 \tau^2 + \sum_i \mu_i \sigma_0^2) \mu + \mu_0^2 \tau^2}{2\sigma_0^2 \tau^2} \right] \tag{1.24}
\end{aligned}$$

The above formula implies that $\mu \mid \dots \sim N\left(\frac{\mu_0 \tau^2 + \sum_i \mu_i \sigma_0^2}{\tau^2 + m \sigma_0^2}, \frac{\sigma_0^2 \tau^2}{\tau^2 + m \sigma_0^2}\right)$.

Posterior for τ^2

$$\begin{aligned} p(\tau^2 \mid \dots) &\propto (\tau^2)^{-a_1-1} \exp\left[-\frac{b_1}{\tau^2}\right] (\tau^2)^{-\frac{m}{2}} \exp\left[-\frac{\sum_i (\mu_i - \mu)^2}{2\tau^2}\right] \\ &\propto (\tau^2)^{-(a_1 + \frac{m}{2})-1} \exp\left[-\frac{2b_1 + \sum_i (\mu_i - \mu)^2}{2\tau^2}\right] \end{aligned} \quad (1.25)$$

The above formula implies that $\tau^2 \mid \dots \sim IG(a_1 + \frac{m}{2}, b_1 + \frac{\sum_i (\mu_i - \mu)^2}{2})$.

Posterior for σ^2

$$\begin{aligned} p(\sigma^2 \mid \dots) &\propto (\sigma^2)^{-a_2-1} \exp\left[-\frac{b_2}{\sigma^2}\right] (\sigma^2)^{-\frac{\sum_i n_i}{2}} \exp\left[-\frac{\sum_i \sum_j (y_{ij} - \mu_i)^2}{2\sigma^2}\right] \\ &\propto (\sigma^2)^{-(a_2 + \frac{\sum_i n_i}{2})-1} \exp\left[-\frac{2b_2 + \sum_i \sum_j (y_{ij} - \mu_i)^2}{2\sigma^2}\right] \end{aligned} \quad (1.26)$$

The above formula implies that $\sigma^2 \mid \dots \sim IG(a_2 + \frac{\sum_i n_i}{2}, b_2 + \frac{\sum_i \sum_j (y_{ij} - \mu_i)^2}{2})$.

1.7.2 Scenario 2: Formulation (a) and Prior (b)

Keep the setup in Scenario 1 except its prior $p(\theta)$, which will now be replaced by the following

$$\begin{aligned} p(\theta) &= p(\mu)p(\tau^2)p(\sigma^2), \quad \text{where} \\ \mu &\sim N(\mu_0, \sigma_0^2), \quad \sigma_0^2 \sim IG(a_0, b_0), \quad \tau^2 \sim IG(a_1, b_1), \quad \sigma^2 \sim IG(a_2, b_2) \\ \mu_0, a_0, b_0, a_1, b_1, a_2, b_2 &\text{ are all given.} \end{aligned}$$

It can be seen that $\mathcal{Q}$ will stay the same as in (1.22) under this construction of the prior for μ , and hence the full conditionals of $\mu_i, \mu, \tau^2, \sigma^2$ remain unchanged as in (1.23) - (1.26). The full conditional of σ_0^2 is also needed to build the Gibbs sampler, and can be derived as follows.

$$\begin{aligned} p(\sigma_0^2 \mid \dots) &\propto (\sigma_0^2)^{-a_0-1} \exp\left[-\frac{b_0}{\sigma_0^2}\right] (\sigma_0^2)^{-\frac{1}{2}} \exp\left[-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right] \\ &\propto (\sigma_0^2)^{-(a_0 + \frac{1}{2})-1} \exp\left[-\frac{2b_0 + (\mu - \mu_0)^2}{2\sigma_0^2}\right] \end{aligned} \quad (1.27)$$

The above formula implies that $\sigma_0^2 \mid \dots \sim IG(a_0 + \frac{1}{2}, b_0 + \frac{(\mu - \mu_0)^2}{2})$.

1.7.3 Scenario 3: Formulation (b) and Prior (a)

We now assume a different formulation for the distribution of y_{ij} , suggested in [17] among others, which is believed to lead to a better performance of the resulting Gibbs sampler. Let

$$y_{ij} = \mu + a_i + \varepsilon_{ij}, \quad a_i \sim N(0, \tau^2), \quad \varepsilon \sim N(0, \sigma^2)$$

where $i \in \{1, 2, \dots, m\}$, $j \in \{1, 2, \dots, n_i\}$, and the variables in the set $\{a_i\}_i \cup \{\varepsilon_{ij}\}_{i,j}$ are assumed mutually independent. The rest of the setting is the same as in Scenario 1. With these updated notations, the posterior is given by

$$\begin{aligned} p(\theta \mid Data) &\propto \frac{p(\theta)p(y_1, \dots, y_m, a_1, \dots, a_m \mid \mu, \tau^2, \sigma^2)}{p(a_1, \dots, a_m \mid \tau^2)} \\ &= \frac{p(\theta)\mathcal{Q}'}{p(a_1, \dots, a_m \mid \tau^2)} \propto p(\theta)\mathcal{Q}' \end{aligned}$$

where $\mathcal{Q}'$ can be elaborated as below.

$$\begin{aligned} \mathcal{Q}' &= p(y_1, \dots, y_m \mid a_1, \dots, a_m, \mu, \tau^2, \sigma^2)p(a_1, \dots, a_m \mid \mu, \tau^2) \\ &= \prod_{i=1}^m \left((2\pi)^{-\frac{1}{2}} (\tau^2)^{-\frac{1}{2}} \exp \left[-\frac{a_i^2}{2\tau^2} \right] \prod_{j=1}^{n_i} (2\pi)^{-\frac{1}{2}} (\sigma^2)^{-\frac{1}{2}} \exp \left[-\frac{(y_{ij} - \mu - a_i)^2}{2\sigma^2} \right] \right) \\ &= (2\pi)^{-\frac{m - \sum n_i}{2}} (\tau^2)^{-\frac{m}{2}} (\sigma^2)^{-\frac{\sum n_i}{2}} \exp \left[-\frac{a_i^2}{2\tau^2} \right] \exp \left[-\frac{\sum_i \sum_j (y_{ij} - \mu - a_i)^2}{2\sigma^2} \right] \end{aligned} \quad (1.28)$$

The last factor in (1.28) can be further detailed as

$$\exp \left[-\frac{\sum_i \sum_j y_{ij}^2 - 2\mu \sum_i (n_i \bar{y}_i) - 2 \sum_i (n_i a_i \bar{y}_i) + \mu^2 \sum_i n_i + 2\mu \sum_i (n_i a_i) + \sum_i n_i a_i^2}{2\sigma^2} \right] \quad (1.29)$$

Posterior for $a_i, \forall i \in \{1, 2, \dots, m\}$

$$\begin{aligned} p(a_i \mid \dots) &\propto \exp \left[-\frac{a_i^2}{2\tau^2} - \frac{-2(n_i a_i \bar{y}_i) + 2\mu n_i a_i + n_i a_i^2}{2\sigma^2} \right] \\ &\propto \exp \left[-\frac{(\sigma^2 + n_i \tau^2) a_i^2 - 2\tau^2 n_i (\bar{y}_i - \mu) a_i}{2\tau^2 \sigma^2} \right] \end{aligned} \quad (1.30)$$

The above formula implies that $a_i \mid \dots \sim N(\frac{\tau^2 n_i (\bar{y}_i - \mu)}{\sigma^2 + n_i \tau^2}, \frac{\tau^2 \sigma^2}{\sigma^2 + n_i \tau^2})$.

Posterior for μ

$$\begin{aligned}
p(\mu \mid \dots) &\propto \exp \left[-\frac{\mu^2 - 2\mu_0\mu + \mu_0^2}{2\sigma_0^2} - \frac{-2\mu \sum_i (n_i \bar{y}_i) + \mu^2 \sum_i n_i + 2\mu \sum_i (n_i a_i)}{2\sigma^2} \right] \\
&\propto \exp \left[-\frac{(\sigma^2 + \sum_i n_i \sigma_0^2)\mu^2 - 2[\mu_0\sigma^2 + (\sum_i n_i \bar{y}_i - \sum_i n_i a_i)\sigma_0^2]\mu + \mu_0^2\sigma^2}{2\sigma_0^2\sigma^2} \right]
\end{aligned} \tag{1.31}$$

The above formula implies that $\mu \mid \dots \sim N\left(\frac{\mu_0\sigma^2 + (\sum_i n_i \bar{y}_i - \sum_i n_i a_i)\sigma_0^2}{\sigma^2 + \sum_i n_i \sigma_0^2}, \frac{\sigma_0^2\sigma^2}{\sigma^2 + \sum_i n_i \sigma_0^2}\right)$.

Posterior for τ^2

$$\begin{aligned}
p(\tau^2 \mid \dots) &\propto (\tau^2)^{-a_1-1} \exp \left[-\frac{b_1}{\tau^2} \right] (\tau^2)^{-\frac{m}{2}} \exp \left[-\frac{\sum_i a_i^2}{2\tau^2} \right] \\
&\propto (\tau^2)^{-(a_1 + \frac{m}{2})-1} \exp \left[-\frac{2b_1 + \sum_i a_i^2}{2\tau^2} \right]
\end{aligned} \tag{1.32}$$

The above formula implies that $\tau^2 \mid \dots \sim IG(a_1 + \frac{m}{2}, b_1 + \frac{\sum_i a_i^2}{2})$.

Posterior for σ^2

$$\begin{aligned}
p(\sigma^2 \mid \dots) &\propto (\sigma^2)^{-a_2-1} \exp \left[-\frac{b_2}{\sigma^2} \right] (\sigma^2)^{-\frac{\sum_i n_i}{2}} \exp \left[-\frac{\sum_i \sum_j (y_{ij} - \mu - a_i)^2}{2\sigma^2} \right] \\
&\propto (\sigma^2)^{-(a_2 + \frac{\sum_i n_i}{2})-1} \exp \left[-\frac{2b_2 + \sum_i \sum_j (y_{ij} - \mu - a_i)^2}{2\sigma^2} \right]
\end{aligned} \tag{1.33}$$

The above formula implies that $\sigma^2 \mid \dots \sim IG(a_2 + \frac{\sum_i n_i}{2}, b_2 + \frac{\sum_i \sum_j (y_{ij} - \mu - a_i)^2}{2})$.

1.7.4 Scenario 4: Formulation (b) and Prior (b)

Keep the setup in Scenario 3 except its prior $p(\theta)$, which will now be replaced by the prior setting of Scenario 2.

It can be seen that $\mathcal{Q}'$ will stay the same as formula (1.28), and hence the colored full conditionals for $a_i, \mu, \tau^2, \sigma^2$ remain unchanged and are given in (1.30)-(1.33). The full conditional for σ_0^2 is also needed to build the Gibbs sampler, and it is exactly the same as (1.27) in Scenario 2.

1.7.5 Additional Note

Giving the prior setting (b), we assume that the priors of μ, τ^2 and σ^2 are mutually independent, and hence their corresponding posteriors are not conditionally conjugate. The posterior of μ and τ^2 are mutually dependent under formulation (a), while the posterior of μ and σ^2 are mutually dependent under formulation (b). Since we usually have much more information to estimate σ^2 than τ^2 in the data, formulation (b) may perform better.

A regime is to set $\sigma_0^2 = \tau^2 \omega$ under formulation (a) or $\sigma_0^2 = \sigma^2 \omega$ under formulation (b), for some large ω reflecting the non-informativeness. Thus, $p(\mu \mid \dots)$ becomes

$$N\left(\frac{\mu_0 + \omega \sum_i \mu_i}{1 + m\omega}, \frac{\omega \tau^2}{1 + m\omega}\right)$$

in (1.24), and

$$N\left(\frac{\mu_0 + (\sum_i n_i \bar{y}_i - \sum_i n_i a_i) \omega}{1 + \omega \sum_i n_i}, \frac{\omega \sigma^2}{1 + \omega \sum_i n_i}\right)$$

in (1.31).

Chapter 2

Another Parametric Application: Setting Control Limit at Release for Stability Assurance

2.1 Introduction

The term *critical quality attributes* (CQAs) is used to refer to the physical, chemical, biological or microbiological properties of a drug that are considered critical to patient safety and drug efficacy (see ICH Q8-R2). Before a batch of drug products can be released to the market, these CQAs are measured on samples taken from the batch, and it is ascertained that all measured values meet certain criteria, referred to as the *release specification* (**RS**). Per CQA, the sampling and the measurement are performed according to a registered laboratory procedure. Since drugs may deteriorate over time there exist separate criteria to ensure that a drug meets its quality specification until the end of its shelf life referred to as the *stability specification* (**SS**). Failing to meet the SS can lead to recall of the batch from the market. If a CQA of a drug is expected to change profoundly over time, its release and stability specifications are usually different.

Typically, the specification of a CQA consists of an acceptable range for a measured sample (also called a *reportable value*) of a batch, denoted as x , such as $x \geq \gamma_{low}$, $x \leq \gamma_{upp}$, or $\gamma_{low} \leq x \leq \gamma_{upp}$, where γ_{low} and γ_{upp} are registered specifications. Besides the theoretical "safe zone" for a CQA, the specification must incorporate a certain degree of random variation due to imperfections in the manufacturing process and laboratory measurements.

This chapter concerns a common task in stability risk control: given historical data, shelf life and stability specification (SS), determine a control limit for a measurement at release (i.e. time zero) that incorporates potential degrading of the drug over time. We make the objectives of this task more

explicit in the next section. This control limit could be a good reference for the release specification (RS) for a stability-indicating CQA in the pre-commercial phase, or could be used as an internal release limit (*IRL*, also known as *release target* in the literature), after commercialization. The latter is useful in a situation that equal RS and SS were registered, but non-negligible stability trend is observed in the larger dataset available after the product is sold in the market.

The remainder of the chapter is organized as follows. Two formulations to incorporate stability risk are presented in Section 2.2. Then we proceed to review a widely used benchmark approach, its probabilistic objectives and its drawbacks in Section 2.3. The underlying population model in this benchmark approach is oversimplified. We set up a more realistic population model in Section 2.4, and then present two probabilistic formulations of the objectives in Sections 2.5 and 2.6. Inference procedures based on data are also provided. The methods are illustrated on two datasets in Section 2.7, and a brief conclusion is drawn in Section 2.8. Proofs of propositions are collected in Section 2.9.

2.2 Problem formulation

Let Y_{it} be a reportable value (i.e. sample measurement) of a CQA for batch i at time t , where $i = 1, \dots, n$ and n is the number of measured batches. We assume a decreasing trend of the CQA over time, and want to ensure a stability specification (SS) (i.e. $\geq \gamma$) at the end of shelf life $T > 0$ by setting a lower limit ($\geq \eta$) at release ($t = 0$). Letting $Y_{n+1,t}$ denote a future reportable value, we make this precise in one of the following two objectives:

1. (Conditional on Trend/**CoT**) A future batch passes the specification $\geq \gamma$ at time T with at least the probability that it passes the limit $\geq \eta$ at time 0, i.e. $P(Y_{n+1,T} \geq \gamma) \geq P(Y_{n+1,0} \geq \eta)$.
2. (Conditional on Individual/**CoI**) Given that it passes the limit $\geq \eta$ at time 0 a future batch passes the limit $\geq \gamma$ with high probability at time T , i.e. $P(Y_{n+1,T} \geq \gamma | Y_{n+1,0} \geq \eta)$ is large.

It turns out to be an ambitious task to assure stability by controlling just a sample measurement at time 0. The key premises are (a) the random measurement error is not high, and (b) a future batch has a similar stability trend as the historical batches. The second premise is usually satisfied, because the drug batches are all manufactured and stored in the same manner. However, in modelling it is too strong to assume all batches follow exactly the same degradation pattern. More realistically, we expect that there is a typical pattern overall, while each individual batch still has a certain idiosyncrasy. In Section 2.4 we shall incorporate this in a mixed model.

The **CoT** formulation is typically requested when establishing a RS that is consistent with the SS. The riskiness of a batch with respect to the specification $\geq \gamma$ can be quantified by its failing rate $P_{\mu_T}(Y_{n+1,T} < \gamma)$, which depends on its true quality level μ_T (batch mean) at time T and laboratory measurement error. It is usually plausible that the variance of measurement error is constant over time, whence the riskiness of a batch changes only through its true quality level μ_t over time. If RS and SS

are the same and a non-negligible degradation exists, a batch might be risky with respect to the SS but not to the RS.

The **CoI** formulation gives a natural objective of quality assurance. However, when $Y_{n+1,0}$ and $Y_{n+1,T}$ only have weak correlation, and hence $Y_{n+1,0}$ carries not much information about $Y_{n+1,T}$, the **CoI** formulation becomes spurious. A weak correlation can be caused by large measurement error and/or large variation between the individual batches. If the trends of the empirical batches vary vastly, then the true quality level in the future is difficult to predict at time 0 even without measurement error, and the further the future, the less reliable the prediction. We discuss this further in Section 2.6.

In this study we shall only consider a (linearly) decreasing trend and lower-bounded SS intervals. Generalization to increasing trend and upper-bounded SS intervals is obvious. When the trend is in the direction of improving CQA, there is no stability risk.

2.3 Benchmark approach in the industry

The literature on controlling stability risk by setting a specification limit at time 0 is scarce. The most prevalent approach seems to be due to Allen, Dukes and Gerger (henceforth denoted as the ADG approach) [3].

In case that the trend in the CQA is linearly decreasing, they proposed to set the limit of the lower-bounded interval ($\geq \eta_{ADG}$) equal to

$$\eta_{ADG} = \gamma - \hat{b}T + t_{0.95, n^*} \sqrt{T^2 s_b^2 + \frac{\hat{\sigma}_e^2}{k}}, \quad (2.1)$$

where γ is the lower-sided SS, T is the shelf life (in months), $\hat{b}$ is an estimate of the degradation slope across historical batches in the data, s_b is the standard error of the estimator $\hat{b}$, $\hat{\sigma}_e^2/k$ is an estimate of laboratory method (assay) variance, k is the number of replicates to be averaged for a reportable value, and $t_{0.95, n^*}$ is the 95% percentile of the student-t distribution with n^* degrees of freedom. Because in their proposal, the estimates $\hat{b}$ and $\hat{\sigma}_e$ may not come from the same fitted model and may not be based on the same dataset, n^* is calculated by the Satterthwaite approximation.

An intuitive interpretation of (2.1) is straightforward: first increase the current SS ($\geq \gamma$) by the amount of the average degradation during shelf life (i.e. $-\hat{b}T > 0$), and then further tighten it by an error margin reflecting the uncertainty in the estimate of the slope and measurement error. However, the original paper did not present a population model and a clearly defined objective function, which prevents a precise interpretation by an associated probabilistic statement.

Here we present our best understanding. Assume that the observation $Y_{n+1,T}$ of a future batch at time T satisfies $Y_{n+1,T} = a + bT + e$, where a is the unknown true batch mean at time 0, b is the batch slope, and e is a residual error, assumed to possess a normal distribution $N(0, \sigma_e^2/k)$. Assume

that $\hat{b}$ is an unbiased estimator of b with a normal distribution with variance σ_b^2 , and that s_b and $\hat{\sigma}_e$ are estimators of σ_b and σ_e . Seeking the minimum value of a such that $P(Y_{n+1,T} \geq \gamma) \geq 0.95$, is equivalent to solving the following inequality for a :

$$P\left(\frac{Y_{n+1,T} - a - \hat{b}T}{\sqrt{T^2 s_b^2 + \frac{\hat{\sigma}_e^2}{k}}} \geq \frac{\gamma - a - \hat{b}T}{\sqrt{T^2 s_b^2 + \frac{\hat{\sigma}_e^2}{k}}}\right) \geq 0.95, \quad (2.2)$$

where $Y_{n+1,T} - a - \hat{b}T = (b - \hat{b})T + e$ is normally distributed with mean 0 and variance $\sigma_b^2 + \sigma_e^2/k$. Therefore, the statistic $(Y_{n+1,T} - a - \hat{b}T)/\sqrt{T^2 s_b^2 + \hat{\sigma}_e^2/k}$ should follow the Student-t distribution with n^* degrees of freedom. The solution to (2.2) is then $a \geq \gamma - \hat{b}T + t_{0.95, n^*} \sqrt{T^2 s_b^2 + \hat{\sigma}_e^2/k}$, the right side of which is the same as (2.1).

This reasoning suggests that the ADG approach sets a lower limit for the expectation (true mean) at zero of the reportable value of a future batch. As this population quantity is not observable, this may not be an attractive target. From an alternative perspective, we found that the approach is in a similar spirit as the **CoT** formulation, which is elaborated in the coming section. The key differences are that (1) the ADG approach does not explicitly model the batch-to-batch slope variation, the estimation error s_b^2 usually shrinks as the number of empirical batches increases in the data, and (2) the ADG approach tries to engage residual variation, which lumps the variation of the lab measuring method and trend variation among batches.

2.4 Our Population Model

In the remainder of the chapter we adopt the following linear random-coefficients model, which seems realistic in our industrial setting. A reportable value (a sample measurement) Y_{it} from batch i at time t follows

$$Y_{it} = a + \alpha_i + (b + \beta_i)t + \varepsilon_{it}. \quad (2.3)$$

The overall coefficients a and b are fixed effects, showing the average trend across batches. Each batch has a random deviation in its individual trend introduced by

$$\alpha_i \sim N(0, \sigma_\alpha^2), \quad \beta_i \sim N(0, \sigma_\beta^2).$$

The residual errors $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$ reflect the measurement error. For simplicity we assume that all random effects are mutually independent. The model is then parameterized by $\theta := (a, b, \sigma_\alpha, \sigma_\beta, \sigma_\varepsilon)$.

We make two remarks to justify this population model, one on linearity and the other on normality. (1) A linear trend might be the best we can handle in practice, unless some specific nonlinear trend is given strong scientific endorsement or is observed empirically during development. To detect

an arbitrary nonlinear pattern, we would need to measure the reported values on a fine time grid. However, a batch is typically observed at a total of 7 time points in a formal stability study, or at 4 or 5 time points in a follow-up stability study. This may not provide enough resolution to reliably distinguish a moderately nonlinear from a linear trend, let alone to distinguish between two types of nonlinear trends. (2) Only few CQAs are known to follow non-Gaussian distribution. One of the most important examples is the relative potency for biologics, known to follow a log-normal distribution due to its ratio nature. In such a case, one may seek an appropriate transformation of the reportable value, so that it follows (2.3).

2.5 CoT formulation

2.5.1 CoT in population model

The objective of **CoT** is to ensure that the passing rate of a future batch for $\geq \gamma$ at time T is at least the passing rate for $\geq \eta$ at the time of release. This makes that the riskiness of a batch will not be underestimated at time 0 when a trend exists. Under the population model (2.3), the trend of a future batch is random, and hence we cannot expect the objective to hold for every future batch. Instead we require it for at least a large proportion, say $100q\%$, of the future batches. The value of $q \in [0, 1]$ is left to the risk appetite in each individual case.

In formula, the objective of **CoT** is then, given γ , to find η that fulfils

$$P_{\alpha, \beta} \left[\frac{P_{\theta}(Y_{.0} \geq \eta | \alpha, \beta.)}{P_{\theta}(Y_{.T} \geq \gamma | \alpha, \beta.)} \leq 1 \right] = q. \quad (2.4)$$

The probabilities in the quotient are conditional given the random effects, while the outer probability refers to the random effects. By some rearrangements (see Section 2.9), one can show that the batch intercept does not play a role in this inequality, and hence the dependency on α of the outer probability can be dropped. Furthermore, the solution can be found explicitly and is given by

$$\eta_{CoT} = \gamma - [b + \Phi^{-1}(1 - q)\sigma_{\beta}]T, \quad (2.5)$$

where $\Phi^{-1}(1 - q)$ denotes the $(1 - q)^{th}$ quantile of the standard normal distribution.

This formula intuitively means to tighten the specification by the expected degradation up to time T , under the assumption that the decreasing trend follows the q^{th} steepest slope from the batch slope distribution. In the most trivial case when $\sigma_{\beta} \approx 0$, i.e. all batches have almost the same trend, we find that $\eta = \gamma - bT$, which is the intuitive solution to balance the stringency of specifications at time 0 and T .

2.5.2 Sample Inference for CoT

In formula (2.5), b and σ_β are unknown population parameters. The simplest solution would be to replace them by sample estimates $\hat{b}$ and $\hat{\sigma}_\beta$. Alternatively, in a Bayesian framework, the posterior distribution of η_{CoT} can be derived from the joint posterior distribution of (b, σ_β) . One can then take a suitable measure of location of that posterior to be the final estimate, say the posterior mean or median.

It is also possible to make extra effort in addressing the sample distribution, by deriving an estimate of η to meet

$$P \left(P_{\alpha_{n+1}, \beta_{n+1}} \left[\frac{P_\theta(Y_{n+1,0} \geq \eta | \alpha_{n+1}, \beta_{n+1})}{P_\theta(Y_{n+1,T} \geq \gamma | \alpha_{n+1}, \beta_{n+1})} \leq 1 \right] = q \mid \{Y_{it}\} \right) = \xi.$$

This solution to the above equation is simply a posterior quantile of η . When η is a lower limit, one might take $\xi > 0.5$ to be conservative.

2.6 CoI formulation

2.6.1 CoI in population model

Let us slightly generalize to two general time instants $t_1 < t_2$ instead of being specific to release (at time 0) and the end of shelf life (at time T). The objective of **CoI** is to set a control limit $\geq \eta$ at time t_1 , so that the event $Y_{it_1} \geq \eta$ indicates a high probability that the event $Y_{it_2} \geq \gamma$ will occur at a later time point t_2 , for any batch i from the population model (2.3).

In formula, this can be expressed as the requirement on the conditional probability, for given t_1, t_2, γ :

$$P_\theta(Y_{it_2} \geq \gamma | Y_{it_1} \geq \eta_{CoI}) = q, \quad \eta_{CoI} \geq \gamma, \quad (2.6)$$

where $\theta = (a, b, \sigma_\alpha, \sigma_\beta, \sigma_\epsilon)$, and q is the prescribed level of the desired assurance, which is typically set close to 1. The constraint $\eta_{CoI} \geq \gamma$ ensures that the solution η_{CoI} is practically meaningful.

The function $\eta_{CoI} \mapsto P_\theta(Y_{it_2} \geq \gamma | Y_{it_1} \geq \eta_{CoI})$ is increasing from $P_\theta(Y_{it_2} \geq \gamma)$ to (typically) 1 (see Lemma 2.6.1). Hence a solution η_{CoI} will exist for $q \geq P_\theta(Y_{it_2} \geq \gamma)$. The probability $P_\theta(Y_{it_1} \leq \eta_{CoI})$ is the proportion of produced batches that will be rejected at the proposed limit η_{CoI} , and this may be large. See Figure 2.1 for illustration. We shall examine the practicality of the **CoI** formulation from this perspective in the next subsection.

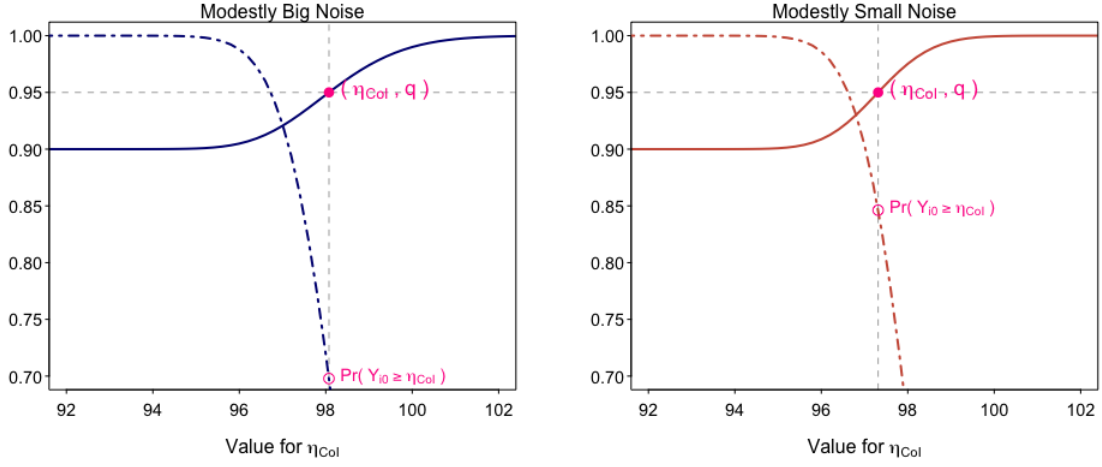


Fig. 2.1 The curves $\eta_{CoI} \mapsto P_{\theta}(Y_{iT} \geq \gamma | Y_{i0} \geq \eta_{CoI})$ (solid) and $\eta_{CoI} \mapsto P_{\theta}(Y_{i0} \geq \eta_{CoI})$ (Dot dash), with η_{CoI} on the horizontal axis, for two different parameter settings and $\gamma = 95$. To obtain stability assurance at $q = 0.95$ the value of η_{CoI} must be set to the value at the vertical dotted line. The height of the Dot dash curve on this line gives the proportion of accepted batches at time 0. Parameter settings: Shelf life $T = 24$ months. (Case I, Modestly Big Noise) : $a = 98.69, b = -0.0635, \sigma_{\alpha} = 1, \sigma_{\beta} = 0.05, \sigma_{\epsilon} = 0.655$, which gives intra-correlation $\rho = \sigma_{\alpha}^2 / (\sigma_{\alpha}^2 + \sigma_{\epsilon}^2) = 0.7$, $P(Y_{iT} \geq \gamma) = 0.9$ and $P(Y_{i0} \geq \gamma) = 0.999$. (Case II, Modestly Small Noise): $a = 98.45, b = -0.0729, \sigma_{\alpha} = 1, \sigma_{\beta} = 0.03, \sigma_{\epsilon} = 0.5$, which gives intra-correlation $\rho = 0.8$, $P(Y_{iT} \geq \gamma) = 0.9$ and $P(Y_{i0} \geq \gamma) = 0.999$.

We now derive the solution η_{CoI} to (2.6), assuming that all the parameters in (2.3) are known. One can show that the pair (Y_{it_1}, Y_{it_2}) follows a bivariate Gaussian distribution.

$$\begin{pmatrix} Y_{it_1} \\ Y_{it_2} \end{pmatrix} \sim N \left(\begin{pmatrix} a + bt_1 \\ a + bt_2 \end{pmatrix}, \begin{pmatrix} \sigma_{\alpha}^2 + \sigma_{\beta}^2 t_1^2 + \sigma_{\epsilon}^2 & \sigma_{\alpha}^2 + \sigma_{\beta}^2 t_1 t_2 \\ \sigma_{\alpha}^2 + \sigma_{\beta}^2 t_1 t_2 & \sigma_{\alpha}^2 + \sigma_{\beta}^2 t_2^2 + \sigma_{\epsilon}^2 \end{pmatrix} \right). \quad (2.7)$$

Equation (2.6) is identical to

$$\frac{P_{\theta}(Y_{it_2} \geq \gamma, Y_{it_1} \geq \eta_{CoI})}{P_{\theta}(Y_{it_1} \geq \eta_{CoI})} = \frac{\int_{\eta_{CoI}}^{\infty} \int_{\gamma}^{\infty} \varphi_{*}(s_1, s_2) ds_2 ds_1}{\int_{\eta_{CoI}}^{\infty} \varphi_{t_1}(s) ds} = q, \quad (2.8)$$

or, in a different format,

$$\frac{\int_{\eta_{CoI}}^{\infty} P_{\theta}(Y_{it_2} \geq \gamma | Y_{it_1} = s) \cdot \varphi_{t_1}(s) ds}{P_{\theta}(Y_{it_1} \geq \eta_{CoI})} = q, \quad (2.9)$$

where $\varphi_{*}(\cdot)$ is the density function of the bivariate Gaussian (2.7), and φ_{t_1} is the marginal density function of $Y_{it_1} \sim N(a + bt_1, \sigma_{\alpha}^2 + \sigma_{\beta}^2 t_1^2 + \sigma_{\epsilon}^2)$.

Equation (2.8) has no apparent analytical solution, but the numerator of the left side of the equation can be approximated efficiently via a numerical algorithm. We use *pmvnorm* from the R package {mvtnorm, 1.0-11} for this task. As for equation (2.9), it is known that the variate Y_{it_2} given $Y_{it_1} = s$

follows the $N(\mu_*, \sigma_*^2)$ -distribution with

$$\mu_* = (a + bt_2) - \frac{\sigma_\alpha^2 + \sigma_\beta^2 t_1 t_2}{\sigma_\alpha^2 + \sigma_\beta^2 t_1^2 + \sigma_\varepsilon^2} (a + bt_1 - s), \quad (2.10)$$

$$\sigma_*^2 = \sigma_\alpha^2 + \sigma_\beta^2 t_2^2 + \sigma_\varepsilon^2 - \frac{(\sigma_\alpha^2 + \sigma_\beta^2 t_1 t_2)^2}{\sigma_\alpha^2 + \sigma_\beta^2 t_1^2 + \sigma_\varepsilon^2}. \quad (2.11)$$

Hence the left side of equation (2.9) can be rewritten as $\int_\eta^\infty (1 - \Phi(\frac{\gamma - \mu_*}{\sigma_*})) \phi_{t_1}(s) ds$, which involves an incomplete Gaussian integral which also has no analytical solution. Nevertheless, (2.9) is a useful expression, because $P_\theta(Y_{it_2} \geq \gamma | Y_{it_1} = s)$ is the answer to a common inquiry in quality assessment: given the release data of batch i , what is the chance for this batch to pass the SS at time T ?

We have found that Wei [64] approached the problem in a similar way as (2.8) and called this the “unconditional rule”. However, Wei formulated the objective differently from (2.6) by focusing on the control of the 95% confidence limit of the predicted Y_{iT} , instead of Y_{iT} .

2.6.2 Practicality of CoI

Unlike the **CoT** solution (2.5), the **CoI** solution has no clear analytical form that can be easily interpreted. Therefore we make some extra effort to analyze its characteristics. This also facilitates to understand when the solution might not be helpful in practice, even if its formulation is an honest translation of our interests.

We first present two lemmas and next three insights that highlight key characteristics of the **CoI** solution. Proofs of the lemmas can be found in Section 2.9.

Lemma 2.6.1. *Suppose that (y_1, y_2) is a bivariate normal random vector with positive correlation. Then the map $\eta \mapsto P(y_2 \geq \gamma | y_1 \geq \eta)$ is increasing for any $\gamma \in \mathbb{R}$, from $P(y_2 \geq \gamma)$ to 1.*

Lemma 2.6.2. *Suppose that (y_1, y_2) is a bivariate normal vector with positive correlation. Then the map $\eta \mapsto P(y_2 \geq \gamma | y_1 = \eta)$ is increasing, for any $\gamma \in \mathbb{R}$. Consequently $P(y_2 \geq \gamma | y_1 \geq \eta) \geq P(y_2 \geq \gamma | y_1 = \eta)$, for any γ, η .*

[Insight 1]: As the correlation $\rho(t_1, t_2)$ between Y_{it_1} and Y_{it_2} tends to 0, the conditionally probability $P_\theta(Y_{it_2} \geq \gamma | Y_{it_1} \geq \eta)$ approaches the unconditional probability $P_\theta(Y_{it_2} \geq \gamma)$, and hence becomes independent of Y_{it_1} . In plain terms, this says that for small correlation there is almost no information in Y_{it_1} about Y_{it_2} , making it pointless to set a control for Y_{it_1} with the purpose to seek assurance that $Y_{it_2} \geq \gamma$. When $t_1 = 0$ and $t_2 = T$, the correlation is equal to $\rho(0, T) = \sigma_\alpha / \sqrt{(\sigma_\alpha^2 + \sigma_\varepsilon^2)(\sigma_\alpha^2 + \sigma_\beta^2 T^2 + \sigma_\varepsilon^2)}$. It will approach 0 if any of, or a combination of, the following occurs: (1) predict a far future, i.e. large T , (2) “unpredictable” individual batch trend, i.e. large σ_β , (3) big random noise from measurement error, i.e. large σ_ε relative to the process variation σ_α .

[Insight 2]: If the passing rate $P(Y_{iT} \geq \gamma)$ at time T is already higher than the desired assurance q , there is no need to set η_{Col} . This is because $P(Y_{iT} \geq \gamma | Y_{i0} \geq \eta)$ is lower bounded by $P(Y_{iT} \geq \gamma)$, for every η . Since $\eta < \gamma$ is practically meaningless, we may also first check if $q < P(Y_{iT} \geq \gamma | Y_{i0} \geq \gamma)$. Otherwise no practical solution of η_{Col} can be found.

[Insight 3]: An alternative narrative in the spirit of **CoI** is to find the minimum reportable value of a future batch at time t_1 that assures a passing rate of at least 100% at time t_2 , i.e.

$$\eta_{Alt} = \underset{\eta_*}{\operatorname{argmin}} P_\theta(Y_{it_2} \geq \gamma | Y_{it_1} = \eta_*) \geq q, \quad \eta_{Alt} \geq \gamma \quad (2.12)$$

This leads to a different solution than (2.6). By the first part of Lemma 2.6.2, the solution η_{Alt} is also simply the solution of the equation $P_\theta(Y_{it_2} \geq \gamma | Y_{it_1} = \eta_{Alt}) = q$. The second part of Lemma 2.6.2 shows that the resulting lower limit will be tighter than the solution of (2.6) for a given q (i.e. $\eta_{Alt} \geq \eta_{Col}$). For $t_1 = 0$ and $t_2 = T$, we find the conditional distribution of Y_{iT} given $Y_{i0} = \eta_*$ to be $N(\mu_{*0}, \sigma_{*0}^2)$, where $\mu_{*0} = (a + bT) - \rho(a - \eta_*)$, $\sigma_{*0}^2 = \sigma_\alpha^2 + \sigma_\beta^2 T^2 + \sigma_\epsilon^2 - \rho \sigma_\alpha^2$, and $\rho = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\epsilon^2)$. Therefore the solution to (2.12) is

$$\eta_{Alt} = \frac{\gamma - bT - (1 - \rho)a + \Phi^{-1}(q)\sigma_{*0}}{\rho}.$$

This inflates quickly as $\rho \downarrow 0$.

Insights 1 and 2 are relevant for implementation in practice. The former reveals that the **CoI** formulation may be spurious, while the latter exhibits situations where setting the control η_{Col} is not necessary. One can make a preliminary check (e.g. one for $\rho(0, T)$ and one for $P(Y_{iT} \geq \gamma)$) to decide the suitability and necessity of calculating η_{Col} .

We now present an example analysis to illustrate how it can be assessed when the **CoI** formulation may be helpful. A key consideration for implementation is the trade-off between the desired quality assurance, governed by q , and the business sustainability, reflected by $P(Y_{i0} \geq \eta_{Col})$. Low $P(Y_{i0} \geq \eta_{Col})$ means that at the determined control limit η_{Col} the majority of the manufactured batches will be rejected. From a business point of view this control is disastrous, as the true quality of these rejected batches may be sufficiently high (at least during the beginning of the period). As discussed, the tight limit in the **CoI** formulation could be due to the low correlation between Y_{i0} and Y_{iT} , which makes controlling Y_{iT} via Y_{i0} practically impossible.

Consider a hypothetical drug with CQA known to degrade over time, and with a registered SS of $95 \leq Y_{iT} \leq 105$. Consider two different process averages ($a = 100$ or $a = 98.5$), two different degrees of degradation ($bT = -2$ or $b = -3$), a process variation (i.e. variation of between-batch mean) fixed at $\sigma_\alpha = 0.5$, a batch slope variation $\sigma_\beta T$ varying over a grid $[0.3, 1]$, and measurement error σ_ϵ varying over $[0.1, 1.5]$. In practice these settings can come from prior knowledge of the product and the laboratory method during development. For each scenario, we derived η_{Col} for $q = 0.95$ and $q = 0.99$, and calculated and visualized the stringency $P(Y_{i0} \geq \eta_{Col})$. Figure 2.2 gives a heat map of the stringency for the best process average ($a = 100$, in the middle of the SS), with the four panels

referring to the four possible combinations of bT and q . The same results for the less advantageous process average ($a = 98.5$) are given in Figure 2.3.

One can use these heat maps to assess the suitability of deploying the **CoI** approach in a given situation by mapping the parameter values (inferred from preliminary data) to the appropriate square in the heat maps. If this square is in a column of deep-pink color, this indicates that the measurement error is too high, and we cannot effectively control the risk of Y_{iT} via the reportable value Y_{i0} . As a remedy, one may improve the precision of the laboratory measurement, so that the status shifts to a column that has at least some green cells. If the square is deep-pink, but in a column with a mix of green and deep-pink cells, one may consider shortening the shelf life of the drug.

In general one finds that the **CoI** approach quickly becomes impractical when the noise is large relative to the process variation σ_α , or the process average a is small, as indicated by the abundance of deep-pink squares.

2.6.3 Sample Inference for CoI

In practice, the population parameters θ in the model (2.3) are unknown, and must be estimated from sample data. To simplify the notation onwards, we fix $t_1 = 0$ and $t_2 = T$, where T is the shelf life in months (typically $T \in \{18, 19, \dots, 60\}$). We denote the empirical data by $y = \{Y_{it} : i = 1, 2, \dots, n; t = 0, \dots, T\}$, indexed by batch ID i and observed time t .

From a frequentist perspective, we may utilize an estimator $\hat{\theta}(y)$ of θ , and seek η_{Col} that fulfils

$$E \left[P_{\hat{\theta}(y)}(Y_{n+1,T} \geq \gamma | Y_{n+1,0} \geq \eta_{Col}) | y \right] \geq q, \quad \text{for } \forall \theta.$$

The expectation pertains to the sample distribution of y . However, it seems awfully complicated to solve such an equation exactly. Instead we may determine an approximate solution as follows. Writing the solution to equation (2.6) by $\eta_{Col}(\theta)$ for clarity, we plug in the estimator $\hat{\theta}(y)$ of θ , and use $\eta_{Col}(\hat{\theta}(y))$ as the control limit. As $\theta \mapsto \eta_{Col}(\theta)$ will be continuous, a consistent estimator for θ , i.e. $\hat{\theta}(y) \rightarrow \theta$ asymptotically, will lead to a consistent estimator for η_{Col} .

From a Bayesian perspective, the solution is more straightforward. We seek

$$E \left[P_\theta(Y_{n+1,T} \geq \gamma | Y_{n+1,0} \geq \eta) | y \right] = q.$$

Here the expectation refers to the posterior distribution of θ given y . This can be approximated by the empirical distribution of a sample $\{\theta_k\}_{k=1}^B$ of B values obtained via an MCMC algorithm, for some large number B . The final η_{Col} is the solution to $\frac{1}{B} \sum_{k=1}^B P_{\theta_k}(Y_{n+1,T} \geq \gamma | Y_{n+1,0} \geq \eta_{Col}) \approx q$. Since B is

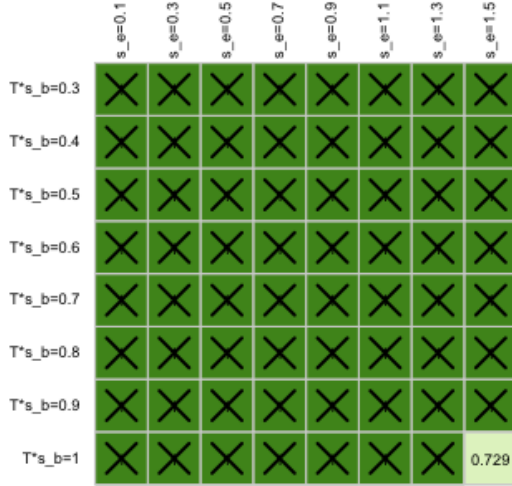
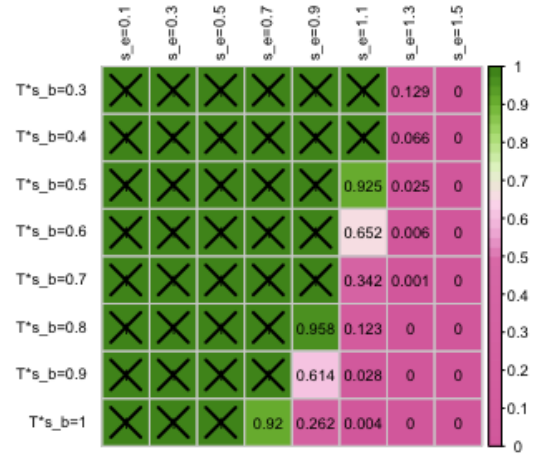
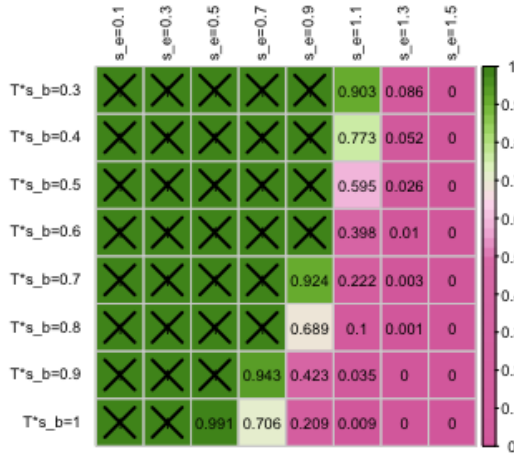
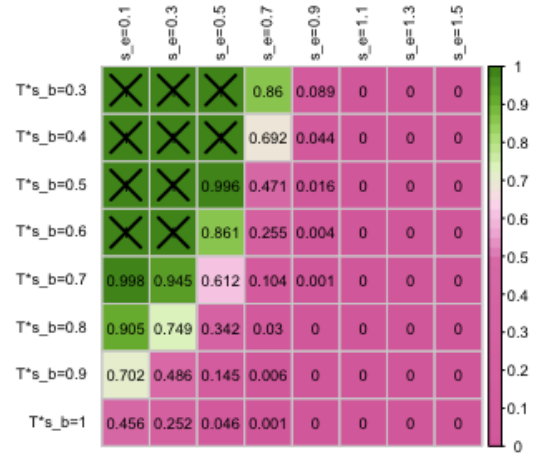
Modest Degradation till T ($bT=2$), Assurance request $q=0.95$ Modest Degradation till T ($bT=2$), Assurance request $q=0.99$ Large Degradation till T ($bT=3$), Assurance request $q=0.95$ Large Degradation till T ($bT=3$), Assurance request $q=0.99$ 

Fig. 2.2 Stringency $P(Y_{i0} \geq \eta_{CoI})$ indicated by color scale under different scenarios. In all cases $a = 100$. The four panels concern the four combinations of the degradation ($bT = -2, -3$) and required quality assurance ($q = 0.95, 0.99$). Each panel gives combinations of $T\sigma_\beta$ (indicated by $T*s_b$) and σ_e (indicated by s_e). Cells marked by a cross are situations in which $P(Y_{iT} \geq \gamma) \geq q$ and hence no η_{CoI} needs to be calculated.

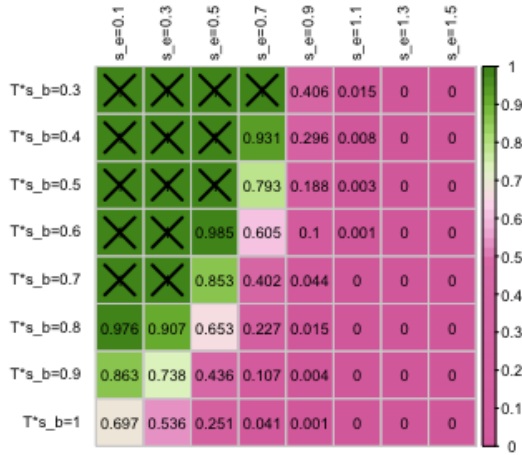
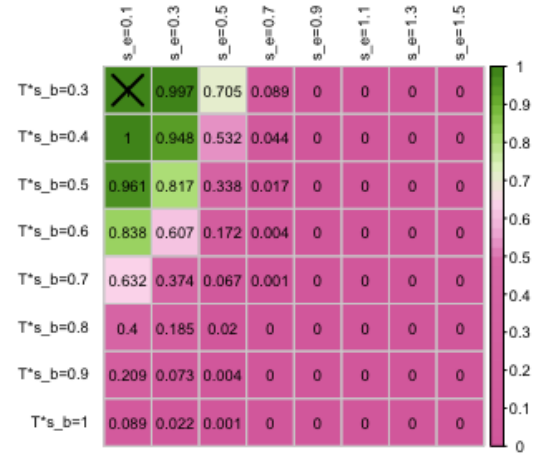
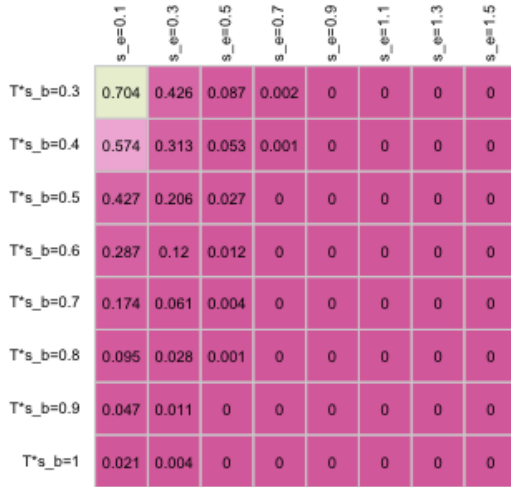
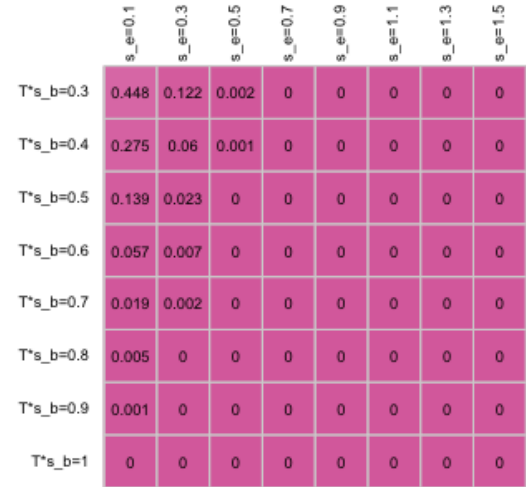
Modest Degradation till T ($bT=2$), Assurance request $q=0.95$ Modest Degradation till T ($bT=2$), Assurance request $q=0.99$ Large Degradation till T ($bT=3$), Assurance request $q=0.95$ Large Degradation till T ($bT=3$), Assurance request $q=0.99$ 

Fig. 2.3 Stringency $P(Y_{i0} \geq \eta_{Col})$ indicated by color scale under different scenarios. In all cases $a = 98.5$. The four panels concern the four combinations of the degradation ($bT = -2, -3$) and required quality assurance ($q = 0.95, 0.99$). Each panel gives combinations of $T\sigma_\beta$ (indicated by T^*s_b) and σ_e (indicated by s_e). Cells marked by a cross are situations in which $P(Y_{iT} \geq \gamma) \geq q$ and hence no η_{Col} needs to be calculated.

a finite number, exact equality may not be obtainable, and instead we take η_{Col} equal to

$$\min \left\{ \eta^*: \frac{1}{B} \sum_{k=1}^B P_{\theta_k}(Y_{n+1,T} \geq \gamma | Y_{n+1,0} \geq \eta^*) \geq q \right\}. \quad (2.13)$$

Denote this solution as **(B1)**.

An alternative solution **(B2)** under the Bayesian framework is as follows. The solution η_{Col} to (2.6) can be viewed as a function $\theta \mapsto \eta_{Col}(\theta)$. Thus the posterior distribution of θ induces a posterior distribution of $\eta_{Col}(\theta)$ given y . Given an MCMC sample $\{\theta_k\}_{k=1}^B$ this can be approximated by the empirical distribution of $\{\eta_{Col}(\theta_k)\}_{k=1}^B$. Next we can proceed to take a desired measure of location of this distribution as the final estimate, for instance the mean or median. The asymptotic properties of these two Bayesian estimator are discussed in section 2.10.

2.7 Calculation with Example datasets

We generated two datasets under the observational model (2.3), using two scenarios of parameter settings, and calculated η via the different approaches. For each scenario we set the shelf life T to 36 months, and SS equal to ≥ 95 , i.e. $r = 95$. We generated 10 batches of data ($i = 1, \dots, 10$), each with one observation at $t = 0, 6, 12, 18, 24$ months.

The parameters governing (2.3) are $a, b, \sigma_\alpha, \sigma_\beta$ and σ_e . We set $\rho = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_e^2)$, and write $\rho(0, T)$ for the correlation between Y_{i0} and Y_{iT} . The two parameter settings were:

- (Modest intra-correlation) $P(Y_{i0} \geq 95) = 0.9999, P(Y_{iT} \geq 95) = 0.9, \sigma_\alpha = 0.5, \sigma_\beta = 0.01, \rho = 0.8$. This results in $a = 97.1, b = -0.034, \sigma_e = 0.25$ and a modest correlation $\rho(0, T) = 0.67$.
- (Weak intra-correlation) $P(Y_{i0} \geq 95) = 0.9999, P(Y_{iT} \geq 95) = 0.9, \sigma_\alpha = 0.5, \sigma_\beta = 0.04, \rho = 0.4$. This results in $a = 100.9, b = -0.068, \sigma_e = 0.6$ and a weak correlation $\rho(0, T) = 0.19$.

For each dataset, η was calculated by four methods:

1. The ADG approach (2.1) with $q = 0.95$. The estimator $\hat{b}$ was taken equal to the ordinary least square estimator based on the $10 \times 5 = 50$ observations Y_{it} with corresponding standard error s_b , ignoring the mixed model structure of the simulation scheme, using the function *lm* from R basic. The error variance σ_e was estimated as the sample variance of the regression residuals after correcting for a linear trend per batch. The value of k was set to 1.
2. The **CoT** solution as in (2.5) with plug-in estimators. We estimated the parameters b and σ_β by fitting a linear mixed model, using the function *lmer* from the R package {lme4}. The parameter q was set to $q = 0.8$, where it is noted that this parameter has a different meaning as in the other two approaches. Setting it to 0.95 would lead to a very conservative **CoT** solution when the batch slope variation is not small.

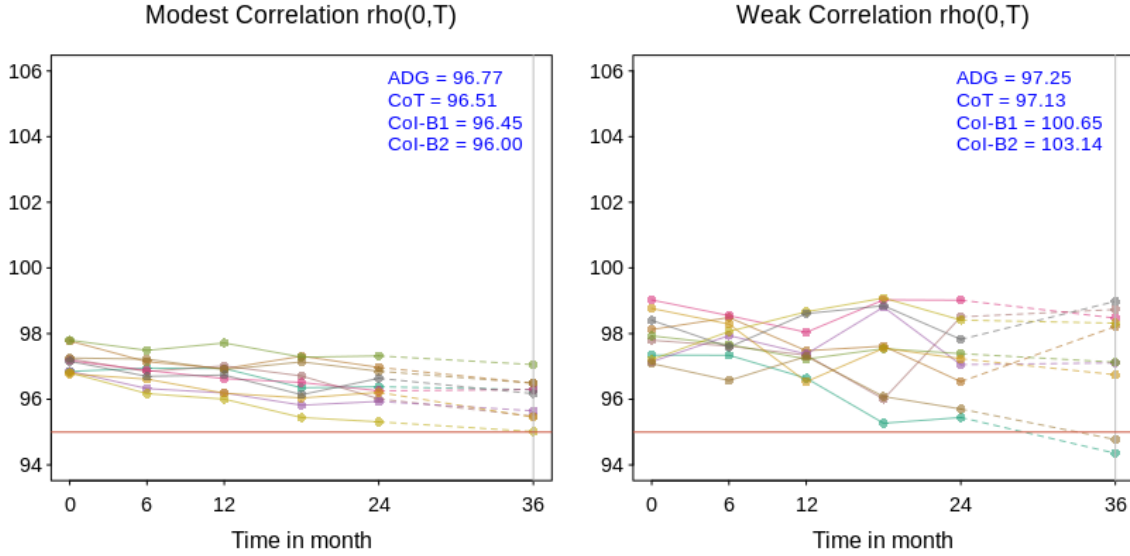


Fig. 2.4 Graphical representation of data simulated using a scenario of modest ($\rho(0,T) = 0.67$) and a scenario of weak ($\rho(0,T) = 0.19$) correlation between reportable values at release and end of shelf life. Shown are also the control limits η at time 0 for four methods, for stability specification set at $\geq \gamma = 95$. Data at 36 months are plotted for illustration, but were not included in the calculation of the control limits.

3. The two Bayesian solutions **B1** and **B2** for the **CoI** formulation, described in Section 2.6.3, with $q = 0.95$. The posterior distribution of the parameters was computed, via R package {rstan}, with 2000 burn-in and $B = 4000$ iterations, using a $N(0, 100^2)$ prior for a and b , and an $InvGamma(0.001, 0.001)$ prior for σ_α^2 , σ_β^2 and σ_ϵ^2 . For **CoI-B1**, we used the average across the MCMC sample in (2.13). For **CoI-B2**, equation (2.6) had no solution for some MCMC iterates θ_k (see Insight 2 in section 2.6.2), those $\eta_{Col}(\theta_k)$ were treated as some synthetic value < 95 , and the final estimate was taken to be the posterior median of $\{\eta_{Col}(\theta_k)\}_k$.

The example datasets and the calculated results are displayed in Figure 2.4. Data at 36 months are plotted for illustration, but were not included in the calculation of the control limits. One can see that with the dataset showing modest $\rho(0,T)$, the three approaches result in implementable control limits at time 0. However, when $\rho(0,T)$ is small, the **CoI** limits become impractically high.

2.8 Conclusion

Within a realistic population model, we have discussed two formulations for setting a control limit for a reportable value at time 0 that incorporates the stability risk of a drug. The two formulations correspond to two different probabilistic objectives. **CoT** seems to be generally appropriate. It determines a RS that tightens the given SS by a conservative estimate of average degradation over the

course of shelf life. The level of conservativeness is controlled by choosing an appropriate quantile from the estimated distribution of batch slope in a linear random-coefficients model. This prevents the RS to underestimate the risk of a batch at time 0 that will challenge the SS at time T . **CoI** is the most natural translation of the objective from a quality assurance perspective. However, the objective to control the stability risk by controlling the reportable value at release becomes unattainable when the correlation $\rho(0, T)$ between Y_{i0} and Y_{iT} is near zero. The control limit $\geq \eta_{CoI}$ may then be so tight that the majority of the batches will be rejected at time 0, as illustrated in Section 2.7. We presented a visualization by heat maps in Section 2.6.2 to help assess when the **CoI** approach will be helpful. In an abstract sense the essence is to assess if the signal-to-noise ratio in the data is sufficiently high. The heat maps can be navigated to infer how to enhance the signal-to-noise ratio, by taking actions that lift the current status from a deep-pink cell to a green cell (e.g. improve the precision of the laboratory measurement to certain level, or shorten the shelf life of the product by certain months).

2.9 Proofs

Proof of (2.10) and (2.11). Let

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

To derive the distribution of $x_1|x_2 = s$, we introduce an auxiliary random variable z , so that $x_1 = Ax_2 + z$, where $\text{cov}(x_2, z) = 0$ and $z \sim N(\mu_z, \Sigma_z)$. Then $x_1|x_2 = s$ follows $N(As + \mu_z, \Sigma_z)$.

To derive μ_z and Σ_z , we have

$$\begin{aligned} E(x_1) &= A\mu_2 + \mu_z & \implies & \mu_z = \mu_1 - A\mu_2 \\ \text{Var}(x_1) &= A\Sigma_{22}A' + \Sigma_z & \implies & \Sigma_z = \Sigma_{11} - A\Sigma_{22}A' \\ \text{Cov}(x_1, x_2) &= A\Sigma_{22} & \implies & A = \Sigma_{12}\Sigma_{22}^{-1} \end{aligned}$$

so $\mu_z = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2$ and $\Sigma_z = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

After clean up, $x_1|x_2 = s \sim N(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(s - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$. ■

Proof (Lemma 2.6.1). For the proof, we can reduce y_1 and y_2 to mean zero and variance 1 without losing generality. Hence assume

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

Then (y_1, y_2) has density

$$f(y_1, y_2) \propto \exp \left\{ -\frac{y_1^2 - 2\rho y_1 y_2 + y_2^2}{2(1 - \rho^2)} \right\}.$$

The density of y_2 given $y_1 \geq \eta$ is given by

$$f_\eta(y) = \frac{\int_\eta^\infty f(s, y) ds}{P(y_1 \geq \eta)} \propto \exp \left\{ -\frac{y^2}{2(1 - \rho^2)} \right\} \int_\eta^\infty \exp \left(\frac{\rho sy}{1 - \rho^2} \right) d\mu(s)$$

for $d\mu(s) = \exp \left\{ -\frac{s^2}{2(1 - \rho^2)} \right\} ds$. Hence

$$\frac{f_\eta(y)}{f_{\eta'}(y)} \propto \frac{\int_\eta^\infty \exp(\tau sy) d\mu(s)}{\int_{\eta'}^\infty \exp(\tau sy) d\mu(s)}, \quad \tau = \frac{\rho}{1 - \rho^2}.$$

We claim that for $\eta > \eta'$, this function is increasing in y . Indeed,

$$\begin{aligned} \frac{d}{dy} \log \frac{f_\eta(y)}{f_{\eta'}(y)} &= \frac{\int_\eta^\infty \tau s \exp(\tau sy) d\mu(s)}{\int_\eta^\infty \exp(\tau sy) d\mu(s)} - \frac{\int_{\eta'}^\infty \tau s \exp(\tau sy) d\mu(s)}{\int_{\eta'}^\infty \exp(\tau sy) d\mu(s)} \\ &= \frac{\tau \int_{\eta'}^\infty \exp(\tau sy) d\mu(s)}{\int_\eta^\infty \exp(\tau sy) d\mu(s)} [E\mathbf{1}_{s \geq \eta} S - E\mathbf{1}_{s \geq \eta'} S] \end{aligned}$$

if S is a random variable with density

$$p(s) = \frac{\exp(\tau sy) \mathbf{1}_{s \geq \eta'}}{\int_{\eta'}^\infty \exp(\tau sy) d\mu(s)}.$$

The term in square bracket is $\text{cov}(\mathbf{1}_{s \geq \eta} S)$ and is non-negative, as $s \mapsto \mathbf{1}_{s \geq \eta}$ and $s \mapsto s'$ are both increasing functions (see Lemma 2.9.1 below).

Now

$$\begin{aligned} P(y_2 \geq \gamma | y_1 \geq \eta) &= \int_\gamma^\infty f_\eta(y) dy \\ &= \int_\gamma^\infty \frac{f_\eta(y)}{f_{\eta'}(y)} f_{\eta'}(y) dy \\ &= E_{\eta'} \mathbf{1}_{y \geq \gamma} \frac{f_\eta(y)}{f_{\eta'}(y)}, \end{aligned}$$

for $y \sim f_{\eta'}$. Again by the covariance inequality, this is bigger than

$$E_{\eta'} \mathbf{1}_{y \geq \gamma} E_{\eta'} \frac{f_\eta(y)}{f_{\eta'}(y)} = P_{\eta'}(y \geq \gamma) \cdot 1.$$

This is the first assertion. ■

Proof (Lemma 2.6.2). The first assertion is immediate from the form of the conditional distribution of $y_2|y_1 = \eta$, which is $N(\lambda\eta + a, b^2)$ for some $\lambda > 0$.

The second assertion follows from

$$\begin{aligned} P(y_2 \geq \gamma | y_1 \geq \eta) &= \frac{\int_{\eta}^{\infty} P(y_2 \geq \gamma | y_1 = s) p_{y_1}(s) ds}{P(y_1 \geq \eta)} \\ &\geq \frac{\int_{\eta}^{\infty} P(y_2 \geq \gamma | y_1 = \eta) p_{y_1}(s) ds}{P(y_1 \geq \eta)} \\ &= P(y_2 \geq \gamma | y_1 = \eta). \end{aligned}$$

■

Lemma 2.9.1. *If f and g are non-decreasing functions $f, g: \mathbb{R} \mapsto \mathbb{R}$, then $\text{Cov}(f(u), g(u)) \geq 0$, for any random variable u .*

Proof. By monotonicity $(f(u) - f(v))(g(u) - g(v)) \geq 0$ for any u, v . Apply this to independent copies u, v of u , to find

$$\begin{aligned} 0 &\leq \mathbb{E}(f(u) - f(v))(g(u) - g(v)) \\ &= \mathbb{E}f(u)g(u) - \mathbb{E}f(u)\mathbb{E}g(v) - \mathbb{E}f(v)\mathbb{E}g(u) + \mathbb{E}f(v)g(v) \\ &= 2\mathbb{E}f(u)g(u) - 2\mathbb{E}f(u)\mathbb{E}g(v) \\ &= 2\text{Cov}(f(u), g(u)). \end{aligned}$$

■

Proof of formula (2.5). Under model (2.3), with $a^* = a + \alpha$ (drop the batch index i for brevity), the equation (2.4) gives

$$\begin{aligned} P_{\alpha, \beta} \left[1 - \Phi \left(\frac{\eta - a^*}{\sigma_e} \right) \leq 1 - \Phi \left(\frac{\gamma - a^* - (b + \beta)t}{\sigma_e} \right) \right] &= q \\ P_{\alpha, \beta} \left[\frac{\eta - a^*}{\sigma_e} \geq \frac{\gamma - a^* - (b + \beta)t}{\sigma_e} \right] &= q \\ P_{\beta} \left[\frac{\gamma - \eta - bt}{\sigma_{\beta}t} \leq \frac{\beta}{\sigma_{\beta}} \right] &= q \\ P_{\beta} \left[\frac{\beta}{\sigma_{\beta}} \leq \frac{\gamma - \eta - bt}{\sigma_{\beta}t} \right] &= 1 - q \\ \gamma - \eta - bt &= \Phi^{-1}(1 - q)\sigma_{\beta}t \end{aligned}$$

Hence, $\eta = \gamma - [b + \Phi^{-1}(1 - q)\sigma_{\beta}]t$. ■

2.10 Annex: Asymptotics of the Bayesian Estimators for CoI Formulation

Let $\Theta \subset \mathbb{R}^d$ be open, and let $H: \Theta \times \mathbb{R} \rightarrow \mathbb{R}$ be a continuously differentiable map such that $\eta \mapsto H(\theta, \eta)$ is strictly increasing from a negative value to a positive value, for every $\theta \in \Theta$. Let $\eta: \Theta \rightarrow \mathbb{R}$ be defined by

$$H(\theta, \eta(\theta)) = 0.$$

We are interested in the plug-in estimator $\eta(\hat{\theta})$ given an estimator $\hat{\theta}$ of θ , in the induced posterior distribution of $\eta(\theta)$ given a posterior distribution for θ , and in the solution to $\int H(\theta, \eta) \Pi(d\theta|X) = 0$, for $\Pi(\theta \in \cdot | X)$ a posterior distribution.

The intended application is the map $H(\theta, \eta) = P_\theta(Y_T \geq \gamma | Y_0 \geq \eta) - q$, for a bivariate normal vector (Y_0, Y_T) with positive correlation, and numbers γ and q such that $P(Y_T \geq \gamma) < q < 1$.

Write $H_\theta(\theta, \eta)$ for the $(d \times 1)$ vector of partial derivatives of H with respect to θ and $H_\eta(\theta, \eta)$ for the partial derivative with respect to η .

Let $\hat{\theta}_n$ be estimators based on some set of data X_n , for $n = 1, 2, \dots$. The following lemma shows that asymptotic normality of these estimators carries over onto the solutions $\eta(\hat{\theta}_n)$.

Lemma 2.10.1. *If $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N_d(0, \Sigma_0)$, for some $\theta_0 \in \Theta$ and positive-definite matrix Σ_0 , then $\sqrt{n}(\eta(\hat{\theta}_n) - \eta_0) \rightsquigarrow N(0, \tau_0^2)$, for $\eta_0 = \eta(\theta_0)$ and $\tau_0^2 = H_\eta(\theta_0, \eta_0)^{-2} H_\theta(\theta_0, \eta_0)^T \Sigma_0 H_\theta(\theta_0, \eta_0)$, provided that $H_\eta(\theta_0, \eta_0) > 0$.*

Proof. Abbreviate $\hat{\eta} = \eta(\hat{\theta}_n)$. By the monotonicity of H in its second argument $P(\hat{\eta} > \eta_0 + \varepsilon) \leq P(H(\hat{\theta}_n, \eta_0 + \varepsilon) \leq 0)$, for any $\varepsilon > 0$. Since $H(\hat{\theta}_n, \eta_0 + \varepsilon) \rightarrow H(\theta_0, \eta_0 + \varepsilon)$ in probability by continuity of H , and $H(\theta_0, \eta_0 + \varepsilon) > 0$, it follows that $P(\hat{\eta} > \eta_0 + \varepsilon) \rightarrow 0$. Combined with a similar bound on the left side, it follows that $\hat{\eta} \rightarrow \eta_0$, in probability.

Define a function R by

$$R(\theta, \eta) = \frac{H(\theta, \eta) - H(\theta_0, \eta_0) - H_\theta(\theta_0, \eta_0)^T(\theta - \theta_0) - H_\eta(\theta_0, \eta_0)(\eta - \eta_0)}{\|\theta - \theta_0\| + |\eta - \eta_0|}.$$

Then $R(\hat{\theta}, \hat{\eta}) \rightarrow 0$ in probability, by the continuous mapping theorem and because H is assumed differentiable. By the definition of $\hat{\eta}$ and η_0 we have

$$\begin{aligned} 0 &= H(\hat{\theta}, \hat{\eta}) - H(\theta_0, \eta_0) \\ &= H_\theta(\theta_0, \eta_0)^T(\hat{\theta} - \theta_0) + H_\eta(\theta_0, \eta_0)(\hat{\eta} - \eta_0) + R(\hat{\theta}, \hat{\eta})(\|\hat{\theta} - \theta_0\| + |\hat{\eta} - \eta_0|). \end{aligned}$$

It follows that, with $0/0 = 0$,

$$\begin{aligned} \sqrt{n}(\hat{\eta} - \eta_0) & \left(H_{\eta}(\theta_0, \eta_0) + o_P(1) \frac{|\hat{\eta} - \eta_0|}{\hat{\eta} - \eta_0} \right) \\ & = -H_{\theta}(\theta_0, \eta_0)^T \sqrt{n}(\hat{\theta} - \theta_0) + o_P(1) \sqrt{n}\|\hat{\theta} - \theta_0\|. \end{aligned}$$

The result follows by Slutsky's lemma and the fact that $H_{\eta}(\theta_0, \eta_0) > 0$. ■

Next assume that for a given prior distribution on Θ , we obtain a posterior distribution $\mathcal{L}(\theta|X_n)$. Let $\|\cdot\|$ denote the total variation norm. Under the Bernstein-von Mises theorem, the posterior distribution permits a normal approximation as assumed in the following lemma.

Lemma 2.10.2. *If $\|\mathcal{L}(\theta|X_n) - N_d(\hat{\theta}_n, n^{-1}\Sigma_0)\| \rightarrow 0$ in probability, for estimators $\hat{\theta}_n$ such that $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N_d(0, \Sigma_0)$, for some $\theta_0 \in \Theta$ and positive-definite matrix Σ_0 , then $\|\mathcal{L}(\eta(\theta)|X_n) - N_d(\eta(\hat{\theta}_n), n^{-1}\tau_0^2)\| \rightarrow 0$ in probability, for τ_0 as in the preceding lemma, provided $H_{\eta}(\theta_0, \eta_0) > 0$.*

Proof. Fix arbitrary $a_2, \dots, a_d \in \mathbb{R}^d$ such that $H_{\theta}(\theta_0, \eta_0), a_2, \dots, a_d$ are orthogonal in $\mathbb{R}^d$ and $a_2, \dots, a_d$ have norm 1. Define a map $\bar{H}: \Theta \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ by

$$\bar{H}(\theta, \eta) = \begin{pmatrix} H(\theta, \eta_1) \\ a_2^T \theta - \eta_2 \\ \vdots \\ a_d^T \theta - \eta_d \end{pmatrix}.$$

Then $\bar{H}$ is continuously differentiable with

$$\frac{\partial \bar{H}(\theta, \eta)}{\partial \theta_1 \cdots \partial \theta_d} = \begin{pmatrix} H_{\theta}(\theta, \eta_1)^T \\ a_2^T \\ \vdots \\ a_d^T \end{pmatrix}, \quad \frac{\partial \bar{H}(\theta, \eta)}{\partial \eta_1 \cdots \partial \eta_d} = \begin{pmatrix} H_{\eta}(\theta, \eta_1) & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

The second matrix at $(\theta_0, \bar{\eta}_0)$, where $\bar{\eta}_0 = (\eta(\theta_0), a_2^T \theta_0, \dots, a_d^T \theta_0^T)$, is invertible. By the implicit function theorem there exists an open neighbourhood G of θ_0 and a diffeomorphism $\bar{\eta}: G \rightarrow \bar{\eta}(G) \subset \mathbb{R}^d$ such that $\bar{H}(\theta, \bar{\eta}(\theta)) = 0$, for $\theta \in G$. Clearly $\bar{\eta}_1(\theta) = \eta(\theta)$. Abbreviate $\hat{\eta} = \bar{\eta}(\hat{\theta})$ and $\bar{\eta}_0 = \bar{\eta}(\theta_0)$.

By the same method as in the preceding lemma we find that $\sqrt{n}(\hat{\eta} - \bar{\eta}_0)$ tends in distribution to a normal distribution. For a given bounded set $B \subset \mathbb{R}^d$, the probability of the event $\hat{\eta} + B/\sqrt{n} \subset G$ therefore tends to one. On this event we have that $\sqrt{n}(\bar{\eta}(\theta) - \hat{\eta}) \in B$ if and only if $\theta \in \bar{\eta}^{-1}(\hat{\eta} + B/\sqrt{n})$, where $\bar{\eta}^{-1}$ is the (ordinary) inverse function of $\bar{\eta}$. By assumption $\theta|X_n$ is asymptotically distributed as $\hat{\theta} + \Sigma_0^{1/2}Z/\sqrt{n}$, for a standard normal vector Z . We thus have that, with ϕ the density of

Z ,

$$\begin{aligned} \mathbb{P}\left(\sqrt{n}\left(\bar{\eta}(\theta) - \hat{\eta}\right) \in B | X_n\right) &= \mathbb{P}\left(\theta \in \bar{\eta}^{-1}(\hat{\eta} + B/\sqrt{n}) | X_n\right) \\ &= \mathbb{P}\left(\hat{\theta} + \Sigma_0^{1/2} Z/\sqrt{n} \in \bar{\eta}^{-1}(\hat{\eta} + B/\sqrt{n})\right) + o_P(1) \end{aligned} \quad (2.14)$$

$$\begin{aligned} &= \int_{z: \sqrt{n}(\bar{\eta}(\hat{\theta} + \Sigma_0^{1/2} z/\sqrt{n}) - \hat{\eta}) \in B} \phi(z) dz + o_P(1) \\ &= \int_B \phi\left(\sqrt{n}\Sigma_0^{-1/2}\left(\bar{\eta}^{-1}(\hat{\eta} + y/\sqrt{n}) - \hat{\theta}\right)\right) \left|\frac{dz}{dy}\right| dy + o_P(1), \end{aligned} \quad (2.15)$$

where we have made the substitution $\sqrt{n}(\bar{\eta}(\hat{\theta} + \Sigma_0^{1/2} z/\sqrt{n}) - \hat{\eta}) = y$, and $|dz/dy|$ is (the determinant of) the Jacobian of the transformation. Here

$$\begin{aligned} \sqrt{n}\Sigma_0^{-1/2}\left(\bar{\eta}^{-1}(\hat{\eta} + y/\sqrt{n}) - \hat{\theta}\right) &\rightarrow \Sigma_0^{-1/2}(\bar{\eta}^{-1})'(\eta_0)y = \Sigma_0^{-1/2}\bar{\eta}'(\theta_0)^{-1}y, \\ \frac{dz}{dy} &= \Sigma_0^{-1/2}(\bar{\eta}^{-1})'(\hat{\eta} + y/\sqrt{n}) \rightarrow \Sigma_0^{-1/2}\bar{\eta}'(\theta_0)^{-1}. \end{aligned}$$

The convergence is uniform in y ranging over bounded sets. It follows that the integral tends to

$$\int_B \phi\left(\Sigma_0^{-1/2}\bar{\eta}'(\theta_0)^{-1}y\right) \left|\Sigma_0^{-1/2}\bar{\eta}'(\theta_0)^{-1}\right| dy.$$

We recognize this as the probability of the set B under the normal distribution $N_d(0, \bar{\eta}'(\theta_0)\Sigma_0\bar{\eta}'(\theta_0)^T)$. The first marginal of this distribution is the normal distribution in the lemma.

To complete the proof we argue that the convergence is valid in the sense of the total variation distance. The approximation with a Gaussian distribution is valid uniformly in B by assumption. Thus it suffices to show that (2.14) tends to the final equation, uniformly in B . Because the integrand in (2.15) converges uniformly for y in bounded sets, the Gaussian probability in (2.14) converges uniformly in bounded sets B . Since we can find a sufficiently large compact set that contains most of the mass, this suffices. ■

As a consequence of the preceding lemma, the posterior median or posterior trimmed mean $\hat{\eta}$ of $\mathcal{L}(\eta(\theta)|X_n)$ satisfies $\sqrt{n}(\hat{\eta} - \eta(\theta_0)) \rightsquigarrow N(0, \tau_0^2)$. (Convergence of the posterior mean would require some extra conditions, as the range of $\eta(\theta)$ may be unbounded, depending on H , presumably also beyond the assumptions of the next lemma.)

Finally given a posterior distribution $\mathcal{L}(\theta|X_n)$, consider the solution $\tilde{\eta}(X_n)$ to $\mathbb{E}(H(\theta, \eta)|X_n) = 0$. Thus we first take the average of the curves $\eta \mapsto H(\theta, \eta)$ relative to the posterior distribution, and then determine a zero. Although in general $\tilde{\eta}_n$ will be different from any of the preceding, the following lemma shows that it is asymptotically very close to the plug-in estimator using the posterior mean $\mathbb{E}(\theta|X_n)$ as the estimator of θ .

Strengthen the assumptions on H to: H is bounded and twice continuously differentiable. Write $H_{\theta\theta}$ for the second derivative matrix of H relative to θ .

Lemma 2.10.3. *Suppose that $n \text{Cov}(\theta|X_n) \rightarrow \Sigma_0$ in probability, and $n^2 E(\|\theta - \hat{\theta}_n\|^4|X_n) = O_P(1)$, for $\hat{\theta}_n = E(\theta|X_n)$. Assume that $(\hat{\theta}_n, \eta(\hat{\theta}_n)) \rightarrow (\theta_0, \eta_0)$ in probability. Then $n(\tilde{\eta}_n - \eta(\hat{\theta}_n)) \rightarrow -\frac{1}{2}H_\eta(\theta_0, \eta_0)^{-1} \text{tr}(H_{\theta\theta}(\theta_0, \eta_0)\Sigma_0)$, in probability.*

Proof. Write $\hat{\eta} = \eta(\hat{\theta})$. By the monotonicity of $\eta \mapsto E(H(\theta, \eta)|X_n)$ and the definition of $\tilde{\eta}$ we have $P(\tilde{\eta} > \eta_0 + \varepsilon) \leq P(E(H(\theta, \eta_0 + \varepsilon)|X_n) \leq 0)$, for every $\varepsilon > 0$, which tends to zero since $E(H(\theta, \eta_0 + \varepsilon)|X_n) \rightarrow H(\theta_0, \eta_0 + \varepsilon) > 0$, in probability. Combined with a similar argument for the left side, this shows that $\tilde{\eta} \rightarrow \eta_0$, in probability.

By the assumption on the fourth moments, we have for, $k = 0, 1, 2$, that $E(\|\theta - \hat{\theta}\|^k 1_{\|\theta - \hat{\theta}\| > \delta} | X_n) = O_P(n^{-(2-k/2)})$, for every $\delta > 0$. In view of the assumed boundedness of H , this property for $k = 0$ and the definition of $\tilde{\eta}$ show that

$$E(H(\theta, \tilde{\eta}) 1_{\|\theta - \hat{\theta}\| \leq \delta} | X_n) = O_P(n^{-2}).$$

The property for $k = 1$ and the definition of $\hat{\theta}$ show that $E((\theta - \hat{\theta}) 1_{\|\theta - \hat{\theta}\| \leq \delta} | X_n) = O_P(n^{-3/2})$.

Define $\hat{R}(\theta, \eta)$ as

$$\hat{R}(\theta, \eta) = \frac{H(\theta, \eta) - H(\hat{\theta}, \hat{\eta}) - P_2(\theta, \eta)}{\|\theta - \hat{\theta}\|^2 + |\eta - \hat{\eta}|^2},$$

where $P_2(\theta, \eta)$ is the second order Taylor polynomial of H around $(\hat{\theta}, \hat{\eta})$, i.e.

$$\begin{aligned} P_2(\theta, \eta) &= H_\theta(\hat{\theta}, \hat{\eta})^T (\theta - \hat{\theta}) + H_\eta(\hat{\theta}, \hat{\eta}) (\eta - \hat{\eta}) \\ &\quad + \frac{1}{2} (\theta - \hat{\theta})^T H_{\theta\theta}(\hat{\theta}, \hat{\eta}) (\theta - \hat{\theta}) + (\theta - \hat{\theta})^T H_{\theta\eta}(\hat{\theta}, \hat{\eta}) (\eta - \hat{\eta}) + \frac{1}{2} H_{\eta\eta}(\hat{\theta}, \hat{\eta}) (\eta - \hat{\eta})^2. \end{aligned}$$

Since $H(\hat{\theta}, \hat{\eta}) = 0$ by definition of $\hat{\eta}$, we have

$$\begin{aligned} O_P(n^{-2}) &= E(H(\theta, \tilde{\eta}) 1_{\|\theta - \hat{\theta}\| \leq \delta} | X_n) \\ &= E(P_2(\theta, \tilde{\eta}) 1_{\|\theta - \hat{\theta}\| \leq \delta} | X_n) + E(R(\theta, \tilde{\eta}) (\|\theta - \hat{\theta}\|^2 + |\tilde{\eta} - \hat{\eta}|^2) 1_{\|\theta - \hat{\theta}\| \leq \delta} | X_n). \end{aligned}$$

Since R is bounded in a neighbourhood of (θ_0, η_0) , we can choose δ small enough so that R can be bounded out of the final term, and next the indicator, which is then seen to be $O_P(n^{-2}) + O_P(|\tilde{\eta} - \hat{\eta}|^2)$. On substituting the definition of P_2 and taking the conditional expectation with respect to θ given X_n over the set $\|\theta - \hat{\theta}\| \leq \delta$, we see that term that is linear in $\theta - \hat{\theta}$ is $O_P(n^{-3/2})$, while the mixed term

in $\theta - \hat{\theta}$ and $\tilde{\eta} - \hat{\eta}$ is $O_P(n^{-3/2})|\tilde{\eta} - \hat{\eta}|$. Rearranging the other terms yields

$$\begin{aligned} & (\tilde{\eta} - \hat{\eta}) \left(H_{\eta}(\theta_0, \eta_0) + o_P(1) \right) \\ &= -\frac{1}{2} \mathbb{E} \left((\theta - \hat{\theta})^T H_{\theta\theta}(\hat{\theta}, \hat{\eta}) (\theta - \hat{\theta}) 1_{\|\theta - \hat{\theta}\| \leq \delta} \mid X_n \right) + O_P(n^{-3/2}). \end{aligned}$$

The indicator within the conditional expectation can be removed, and $H_{\theta\theta}(\hat{\theta}, \hat{\eta})$ replaced by $H_{\theta\theta}(\theta_0, \eta_0)$, at the cost of an $o_P(n^{-1})$ -term. The expectation can next be computed as the trace of the matrix $H_{\theta\theta}(\theta_0, \eta_0) \text{Cov}(\theta \mid X_n)$. This matrix times n tends to $H_{\theta\theta}(\theta_0, \eta_0) \Sigma_0$, by assumption. ■

The intended application is to find a “control limit” η such that $P_{\theta}(Y_T \geq \gamma \mid Y_0 \geq \eta) = q$. If θ is unknown, it is replaced by a data-based approximation. The following corollary shows that estimation leads to a random error of order $O_P(n^{-1/2})$.

Corollary 2.10.1. *Assume that the conditions of the preceding lemmas hold and let $\hat{\eta}$ be either the plug-in estimator $\eta(\hat{\theta})$, or the median or trimmed mean of the posterior distribution of $\eta(\theta)$, or the solution to $\mathbb{E}(H(\theta, \eta) \mid X_n) = 0$. Then $\sqrt{n}(H(\theta_0, \hat{\eta}) - H(\theta_0, \eta_0)) \rightsquigarrow N(0, H_{\theta}(\theta_0, \eta_0)^T \Sigma_0 H_{\theta}(\theta_0, \eta_0))$.*

Proof. This is an immediate consequence of the three lemmas and the Delta-method. ■

Chapter 3

Nonparametric Prediction Regions

3.1 Introduction

In the absence of information on grouping or covariates, observed data $X = \{X_i \in \mathbb{R}^d : i = 1, \dots, n\}$ are often assumed to be an independent and identically distributed (i.i.d.) sample from some distribution F . If we know little about this distribution, and the empirical distribution of X shows features that invalidate common parametric assumptions, we may resort to nonparametric statistical approaches. In this chapter we shall be concerned with nonparametric tolerance sets.

If case F is known, a *prediction set* $\mathcal{R}_F$ of level $1 - \delta$ for a future observation $X_{n+1} \sim F$ is a subset of the sample space such that

$$F(\mathcal{R}_F) \equiv P_F(X_{n+1} \in \mathcal{R}_F) \geq 1 - \delta. \quad (3.1)$$

The set $\mathcal{R}_F$ can (and usually will) depend on F , as indicated in the notation.

A prediction set can take many shapes, in particular in higher dimensions. A natural candidate, which we shall study in this chapter, is a *density level set*. For f a density of F and $L(t) = \{x \in \mathbb{R}^d : f(x) \geq t\}$, we define a threshold by $t_F(\delta) = \inf\{t : F(L(t)) \geq 1 - \delta\}$, and use $\mathcal{R}_F = L(t_F(\delta))$ as prediction set. Such a construction is intuitive and leads to prediction sets of minimum volume. Other methods to construct prediction sets are based on some form of ranking the observations.

In practice, the distribution F or its density f are almost never known, and the prediction set must be estimated from the observed sample X . In the nonparametric situation, in particular for a density level set, the sample size n should be adequately large, so that the estimators can be reliable. This is especially true when δ is small, and the prediction set must incorporate aspects of the tail of the distribution.

We denote a data-based prediction set by $\mathcal{R}_n(X)$, or $\mathcal{R}_n$ for brevity, instead of $\mathcal{R}_F$, as a reminder that it is estimated via $X \sim F^n$. The set ought to satisfy an analogue of (3.1), which allows for the

uncertainty due to estimation. There are two common ways to make this precise:

$$P_{F^n} \left(F(\mathcal{R}_n) \geq 1 - \delta \right) \geq 1 - \beta, \quad \text{for all } F, \quad (3.2)$$

or

$$E_{F^n} \left(F(\mathcal{R}_n) \right) \geq 1 - \delta, \quad \text{for all } F. \quad (3.3)$$

A prediction set $\mathcal{R}_n$ satisfying one of these requirements is commonly known as a *tolerance region* (TR). The values of β and δ are typically chosen small, say, 5%.

In this chapter our main focus is on the construction of tolerance regions within a framework of nonparametric Bayesian density estimation, but we also review methods that build tolerance sets $\mathcal{R}_n$ directly based on X . Here by “directly” we mean that the set is the convex hull of some subset of X , typically based on some ranking of the observations. Such rank-based approaches were developed for the nonparametric situation (e.g. [61], [69], [30]) to alleviate restrictive assumptions behind tolerance intervals (in dimension $d = 1$) or ellipsoids (for $d \geq 2$) centered at the sample mean. The ranking of the data is straightforward for one-dimensional data, but needs care when $d \geq 2$. Liu et. al. [31] provided a rich methodology called *data depth* to this purpose. An alternative approach based on “conformity scores” was discussed by Lei et al. [29], inspired by [60]. However, we found the setup of conformity scores to be very similar to the likelihood depth in [31].

The chapter is organized as follows. In Section 3.2 we start with a brief review of the classical nonparametric prediction sets based on ranks in the univariate case. Then in Section 3.3 we proceed with the key concepts of data depth (DD) and DD-based prediction sets (DDPS), which apply also in more than one dimension. A brief discussion about conformity scores and the associated conformal prediction sets is provided in Section 3.4. Our focus is to build tolerance sets in a Bayesian nonparametric framework. We derive prediction sets that meet Bayesian analogues of (3.2) or (3.3) when F is assumed a priori to follow a Dirichlet Process Mixture (DPM) of Gaussians. The setup of a DPM model and algorithms for posterior inference are discussed in Section 3.6. The construction of a prediction set in the DPM model framework is elaborated in Section 3.5. We evaluate the performance of the different approaches via simulation, and present the results in Section 3.7.

3.2 Univariate Nonparametric Prediction Intervals

The classical nonparametric prediction set in the univariate case dates back to [66]. It is based on the (conventional) order statistics of the sample, i.e. the ranking of the observations from small to large values. Because in dimensions $d \geq 2$ expressions such as $X_i > X_j$ lose their meaning, the method is limited to univariate observations.

We consider only the two-sided case here, as later on in the multivariate case we shall also be mostly concerned with bounded prediction sets. Let $X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$ be the ordered values of

the sample of observations X , and for some positive integer $r < (n+1)/2$, consider

$$\mathcal{R}_n(X) = (X_{(r)}, X_{(n-r+1)}].$$

If the distribution function F of the X_i is continuous, then the coverage $F(\mathcal{R}_n(X)) = F(X_{(n-r+1)}) - F(X_{(r)})$ of this prediction set can be shown to follow a $Beta(n-2r+1, 2r)$ -distribution (cf. Lemma 3.3.1 below). Therefore, to make $\mathcal{R}_n(X)$ meet (3.2), we seek a value of r that satisfies, for $Z_r \sim Beta(n-2r+1, 2r)$,

$$P(Z_r \geq 1 - \delta) \geq 1 - \beta, \quad (3.4)$$

and to meet (3.3), we seek a value of r that satisfies

$$E(Z_r) \geq 1 - \delta. \quad (3.5)$$

The maximal solution to (3.5) is simply $r = \lfloor \delta(n+1)/2 \rfloor$, while (3.4) can be easily solved by a numerical algorithm.

This construction of a nonparametric prediction interval ensures that equal numbers of data points are left outside the interval past each of its end points. There are various ways to generalize this approach. We refer to [30] for a good summary on this topic.

3.3 Data Depth and DD-based Prediction Sets

3.3.1 Data Depth

A "data depth" is a function from the sample space (here taken equal to $\mathbb{R}^d$) to $\mathbb{R}^+$ that attaches to every point a positive number that measures its centrality within a data set or distribution. It can take many forms, expressing different notions of 'centrality'. Larger depth indicates that the point sits more centrally within the distribution or dataset. Data depth is by definition positive, but its theoretical range varies with its chosen form.

For a given notion of data depth, there are both a population and a sample version. For a probability distribution F , we denote the population data depth of a point x by $D(x; F)$ or $D_F(x)$. Given a sample $X = \{X_i\}_{i=1}^n$, a corresponding sample version of data depth is obtained by plugging in the empirical distribution F_n of the sample, yielding $D(x; F_n)$ or $D_{F_n}(x)$. This gives 'the depth of x in the sample', and can serve as an estimate of the population depth of x . Sometimes sample values are also used in a different way than simple plug-in to form a data-based version, for instance if D_F depends on F through a density f .

For a thorough introduction to data depth and many examples, see [30]. Here we list four commonly used data depths. We denote these by adding a pre-fix letter to $D_F(x)$ (e.g. giving $MD_F(x)$)

or $SD_F(x)$) to differentiate between the depth measures, while we shall continue to use $D_F(x)$ or $D_{F_n}(x)$ to refer to data depth in general. In all cases the sample space is $\mathbb{R}^d$.

1. (Mahalanobis depth) $MD(x; F) = [1 + (x - \mu_F)^T \Sigma_F^{-1} (x - \mu_F)]^{-1}$, where μ_F and Σ_F are the mean vector and covariance matrix of F . Replacing μ_F and Σ_F by the sample mean and covariance, we obtain $MD(x; F_n)$.
2. (Half-space depth) $HD(x; F) = \inf_H \{F(H) : x \in H\}$, where H ranges over all closed half-spaces in $\mathbb{R}^d$, also known as Tukey depth. When $d = 1$, closed half spaces are half intervals $(-\infty, h]$ or $[h, \infty)$, and the half space depth can be calculated explicitly as $\min(F(x), 1 - F(x-))$; the point of maximum depth is the median.
3. (Simplicial depth) $SD(x; F) = P_F(x \in \Delta_{(Z_1, \dots, Z_{d+1})})$, where $\Delta_{(Z_1, \dots, Z_{d+1})}$ denotes the closed simplex with vertices $Z_1, \dots, Z_{d+1}$, taken to be an i.i.d. sample from F . Even for standard distributions F , there seems to exist no analytical solution to $SD(x; F)$, but one may approximate the right side of the equation by Monte Carlo simulation. For instance, we may generate many i.i.d. samples of size $d + 1$, form the simplex each time, and compute the proportion of simulated simplices that contain the value x . More efficiently, we may generate a single large i.i.d. sample $\{X_i\}_{i=1}^n$ from F and estimate the simplicial depth by the U-statistic

$$SD(x; F_n) = \binom{n}{d+1}^{-1} \sum_{(*)} \mathbf{1}_{x \in \Delta(X_{i_1}, \dots, X_{i_{d+1}})}, \quad (3.6)$$

where $(*)$ denotes all possible subsets of $\{X_i\}_{i=1}^n$ of size $d + 1$. When $d = 1$, the closed simplex spanned by two points is the interval between the points, and for $x \notin X$ we have $SD(x; F_n) = \sum_{i=1}^n \mathbf{1}_{\{X_i < x\}} \cdot \sum_{i=1}^n \mathbf{1}_{\{X_i > x\}} / \binom{n}{2}$.

4. (Likelihood depth) $LD(x; F) = f(x)$, where f is a density of F . In the sample version, f is replaced by a density estimator $\hat{f}$ based on X .

For a given data depth measure and sample $\{X_i\}_{i=1}^n$, we can compute the depth $D_F(X_i)$ of every sample value and sort these values by size. We denote by $X_{[1]}, \dots, X_{[n]}$ the sample values ordered by data depth, from large to small. Thus $D_F(X_{[1]}) \geq D_F(X_{[2]}) \geq \dots \geq D_F(X_{[n]})$, and $X_{[1]}$ is the deepest or most central point, while $X_{[n]}$ is the most outlying one, relative to the distribution F . If instead of the population version, we use the sample version D_{F_n} , then this procedure yields an ordering by centrality ‘in the observed data cloud’.

The ordering of the depth values gives rise to a ranking of the observations, the rank of X_i being the rank of $D_F(X_i)$ (or $D_{F_n}(X_i)$) in the ordered sample of depth values. The empirical Mahalanobis, half space and simplicial data depths have the desirable property of being affine invariant: they remain the same if all sample values are transformed by an affine map $X_i \mapsto AX_i + b$. The same is then true

for the corresponding depth ranks. For the likelihood depth, this also holds if the density estimator used is affine equivariant.

The ranks thus obtained may be used for descriptive statistical summaries, outlier detection, and many other purposes. Here, we use them to define prediction sets.

3.3.2 DD-based Prediction Sets

For $Y_i = D_F(X_i)$, and $Y_{[1]} \geq Y_{[2]} \geq \dots \geq Y_{[n]}$ the ordered depth values of the observations, we consider prediction sets of the form:

$$\mathcal{R}_F = \{x: D_F(x) \geq Y_{[r]}\}. \quad (3.7)$$

The value of r determines the size and coverage of these sets. It can be chosen based on the following lemma, due to [30].

Lemma 3.3.1. *Let $X_1, \dots, X_n \stackrel{iid}{\sim} F$, where F is a probability distribution over $\mathbb{R}^d$, and let $D_F: \mathbb{R}^d \mapsto \mathbb{R}^+$ be an arbitrary measurable map. Let $Y_{[1]} \geq Y_{[2]} \geq \dots \geq Y_{[n]}$ be the reverse order statistics of the variables $Y_i = D_F(X_i)$. If the cumulative distribution function of Y_i is continuous, then the random variable $F(x: D_F(x) \geq Y_{[r]})$ follows a $\text{Beta}(r, n - r + 1)$ -distribution.*

Proof. Let $U(y) = P_F(D_F(X_i) \geq y)$ be the survival function of $Y_i = D_F(X_i)$. If U is continuous, then $U(Y_1), \dots, U(Y_n) \stackrel{iid}{\sim} \text{Unif}[0, 1]$, by the ‘probability integral transformation’. Since U is decreasing, the order statistics of these variables are $U(Y_{[n]}) \leq U(Y_{[n-1]}) \leq \dots \leq U(Y_{[1]})$, and the corresponding spacings are $U(Y_{[n]}), U(Y_{[n-1]}) - U(Y_{[n]}), \dots, U(Y_{[1]}) - U(Y_{[2]}), 1 - U(Y_{[1]})$. It is well known that the spacings of a uniform sample follow a $\text{Dirichlet}(n + 1; 1, \dots, 1)$ -distribution. The variable $U(Y_{[r]})$ is the sum of the r left-most spacings and hence possesses a $\text{Beta}(r, n - r + 1)$ -distribution, by the aggregation properties of the Dirichlet distribution. By definition $U(Y_{[r]}) = F(x: D_F(x) \geq Y_{[r]})$. ■

If F were known, then we could use a DD-based prediction set $\mathcal{R}_F$ defined in (3.7) with a suitable value r . Based on Lemma 3.3.1, to meet requirements (3.2) or (3.3) we would choose r to satisfy, for $Z_r^* \sim \text{Beta}(r, n - r + 1)$, either

$$P(Z_r^* \geq 1 - \delta) \geq 1 - \beta, \quad (3.8)$$

or

$$E(Z_r^*) \geq 1 - \delta. \quad (3.9)$$

The minimal solution to (3.9) is $r = \lceil (1 - \delta)(n + 1) \rceil$. Equation (3.8) may be solved numerically, or approximately, for large n , by the normal approximation to the beta distribution (see [30]):

$$r = \lceil n(1 - \delta) + \xi_{1-\beta} \sqrt{n\delta(1 - \delta)} \rceil, \quad (3.10)$$

where ξ_a is the a^{th} quantile of the standard normal distribution.

For a data-based prediction set, we replace the population data depth D_F in (3.7) (also as hidden within the definition of $Y_{[r]}$) by its data-based version D_{F_n} . For the given values of r the equations (3.2) or (3.3) will not be met exactly, but in [30] the authors give conditions that the equations will be asymptotically correct as $n \rightarrow \infty$.

The equation $D_{F_n}(x) \geq Y_{[r]}$ may not be analytically solvable in x , and it may be infeasible to outline the prediction set $\mathcal{R}_{F_n}$ of (3.7) exactly, especially when d is large. A common shortcut is to determine the set $\{X_i: D_{F_n}(X_i) \geq Y_{[r]}, i = 1, \dots, n\}$ of sample points contained in $\mathcal{R}_{F_n}$, and take the tolerance set equal to the convex hull of these points, or at least visualize the tolerance set as this convex hull.

3.4 Conformity Scores and Conformal Prediction Sets

Given a sample $\{X_i\}_{i=1}^n$, a *conformity score* of a point x is a number that measures the similarity of the point x to the sample. For convenience of notation, we consider the value x as our $(n+1)^{\text{th}}$ observation, i.e. $X_{n+1} = x$, and denote the conformity score for the sample value X_i by $\kappa_i = \kappa(X_i; \{X_j\}_{j=1}^{n+1})$. The authors of [29] then require that $\kappa(\cdot; \{X_j\}_{j=1}^{n+1})$ be invariant to the ordering in $\{X_j\}_{j=1}^{n+1}$. They choose $\kappa_k = \hat{f}(X_i)$ for illustration, where $\hat{f}$ is a kernel density estimator based on $\{X_j\}_{j=1}^{n+1}$. In this case it is similar to the sample version of likelihood depth. In fact, other depth forms are also legitimate to set up a conformity score.

The key difference between a data depth $D_{F_n}(x)$ and conformity score $\kappa(x; \{X, x\})$ is that the former is based on the observed n observations X_i , whereas the latter is based on the observed sample X augmented with the theoretical value x . If the given theoretical value changes, so will the conformity scores of the observed n samples. This seems computationally intensive. To overcome this the authors of [29] suggested to use an ‘augmented density estimator’, which updates an estimator based on X by a newly “observed” x . One may come up with similar estimators for other chosen depth forms. However, we would expect that the impact of x in the calculation of $\hat{f}$ is small when the observed sample size n is large, unless x will be an outlier relative to the data X , whence the likelihood depth will typically be a good approximation.

Another contribution from [29] is their construction of a prediction set. In their examples ($d \in \{1, 2, 3\}$), prediction sets are outlined by theoretical values that fulfil (3.1). We may expect that the computational burden will quickly evolve to an unbearable level as d increases, even with their proposed efficient algorithm. Although conformal prediction sets may provide valuable alternatives to build nonparametric prediction sets, due to their similarity to likelihood depth and computational considerations, we will not include them in our performance evaluation. However, we will include a 2D numerical example from [29], and compare our results (Figure 3.3) to Figure 2 in that paper.

3.5 Bayesian Nonparametric Prediction Sets

All inference within the Bayesian paradigm is based on the posterior distribution and hence it is natural to replace the requirements (3.2) and (3.3) for a tolerance set by Bayesian versions. In a nonparametric Bayesian approach we put a prior on the density f of the observations, and then obtain a corresponding posterior distribution $\Pi(f \in \cdot | X)$, given the data $X = (X_1, \dots, X_n)$. We discuss an appropriate nonparametric prior for f in the next section. Bayesian tolerance sets in the spirit of definitions (3.2) and (3.3) are then sets $\mathcal{R}_n$ such that

$$\Pi\left(f: \int_{\mathcal{R}_n} f(z) dz \geq 1 - \delta | X\right) \geq 1 - \beta, \quad (3.11)$$

$$\int \int_{\mathcal{R}_n} f(z) dz d\Pi(f | X) \geq 1 - \delta. \quad (3.12)$$

The key difference, compared to (3.2) and (3.3), is that the sampling distribution F^n of X is replaced by the posterior distribution of f given X . This makes interpretation easier; in particular the requirement that (3.2) or (3.3) are met for all possible F is replaced by a single requirement given the observed data.

By Fubini's theorem, (3.12) can be rewritten in the form

$$\int_{\mathcal{R}_n} \int f(z) d\Pi(f | X) dz \geq 1 - \delta.$$

Here the inner integral $\int f(z) d\Pi(f | X)$ as a function of z is the *posterior mean density*. Hence a prediction set $\mathcal{R}_n$ in the sense of (3.12) is a set with probability at least $1 - \delta$ under the posterior mean density. Equation (3.11) is harder to interpret. It requires that a fraction $1 - \beta$ of posterior samples of f give probability at least $1 - \delta$ to the set $\mathcal{R}_n$.

The requirements (3.11) and (3.12) do not determine the shape of the tolerance sets $\mathcal{R}_n$. Often we shall fix a shape and determine the tolerance set by setting a tuning parameter to meet (3.11) or (3.12). For instance, we can choose $\mathcal{R}_n$ to be an exceedance set, with the level as a tuning parameter. A natural choice is to set $\mathcal{R}_n$ equal to a density level set of the posterior mean density, i.e.

$$\mathcal{R}_n(X) = \left\{ z: \int f(z) d\Pi(f | X) \geq t \right\}.$$

Here t is the tuning parameter, which would be set to satisfy (3.11) or (3.12). Monotonicity in the tuning parameter should ensure a solution, although in practice finding the solution could be computationally demanding.

The actual computation of the posterior distribution and Bayesian tolerance sets will typically be based on simulation. Given a sample $f^{(1)}, \dots, f^{(B)}$ from the posterior distribution of f given X ,

possibly obtained via an MCMC-procedure, we may approximate (3.11) and (3.12) by

$$\frac{1}{B} \sum_{i=1}^B 1_{\int_{\mathcal{R}(X)} f^{(i)}(z) dz \geq 1-\delta} \geq 1-\beta, \quad (3.13)$$

$$\int_{\mathcal{R}_n} \frac{1}{B} \sum_{i=1}^B f^{(i)}(z) dz \geq 1-\delta. \quad (3.14)$$

The posterior mean density, which is also the predictive density for a new observation, may also be available in another form than an MCMC average, as we shall see below.

3.6 Bayesian Nonparametric Prediction Sets via Dirichlet Mixtures

Dirichlet Process Mixtures (DPM) of Gaussian distributions are a popular nonparametric prior for a density. As mixtures offer much flexibility in the shapes of the resulting prediction sets, we found these priors useful in the present context.

The DPM of Gaussians prior and data model can be defined by the hierarchy:

$$\begin{aligned} X_i | \theta_i, G &\stackrel{\text{ind}}{\sim} F(\cdot; \theta_i), & i = 1, \dots, n, \\ \theta_i | G &\stackrel{\text{iid}}{\sim} G, \\ G &\sim DP(\alpha G_0). \end{aligned} \quad (3.15)$$

Here $\theta_i = (\mu_i, \Sigma_i)$ is a pair of a mean vector and covariance matrix and $F(\cdot; \theta_i)$ is the multivariate normal distribution $N_d(\mu_i, \Sigma_i)$. The ‘latent’ variables θ_i are unobserved and part of the prior specification. The parameters of the Dirichlet distribution $DP(\alpha G_0)$ are a positive scalar α , known as the prior strength, and a probability distribution G_0 , known as the center measure. In the case of Gaussian mixtures, the center measure is typically chosen to be a Normal-Inverse-Gamma (NIG) distribution when $d = 1$ or a Normal-Inverse-Wishart (NIW) distribution when $d \geq 2$, which are conjugate to the normal distribution and therefore computationally convenient.

A draw G from a Dirichlet distribution is almost surely a discrete distribution with countably many support points. In fact, it can be represented as $G = \sum_{i=1}^{\infty} p_i \delta_{\phi_i}$, where the (random) weights (p_i) follow a so-called stick-breaking distribution (determined by α) and are independent of the (random) support points $\phi_i \stackrel{\text{iid}}{\sim} G_0$. Therefore the prior conditional distribution of X_i given G is a countably infinite mixture of Gaussians, that is, the density of X_i can be written as

$$f(x) = \int \psi(x | \mu, \Sigma) dG(\mu, \Sigma) = \sum_{i=1}^{\infty} p_i \psi(x | \tilde{\mu}_i, \tilde{\Sigma}_i),$$

where $\phi_i = (\tilde{\mu}_i, \tilde{\Sigma}_i)$, and ψ denotes the multivariate Gaussian density

$$\psi(x|\mu, \Sigma) = (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right).$$

The Gaussian distribution is chosen for convenience. There exists a rich literature that demonstrates the good properties of the resulting posterior distribution of f . See Chapters 7.2 and 9.4 in [18] for an overview.

3.6.1 Posterior Sampling and SUGS

The Dirichlet process can be presented in many ways, two common representations being through stick-breaking and the Pólya urn scheme. Accordingly there are multiple possibilities for posterior sampling in the DPM model. An account of the two representations and associated procedures for full posterior inference in a DPM model can be found in Sections 3.8–3.10.

Full posterior inference becomes computationally expensive and therefore impractical when the sample size or dimension increase. Because computational efficiency is a critical limitation to apply such a nice model in practice, several shortcuts have been suggested in the literature. Here we apply an algorithm proposed by [63], named Sequential Updating and Greedy Search (SUGS), which scales well with sample size and dimension and provides online clustering as a by-product.

In reference to (3.15), it suffices to find an approximation to the posterior distribution of the latent variables $(\theta_1, \dots, \theta_n)$ given $X = (X_1, \dots, X_n)$, as given $(\theta_1, \dots, \theta_n)$ the distribution G is independent of X and is known to follow a $DP(\alpha G_0 + \sum_{i=1}^n \delta_{\theta_i})$ -distribution. Because the Dirichlet process produces a discrete distribution G in (3.15), the realizations of $\theta_1, \dots, \theta_n$ will partition in sets of equal values, as the result of ‘having chosen’ the same support point in G . The partition is completely determined by the weights in the Dirichlet distribution, while the common value of the variables within a cluster is drawn from the center measure G_0 , independently from the partition and the common values in the other clusters. The observations $X_1, \dots, X_n$ can be partitioned in the same way, with the observations for a given partitioning set drawn independently from a Gaussian distribution with the same value of θ_i as parameter. Unfortunately, the partitioning cannot be recovered with certainty from the observations, which will be different with probability one due to the addition of Gaussian noise to the values θ_i . However, if the partition were given, then finding the posterior distribution of the common value θ_i would be a simple parametric Bayes problem: given θ_i and some observations $Y_j | \theta_i \stackrel{\text{iid}}{\sim} F(\cdot; \theta_i)$ find the posterior distribution of θ_i given (Y_j) . This can be solved by Bayes’s rule. An algorithm might thus focus on recovering the partition.

A complicating fact is that given the observations $X_1, \dots, X_n$, the partition is random, i.e. the posterior gives a probability distribution over partitions. The SUGS algorithm greatly simplifies the computations by replacing this distribution by a degenerate one, roughly a most probable one. The algorithm further simplifies by assigning observations $X_1, X_2, \dots$ to clusters sequentially, never

re-visiting and changing previous assignments. We discuss the drawbacks of these simplifications in Section 3.6.2.

The partitioning is coded by latent class labels $c_1, \dots, c_n$, one for each observation, taking values in $\mathbb{N}$. If $c_i = c_j$, then observations X_i and X_j belong to the same partitioning set, i.e. they correspond to identical latent values $\theta_i = \theta_j$. The latter latent values are coded $\theta_i = \phi_{c_i}$, where $\{\phi_l: l = 1, 2, \dots\}$ is the set of (distinct) possible values. Thus observation X_i corresponds to partitioning set l if $c_i = l$ and in that case $\theta_i = \phi_l$. The SUGS algorithm progresses by sequentially determining c_i for each i from 1 to n , based on the value of X_i , the previous latent class assignments $(c_1, \dots, c_{i-1})$, and the previous observations $X_1, \dots, X_{i-1}$. Once assigned the class labels are fixed and not updated based on new observations. For convenience the class labels c_i are assigned consecutive values $1, 2, \dots$, meaning that the support points of G are numbered by (estimated) order of appearance. The number of clusters after stage i (number of distinct values among $\theta_1, \dots, \theta_i$) is denoted by m_i .

For brevity, denote $(X_1, \dots, X_i)$ by $X^{(i)}$ and $(c_1, \dots, c_{i-1})$ by $c^{(i-1)}$. To start we set $c_1 = 1$ and determine the posterior of $\theta_1 = \phi_1$ based on X_1 from the model $X_1 | \phi_1 \sim F(\cdot; \phi_1)$ and $\phi_1 \sim G_0$. At step $i \geq 2$, the task is to determine c_i based on $X^{(i)}$ and $c^{(i-1)}$, and to update the posterior of the parameters by incorporating the ‘new’ observation X_i .

The update of the posterior distribution of the parameter is carried out after determining c_i , so that it is known that the new observation joins an existing cluster l (hence $c_i = l \leq m_{i-1}$), or opens a new cluster ($c_i = m_{i-1} + 1$). In the first case the posterior of the parameter ϕ_l is updated to the model $(X_j: j \leq i, c_j = l) \stackrel{\text{iid}}{\sim} F(\cdot; \phi_l)$ and $\phi_l \sim G_0$, while in the second case a new value ϕ_{c_i} is determined as in the case $i = 1$, but now based on $X_i | \phi_{c_i} \sim F(\cdot; \phi_{c_i})$. In the case of the Gaussian model with G_0 the conjugate NIG or NIW prior, the updates are explicit and given by (3.53)–(3.57), or as (3.42)–(3.45). In fact the updates can be computed sequentially as parameter updates from the parameter values at stage $i - 1$ by the formulas below. At stage i an update is needed only for the parameter ϕ_l of the cluster c_i , the posterior distributions of the other parameters remain the same.

The crux of the SUGS algorithm is to determine the new label by maximizing its conditional probability given $(X^{(i)}, c^{(i-1)})$, as

$$c_i = \underset{l \in \{1, \dots, m_{i-1} + 1\}}{\operatorname{argmax}} P(c_i = l | X^{(i)}, c^{(i-1)}).$$

Here $\{1, \dots, m_{i-1}\}$ is the set of class labels of $X_1, \dots, X_{i-1}$, and choosing c_i equal to $m_{i-1} + 1$ would introduce a new label for X_i . By Bayes’ rule the probability on the right is proportional to

$$P(c_i = l | X^{(i)}, c^{(i-1)}) \propto P(c_i = l | X^{(i-1)}, c^{(i-1)}) f(X_i | c_i = l, X^{(i-1)}, c^{(i-1)}) =: \pi_{il} L_{il}(X_i).$$

The first term on the right is the conditional probability $\pi_{il} = P(c_i = l | X^{(i-1)}, c^{(i-1)})$ that the new latent variable θ_i is allocated to cluster l or opens a new cluster. Given the clustering $c^{(i-1)}$ it is independent

of the observations $X^{(i-1)}$ and given by the Pólya urn scheme (see Sections 3.9 and 3.10.2) as

$$\pi_{il} = \begin{cases} \sum_{h=1}^{i-1} \mathbf{1}_{\{c_h=l\}} / (i-1+\alpha), & l = 1, \dots, m_{i-1}, \\ \alpha / (i-1+\alpha), & l = m_{i-1} + 1. \end{cases} \quad (3.16)$$

The second term $L_{il}(X_i)$ is the conditional likelihood of observation X_i given that it is assigned to cluster l and given the preceding observations and their clustering. Given the latent variable θ_i it is independent of $(X^{(i-1)}, c^{(i-1)})$. Therefore, for $l \in \{1, \dots, m_{i-1}\}$, we find

$$L_{il}(X_i) = \int \psi(X_i | \phi_l) \pi(\phi_l | X^{(i-1)}, c^{(i-1)}) d\phi_l, \quad (3.17)$$

where $\phi_l \mapsto \pi(\phi_l | X^{(i-1)}, c^{(i-1)})$ is the posterior density of ϕ_l given $(X^{(i-1)}, c^{(i-1)})$, which was determined in stage $i-1$. When X_i is allocated to a new class (i.e. $c_i = m_{i-1} + 1$), its latent variable is generated from the center measure G_0 , and its conditional density is simply $L_{il}(X_i) = \int \psi(X_i | \phi) dG_0(\phi)$. When G_0 is NIG or NIW, these predictive likelihoods have a closed form, as a Gaussian variable or vector with random mean and (co)variance drawn from a NIG or NIW distribution follows a (multivariate) non-central t -distribution, as given in (3.60), or (3.48) in the multivariate case.

The algorithm can be summarized as follows.

Algorithm 3: SUGS for DPM model

Input: Given the model setup (3.15), observed data $\{X_i\}_{i=1}^n$, and the prior setup (α, G_0)

1 Set $c_1 = 1$ and calculate $\pi(\theta_1 | X_1)$ by updating the prior $\theta_1 \sim G_0$ with the single observation $X_1 \sim F(\cdot; \theta_1)$

2 **for** $i = 2, \dots, n$ **do**

3 Determine the latent class label c_i by

$$c_i = \underset{l \in \{1, \dots, m_{i-1}+1\}}{\operatorname{argmax}} P(c_i = l | X^{(i)}, c^{(i-1)}), \quad m_{i-1} = \max(c^{(i-1)}) \quad (3.18)$$

4 Update

$$\pi(\phi_{c_i} | X^{(i)}, c^{(i)}) \propto \psi(X_i | \phi_{c_i}) \pi(\phi_{c_i} | X^{(i-1)}, c^{(i-1)}) \quad (3.19)$$

Result: $(c_1, \dots, c_n)$ and $\pi(\phi_l | X^{(n)}, c^{(n)})$ for $l = 1, \dots, m_n$

For the conjugate NIG and NIW priors, the posterior distributions of the variables ϕ_l remain in the family and it suffices to update the four parameters of these distributions. To gain extra efficiency, these updates can be computed sequentially. In the univariate case $\phi_l = (\mu_l, \sigma_l^2)$ possesses an NIG distribution with parameters (μ, κ, A, B) . If at stage $i-1$ these parameters are $(\mu_l^{*(i-1)}, \kappa_l^{*(i-1)}, A_l^{*(i-1)}, B_l^{*(i-1)})$, then after assigning X_i to class $c_i = l$ (for $i > 1$) at stage i of the

algorithm, the new parameters are given by

$$\mu_l^{*(i)} = \frac{\kappa_l^{*(i-1)} \mu_l^{*(i-1)}}{\kappa_l^{*(i-1)} + 1} + \frac{X_i}{\kappa_l^{*(i-1)} + 1}, \quad (3.20)$$

$$\kappa_l^{*(i)} = \kappa_l^{*(i-1)} + 1, \quad (3.21)$$

$$A_l^{*(i)} = A_l^{*(i-1)} + \frac{1}{2}, \quad (3.22)$$

$$B_l^{*(i)} = B_l^{*(i-1)} + \frac{1}{2} \left((\mu_l^{*(i-1)})^2 \kappa_l^{*(i-1)} + X_i^2 - (\mu_l^{*(i)})^2 \kappa_l^{*(i)} \right). \quad (3.23)$$

In the multivariate case $\phi_l = (\mu_l, \Sigma_l)$ possesses a NIW distribution with again four parameters. If the parameters at stage $i-1$ are $(\mu_l^{*(i-1)}, \kappa_l^{*(i-1)}, \nu_l^{*(i-1)}, \Lambda_l^{*(i-1)})$, then the first two parameters $\mu_l^{*(i)}$ and $\kappa_l^{*(i)}$ are updated by the same formulas (3.20) and (3.21) as before, while

$$\nu_l^{*(i)} = \nu_l^{*(i-1)} + 1, \quad (3.24)$$

$$\Lambda_l^{*(i)} = \Lambda_l^{*(i-1)} + \frac{\kappa_0}{\kappa_0 + 1} \left(X_i - \mu_l^{*(i-1)} \right) \left(X_i - \mu_l^{*(i-1)} \right)'. \quad (3.25)$$

The parameters for other clusters than c_i are unchanged at stage i .

The joint posterior distribution of $\phi = (\phi_1, \dots, \phi_{m_n})$ given the class vector $c = (c_1, \dots, c_n)$ by SUGS is

$$\pi(\phi|X, c) = \left(\prod_{l=1}^{m_n} \pi(\phi_l | \{X_i: c_i = l\}) \right) \int \left(\prod_{j=m_n+1}^{\infty} dG_0(\phi_j) \right), \quad (3.26)$$

where $\{\phi_l\}_{l=1}^{m_n}$ is the set of distinct values among $\{\theta_i\}_{i=1}^n$ and $\phi_{c_i} = \theta_i$. The unobserved variables ϕ_j ($j = m_n + 1, m_n + 2, \dots$) are independent from $X_1, \dots, X_n$ and hence *i.i.d.* from G_0 under both prior and posterior distributions. To numerically derive a (δ, β) -type of tolerance set as (3.11), one needs a posterior sample of f . However, since SUGS gives a deterministic partition of the data, it is not clear or even possible to draw a proper posterior sample from true $f|X$. One may use the following procedure to generate a pseudo posterior sample, given the class vector c :

1. generate $\phi^* = (\phi_1^*, \dots, \phi_{m_n}^*)$ from (3.26), and in turn we have $\theta_i^* = \phi_{c_i}^*$, for $i = 1, \dots, n$;
2. generate

$$f^*(x) = \int \psi(x|\phi) dG^*(\phi) = (\alpha + n)^{-1} \left(\int \psi(x|\phi) dG_0(\phi) + \sum_{i=1}^n \psi(x|\theta_i^*) \right),$$

3. repeat the previous steps a large number of times.

The motivation for Step 2 is that $G|\theta_1, \dots, \theta_n \sim DP(\alpha G_0 + \sum_{i=1}^n \delta_{\theta_i})$, and a linear function of a Dirichlet measure is the integral of the function with respect to the center measure.

The posterior mean density is also the predictive density of a new observation, i.e. the conditional density of X_{n+1} given $X_1, \dots, X_n$. In the SUGS framework this can be conveniently approximated by

one more iteration of the algorithm (3.18)-(3.19) (and (3.16)-(3.17)), for $i = n + 1$, giving

$$\hat{f}(x) = \sum_{l=1}^{m_n+1} \pi_{n+1,l} \int \psi(x|\phi_l) \pi(\phi_l|X^{(n)}, c^{(n)}) d\phi_l. \quad (3.27)$$

As was noticed after (3.17), the integral on the right hand side can be computed analytically and is a t -density, which makes (3.27) a finite mixture of t -densities. When the sample size n is sufficiently large, the probability $\pi_{n+1,m_n} = \alpha/(\alpha + n)$ that X_{n+1} opens a new cluster may be small and the term $l = m_n + 1$ in (3.27) neglected. Furthermore, if one of the other components ($l = 1, \dots, m_n$) has sufficiently many members, then the corresponding t -distribution has many degrees of freedom and can be approximated by a Gaussian.

3.6.2 SUGS Drawbacks and Remedies

The main drawback of SUGS algorithm is that it does not give a posterior distribution of the partitioning. There appears to be no easy fix of this, short of performing a computationally more demanding MCMC procedure. One may expect the posterior mean produced by the SUGS algorithm to be a good approximation to the true posterior mean, but the variation in posterior samples will likely underestimate the variation in the true posterior.

Two other drawbacks in the algorithm as presented are the dependency on the ordering of the input data and the fixed prior strength α .

To address the first drawback, [63] proposed to randomly permute the data points $X_1, \dots, X_n$ many times, each time run the SUGS algorithm and calculate the ‘pseudo-marginal likelihood (PML)’

$$PML(X|c_1, \dots, c_n) = \prod_{i=1}^n \sum_{l=1}^{m_n+1} P(c_i|X^{(-i)}, c^{(-i)}) \int \psi(X_i|\phi_l) \pi(\phi_l|X^{(-i)}, c^{(-i)}) \quad (3.28)$$

where $X^{(-i)}$ and $c^{(-i)}$ are the set of all data and assigned class labels except sample i . Next one chooses the output with the largest PML value.

In practice, we found that a pruning procedure can also alleviate the dependence of SUGS on the ordering of the input sequence. After running the SUGS algorithm to completion one retains all clusters of size bigger than some threshold, but reassigns all observations belonging to a smaller cluster by repeating steps 3-4 of Algorithm 3, treating the observations from the bigger clusters as assigned previously. Alternatively, one may directly reassign observations from very small clusters (relative to the sample size, say, single member clusters when n is big) to the remaining clusters.

The second concern is that we treated the prior strength α in Algorithm 3 as a fixed parameter, even though this parameter is known to be influential. Next to fixing the informativeness of the prior relative to the data (as expressed in the name ‘prior strength’), it has a strong influence on the number of latent classes. The authors [63] extended the algorithm by including a discrete prior for α supported

on a pre-specified grid. An alternative, suggested in [54], is to tune α adaptively as the number of data points grows, so as to stabilize the number of clusters.

3.6.3 Derive a Prediction Set

For the purpose of illustration we derive prediction sets in the sense of (3.13) or (3.14). We restrict to the case that the (mean) posterior density is a finite mixture of Gaussian distributions. This essentially covers the posterior mean density and individual posterior densities from a Gibbs sampler, as well as the output of the SUGS algorithm. As noted, the posterior mean density (3.27) output by the SUGS algorithm is a mixture of t -densities, but this is close to a mixture of Gaussian if the degrees of freedom are large. Alternatively, the arguments below could be adjusted to t -densities.

Consider a prediction set $\mathcal{R}_n$ that is parameterized by a tuning parameter t in a monotone manner. We then seek an appropriate value of t to cover a prescribed probability mass.

BNPS-1

For a density level set $\mathcal{R}_n$ (as defined in Section 3.1) the tuning parameter t is the minimum level that ensures some target area or volume under the density curve above the exceedance set. For a given level the latter volume can be estimated via Monte Carlo simulation. The cutoff value can then be found via a numerical search algorithm. See Algorithm 4 for an example.

Sampling from a mixture $f = \sum_{l=1}^K \eta_l \psi(\cdot | \mu_l, \Sigma_l)$ of Gaussians can be done in two steps: (1) draw a latent class from a multinomial distribution with K categories and probabilities $(\eta_1, \dots, \eta_K)$, and then (2) draw a value from the corresponding Gaussian component. Given the level it can be ascertained whether a point x belongs to the density level set by computing the value of the mixture density at the point, which is straightforward.

Unfortunately, when K or d are large, the computational burden may be significant. This is a relevant concern in the case of full posterior inference in the DPM model. In the blocked Gibbs sampler under the stick breaking scheme, as elaborated in Section 3.10.1, one might perform B scans/iterations, with each scan resulting in a mixture of K different Gaussians. The posterior mean density $\int f d\Pi(f|X)$ will then be approximated by a mixture of $B \times K$ Gaussian components, where $B \times K$ can run in the millions. The resulting mixture will be a good estimator of the true density function, but Monte Carlo simulation and density level calculation will be impractical, or even infeasible. To a large extent the problem is caused by the presence of many components with small weights. A possible remedy is to rank the mixing coefficients from large to small, and retain only the largest ones with cumulative weight at least, say, 0.9, renormalized to 1.

Algorithm 4: Determine a univariate tolerance interval based on likelihood depth

Input: *Provided the posterior mean density $\bar{p}$ is known via estimation, also $B = 1000$ posterior sample densities $\{p_j^*\}_{j=1}^B$ are collected. $\bar{p}, p_1^*, \dots, p_B^*$ are all mixture of Gaussians.*

- 1 Generate a large number say $N_{mc} = 5000$, of sample values $\{x_h^*\}_{h=1}^{N_{mc}}$ from $\bar{p}$;
- 2 Obtain the likelihood depths $\{y_h^* = \bar{p}(x_h^*)\}_{h=1}^{N_{mc}}$;
- 3 **if** target is to derive a $(1 - \delta)$ -expectation tolerance interval as (3.12) **then**
- 4 Determine the cutoff depth value t such that $\sum_{h=1}^{N_{mc}} \mathbf{1}_{\{y_h^* \leq t\}} / N_{mc} = 1 - \delta$;
- 5 Create the group index for all $\{x_h^*\}_{h=1}^{N_{mc}}$, let $K_h = 1$ if $y_h^* \leq t$, otherwise $K_h = 0$;
- 6 Sort $\{K_h\}_{h=1}^{N_{mc}}$ based on $\{x_h^*\}_{h=1}^{N_{mc}}$ from small to large, giving $K_{[1]}, \dots, K_{[N_{mc}]}$ and $x_{[1]}^*, \dots, x_{[N_{mc}]}^*$;
- 7 Transform $\{K_{[h]}\}_{h=1}^{N_{mc}}$ to be the run length encoding (RLE) (for example, $\{1, 1, 1, 0, 0, 1\}$ becomes $\begin{Bmatrix} \text{value} & 1 & 0 & 1 \\ \text{length} & 3 & 2 & 1 \end{Bmatrix}$);
- 8 Define a segment as a subset of $\{x_{[h]}^*\}_{h=1}^{N_{mc}}$ with value “1” in the previous RLE. Calculate the range per segment. The union of these ranges is the final tolerance interval.
- 9 **if** target is to derive (β, δ) -type tolerance interval as (3.11) **then**
- 10 Set a grid of increasing values $t_1, \dots, t_W$ as the candidate cutoff values, t_1 can be a value slightly lower than the cutoff value suggested by (3.10), and $t_W = \max(\{y_h^*\}_{h=1}^{N_{mc}})$;
- 11 For each candidate cutoff value t_w ($w = 1, \dots, W$), a prediction set R_w can be derived as in steps 5–8;
- 12 Evaluate the coverage of each prediction set $R_1, \dots, R_W$ under the posterior sample densities $\{p_j^*\}_{j=1}^B$, and denote these coverages as $\{C_{w,j}\}_{w=1, j=1}^{W,B}$;
- 13 Calculate the confidence for each candidate cutoff value by $q_w = \sum_{j=1}^B \mathbf{1}_{\{C_{w,j} \geq 1 - \delta\}} / B$;
- 14 Given $\{t_1, \dots, t_W\}$ and $\{q_1, \dots, q_W\}$, find the right cutoff value t that has a confidence level of $1 - \beta$ via a simple interpolation method.
- 15 Determine the final tolerance set as in steps 5–8 given t .

Result: A tolerance set, as an interval or a union of several disjoint intervals

BNPS-2

The second approach is to take the prediction set equal to the union of prediction sets per component. We choose the latter of a fixed shape, with volumes proportional to the mixing weights η_l and depending on a tuning parameter t . Components of heavier weights will contribute more to the final prediction set, and grow faster with increasing t . A density level set is a natural choice for the component prediction sets.

In the one-dimensional case density level sets of Gaussian components are intervals centred at the mean μ_l with length proportional to the standard deviation σ_l of the component. Denote the component prediction sets as $\mu_l \pm g_l(t)\sigma_l$, for $l = 1, \dots, K$, where the $g_l(t)$ are to be determined so that the union of the intervals has the desired coverage. If the components are distanced from each other, then the density function of the mixture can be approximated by $\max [\eta_l \psi(x | \mu_l, \sigma_l) : l = 1, \dots, K]$,

and the componentwise prediction sets correspond to an overall level t^* for the mixture if

$$\eta_l(2\pi\sigma_l^2)^{-\frac{1}{2}} \exp\left(-\frac{(\mu_l - \mu_l \pm g_l(t)\sigma_l)^2}{2\sigma_l^2}\right) = t^*, \quad (3.29)$$

$$\log\left(\frac{1}{\sqrt{2\pi}\sigma_l}\right) - \frac{(g_l(t))^2\sigma_l^2}{2\sigma_l^2} = \log\frac{t^*}{\eta_l},$$

$$(g_l(t))^2 = -2\log\frac{t^*}{\eta_l} + 2\log\left(\frac{1}{\sqrt{2\pi}\sigma_l}\right),$$

$$g_l(t) = \sqrt{2\log\left(\frac{\eta_l}{\sqrt{2\pi}\sigma_l}\right) - 2\log(t^*)}. \quad (3.30)$$

For simplicity, set $t = -2\log(t^*)$ in (3.30). Given $\{\phi_l = (\mu_l, \sigma_l)\}_{l=1}^K$, the prediction set is then

$$\mathcal{R}_n = \bigcup_{l=1}^K \left(\mu_l \pm \sigma_l \sqrt{2\log(\eta_l) - \log(2\pi\sigma_l^2) + t} \right). \quad (3.31)$$

Graphically, this comes down to finding the intersected segments of the rescaled density functions $\eta_l\psi_{\sigma_l}(\cdot; \mu_l)$ and a horizontal line at t^* . We now determine t so that a set of this form meets either (3.11) or (3.12).

When K is a large, we might drop the components of smallest weights, similarly as discussed in BNPS-1. When a numerical algorithm, for instance grid search, is deployed to find the right t in (3.31), it is important to ensure that $2\log(\eta_l) - \log(2\pi\sigma_l^2) + t$ remains nonnegative. After collecting intervals from all components, Algorithm 5 can be used to find the union of the intervals.

Algorithm 5: Determine the Union of Intervals

Input: Given a $h \times 2$ matrix Y that represents $h \geq 2$ intervals and with the left endpoint in the first column

- 1 Sort the rows of the matrix by their values in the first column from small to large.
- 2 Let IND be a logical vector of length h and all elements set to $TRUE$. We use $IND[i]$ to denote the i^{th} element of the vector.
- 3 **for** $i = 2, \dots, h$ **do**
- 4 **if** $Y[i, 1] < Y[i-1, 2]$ **then**
- 5 $Y[i-1, 2] = \max(Y[i-1, 2], [i, 2])$
- 6 $Y[i,] = Y[i-1,]$
- 7 $IND[i-1] = FALSE$
- 8 **else**
- 9 **next**
- 10 Keep only the rows of Y that have the corresponding elements in IND to be $TRUE$.

Result: the processed Matrix

In higher dimensions one can proceed in a similar way, but the componentwise prediction sets will be ellipsoids, with centres μ_l and shapes given by the covariance matrices Σ_l .

Other Possibilities

Rather than by density level we may determine the shape of the prediction set through a general data-depth measure. We may simulate a large sample from an estimated density, and set the prediction set equal to $\mathcal{R}_n = \{x: D_{F_n}(x) \geq t\}$, where F_n is the empirical measure of the simulated sample and the threshold t is set to obtain the desired coverage. The latter may be approximated based on the same sample of simulated values. When the number of original observations is small, a data-depth prediction set based on an estimated density will have a smoother hull, compared to a set based directly on the raw data. If the density estimator is consistent, then the resulting prediction sets will be consistent for the true data-depth prediction sets.

The density level set approximated via simulation in BNPS-1 can be seen as the special case, based on the likelihood depth.

Final Remark

To construct (β, δ) prediction sets as in (3.13) both a posterior mean density and densities sampled from the posterior distribution are needed (for instance MCMC scans), as explained in Algorithm 4 for BNPS-1. One might determine the shape of the prediction set through the posterior mean density, and next vary a tuning parameter so that the resulting prediction set has coverage at least $1 - \delta$ for at least $100(1 - \beta)\%$ of the sample densities.

The SUGS algorithm provides an adequate approximation to the posterior mean, and one can generate pseudo sample densities as discussed in Section 3.6.1. However, due to the greedy nature of the algorithm, the variability in this pseudo sample will typically underestimate the variability in the true posterior distribution. Hence the preceding procedure may result in overly tight (β, δ) prediction sets. This expectation was borne out in our experiments, when applying Algorithm 4. A remedy is to use a nominal level of $(1 - \beta) + \xi_{0.95} \cdot \sqrt{(1 - \beta)\beta/B}$ in Step 13 of the algorithm, instead of the nominal confidence level $1 - \beta$. Here $\xi_{0.95}$ is the 95% percentile of the standard normal distribution, and B is the number of posterior scans. We found that this procedure resulted in better performance and adopted it for the numerical evaluations.

3.7 Numerical Evaluation

We investigate the performance of the aforementioned methods in two small studies.

3.7.1 Univariate Case

Assume that the population distribution is the mixture $0.5 \cdot N(0, 1) + 0.5 \cdot N(8, 1)$ of two univariate normal distributions, located so that the density has a clear bimodal shape. We ran 1000 simulations at two sample sizes: $n = 60$ and $n = 200$. For each sample drawn from the mixture distribution, we used five methods to calculate four tolerance sets: two of type (3.3) at expectation levels $1 - \delta$ equal to 90% and 95%, and two of type (3.2) at content and confidence levels $(1 - \delta, 1 - \beta)$ equal to (90%, 95%) and (95%, 90%). We evaluated the true coverage of the fitted tolerance sets at each run under the known population distribution, and recorded the lengths of the tolerance sets.

The average coverage of a tolerance set of $(1 - \delta)$ -type over the 1000 simulations should be close to its nominal level $(1 - \delta)$, while for a fraction of $100(1 - \beta)\%$ of the simulations the tolerance sets of $(1 - \beta, 1 - \delta)$ -type should have coverage larger than or equal to $(1 - \delta)$.

The following five methods were involved in the evaluation:

1. (Conv.TI) Conventional nonparametric tolerance interval as described in Section 3.2.
2. (Sim-DD.TI) DD-based tolerance interval as described in Section 3.3.2 with simplicial depth (Section 3.3.1).
3. (BGTI.DD) Bayesian tolerance set as BNPS-1 based on 5000 simulated points from the fitted posterior mean density and with likelihood depth. Posterior inference is done via the block Gibbs sampler, described in Section 3.10.1, with 2000 burn-in and 1000 extra scans. The prior setup was: $K = 30$, $\mu_0 = \bar{x}$, $\kappa_0 = 0.1$, $A_0 = 0.1 \cdot \hat{s}^2$, $B_0 = 0.1$, where $\bar{x}$ and $\hat{s}^2$ are the sample mean and variance of the (simulated) data.
4. (SUGS.TI) Bayesian tolerance set as BNPS-2. Posterior inference is done via the SUGS algorithm as described in Section 3.6.1. Prior setup: $\mu_0 = \bar{x}$, $\kappa_0 = 0.1$, $A_0 = 0.1 \hat{s}^2$, $B_0 = 0.1$. For a given dataset, the SUGS algorithm was repeated 50 times with randomly permuted input sequence, and small clusters (≤ 8 in case of sample size 60 or ≤ 12 in case of sample size 200) were merged into existing clusters. The final model was selected based on maximizing PML (3.28).
5. (SUGS.TI.DD) Bayesian tolerance set as BNPS-1 based on 5000 simulated points from the fitted posterior mean density and with likelihood depth. Posterior inference was done via the SUGS algorithm, with the same execution details as in SUGS.TI.

The results of the simulation are summarized in Table 3.1. In general, the performance enhances with bigger sample size, and Bayesian nonparametric estimators seem more competitive than the conventional univariate nonparametric approach and the purely DD-based approach. The tolerance set constructed as BNPS-1 with posterior samples from a full-Bayesian, block Gibbs sampler has the best overall performance across both types of tolerance sets, and its performance seems to be adequate already at sample size 60. Bayesian tolerance sets obtained from a posterior sample generated through the SUGS algorithm perform well for the $(1 - \delta)$ -expectation type, but are less competitive for the

(β, δ) -type. The latter is expected as the greedy nature of SUGS tends to underestimate the posterior variance.

Table 3.1 Performance evaluation in univariate case. The table value is the average true coverage of the simulated tolerance sets in the case of the $(1 - \delta)$ -expectation type, and the fraction of simulated tolerance sets with true coverage at least $1 - \delta$ in the $(1 - \delta)$ -content, $(1 - \beta)$ -confidence type tolerance sets.

	Sample Size = 60					Sample Size = 200				
	Conv.TI	Sim-DD.TI	BGTI.DD	SUGS.TI	SUGS.TI.DD	Conv.TI	Sim-DD.TI	BGTI.DD	SUGS.TI	SUGS.TI.DD
$(1 - \delta)$ -expectation type										
$\delta = 0.1$	0.868	0.851	0.884	0.883	0.883	0.890	0.885	0.894	0.892	0.894
$\delta = 0.05$	0.934	0.918	0.940	0.937	0.936	0.940	0.935	0.946	0.944	0.943
$(1 - \delta)$ -content, $(1 - \beta)$ -confidence										
$\delta = 0.1, \beta = 0.05$	0.99	0.72	0.92	0.80	0.81	0.98	0.77	0.93	0.83	0.87
$\delta = 0.05, \beta = 0.05$	(0.80)	0.58	0.94	0.79	0.79	0.99	0.87	0.95	0.80	0.83

The under-performance for small sample size of the conventional nonparametric method (method (Conv.TI)) in fitting (β, δ) -type tolerance sets is notable. At sample size 60 the true confidence level of this method is 0.99, higher than the nominal level 0.95, when targeting coverage $1 - \delta = 0.90$, while it drops to 0.80 at the same nominal level, when targeting coverage 0.95. This appears to be due to discretization when determining the value of $r \in \mathbb{N}$ meeting (3.4) used to select the order statistics that give the interval $(X_{(r)}, X_{(n-r+1)})$. For $n = 60$ and $r = 1$ the probabilities that a $Beta(n - 2r + 1, 2r)$ -variable exceeds 0.95 or 0.9 are 0.987 and 0.816, respectively, while for $r = 2$ these probabilities drop to 0.86 and 0.35. Thus there may be no solution to r that can meet (3.4) when sample size is small but the desired coverage is high, for example, $n = 60$ and $\delta = 0.05, \beta = 0.05$ in the above numerical study. In such case, we take $r = 1$ so that the prediction interval is just the range of the data, which is similar to the default setting of R package `{tolerance}`. The evaluation result is marked with parentheses in Table 3.1.

This warns us not to pursue high coverage and high nominal confidence level when the sample size is small when using such a nonparametric approach.

Figure 3.1 shows the total lengths of the tolerance sets. In all cases the Bayesian tolerance sets have shorter length. This is explained by the fact that methods (Conv.TI) and (Sim-DD.TI) produce intervals by design, whereas the Bayesian method can yield unions of multiple intervals as outcome.

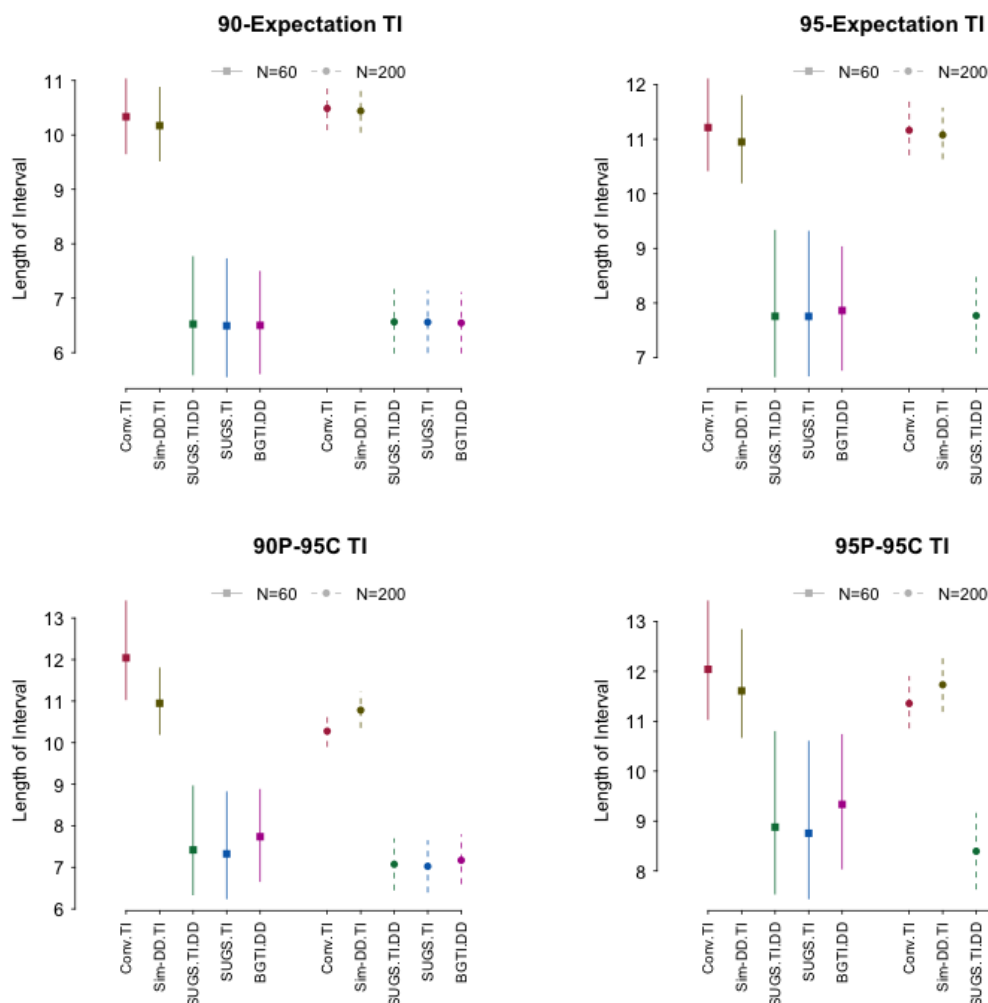


Fig. 3.1 Total lengths of tolerance sets in the simulated datasets by different methods. The square points are the median, and each bar marks the 0.05 and 0.95 quantile across the 1000 simulations.

3.7.2 Multivariate case

In this section we examine the performance of the $(1 - \delta)$ -expectation type nonparametric prediction sets in examples in dimensions 2 or 5. We omit a discussion of the (β, δ) -type, since it is rare to see it used in pharmaceutical practice for multivariate cases. For the DD-based approach, we consider the Mahalanobis depth (MD) and the Likelihood depth (LD), where the latter is based on a nonparametric Bayesian density estimator. We found other types of depth implemented in relevant R packages (e.g. depth, DepthProc, ddalpha) to be burdensome for use in a simulation exercise in dimension 5; in particular, Simplicial depth is prohibitive in more than 3 dimensions. Because the full Bayesian approach via the block Gibbs sampler also becomes computationally expensive, we only considered the approximation through the SUGS algorithm, which remained competitive. As observed in the

univariate case, we would expect that (β, δ) -type tolerance set via the SUGS algorithm, if needed, will attain smaller than the nominal confidence, due to the underestimation of the true spread of the posterior distribution.

The true population distribution for the simulation in dimension 2 was taken from [29], and is a mixture of two bivariate normal distributions:

$$0.5 \times N \left(\begin{pmatrix} \sqrt{2 \log n} - 2 \\ 0 \end{pmatrix}, \begin{pmatrix} 4 & 0 \\ 0 & 1/4 \end{pmatrix} \right) + 0.5 \times N \left(\begin{pmatrix} 0 \\ \sqrt{2 \log n} - 2 \end{pmatrix}, \begin{pmatrix} 1/4 & 0 \\ 0 & 4 \end{pmatrix} \right), \quad (3.32)$$

where n is the sample size. The special design of the mean vectors as a function of n results in an overlap between the two components that is challenging for inference at the given sample size. The population distribution in dimension 5 had the distribution (3.32) as its marginal in the first two dimensions, and independent centered univariate normal distributions with variance 1/4 for the marginals in dimensions 3 to 5. Per setup (dimensions 2 or 5), we considered two sample sizes, namely $n = 100$ and $n = 500$. The values of δ ranged over 0.1, 0.05, 0.01 and 0.003, corresponding to 90%, 95%, 99% and 99.7%-expectation tolerance sets.

The true coverage of a given prediction set was evaluated via Monte Carlo simulation from the corresponding population model. The procedure of deriving the expectation coverage is as follows: (1) generate a sample of size n from the population model and derive the prediction sets (MD-based, LD-based, Bayesian prediction set via SUGS), (2) generate 5000 points from the population model, calculate the proportion of simulated points that fall within each prediction set, (3) repeat steps (1) and (2) 100 times, resulting in 100 calculated proportions per method for determining a prediction set, (4) per method, take the average of the 100 calculated proportions; compare this to the nominal value $1 - \delta$. The results of the evaluation are summarized in Table 3.2. All approaches performed reasonably well, even when $n = 100$, and performance improved with increasing sample size. The performance of the MD-based method was almost uniformly the best, except for sample size $n = 100$ and over-ambitious coverage ($\delta = 0.003$). The Bayesian prediction sets obtained via SUGS tend to yield coverage below the nominal value.

Table 3.2 Performance evaluation in 5 dimensions: $(1 - \delta)$ -expectation type

	Sample Size = 100				Sample Size = 500			
	$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.01$	$\delta = 0.003$	$\delta = 0.1$	$\delta = 0.05$	$\delta = 0.01$	$\delta = 0.003$
2-dimension								
MD-based	0.893	0.942	0.988	0.988	0.900	0.949	0.989	0.998
LD-based	0.872	0.928	0.980	0.980	0.896	0.945	0.988	0.997
SUGS (BNPS-1)	0.830	0.901	0.969	0.986	0.878	0.934	0.983	0.994
5-dimension								
MD-based	0.866	0.925	0.980	0.980	0.893	0.946	0.988	0.998
LD-based	0.839	0.902	0.971	0.971	0.886	0.939	0.985	0.997
SUGS (BNPS-1)	0.845	0.913	0.976	0.990	0.881	0.937	0.985	0.995

These results motivated to refine the way in which the SUGS algorithm selects a permutation of the input data. It was proposed by [63] to randomly permute the input sequence and select the model

with the highest PML (3.28). In our study we found that this proposal tends to favour more complex models (i.e. a mean density estimator with a higher number of components), as illustrated by the top left plot in Figure 3.2. We considered three alternatives: (1) Using the Silhouette metric, which is popular in machine learning community for assessing clustering results, (2) Using the first 70% sample values for training SUGS and calculating the predicted likelihood on the remaining 30% of the sample. (3) Taking the minimum score between (1) and (2).

The Silhouette metric is calculated from the data points $x_1, \dots, x_n$ and their cluster assignments $c_1, \dots, c_n$ as follows. For $i \in \{1, \dots, n\}$ and $k_{c_i} = \#\{l: c_l = c_i\}$ the number of points in its cluster,

$$s_a(i) = \frac{1}{k_{c_i} - 1} \sum_{l \neq i: c_l = c_i} \|x_i - x_l\|, \quad (3.33)$$

$$s_b(i) = \min_{h \neq c_i} \frac{1}{k_h} \sum_{l: c_l = h} \|x_i - x_l\|, \quad (3.34)$$

$$s(i) = \frac{s_b(i) - s_a(i)}{\max\{s_a(i), s_b(i)\}}, \quad \text{if } k_{c_i} > 1. \quad (3.35)$$

The definition implies that every $s(i)$ is contained in the interval $[-1, 1]$. A value close to -1 means that point x_i is on the average (much) closer to another cluster than to its own cluster, whereas for a value close to 1 the opposite is true. The overall Silhouette metric is the summation, i.e. $\sum_{i=1}^n s(i)$. A large value indicates a good clustering relative to distance.

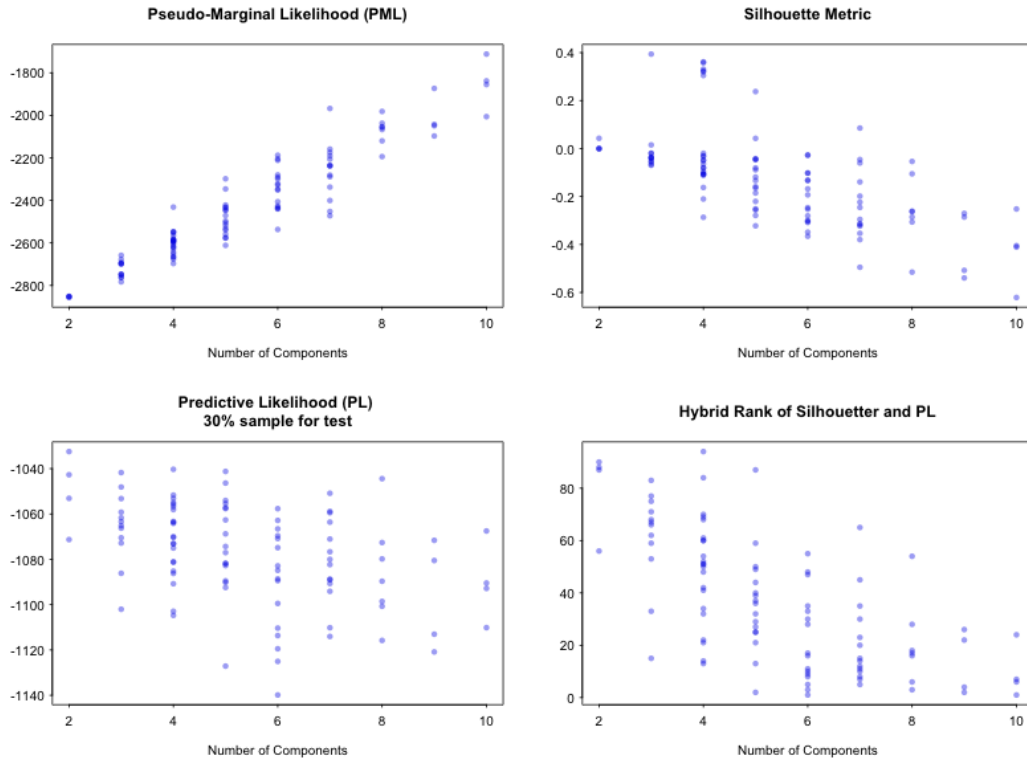


Fig. 3.2 Plot of the values of 4 possible selection criteria computed on SUGS outputs after 100 random permutations of a dataset of size 1000 simulated from the mixture distribution (3.32). The criteria are on the vertical axis and plotted versus the number of components estimated by SUGS on the horizontal axis.

We implemented these refinements in a simulation study with as population density the 2-dimensional distribution in (3.32), taking a sample of size 500 from each component. We calculated the PML criterium, the Silhouette metric and the alternatives (2) and (3) on the SUGS outputs of every of 100 random permutations of the input data. For each of the four criteria the 100 values are plotted vertically in Figure 3.2, against the number of components on the horizontal axis. The first panel in the figure shows that the PML criterion favours higher number of components, whereas the other three criteria prefer a less complicated models.

Figure 3.3 shows the 1000 simulated points, and the contours of the density level sets based on the mean density estimator of the SUGS output and on the Simplicial depth values. The contour of the true density level (the population model) is added in blue as a reference. The SUGS level set matches the true density well, whereas the Simplicial depth set seems to fail in capturing the two-component nature. The 90% prediction sets are shown in Figure 3.4. Although the estimated coverage of the SUGS and Simplicial depth sets are close, their volumes differ a lot.

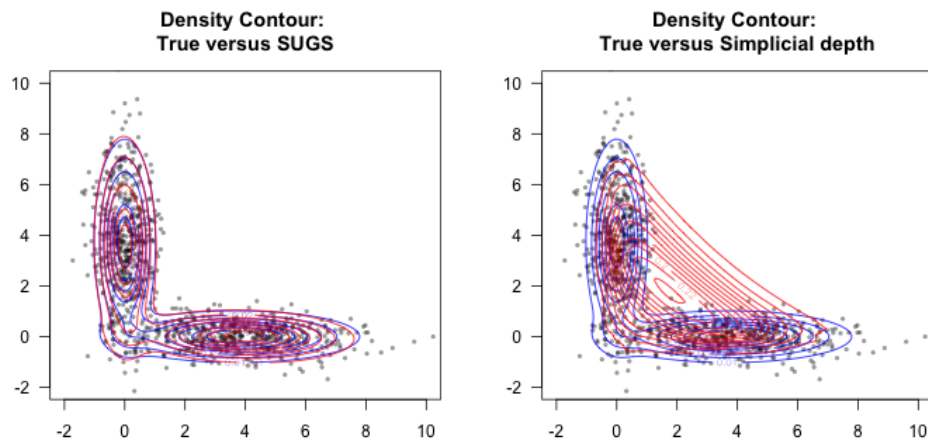


Fig. 3.3 Plot of the 1000 simulated sample values, the true density level contour (in blue), the density level contour based on the mean density estimator of SUGS (red in the left plot), and contour of the Simplicial depth values relative to the simulated data (red in the right plot).

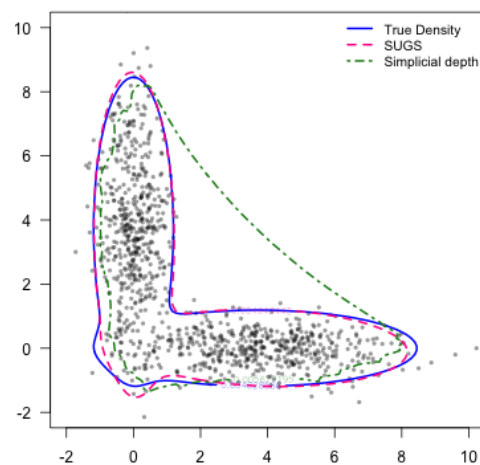


Fig. 3.4 Plot of the 1000 simulated sample values with 90% density level set of the true distribution, and 90% prediction sets based on SUGS density estimator and Simplicial depth value.

3.8 Annex: Stick-Breaking Process

In the notation of [23], a stick-breaking process can be defined as:

$$\begin{aligned}
 P &= \sum_{k=1}^K p_k \delta_{\phi_k}, \\
 \phi_k &\stackrel{\text{iid}}{\sim} G, \quad \text{for } k \geq 1, \\
 p_1 &= V_1; \quad p_k = (1 - V_1) \cdots (1 - V_{k-1}) V_k, \quad \text{for } k \geq 2, \\
 V_k &\stackrel{\text{iid}}{\sim} \text{Beta}(a_k, b_k).
 \end{aligned} \tag{3.36}$$

Here $K = \infty$, δ_ϕ is a degenerated measure that concentrates all its mass on ϕ , and $\{\phi_k\}_k$ and $\{p_k\}_k$ are mutually independent random variates. The mechanism of generating the random weights $\{p_k\}_{k=1}^K$ coins its name of “stick-breaking”. Special cases include:

- Pitman-Yor process: $a_k = 1 - a$ and $b_k = b + ka$, where $a \in [0, 1)$ and $b > -a$.
- Dirichlet process: $a_k = 1$ and $b_k = \alpha$, where $\alpha > 0$.
- For a stable law process: $a_k = 1 - \alpha$ and $b_k = k\alpha$, where $\alpha \in (0, 1)$.

Let $PY(a, b, G)$ denotes a Pitman-Yor process. Both a Dirichlet process and a stable law process are special cases of the Pitman-Yor process: a Dirichlet process $DP(\alpha G)$ is $PY(0, \alpha, G)$, and a stable law process is $PY(\alpha, 0, G)$.

There are many other ways to set up the parameters a_k and b_k . In theory, the stick-breaking process is well defined as long as $\sum_{k=1}^\infty p_k = 1$, which can be shown to be equivalent to $\sum_{k=1}^\infty \log(1 + a_k/b_k) = +\infty$. In practice, K must be finite; we can change the definition of the last stick length to $V_K = 1$ to ensure that this clause is met. The approximation to (3.36) with $K < \infty$ is crucial to the block Gibbs sampler proposed in [23], which takes advantage of the fact that a finite Dirichlet distribution is a conjugate prior for the multinomial distribution. Other than generating the random weights $\{p_k\}_{k=1}^K$ via a sequence of beta random variates as in (3.36), one may also define $p_k = J_k / \sum_{i=1}^K J_i$, where J_k is a Gamma random variable in the case of the Dirichlet process. See [49] and [43] for details, or [18] for a comprehensive summary.

3.9 Annex: Generalized Pólya Urn Scheme

In this section we restrict attention to the Pitman-Yor and Dirichlet processes, and set $a_k = 1 - a$ and $b_k = b + ka$, for given $a \in [0, 1)$ and $b > -a$. For $\theta_i | P \sim P$ and $P \sim PY(a, b, G)$, the marginal distribution of the (θ_i) can be shown to form a *Pólya sequence*:

$$\begin{aligned}
 \theta_1 &\sim G, \\
 \theta_i | \theta_1, \dots, \theta_{i-1} &\sim \frac{b + am_{i-1}}{b + i - 1} G + \sum_{l=1}^{m_{i-1}} \frac{n_{l,i-1} - a}{b + i - 1} \delta_{\phi_l}, \quad i \geq 2.
 \end{aligned} \tag{3.37}$$

Here $\phi_1, \phi_2, \dots$ are the distinct values in $\theta_1, \theta_2, \dots$ in order of appearance, m_{i-1} is the number of different values among $\theta_1, \dots, \theta_{i-1}$, c_i is the class label of θ_i , so that $c_i = l$ if and only if $\theta_i = \phi_l$, and $n_{l,i-1} = \sum_{j=1}^{i-1} \mathbf{1}_{\{c_j=l\}}$ is the multiplicity of ϕ_l in $\theta_1, \dots, \theta_{i-1}$. Of course $m_i \leq i$ and $\sum_{l=1}^{m_i} n_{l,i} = i$.

For a Dirichlet Process $DP(\alpha G) = PY(0, \alpha, G)$, this indicates the following prediction rule for the labels c_i :

$$\begin{aligned} P(c_i = l | c_1, \dots, c_{i-1}) &= \frac{n_{l,i-1}}{\alpha + i - 1}, \quad l = 1, \dots, m_{i-1} \\ P(c_i = m_{i-1} + 1 | c_1, \dots, c_{i-1}) &= \frac{\alpha}{\alpha + i - 1} \end{aligned} \quad (3.38)$$

This is also known as the *Blackwell and MacQueen prediction rule* [4]. By the law of large numbers for exchangeable sequences, the empirical distribution of a sample $\theta_1, \dots, \theta_n$ generated according to (3.37) will converge almost surely to the (random) distribution P . (Although this is not clear from (3.37), the construction as an i.i.d. sample from P shows that such a sequence will be exchangeable.) Next to having computation merit in some scenarios, this prediction rule (3.38) also provides a hint on how the number of latent classes (i.e. m_i) evolves, in relation to the prior strength α . It can be shown that the expected number of latent classes is $\sum_{j=1}^i \frac{\alpha}{\alpha + j - 1}$, and the probability that any two sample units are in the same latent class is $P(c_i = c_j) = 1/(1 + \alpha)$, for $\forall i \neq j$.

3.10 Annex: Pitman-Yor and Dirichlet Process Mixture Model

Because the random measure P produced in (3.36) is discrete, it is not useful as a prior for density estimation. This is solved by adding another layer to the hierarchy, as follows:

$$\begin{aligned} X_i | (\theta_i), G &\sim F(\cdot; \theta_i), \quad i = 1, \dots, n \\ \theta_i | G &\stackrel{\text{iid}}{\sim} G, \\ G &\sim PY(a, b, G_0), \end{aligned} \quad (3.39)$$

where $F(\cdot; \theta)$ is a distribution that is completely determined by the parameter θ . If this distribution is continuous, then so will be the (marginal) distribution of the variables X_i , which will be a countably infinite mixture of $F(\cdot; \theta)$ distributions. This is called Pitman-Yor Process Mixture (PYPM) model, with the Dirichlet Process Mixture (DPM) model as the special case when $G \sim PY(0, \alpha, G_0) = DP(\alpha G_0)$.

Full posterior inference for such model is usually performed in two ways: using the stick-breaking representation in Section 3.8; or the Pólya urn scheme as in Section 3.9. We give an overview of these methods in the next two sections, and refer to [34], [37], [23], and [18] for details. Other algorithms and refinements, beyond our scope here, have been developed (i) to overcome the finite approximation of a stick breaking process, by [62] and [26] via slice sampling, or by [40] via retrospective sampling,

(ii) to improve the efficiency of Gibbs sampler against local modes, by [25], [24], and [13] via split-and-merge, or by [5] via variational approximation.

3.10.1 Block Gibbs Sampler in Stick-Breaking Scheme

By the stick-breaking representation, the PYM model (3.39) can be re-written as follows.

$$\begin{aligned}
 X_i | \phi, c_i &\stackrel{\text{iid}}{\sim} F(\cdot; \theta_i = \phi_{c_i}), \quad i = 1, \dots, n, \\
 c_i | p &\stackrel{\text{iid}}{\sim} \sum_{l=1}^K p_l \delta_l, \\
 \phi &\stackrel{\text{iid}}{\sim} G_0, \\
 p_1 = V_1; p_l &= (1 - V_1) \cdots (1 - V_{l-1}) V_l, \quad \text{for } l = 2, \dots, K \\
 V_l &\sim \text{Beta}(1 - a_0, b_0 + a_0 l),
 \end{aligned} \tag{3.40}$$

where $p = (p_1, \dots, p_K)$, $c = (c_1, \dots, c_n)$, $\phi = (\phi_1, \dots, \phi_K)$, $a_0 \in [0, 1)$ and $b_0 > -a_0$, and K is set to be some pre-specified large integer for approximation. The Gibbs sample of such setup consists of repeatedly sampling from the full conditional distributions $(\phi | c, X)$, $(c | \phi, p, X)$ and $(p | c)$.

More specifically, we consider PYM of Gaussians, that is, $F(\cdot; \theta_i) = N(\mu_i, \Sigma_i)$ and so $\{\phi_l = (\mu_l, \Sigma_l)\}_{l=1}^K$. Recall $X_i \in \mathbb{R}^d$, and hence Σ_l is $d \times d$ matrix here. For computational efficiency, conjugacy can be introduced by presuming that G_0 is a Normal-Inverse-Wishart distribution when $d \geq 2$, that is, $\Sigma_l \stackrel{\text{iid}}{\sim} \text{Inv-Wishart}(v_0, \Lambda_0^{-1})$ and $\mu_l | \Sigma_l \stackrel{\text{iid}}{\sim} N(\mu_0, \Sigma_l / \kappa_0)$, where $v_0 > d - 1$, mode of $\text{Inv-Wishart}(v_0, \Lambda_0^{-1})$ is $\Lambda_0^{-1} / (v_0 + d + 1)$ and mean is $\Lambda_0^{-1} / (v_0 - d - 1)$ for $v_0 > d + 1$. Equivalently,

$$\pi(\mu, \Sigma) = dG_0(\mu, \Sigma) \propto |\Sigma|^{-\frac{v_0+d}{2}+1} \exp\left(-\frac{1}{2} \text{tr}(\Lambda_0 \Sigma^{-1}) - \frac{\kappa_0}{2} (\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0)\right). \tag{3.41}$$

The steps of a Gibbs sampler at h^{th} scan can then be elaborated as below:

- **Step:** $\phi^{(h)} = (\mu^{(h)}, \Sigma^{(h)}) | c^{(h-1)}, X$

Given $c^{(h-1)} = (c_1^{(h-1)}, \dots, c_n^{(h-1)})$, for every $l \in \{1, \dots, K\}$, let $C_l^{(h-1)} = \{i: c_i^{(h-1)} = l, \text{ and } i =$

$1, \dots, n\}$, and

$$\phi_l^{(h)} | c^{(h-1)}, X \sim \text{NIW}(\mu_{n_l}^*, \kappa_{n_l}^*, \nu_{n_l}^*, \Lambda_{n_l}^*),$$

$$\mu_{n_l}^* = \frac{\kappa_0}{\kappa_0 + n_l^{(h-1)}} \mu_0 + \frac{n_l^{(h-1)}}{\kappa_0 + n_l^{(h-1)}} \bar{X}_l^{(h-1)}, \quad (3.42)$$

$$\kappa_{n_l}^* = \kappa_0 + n_l^{(h-1)}, \quad (3.43)$$

$$\nu_{n_l}^* = \nu_0 + n_l^{(h-1)}, \quad (3.44)$$

$$\Lambda_{n_l}^* = \Lambda_0 + S_l^{(h-1)} + \frac{\kappa_0 n_l^{(h-1)}}{\kappa_0 + n_l^{(h-1)}} (\bar{X}_l^{(h-1)} - \mu_0)(\bar{X}_l^{(h-1)} - \mu_0)', \quad (3.45)$$

where $n_l^{(h-1)} = |C_l^{(h-1)}| = \sum_i \mathbf{1}_{\{c_i^{(h-1)} = l\}}$, $\bar{X}_l^{(h-1)}$ is the sample average of $\{X_i : i \in C_l^{(h-1)}\}$, $S_l^{(h-1)} = \sum_{i \in C_l^{(h-1)}} (X_i - \bar{X}_l^{(h-1)})(X_i - \bar{X}_l^{(h-1)})'$. The marginals are

$$\Sigma_l | c^{(h-1)}, X \sim \text{IW}(\nu_{n_l}^*, \Lambda_{n_l}^{*-1}), \quad (3.46)$$

$$\mu_l | c^{(h-1)}, X \sim t_{\nu_{n_l}^* - d + 1} \left(\mu_{n_l}^*, \frac{\Lambda_{n_l}^*}{\kappa_{n_l}^* (\nu_{n_l}^* - d + 1)} \right). \quad (3.47)$$

The posterior predictive likelihood

$$p_l(x | X) \propto \int p(x | \mu_l, \Sigma_l) \pi(\mu_l, \Sigma_l | X) = t_{\nu_{n_l}^* - d + 1} \left(\mu_{n_l}^*, \frac{\Lambda_{n_l}^* (\kappa_{n_l}^* + 1)}{\kappa_{n_l}^* (\nu_{n_l}^* - d + 1)} \right). \quad (3.48)$$

• **Step:** $c^{(h)} | \phi^{(h)}, p^{(h-1)}, X$

With the pre-specified larger integer K , for every sample $i \in \{1, \dots, n\}$, the posterior distribution is given by

$$c_i^{(h)} | \phi^{(h)}, p^{(h-1)}, X_i \stackrel{\text{ind}}{\sim} \sum_{l=1}^K p_{l,i}^* \delta_l, \quad (3.49)$$

$$p_{l,i}^* \propto p_l^{(h-1)} f(X_i | \phi_l^{(h)}), \quad l = 1, \dots, K.$$

Here f is the density function of a normal distribution. Normalization is needed to ensure that $\sum_{l=1}^K p_{l,i}^* = 1$.

• **Step:** $p^{(h)} | c^{(h)}$

Due to the conjugacy of a generalized Dirichlet distribution to a multinomial distribution,

the posterior is

$$p_l^{(h)} | c^{(h)} = \begin{cases} V_1^*, & l = 1, \\ (1 - V_1^*) \cdots (1 - V_{l-1}^*) V_l^*, & l = 2, \dots, K-1, \\ (1 - V_1^*) \cdots (1 - V_{K-1}^*), & l = K, \end{cases} \quad (3.50)$$

where

$$V_l^* \stackrel{\text{ind}}{\sim} \text{Beta} \left(1 - a_0 + n_l^{(h+1)}, b_0 + a_0 l + \sum_{j=l+1}^K n_j^{(h+1)} \right), \quad \text{for } l = 1, \dots, K-1. \quad (3.51)$$

These steps are run cyclically. The choice of the initial step is not expected to have a strong impact.

To facilitate the content in the main chapter, we shall make two special cases more explicit here. The first case is when $X \in \mathbb{R}$ and $\{\phi_l = (\mu_l, \sigma_l^2)\}_{l=1}^K$. To retain conjugacy, we set

$$G_0 := \text{NIG} \left((\mu, \sigma^2) | \mu_0, \kappa_0, A_0, B_0 \right) = N(\mu | \mu_0, \sigma^2 / \kappa_0) IG(\sigma^2 | A_0, B_0), \quad (3.52)$$

where $dIG(x | A, B) = \Gamma(A)^{-1} B^A x^{-A-1} e^{-B/x}$ and $x, A, B > 0$. Formulas (3.42)–(3.45) can then be updated as follows:

$$\begin{aligned} \phi_l^{(h)} | c^{(h-1)}, X &\sim \text{NIG}(\mu_{n_l}^*, \kappa_{n_l}^*, A_{n_l}^*, B_{n_l}^*), \\ \mu_{n_l}^* &= \frac{\kappa_0}{\kappa_0 + n_l^{(h-1)}} \mu_0 + \frac{n_l^{(h-1)}}{\kappa_0 + n_l^{(h-1)}} \bar{X}_l^{(h-1)}, \end{aligned} \quad (3.53)$$

$$\kappa_{n_l}^* = \kappa_0 + n_l^{(h-1)}, \quad (3.54)$$

$$A_{n_l}^* = A_0 + \frac{n_l^{(h-1)}}{2}, \quad (3.55)$$

$$B_{n_l}^* = B_0 + \frac{1}{2} \left[\mu_0^2 \kappa_0 + \sum_{i \in C_l^{(h-1)}} X_i^2 - (\mu_{n_l}^*)^2 \kappa_{n_l}^* \right], \quad (3.56)$$

Equivalently, we can work with precision $\lambda_l = 1/\sigma_l^2$, instead of variance σ_l^2 , and then set G_0 to be $\text{NG}((\mu, \lambda) | \mu_0, \kappa_0, A_0, B_0) = N(\mu | \mu_0, 1/\lambda \kappa_0) \text{Gamma}(\lambda | A_0, B_0)$, where

$$d\text{Gamma}(x | A, B) = \Gamma(A)^{-1} B^A x^{A-1} e^{-Bx}, \quad x, A, B > 0.$$

The posterior distribution $(\mu_l^{(h)}, \lambda_l^{(h)} | c^{(h-1)}, X)$ is $\text{NG}(\mu_{n_l}^*, \kappa_{n_l}^*, A_{n_l}^*, B_{n_l}^*)$ due to conjugacy, where $\mu_{n_l}^*, \kappa_{n_l}^*$ and $A_{n_l}^*$ are the same as (3.53)–(3.55), and

$$B_{n_l}^* = B_0 + \frac{1}{2} \left[\sum_{i \in C_l^{(h-1)}} \left(X_i - \bar{X}_l^{(h-1)} \right)^2 + \frac{n_l^{(h-1)} \kappa_0}{n_l^{(h-1)} + \kappa_0} \left(\bar{X}_l^{(h-1)} - \mu_0 \right)^2 \right]. \quad (3.57)$$

The marginal posteriors are

$$\lambda_l^{(h)} | c^{(h-1)}, X \stackrel{\text{ind}}{\sim} \text{Gamma}(A_{n_l}^*, B_{n_l}^*), \quad (3.58)$$

$$\mu_l^{(h)} | c^{(h-1)}, X \stackrel{\text{ind}}{\sim} t_{2A_{n_l}^*} \left(\mu_{n_l}^*, B_{n_l}^* / (A_{n_l}^* \kappa_{n_l}^*) \right). \quad (3.59)$$

The predictive posterior likelihood under both parameterizations (with a NIG or NG prior) are as follows.

$$p(x|X) = t_{2A_{n_l}^*} \left(\mu_{n_l}^*, B_{n_l}^* (\kappa_{n_l}^* + 1) / (A_{n_l}^* \kappa_{n_l}^*) \right). \quad (3.60)$$

The second case to be made explicit is the DPM of Gaussians. As said, DP is a special case of a Pitman-Yor process. Hence, it is identical to have the prior setting $PY(a_0 = 0, b_0 = \alpha, G_0)$ in (3.40). In this case, formula (3.51) can be simplified to

$$V_l^* \stackrel{\text{ind}}{\sim} \text{Beta} \left(1 + n_l^{(h)}, \alpha + \sum_{j=l+1}^K n_j^{(h)} \right), \quad \text{for } l = 1, \dots, K-1. \quad (3.61)$$

A hyper-prior may be assumed for the prior strength α if one wants to derive the model in a full Bayesian way. The conjugate prior for α is a $\text{Gamma}(\gamma_a, \gamma_b)$; see section 4.8 (4.16) in [18] for details.

3.10.2 Gibbs Sampler in Pólya Urn Scheme

Recall the setup of (3.39) where $\theta_i = \phi_{c_i}$. The efficiency of Gibbs sampling in the Pólya urn scheme can be increased by marginalizing $\theta | G$ over its prior $PY(a, b, G_0)$. The model can be rewritten as

$$\begin{aligned} X_i | \theta_i &\stackrel{\text{ind}}{\sim} F(\theta_i), \quad i = 1, \dots, n, \\ (\theta_1, \dots, \theta_n) &\sim \pi(\theta_1, \dots, \theta_n), \end{aligned} \quad (3.62)$$

where $\pi(\theta_1, \dots, \theta_n)$ denotes the joint distribution by sequentially multiplying (3.37) for $i = 1, \dots, n$.

The full conditional prior $\theta_i | \theta_{-i}$ can be expressed by (3.37) and treating sample i as the last observation. This is legitimate due to the full exchangeability of samples under (3.37).

$$\theta_i | \theta_{-i} \sim \frac{b + am}{b + n - 1} G_0 + \sum_{l=1}^m \frac{n_l - a}{b + n - 1} \delta_{\phi_{c_i}}, \quad i = 1, \dots, n \quad (3.63)$$

where $\theta_{-i} = \{\theta_j\}_{j=1}^n \setminus \{\theta_i\}$, $\phi = \{\phi_l\}_{l=1}^m$ is the set of distinct values among $\{\theta_i\}_{i=1}^n$, c_i is the latent class label of sample i , and n_l is $\sum_i \mathbf{1}_{\{c_i=l\}}$. To sample from the posterior distribution $\theta|X$, we can iteratively sample from $\theta_i|\theta_{-i}, X_i$ for $i = 1, \dots, n$, where θ_{-i} includes ϕ and c_{-i} , and

$$\theta_i|\theta_{-i}, X \sim w_0^* \pi(\theta_i|X_i) + \sum_{l=1}^m w_l^* \delta_{\phi_l}, \quad (3.64)$$

$$w_0^* \propto (b + am) \int f(X_i|\theta) dG_0(\theta), \quad (3.65)$$

$$w_l^* \propto (n_l - a) f(X_i|\phi_l). \quad (3.66)$$

Here $\pi(\theta_i|X_i)$ is the posterior of θ_i given prior G_0 and single observation X_i , and $w_0^* + \sum_{l=1}^m w_l^* = 1$. This Gibbs sampling scheme is relatively effective if the integral on the right side of (3.65) has a closed analytical form, see [37] for a more comprehensive overview. However, it is still very computationally expensive to be applied when the sample size n is big.

Chapter 4

Nonparametric Tests and Prediction Sets for Functional Data

4.1 Introduction

This chapter studies data depth methodology in functional data analysis, for (i) detecting “outliers”, (ii) constructing prediction sets, and (iii) performing two-sample tests. The motivation comes from emerging types of functional data in biotech manufacturing. The following are examples of routine tasks:

1. assess if the spectroscopy profile of a newly-manufactured batch is ‘within expectation’ relative to the profiles of historical batches. A spectroscopy profile captures chemical and/or biophysical characteristics of a material. A typical example is the absorption of light of different wavelengths over a prefixed range.
2. assess the difference or similarity between two sets of time series representing a key process attribute measured on a cell culture (or purification) before and after a critical process change.
3. compare particle size distributions before and after a change in starting material or process.

Currently these tasks are performed visually by well-trained specialists following a detailed protocol. A more data-driven approach would remove human bias.

In functional data analysis the observations are functions $t \mapsto X_i(t)$ of a continuous argument t . This setup fits the preceding examples, in which wavelength, time or particle size are conceptually continuous variables. In practice, one observes the values of these functions only at a finite grid of values t , and hence the observations are vectors $\mathbf{X}_i = (X_i(t_1), \dots, X_i(t_d))$. If we observe N curves at d grid points, then the full data will be an $N \times d$ data matrix $\mathbf{X}$. A data matrix resulting from functional data is different than a data matrix resulting from just any N observational vectors, as the columns are ordered and usually have a numerical reference. When visualizing the datamatrix, it is more intuitive to plot N curves against this numerical index, than to think of N points in d -dimensional space. Also the

number of columns, resulting from the sampling frequency of the grid points $t_1, \dots, t_d$, is to a certain extent a subjective choice. Ideally it should reflect our expectation on the variability of the underlying functions. Less smooth functions should be observed on a finer grid, as otherwise important local features may be missed. To make the distinction explicit, we shall refer to the observations $\mathbf{X}_i$ as “functional vectors”.

The chapter is organized as follows. We start with an overview of functional data depth (FDD) in Section 4.2, and propose two new constructions of FDD, with a view to outlier detection when the population is multi-modal. Next we examine the sensitivity of different FDDs to different types of outliers in Section 4.3, and show the potential advantage of centering the curves. In Section 4.4 we discuss three approaches to construct a prediction set with functional data. We end the chapter in Section 4.5 with a discussion of three approaches to two-sample testing with functional data.

Pre-processing of the functional data, such as landmark registration, may be needed to remove contamination (e.g. a slight horizontal shift or stretching of a curve), and to ensure that the inference can be focused on functional features of interest. This is usually case specific; it is beyond the scope of this chapter. One may consult Chapters 3-5 in [45] for a detailed account.

4.2 Functional Data Depth (FDD)

4.2.1 Definition and Overview

We restrict to functions $x: V \mapsto \mathbb{R}$ on an interval $V = [a, b] \subset \mathbb{R}$. If ζ is the set of all functions of interest and $\mathcal{P}$ is the set of all probability measures on ζ , then a statistical depth functional is a mapping $D: \zeta \times \mathcal{P} \mapsto \mathbb{R}$. A value $D(x, P)$ is meant to measure centrality of the function x within the distribution P , a higher value indicating that x is closer to the center of P . The ‘population depth’ $D(x, P)$ will be estimated with the help of a random sample $x_1, \dots, x_n$ of functions from some $P \in \mathcal{P}$, and then leads to a measure of centrality ‘within a sample’.

For a discussion on functional depth when the functions x have more than one argument, we refer to [11].

By discretizing a function x over a grid with d points, it becomes a vector in $\mathbb{R}^d$, to which we can apply any notion of data depth for random vectors. Several fundamental differences between such a functional vector and a general multivariate observation need to be taken into account. Following the summary in [48], we note:

1. A functional vector $\mathbf{X}_i$, when visualized as a curve divides the two-dimensional plane into upper and lower regions. This provides intuition on possible definitions of a “central” curve in a sample. For instance, roughly 50% of the sample curves should lay entirely or mostly in the upper or lower region defined by the “central” curve.

2. The notion of *centrality* (or *outlyingness*) of a curve should not solely focus on addressing its *location* relative to the population, but also on its *shape*. The vertical location of a curve x over support $[a, b]$ is commonly defined as

$$\frac{1}{b-a} \int_a^b x(t) dt. \quad (4.1)$$

For depicting a functional pattern there are many possibilities.

3. A notion of centrality of a curve can involve a trade-off between global aspects across its full support V and local aspects in the neighborhood of points in V .
4. Not all desirable properties of multivariate data depth (see [72]) are meaningful for functional depth, such as full affine invariance. See [38], [20], [48] for a discussion. A list of desirable properties of functional data depth was given in [38] and further refined in [20].

The prevalent functional depths may be classified into:

- (i) Integrated depth (ID), proposed by [16],
- (ii) Projection-based depth (PBD), proposed in [71] and in [12],
- (iii) Geometrical depths, including band depth (BD) from [32], half-region depth (HRD) from [33] and spatial depth (SD) for functional data proposed in [10], [9] and [50].

Consistency of the sample estimators of these depths and other desired statistical properties have been proven by [16] for ID, [71] for PBD, [19] for BD and HRD, and [9] and [48] for SD. See [20] for a comprehensive review.

Some multivariate depths are degenerate in the functional case, that is, the resulting $D(x, P)$ is the same for any x in the set ζ (see [9] and [20]). This problem is known to burden BD and HRD when the observed curves are very wiggly and hence cross each other frequently. In practice, the degeneracy issue may be avoided by using the depths of the smoothed functions, instead of the raw "fuzzy" curves. Such pre-processing is best left to the specific needs in the application context. To alleviate this problem, modified versions MBD and MHRD were proposed. However, MBD and MHRD are not good in detecting a spiky outlier curve (see Section 4.3 for illustration).

Besides being useful for the purposes described in the next sections, functional data depth can be applied (a) to define a robust estimate of a typical (or "center") pattern, using the median or a trimmed average ([16]); (b) to build a confidence band around the estimated "center" curve ([8], [7]); (c) to define surrogate responses in a classification problem ([12], [44]).

4.2.2 ID and WID

Given a univariate depth $D(x(t), F_t)$ of a functional value $x(t)$ relative to its marginal distribution F_t , one can form an *integrated depth* (ID) as $ID = \int_a^b D(x(t), F_t) dt$. The univariate depth can take any form, such as Tukey depth or simplicial depth. For arithmetic simplicity [16] proposed the median

depth $D(x(t), F_t) = 1 - |1/2 - F_t(x(t))|$, where F_t also denotes the marginal cumulative distribution function, which results in

$$ID(x, F) = b - a - \int_a^b \left| \frac{1}{2} - F_t(x(t)) \right| dt.$$

In the sample version the population distribution function F_t is replaced by the empirical distribution function of the observed data points $x_1(t), \dots, x_n(t)$ at t . The function *depth.FM* from the R package *fda.usc* will be used to calculate this integrated depth in later sections.

In our practical situation of interest, the population usually is a mixture involving multiple stereotypical functions. We propose a measure of functional depth, weighted integrated depth (WID), that is appropriate for this situation. A useful population model is a mixture of Gaussian processes (GPs), defined structurally as

$$\begin{aligned} X(t) &= \mu + Z(t), \quad t \in [a, b], \\ \mu &\sim N(m_\mu, \varepsilon_\mu^2), \\ Z|c &\sim GP_c(m_c, k_{\theta_c}), \\ c &\sim (w_c)_{c=1, \dots, C}, \\ \kappa_{\theta_c}(s, t) &= \text{Cov}(Z(s), Z(t)|c) = l_c^2 \exp\left(-\frac{(s-t)^2}{\tau_c^2}\right) + \varepsilon_c^2 \delta_{s,t}, \end{aligned} \tag{4.2}$$

where the scalar μ reflects the overall location of the curves, and the functions m_c , for $c = 1, \dots, C$, are the different stereotypical curves after centering. In the model (4.2) the functions m_c are the mean functions of Gaussian processes, which are further specified by their covariance functions k_{θ_c} . Given a latent variable c with distribution (w_c) , the process Z is distributed as a Gaussian process with parameters (m_c, k_{θ_c}) . The covariance functions are specified by vectors of three positive numbers $\theta_c = (l_c, \tau_c, \varepsilon_c)$, and are the sum of the Gaussian kernel, with spread and smoothness determined by l_c and τ_c , and a white noise function with variance ε_c^2 , reflecting observational noise. (The latter process could be restricted to the finite grid at which the processes are observed.) The notation $\delta_{s,t}$ is used for the Kronecker delta, i.e. $\delta_{s,t} = 1$ if $s = t$ and $\delta_{s,t} = 0$ otherwise. The Gaussian kernel is chosen for illustration, and because the resulting Gaussian process can adapt to any smoothness by proper choice of τ_c . The distribution of Z is a mixture $F = \sum_{c=1}^C w_c GP_c(m_c, k_{\theta_c})$ of Gaussian distributions.

We define the WID of a *centered* function x (i.e. a function x with $\int_a^b x(t) dt / (b-a) = 0$) as

$$WID(x, F) = \sum_{c=1}^C w_c P_{Z_c \sim GP_c} \left(O(x, m_c, k_{\theta_c}) \leq O(Z_c, m_c, k_{\theta_c}) \right), \tag{4.3}$$

where $O(x, m_c, k_{\theta_c})$ is a weighted square distance of x relative to m_c :

$$O(x, m_c, k_{\theta_c}) = \int_a^b \frac{(x(t) - m_c(t))^2}{\kappa_{\theta_c}(t, t)} dt. \quad (4.4)$$

The weights in (4.4) are chosen so that the distances $O(x, m_c, k_{\theta_c})$ are on the same scale across the component Gaussian processes. The square distance (4.4) will be larger if x is further from the mean curve of the c^{th} cluster, and hence the corresponding probability in the definition of $WID(x, F)$ will be larger if x is closer to m_c . All probabilities in (4.3) will be small if x is far from every m_c . Thus $WID(x, F)$ measures data depth of x relative to the mixture distribution.

In practice F will be unknown and replaced by an estimator $F_n = \sum_{c=1}^C \hat{w}_c GP_{c,n}$, based on a random sample $\{x_i\}_{i=1}^n$ from the distribution of X in (4.2). This involves estimators $\{\hat{w}_c\}_{c=1}^C$, $\{\hat{m}_c\}_{c=1}^C$ and $\{\hat{\theta}_c\}_{c=1}^C$ of the parameters w_c , m_c and θ_c . A rough, but computationally efficient, method of estimation could be to cluster first the curves $x_1, \dots, x_n$, after centering, next let the weights $\hat{w}_c$ be proportional to the sizes of the clusters, and estimate the remaining parameters by fitting the Gaussian process model to the curves in the corresponding cluster. As a final note, a pruning procedure may be needed to remove very small clusters. Centering of a curve x_i can be done by subtracting the trapezium-rule approximation $v_i = 0.5 \sum_{j=1}^{d-1} (t_{j+1} - t_j) (x_i(t_j) + x_i(t_{j+1})) / (b - a)$ to its weighted integral. The numbers $\{v_i\}_{i=1}^n$ may be ranked separately using a univariate depth measure. In practice we may observe curves of similar patterns but with varying vertical locations, and decomposing depth in a measure for curve pattern and vertical height will provide more insight.

The probabilities in (4.3) may be approximated by simulating large samples $\{x_b^*\}_{b=1}^B$ from the estimated Gaussian process $GP_{c,n}$ and computing

$$B^{-1} \sum_{b=1}^B \mathbf{1}_{\{O(x, \hat{m}_c, k_{\hat{\theta}_c}) \leq O(x_b^*, \hat{m}_c, k_{\hat{\theta}_c})\}}.$$

If the chosen estimators $\hat{w}_c$, $\hat{m}_c$ and $\hat{\theta}_c$ converge to their population counterparts, then the corresponding estimator of $WID(x, F)$ will also be asymptotically consistent.

4.2.3 PBD, RPD and MPD

Projection-based depth (PBD) defines the depth of a vector as the smallest univariate depth of its one-dimensional projections. It is developed for multivariate data in [71]. The outlyingness of $x \in \mathbb{R}^d$ relative to a distribution F on $\mathbb{R}^d$ is defined as

$$O(x, F) = \sup_{\|u\|=1} \frac{|u'x - \mu(F_u)|}{\sigma(F_u)},$$

where F_u is the univariate distribution of $u'X$ if $X \sim F$, for a given $u \in \mathbb{R}^d$, and $(\mu(F_u), \sigma(F_u))$ are any pair of location and scale functionals of F_u , e.g. (mean, standard deviation) or (median, median

absolute deviation). (If $\sigma(F_u) = 0$, the quotient is read as 0.) The population depth of x is then defined to be $1/(1 + O(x, F))$. For a sample version, the distributions F_u are replaced by their empirical counterparts.

Multivariate projection-based depth was generalized to functional data by [12]. The idea is to replace the vectors u by functions and $x'u$ by the integrals $\int_a^b x(t)u(t) dt$. Because there are too many possible choices for function u , [12] proposed a random projection depth (RPD) based on functions $u_1, \dots, u_B$ generated from a Gaussian distribution and standardized to norm 1, given by

$$RPD(x, F) = \frac{1}{1 + O(x, F)}, \quad (4.5)$$

where

$$O(x, F) = \frac{1}{B} \sum_{u \in \{u_1, \dots, u_B\}} \left| \frac{\int_a^b u(t)x(t) dt - \mu(F_u)}{\sigma(F_u)} \right|, \quad (4.6)$$

and where F_u denotes the distribution of $\int_a^b u(t)X(t) dt$ for $X \sim F$, for a given u , with location and scale $\mu(F_u)$ and $\sigma(F_u)$. (The functions u could be normed, but this will be irrelevant if the μ and σ are scale invariant.) The properties of this RPD will not only depend on the choice of (μ, σ) , but also on B and the assumed stochastic process for u . In practice the functions will only be evaluated on a grid and the integrals will be approximated by sums. The function *depth.RP* from the R package *fda.usc* will be used to calculate RPD.

We investigate an alternative, which is relevant under the population model given by the mixture of Gaussian processes as in (4.2). Rather than the random directions we use the mean functions m_c of the components, leading to

$$MPD(x, P) = \left(1 + \max_c w_c \left| \frac{\int_a^b m_c(t)x(t) dt - \mu(F_{m_c})}{\sigma(F_{m_c})} \right| \right)^{-1}. \quad (4.7)$$

Here F_{m_c} is the distribution of $\int_a^b m_c(t)X(t) dt$ if $X \sim GP_c$. We choose (μ, σ) to be the median and median absolute deviation to avoid the impact of outliers.

In the sample version, we first estimate w_c and m_c from data $\{x_i(t): t = t_1, \dots, t_d\}_{i=1}^n$, by the same procedure as described for for WID. Furthermore, we evaluate $\mu(F_{m_c})$ and $\sigma(F_{m_c})$ based on large samples from GP_c .

4.2.4 BD and MBD

For an arbitrary subset $I \subset \{1, \dots, n\}$, we define the ‘band’ formed by the set of sample functions $x_I = \{x_i: i \in I\}$ as

$$B(x_I) = \{(t, y): t \in V, \min_{k \in I} (x_k(t)) \leq y \leq \max_{k \in I} (x_k(t))\}.$$

Next for $G(x) = \{(t, x(t)) : t \in [a, b]\}$ the graph of an arbitrary function x , define the sample band depth (BD) of x relative to $x_1, \dots, x_n$ as

$$BD_{n,J}(x) = \sum_{j=2}^J \left[\frac{1}{\binom{n}{j}} \sum_{I \subset \{1, \dots, n\}; |I|=j} \mathbf{1}_{\{G(x) \subseteq B(x_I)\}} \right]. \quad (4.8)$$

Here J is a fixed integer with $2 \leq J \leq n$, which is commonly recommended to be chosen equal to $J = 3$, in view of the computational burden.

A modified version (MBD) is obtained by not requiring that the full graph be contained in the band, but instead measure the length of the interval on which x falls within the band, i.e. the length of the set

$$A(x_I) = \{t \in V : \min_{k \in I} (x_k(t)) \leq x(t) \leq \max_{k \in I} (x_k(t))\}.$$

Next we replace the indicator function $\mathbf{1}_{(\cdot)}$ by the Lebesgue measure of $A(x_I)$, and define

$$MBD_{n,J}(x) = \sum_{j=2}^J \left[\frac{1}{\binom{n}{j}} \sum_{I \subset \{1, \dots, n\}; |I|=j} \frac{\lambda(A(x_I))}{\lambda(V)} \right]. \quad (4.9)$$

The functions BD and MBD from the R package *roahd* will be used to calculate BD and MBD in later analyses.

4.2.5 HRD and MHRD

The half-region depth measures the proportion of sample curves that are above or below a given function. Define the hypograph (*hyp*) and epigraph (*epi*) of a function x with support V as

$$\begin{aligned} hyp(x) &= \{(t, y) \in V \times \mathbb{R} : y \leq x(t)\}, \\ epi(x) &= \{(t, y) \in V \times \mathbb{R} : y > x(t)\}. \end{aligned}$$

Then for $G(x_i)$ the graph of the sample function x_i , the sample half-region depth (HRD) of x is defined as

$$HRD_n(x) = \min\{K_1(x), K_2(x)\}, \quad (4.10)$$

where

$$\begin{aligned} K_1(x) &= \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{G(x_i) \subseteq hyp(x)\}}, \\ K_2(x) &= \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{G(x_i) \subseteq epi(x)\}}, \end{aligned}$$

are the proportions of sample values with graph fully below or above x .

This depth measure can be modified by not requiring that a sample function is everywhere below of above x , but by measuring the length of the sets where this is the case, in the same spirit as MBD relaxes BD. we define MHRD as

$$MHRD_n(x) = \min\{K_1^*(x), K_2^*(x)\}, \quad (4.11)$$

where

$$K_1^*(x) = \frac{1}{n\lambda(V)} \sum_{i=1}^n \lambda\{t \in V: x(t) \leq x_i(t)\},$$

$$K_2^*(x) = \frac{1}{n\lambda(V)} \sum_{i=1}^n \lambda\{t \in V: x(t) > x_i(t)\}.$$

The functions *HRD* and *MHRD* from the R package *roahd* will be used to calculate HRD and MHRD in later analyses.

4.3 Topic I - Visualization of Outliers

Given a sample of curves, a future curve can be viewed as an outlier if (a) it is at an abnormal location, e.g. the entire curve lays above or below the majority of the sample curves, (b) its pattern is not similar to any of the sample curves, or (c) a combination of (a) and (b). Additionally, a deviation in pattern in the sense of (b) may be global or a local. In this section, we use a simulated dataset to examine the sensitivity of the different FDDs with respect to outlier functions of different designs, and illustrate the potential advantage to monitor outliers via the depth values of curves after centering. A bivariate plot is presented as a useful tool to separate outlyingness due to (a) or (b).

We consider a population that is a mixture of two stochastic components (M1) and (M2), specified below, with weights p_{M1} and $p_{M2} = 1 - p_{M1}$:

- (M1) $X(t) = \mu + 4t + \varepsilon(t)$, where $\mu \sim N(0, 0.9^2)$ and ε is a centered Gaussian process with covariance kernel $\kappa(s, t) = 0.1 \exp\left(-(t-s)^2/0.07\right)$, for $0 \leq s, t \leq 1$.
- (M2) $X(t) = \mu + 0.9 \sin(3\pi t) + 4t + \varepsilon(t)$, where $\mu \sim N(0, 0.4^2)$ and ε is a centered Gaussian process with covariance kernel $\kappa(s, t) = 0.05 \exp\left(-(t-s)^2/0.02\right)$, for $0 \leq s, t \leq 1$.

The curves are observed at 50 equally spaced points in $[0, 1]$.

We design the following three outlier curves:

- (Outlier-1: Global pattern deviation): a sample curve from $X(t) = -1.5 + 7t + \varepsilon(t)$, where ε is the same as (M1).
- (Outlier-2: Sharp local spike): the sum of a sample curve from (M1) and the function that is 4.5 when $t \in (0.7, 0.75)$ and zero otherwise.

- (Outlier-3: Modest local spike): the sum of a sample curve from (M1) and the function that is 2.5 when $t \in (0.7, 0.75)$ and zero otherwise. In addition, the curve location is chosen to be relatively low so that the modest spike is among the sample curves.

We simulated 250 curves from (M1), and 250 curves from (M2) and added the three outlier curves, giving a dataset of 503 curves in total. The 503 curves are shown in the top left panel of Figure 4.1. The top right panel gives the same curves but centered vertically by subtracting the constant

$$FL(x) = \frac{1}{d} \sum_{j=1}^d x(t_j) \quad (4.12)$$

from each curve x . These are estimates for the ‘functional locations’ $\int_0^1 x(t) dt$ of the curves.

We calculated the depth of all 503 individual curves within the total sample using every of the methods ID, WID, BD, MBD, HRD, MHRD, RPD and MPD. To facilitate visual comparison, we rescaled the depth values to the interval $[0, 1]$, by subtracting per method the minimum depth over the 503 curves and dividing by the range of the depth values. These rescaled values are plotted in the bottom left panel of Figure 4.1, while the bottom right panel shows the same but now for the centered curves. The depth values of WID are absent in the left plot, since this method is designed for centered curves, while BD and HRD are absent in the right plot in view of the degeneracy of these methods for the original curves (the curves cross each other too often).

We may expect that the FL-centered curves are more suitable for discovering functional patterns. Besides, we found that clustering the centered curves lead to more accurate recovery of the two clusters in the data (needed for WID and MPD calculation).

Details of the calculations are as follows. For WID and MPD we clustered the curves by the K -means method, with $K = 2$, the correct number of clusters in this case. In practice, we would determine a suitable value of K by a preliminary analysis. Given the clustering the center curve per cluster was estimated by pointwise α -trimmed means with $\alpha = 0.05$, to avoid the influence of the extreme values. For WID, we set the value of $\kappa_{\theta_i}(t, t) + \varepsilon_{\mu}^2$ in (4.4) for every t equal to the pointwise variance of the functions in a cluster averaged over the 50 grid points. We used here the fact that in the present data the variance is constant in t . In practice we would make full inference on the parameters of the Gaussian processes.

Ideally the depth values of the three outlying curves would be smallest among all depth values. In the left bottom panel of Figure 4.1, we see that only BD is sensitive enough to detect all the three outliers in this way. This comes at a high computational burden as the number of the curves increases. The HRD method seems to be sensitive to a sharp local spike, while the other FDDs all perform poorly overall.

Centering the curves appears to enhance the sensitivity to outliers for all FDDs. All FDDs except MHRD and RPD can detect the global pattern deviation. MBD and MPD are not sensitive to the

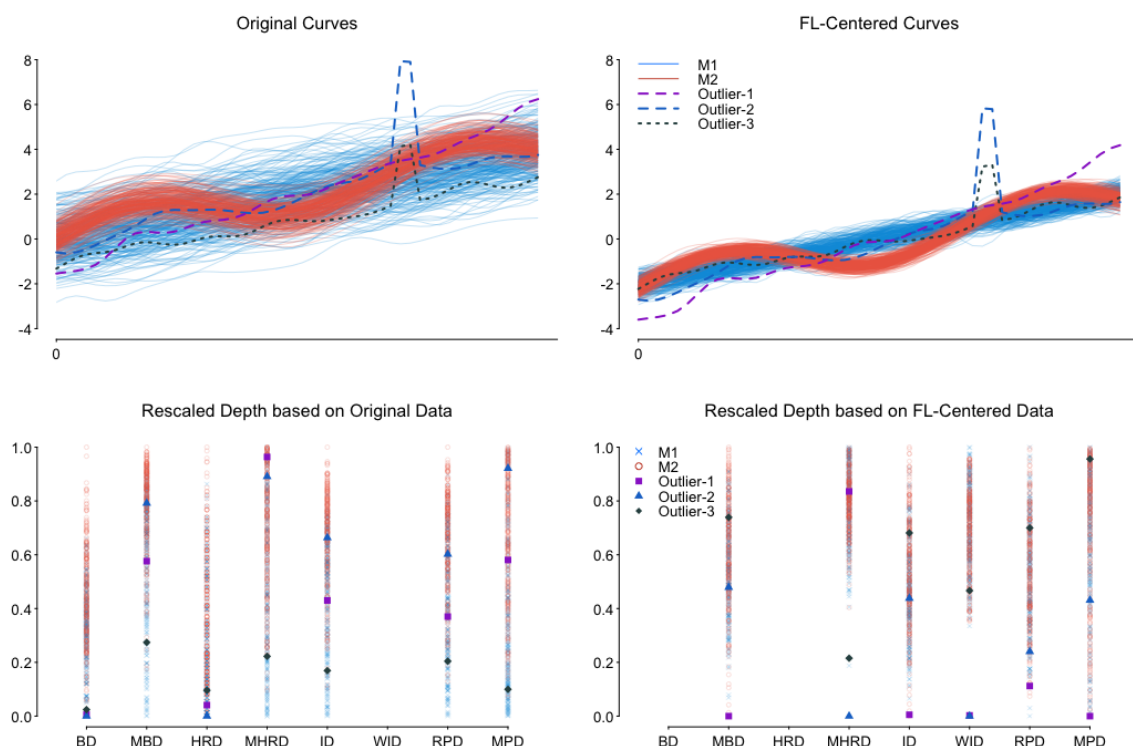


Fig. 4.1 Top: the complete set of 503 simulated curves (top left) and the corresponding centered curves (top right): 250 sampled from (M1), 250 from (M2) and 3 outlier curves. Bottom: scaled FDDs of eight methods based on the original curves (bottom left) and on the centered curves (bottom right). Outlier curves are marked with different colors, line and point types.

local spike, as can be expected from their design through integrating a deviation metric over the full support, while WID has a higher sensitivity by integrating a squared deviation metric. MHRD is good at detecting a local spike, even if this is of modest strength. RPD fails in assigning any of the three outliers the lowest depth value.

The advantage of using the centered curves is clear in the preceding. However, centering does neglect the vertical location of the curve, which itself can be a source of the outlyingness. To remedy this one can calculate the univariate depths of the locations $FL(x)$ of the curves. In Figure 4.2 these are plotted against the corresponding FDD values based on the centered curves. If a future curve shows up at the left side of this bivariate plot, then it is likely a shape outlier, while a location outlier will show up at the bottom of the point cloud. Points in the bottom left corner are outlying relative to both aspects.

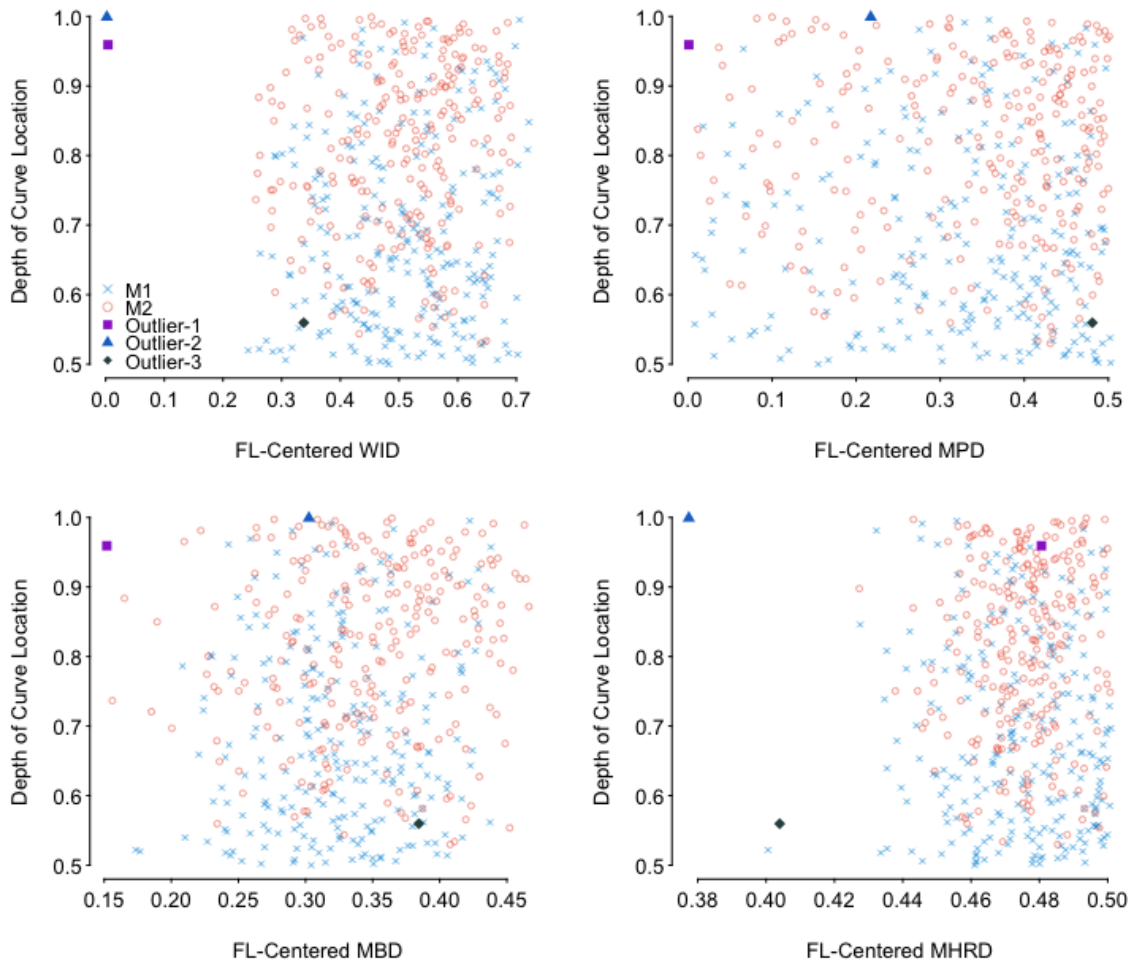


Fig. 4.2 Plot of univariate depth of curve locations versus FDD value of the centered curves. The four panels show the four types of FDD: WID, MPD, MBD and MHRD. In every case the univariate depth is the median depth $1 - |1/2 - F_n(FL(x))|$, where F_n is the empirical cumulative distribution function of the sample of curve locations $LF(x_i)$.

4.4 Topic II - Prediction Sets

In this subsection, we discuss two approaches to build a prediction set for functional data using depth values. In our practice, constructing a prediction set is needed for monitoring a future abnormal curve or assessing comparability of two groups of curves. In the latter application, one may be requested to compare a test group with only 3–5 individuals to a large reference group. A rigorous statistical test might then lack power, but we could build a prediction set based on the reference data, and conclude that the groups are “comparable” if the samples from the test group are within this prediction set.

Practitioners often build a prediction set for functional data by combining pointwise prediction intervals into a prediction band for a curve. A future curve is considered abnormal if any point of its graph falls outside the band. Although this has an intuitive visual interpretation, a drawback is that the coverage of the band may be much smaller than the coverage of the individual intervals, and this may be hard to correct. In addition, this approach neglects curves with abnormal patterns, as long as their values are within the band.

An alternative is to build a prediction set using functional depth values. A first possibility is to use the depth values directly, and define a prediction set as the set of curves with depth higher than some threshold. This is reasonable, since FDD ranks curves in a center-outwards manner, where a higher depth value indicates that a curve is closer to the “center” (and hence more “typical”). Given a FDD, it suffices to determine the threshold to ensure a desired coverage. The sensitivity of such a prediction set to different types of abnormality depends on the chosen FDD and how it is calculated (based on the original curves or the centered curves). For instance, in Section 4.3 Outlier-2 and 3 were seen to be within the 95%, or even 90%, prediction sets for most choice of data-depth (see Figure 4.1).

Another approach is to project the curves to a set of basis functions, and to build a prediction set based on the multivariate depth measure of the projection coefficients. We shall use the (functional) eigenbasis of the reference sample, and likelihood depth based on a Bayesian density estimator for the distribution of the coefficient vectors, as described in Section 3.6, to build a prediction set. Since an outlier curve may be badly approximated by the eigenfunctions, but still have similar projection coefficients, we also perform a preliminary test on the distance between the curve and its projection. If this distance is larger for the new curve than for any of the reference curves, or exceeds a large quantile of the latter distances, then the new curve is classified as an outlier. By using a very large quantile, this test will be stringent, and the preliminary check will hardly change the coverage of the prediction set.

The eigenfunctions $\xi_1, \xi_2, \dots$, or ‘functional principal components’, of a sample $x_1, \dots, x_n$ of curves sequentially maximize the sum of squares $\sum_i v_{il}^2$ of the loadings $v_{il} = \int_a^b (x_i(t) - \bar{x}(t)) \xi_l(t) dt$ of the centered x_i on ξ_l , under the constraints that $\int_a^b \xi_l(t)^2 dt = 1$ and $\int_a^b \xi_l(t) \xi_j(t) dt = 0$, for $l > j$.

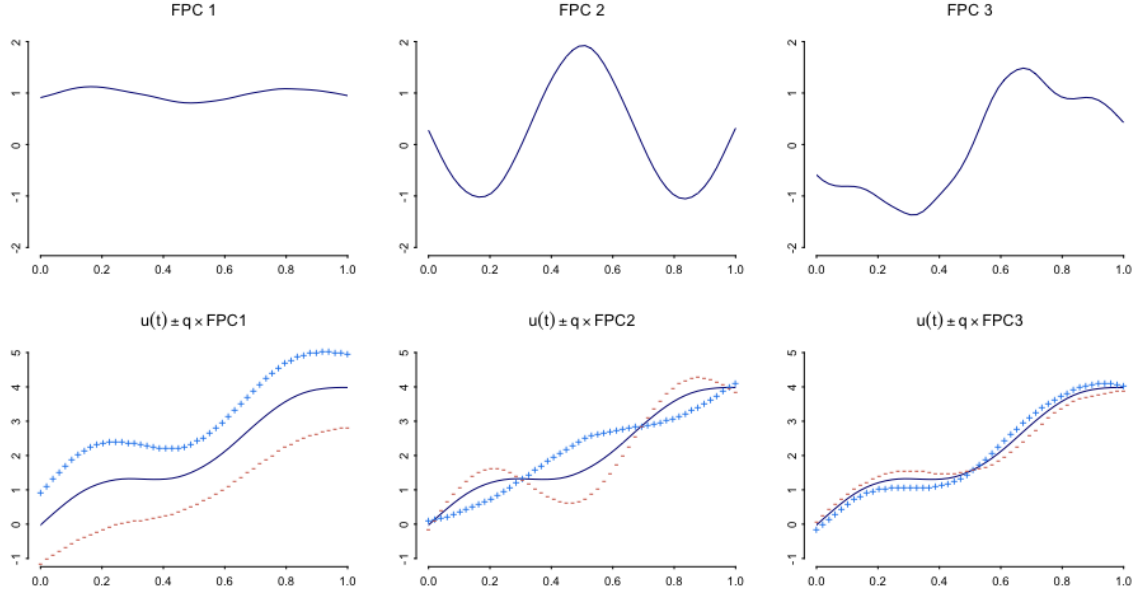


Fig. 4.3 Top row: first three eigen-functions ξ_1, ξ_2, ξ_3 corresponding to the centered set of 500 reference curves. Bottom row: average curve $u(t) = \bar{x}(t)$ (solid line), and average curve $+q_{upp} \times \xi_l(t)$ (line made of '+') and $-q_{low} \times \xi_l(t)$ (line made of '-'), where q_{upp} is the 0.95 quantile of the coefficients $\{v_{il}\}_{i=1}^n$ and q_{low} is the 0.05 quantile, for $l \in \{1, 2, 3\}$.

Given the first L eigen-functions, a sample curve x_i can be approximated by

$$\hat{x}_i(t) = \bar{x}(t) + \sum_{l=1}^L v_{il} \xi_l(t) \quad (4.13)$$

A new function can be similarly projected onto the span of $\xi_1, \dots, \xi_L$. In practice the functions will be discretized on a grid and the integrals will be approximated by sums. The number L of basis functions will typically be chosen much smaller than the number of grid points. Details on functional principal component analysis can be found in Chapter 8 of [45]. We use the R package `{fda.usc, version 2.0.1}` for FPCA in this study.

We implemented this method on the dataset generated in Section 4.3, consisting of a reference sample of 500 curves and three outlying curves, visualized in the top-left panel of Figure 4.1. We decided to retain the first three basis functions, which explained 94% of the total variation in the 500 reference curves. They are shown in the top panels of Figure 4.3, and superimposed as deviations on the mean function $u = \bar{x}$ in the bottom panels of the same figure.

The coefficients of the three outlying curves x_{n+1} were calculated as $v_{n+1,l} = \int_a^b (x_{n+1}(t) - \bar{x}(t)) \xi_l(t) dt$, and their projections similarly to (4.13) (shown in Figure 4.4). We calculated the square distances between all x_i and their projections $\hat{x}_i$ as $\Delta_i = \int_a^b (x_i(t) - \hat{x}_i(t))^2 dt$. For the reference sample

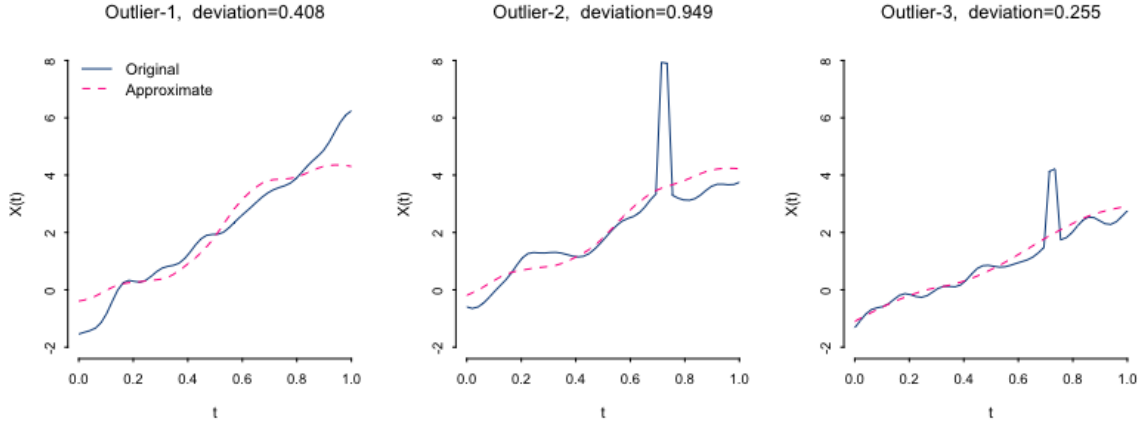


Fig. 4.4 The three outlier curves (solid lines) and their approximation based on $\{\xi_l\}_{l=1}^3$ (visualized in Figure 4.3). Deviations Δ between the observed and the approximated are shown in the title.

these ranged from 0.002–0.184, while for the three outlying curves these values were 0.408, 0.949 and 0.255, all exceeding 0.184.

We computed the likelihood data depth based on fitting a Bayesian nonparametric density estimator to the vectors $\{(v_{i1}, v_{i2}, v_{i3})\}_{i=1}^n$ of coefficients relative to the first three eigenfunctions. This density estimator assumes a Dirichlet process mixture of Gaussians as a prior. We used the SUGS algorithm, described in Section 3.6, to approximate the posterior mean density, with the prior hyper-parameters set to: $\kappa_0 = 0.1$, $v_0 = 7$, and μ_0 and Λ_0 set to be the sample mean and the inverse of the sample covariance matrix of $\{(v_{i1}, v_{i2}, v_{i3})\}_{i=1}^n$. We used 50 random permutations of the input sequence for SUGS, and selected the sequence with the highest predictive likelihood to perform the final fit. Given a sequence of input data, the predictive likelihood is the log-likelihood of the last 20% of the data based on the density estimator trained on the first 80% of the data via SUGS. The prediction set was then set equal to functions with a likelihood depth greater than the threshold set to achieve Bayesian $1 - \delta$ -expectation coverage, as in (3.12). This threshold was 1.8×10^{-5} for $\delta = 0.01$ (99% coverage), and 6.5×10^{-5} for $\delta = 0.05$. The likelihood depths of the three outlying curves were 2.9×10^{-11} for Outlier-1 (global deviation), 2.0×10^{-6} for Outlier-2 (sharp local spike), and 6.8×10^{-4} for Outlier 3 (modest local spike). Hence only the Outlier-3 sits in the prediction set.

We confirmed the coverage of the prediction set in a simulation study, using the mixture of Gaussian curves (M1) and (M2) with weights $p_{M_1} = 0.5$, described in Section 4.3, as population model. We generated a validation set of 5000 curves from this population, and checked inclusion in the prediction sets computed from 100 different samples of 500 curves. For each of the 100 samples we computed the first three eigenfunctions, the likelihood depth of the 500 coefficient vectors, and the lower thresholds of the prediction sets for 90%, 95% and 99% coverage. We found that proportions of 0.9019, 0.9486 and 0.9889 of the 5000 validation curves belonged to the prediction sets, all of which are very close to their nominal values.

In all calculations curves were discretized to a grid of 50 equally spaced nodes in $[0, 1]$. The first three eigenfunctions explained between 93-95% of the total variation across the 100 simulated datasets.

4.5 Topic III - Two Functional Sample Testing

Assume given a reference group $\{x_i\}_{i=1}^{n_x}$ of n_x curves and a test group $\{y_j\}_{j=1}^{n_y}$ of n_y curves, assumed to be *i.i.d.* samples from distributions with mean curves μ_X and μ_Y , respectively. An ubiquitous question in practice is whether the two distributions are equal or not. Here the curves may be preprocessed, for instance centered, if it is desired to focus on a difference of particular patterns between the two groups.

We discuss three approaches to conduct two-sample testing with functional data: a rank test from [32], a permutation test, and a bootstrap test. We illustrate the tests on growth curves of groups of boys and girls, originally from [45], used in [32], and compare the results to [32]. The motivation for the new tests is that the rank test proposed in [32] requires to split the reference sample in a seemingly arbitrary way.

4.5.1 Rank Test

The rank test proposed by [32] starts by randomly dividing the reference group into two parts, a ‘baseline’ sample $Z = \{z_k\}_{k=1}^{n_0} \subset \{x_i\}_{i=1}^{n_x}$ and a ‘reference’ sample consisting of the remaining $n_x - n_0$ curves x_i . It is advised to choose n_0 not too small, for instance $n_0 = \max(n_x/2, n_y)$.

For a given functional depth measure $D(x, P)$ and P_{n_0} the empirical distribution of Z , a score of a function x is defined as $\sum_k \mathbf{1}_{\{D_n(z_k, P_{n_0}) \leq D_n(x, P_{n_0})\}}$. The scores of all functions in the pooled sample of $n_x - n_0$ remaining ‘non-baseline’ functions x_i and the n_y functions y_j are computed, ordered, and $R(x_i)$ and $R(y_j)$ are set to be their ranks in this pooled sample. If there are ties among the scores, the rank is defined as the midpoint of the unadjusted ranks, as usual. The proposed test statistic W is the sum of the ranks $\{R(y_j)\}_{j=1}^{n_y}$ of the test group.

The scores are a measure of centrality relative to the distribution P_{n_0} , which should be close to the distribution of the x_i . Thus we may expect that the scores in the test group will be smaller, and hence their ranks in the pooled sample to be smaller as well. Thus the null hypothesis that there is no difference between the reference and test group is rejected for small values of W .

Under the null hypothesis that there is no difference, the ranks in the pooled sample should be a uniform permutation of the numbers $\{1, 2, \dots, n_x + n_y - n_0\}$, adjusted for ties if there are any. Thus the test statistic W has the same null distribution as the sum of a random sample of size n_y without replacement drawn from $\{1, \dots, n_x + n_y - n_0\}$, adjusted for ties if there are any.

4.5.2 Permutation Test

Given any two-sample test statistic, a permutation test compares the value of the statistic on the two observed samples to the values on all pairs of samples obtained by splitting the pooled sample $x_1, \dots, x_{n_x}, y_1, \dots, y_{n_y}$ arbitrarily in groups of sizes n_x and n_y . If the sample sizes are large, one often uses a randomly chosen set of splits rather than all splits, together with the original split. Because under the null hypothesis all splits are equally likely to give the original grouping, the P-value of the permutation test can be computed as the proportion of splits giving a larger value of the test statistic than the observed value.

If the test statistic is a distance $\Delta(\hat{\mu}_X, \hat{\mu}_Y)$ between estimated mean functions in the two samples, then this P-value becomes $\sum_{b=1}^B \mathbf{1}_{\{\Delta(\hat{\mu}_X^{(b)*}, \hat{\mu}_Y^{(b)*}) \geq \Delta(\hat{\mu}_X, \hat{\mu}_Y)\}} / B$, where b ranges over all splits. The null hypothesis is rejected when this is smaller than the significance level. The number of possible splits is $\binom{n_x+n_y}{n_x}$. If this is too large, then B will be chosen a preset value and a random selection of splits will be used.

An alternative test statistic is the centrality of the mean difference curve $\hat{\mu}_X - \hat{\mu}_Y$ within an estimate of the distribution of this difference under the null hypothesis. For $\{\hat{\mu}_X^{(b)*} - \hat{\mu}_Y^{(b)*}\}_{b=1}^B$ the mean difference curves after B reassignments of the group labels, the latter distribution can be estimated as the empirical distribution F_B of these difference curves. The test statistic on the original sample then becomes $D(\hat{\mu}_X - \hat{\mu}_Y, F_{B+1})$ and the P-value is evaluated as $\sum_{b=1}^B \mathbf{1}_{\{D(\hat{\mu}_X^{(b)*} - \hat{\mu}_Y^{(b)*}, F_B) \leq D(\hat{\mu}_X - \hat{\mu}_Y, F_B)\}} / B$.

4.5.3 Bootstrap Test

The bootstrap procedure is the same as the permutation procedure, except that the new samples are now formed by drawing n_x and n_y curves with replacement from the pooled sample of curves.

When the sample sizes n_x and n_y of the two groups are highly unbalanced, some caution is necessary. For instance, the chance of redrawing all curves from the testing group is $(\frac{n_x}{n_x+n_y})^{n_x+n_y}$, which can be larger than desired if $n_y/(n_x+n_y)$ is extremely small.

4.5.4 Example Analysis

The example dataset, shown in Figure 4.5, consists of the growth curves of 39 boys and 54 girls, In [32] the P-value of the rank testing procedure was found to be 0.0001 with BD, 0.1199 with MBD, and 0.1636 with ID. We applied the permutation and bootstrap test with $B = 5000$ and test statistics (1) **L1-Delta** $\Delta(x, y) = \int_a^b |x(t) - y(t)| dt$, (2) **L2-Delta** $\Delta(x, y) = \int_a^b (x(t) - y(t))^2 dt$, (3) MBD, or (4) ID. We found that BD was too computationally intensive to be calculated at such large sample size.

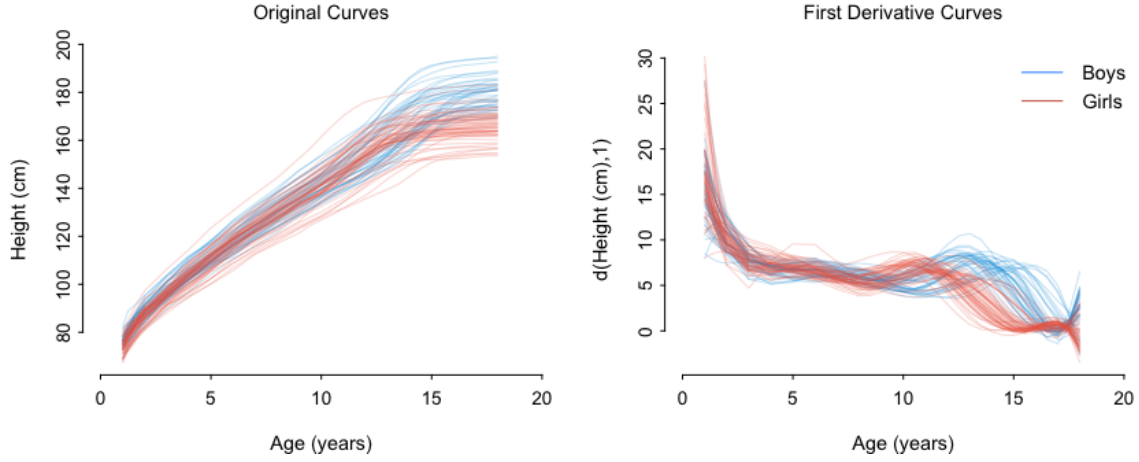


Fig. 4.5 Growth curves (left subplot) of boys (in blue) and girls (in red), and their 1st derivative curves (right subplot).

Table 4.1 P-values of the different test procedures with the growth curve dataset

Setting	Rank Test ¹	Permutation	Bootstrap
L1-Delta	-	0.0010	0.0018
L2-Delta	-	<0.0001	<0.0001
MBD	0.1199	0.0836	0.0744
ID	0.1636	0.1098	0.0984

The results are summarized in Table 4.1. For illustration the redrawn curves $\{\hat{\mu}_X^{(b)*} - \hat{\mu}_Y^{(b)*}\}_{b=1}^{B+1}$ are shown in Figure 4.6.

Figure 4.5 shows that there is a clear difference between the curves of the two groups, with a change in pattern starting around age 12. The pointwise difference curve $\hat{\mu}_{boys} - \hat{\mu}_{girls}$ is also clearly different from the difference curves simulated by the permutation or bootstrap procedure, shown in Figure 4.6. Permutation and Bootstrap tests with L1 and L2-Delta all return very small p-values. However, MBD and ID are not sensitive enough to flag this obvious deviation, due to their global definition, while the deviation is local. Actually, if we perform the two-sample testing based on the first derivative of the original curves (shown in the right subplot of Figure 4.5), all instances of the permutation and the bootstrap tests in Table 4.1 return p-values <0.0001. This is explained by the fact that the first derivative curves of girls starts to show different pattern already around age 9 comparing to boys, while the full range of observation is from age 1 to 18.

As a visual instrument, one may build a prediction band based on the data from the girls, and see starting from what age the growth curves of the boy increase outside the band (see Figure 4.7 for

¹results are cited from [32]

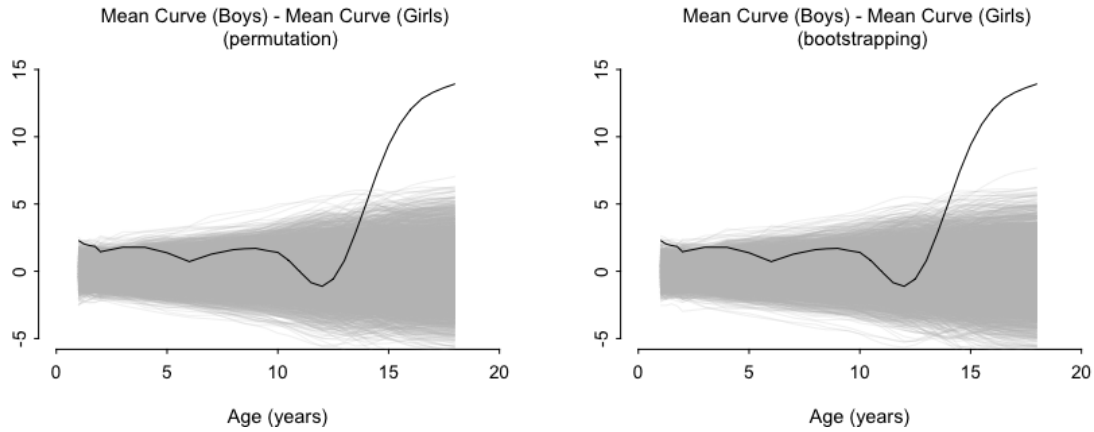


Fig. 4.6 Pointwise difference curve $\hat{\mu}_{boys} - \hat{\mu}_{girls}$ between original mean curves of boys and girls (black line), and 5000 simulated difference mean curve $\hat{\mu}_{boys}^{(b)*} - \hat{\mu}_{girls}^{(b)*}$ with the permutation procedure (grey lines in the left subplot) or the bootstrap procedure (grey lines in the right subplot).

illustration). We constructed such a prediction band by performing FPCA on the observed growth curves $\mathbf{X} = \{x_i(t)\}_{i=1}^{54}$, $t \in [1, 18]$ for girls, retaining the first 7 eigen-functions, which explained 99.5% of the total variation. We simulated 1000 vectors v_{bl}^* from the multivariate normal distribution with mean and covariance matrix being the sample mean and covariance matrix of the loading vectors of the 54 girl's curves, and then calculated 1000 curves via $x_b^*(t) = u(t) + \sum_l v_{bl}^* \xi_l(t)$, where u is the mean of the girl's curves and ξ_l are the eigenfunctions. Determine a 95% prediction set S_{95}^* based on $\{v_b^*\}_{b=1}^{1000}$ and their likelihood depths, and denote $B_{95} = \{b: v_b^* \in S_{95}^*\}$. The representative band is $\{(t, y): t \in [1, 18], \min_{b \in B_{95}}(x_b^*(t)) \leq y \leq \max_{b \in B_{95}}(x_b^*(t))\}$.

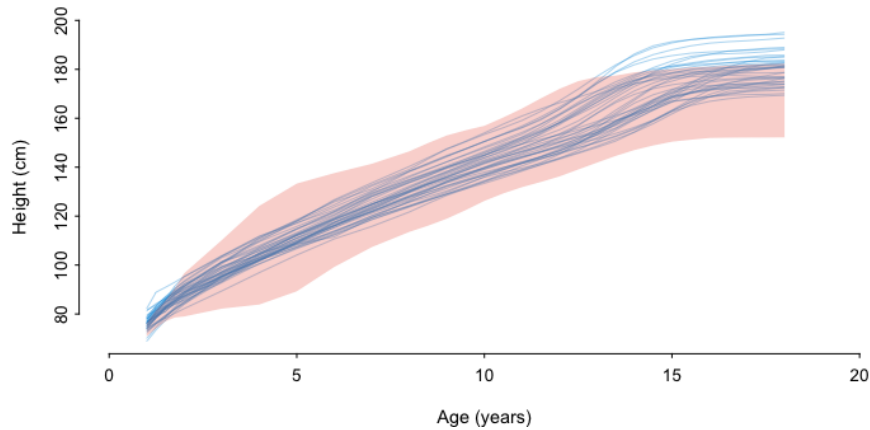


Fig. 4.7 Growth curves of the boys (in blue) and the prediction band (in red) of the girls, the band intended to cover at least 95% of the girl population.

Chapter 5

Application: Image-Based Filter for Silicon Oil Particles¹

5.1 Introduction

The objective of this chapter is to classify images as showing either a silicon oil (SO) or a non-silicon oil (NSO) particle. Silicon oil particles have a clear profile, but non-silicon oil particles may have many, also unknown forms. For this reason our chosen method of classification is based on agreement of a future image with a data-base of images of SO particles, rather than on training a classifier on samples of SO and NSO particles. In machine learning terms our approach is unsupervised, rather than supervised. We measure the agreement by data depth, whence this chapter also presents an application of data depth as described in the preceding chapter.

The images of interest are acquired from *flow imaging microscopy* (FIM), more precisely Micro-Flow Imaging (MFI) 5200 in this study, which is prevalent in the pharmaceutical industry. Their resolution varies by the total number of pixels, and by the average and range of brightness across the pixels. Since our proposed SO filter/classifier is image based, rather than on the metadata of the images, we expect that the methods developed here will lend themselves also to other FIM systems with an output of 2D images in gray scale, such as FlowCAM.

The industrial background of the application is described in Section 5.2. In Section 5.3 we review the state-of-the-art classification methods (or ‘filters’) currently used in the industry. These are all

¹This chapter is based on joint work with Björn Boll and Migle Grauzinyte in Novartis Global Drug Development. A significant part of the chapter is reworked into a paper, "Applying pattern recognition as a robust approach for silicone oil droplet identification in flow-microscopy images of protein formulations", that has been accepted and published by Journal of Pharmaceutical Sciences, DOI:<https://doi.org/10.1016/j.xphs.2020.10.044>.

based on metadata of the images, in contrast to our new approach, which takes the full images as inputs. Section 5.4 presents the data used, and the methods used to prepare the raw images for analysis. The construction of our SO classifier is described in Section 5.5.

Section 5.6 compares the new filter against three benchmark approaches (summarized in Section 5.3) on three groups of test data, consisting of images of particles that are clearly SO (I), likely SO (II), or clearly NSO (III). The selection of these test datasets is motivated at the beginning of the section. Evaluation criteria are the false negative rate for datasets (I) and (II), and the false positive rate for dataset (III). The newly proposed filter is overall an improvement on the benchmark approaches. Detailed conclusions and thoughts for future improvements are provided in Sections 5.7 and 5.8.

5.2 Industrial Background

Injections and parenteral infusions are typically contaminated by extraneous, mobile and undissolved particulates, which are unintentionally present in the solutions [39]. Sources of contamination in sterile and/or biologic products include the solution itself, its ingredients, the manufacturing process, environmental factors, packaging, and the preparation of the product for administration. See [55] for discussion and further sources.

Particles can be categorized by size, as visible ($>100\ \mu\text{m}$), micron ($1\text{--}100\ \mu\text{m}$), sub-micron ($100\text{--}1000\ \text{nm}$) and nanometer ($<100\ \text{nm}$) particles [65]. In a vaguer sense, all particles of size $<100\ \mu\text{m}$ can be called subvisible. Currently the US and European pharmacopoeia only have clear requirements on the number of subvisible particles of sizes ≥ 10 and of sizes $\geq 25\ \mu\text{m}$, which are usually quantified by light obscuration in the parenteral solution. However, regulatory changes are expected ([15] and USP<1787>), spurring industrial interest in the quantification of subvisible particles with techniques that are orthogonal to light obscuration. A recent option is *flow imaging microscopy* (FIM), which uses a camera with high magnification power to capture images of a liquid solution passing through a thin flow cell. A detailed account of FIM, with a comparative evaluation of several instruments, is offered in [70].

It is presently not uncommon to see FIM applied in quality control of protein therapeutics, which takes an increasing share in the product portfolio of many pharmaceutical companies. Distinguishing proteinaceous from other particles is important for assessing stability and for controlling other potential risk factors (see [52]). However, the irregular shape of proteinaceous particles and their aggregation make this task genuinely difficult. As a practical work-around, one tends to identify non-protein particles of known shape, and counts the remaining observed particles as aggregated protein.

A routine task is to identify silicon oil (SO) particles and distinguish them from all other, non-silicon oil (NSO) particles. SO particles are ball-shaped and are commonly present in prefilled syringes, cartridges or vials, as lubrication for the glass barrel, the plunger or the stopper. Because they look spherical from any viewing angle, classifying SO particles is considered feasible even when

common FIM only captures 2D projections. Air bubbles are among the few kinds of entities, if not the only one, that also have spherical projections and can occur in the pharmaceutical products under consideration. However, their presence is avoided or minimized by sample preparation.

There are many types of possible NSO particles, which would make it difficult to compile a representative data-base of such particles for training purposes. On the other hand, images of SO particles are very similar, in particular after some pre-processing. This makes it attractive to classify a future image by similarity to SO pictures, rather than train a supervised classifier, which would necessarily have to be trained on a non-representative set of example images.

From an industrial background preventing false positives (unjustly labeling particles as SO) is more serious than false negatives (labelling SO as NSO). This is because SO particles are not immunogenic, in contrast to some NSO particles.

In the industrial context classification algorithms are also referred to as *filters*.

5.3 Three Benchmark Approaches

The standard output from the MFI system (and similar systems) includes images and their corresponding metadata. The images consist of the intact picture frames taken in a defined frequency and sub-frames of “individual” particles. The sub-frames are captured in the intact frames by some generic algorithm in the MFI system. Below we take this capturing as given and refer to the sub-frames when using the word “images”, unless specified otherwise. The metadata contains features of each sub-frame, including morphological (e.g. aspect ratio, perimeter and circularity) and optical aspects (e.g. intensity mean, SD, min and max across pixels), together with some descriptive information (e.g. from which intact frame, ECD, area, coordinates of the sub-frame in an intact frame).

Existing filters of SO particles ([53],[46]) are all based on the metadata. We discuss two industry standards, and a filter based on random forests.

Filter 1 (Benchmark-SRD). *Strehl et al. [53] propose a filter based on five attributes from the metadata. They first calculate an ‘S-factor’ as the product*

$$\mathbf{S\text{-}factor} = \text{Circularity} \times \text{AspectRatio} \times \text{IntensityMax} \times \text{IntensitySD}.$$

They next determine a cut-off for the S-factor that depends on the particle size (ECD). But since their original work were for the MFI 4100 system, we use the following cut-offs that are suited for the MFI

5200 system, updated by as Torres et al. [47].

$$\text{Cut-Off}(ECD) = \begin{cases} 29732 \times \ln(ECD) + 3913, & 1.125 \leq ECD < 2.875 \\ 46609 \times \ln(ECD) - 10464, & 2.875 \leq ECD < 6.125 \\ 3368 \times ECD + 49124, & ECD \geq 6.125 \end{cases}$$

If the S-factor of a sub-frame is larger than the Cut-Off value, then the image is classified as "SO", and otherwise as "NSO".

Filter 2 (Benchmark-AR). A sub-frame with Aspect ratio (AR) ≥ 0.85 is classified as "SO", and otherwise as "NSO". This simple, empirical rule is sometimes used in practice.

Filter 3 (Benchmark-RF). In [46] the authors propose to use a random forest classifier based on the metadata. As their training data are not accessible and it may be worth while to tune the random forest to typical data in our domain, we fitted a random forest classifier to our own dataset. We compiled the metadata of samples of 1990 NSO and 2316 SO particles, including the 10 predictors: ECD, Area, Perimeter, Circularity, Aspect Ratio, Maximum Feret Diameter, Intensity mean, STD, min and max from MFI 5200. The sample of SO particles included carefully selected SO particles and data from experiments measuring only SO particles. The sample of NSO particles was from a study where the presence of SO was minor and experts had carefully excluded the SO particles. We fitted a random forest model by the function `randomForest()` from the R package `{randomForest, 4.6-14}`, using the default setup, with the maximum number of terminal nodes (maxnodes) as a tuning parameter. The value 4 for maxnodes was selected via a cross-validation procedure, to minimize the chance of overfitting (see Figure 5.1). The prediction accuracy was 99.6% to the whole training set. Two important features in the fitted model, as measured by the mean decrease in Gini index when dropping the predictor, were the maximum intensity across all pixels and the aspect ratio.

5.4 Data and Image Pre-processing

Even though SO images acquired by the MFI system share some key feature, they also display visible differences. Inspection of a set of images (see Appendix 1) reveals that, within their spherical shape, SO particles display an inner white spot. This is due to the reflection of camera light. The position of this spot depends on the position of the SO particle in the intact picture frame. Its size varies with the size of the particle, and its shape is not necessarily spherical. When the white spot is close to the edge of the spherical shape and the resolution of the image is low (which is typical for small particles), the spherical shape may be gently breached. The spot can be very small or blur into the background, in particular in the images of small particles. Thus small SO particles tend to show different patterns than bigger SO particles.

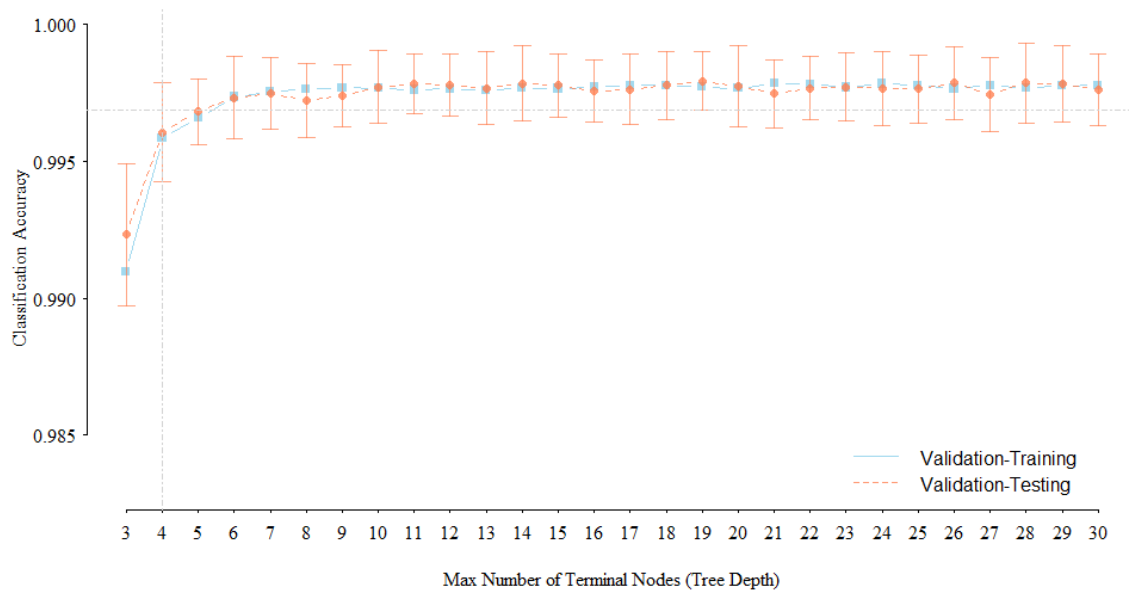


Fig. 5.1 Cross-validation curve to determine the maximum size of terminal nodes in benchmark filter 3. The data were randomly split 70:30 % in training and testing samples, with 50 repeats at each candidate setting of the variable *maxnodes* (shown on the horizontal axis). The vertical axis gives the proportion of correctly classified samples, with a blue square indicating the average in the training set, a red dot the average in the test set with red bar indicating the standard deviation across 50 repeats. The horizontal gray dashed line marks the lower end of the red bar at the candidate setting with the highest average accuracy. The selected value 4 is the minimum integer for which the red bar contains the horizontal gray dashed line.

Algorithm 6: Pre-processing of raw images to a square image**Data:** A $a \times b$ image/matrix X and an integer h .

- 1 Rescale the values in image X to improve the contrast:
 $\forall j, k, \quad X_{jk} := (X_{jk} - \min(\{X_{jk}\}_{j,k}) / (\max(\{X_{jk}\}_{j,k}) - \min(\{X_{jk}\}_{j,k})))$;
- 2 Chop off empty edges: for every row and column, determine its average and range length (maximum-minimum), chop off the row(s) and column(s) if the average is > 0.78 and the range length is < 0.28 ;
- 3 Repeat step 1-2 one more time, resulting in a $a' \times b'$ matrix;
- 4 Pad the image close to a square before resizing to avoid pattern distortion: If $|a' - b'| \leq 1$, no padding is needed. If $a' > b' + 1$, pad $\lfloor |a' - b'|/2 \rfloor$ blank columns to each of the two sides of the matrix. If $b' > a' + 1$, pad $\lfloor |a' - b'|/2 \rfloor$ blank rows to each of the two sides of the matrix. A blank column/row has elements equal to the average of the maximum of the first column/row and the maximum of the last column/row;
- 5 Resize the matrix to an $h \times h$ matrix by linear interpolation, to ensure comparability across images of different sizes: `resize()` from R package {EBImage, 4.22.1} was used in this chapter;

Result: An $h \times h$ matrix

Features such as these are not captured in the metadata of the images, which motivates to build a filter based on the complete images. To train our image-based (IB) filter, we selected 798 SO images from the MFI system, aiming to include a representative set to cover the variation in the aforementioned features (e.g. images of big and small sizes, the inner white spot having various positions, shapes and sizes, the spherical hull gently breached by the white spot).

An image of $a \times b$ pixels can be viewed as an $a \times b$ matrix of numbers with values between 0 and 1, where 0 indicates pure black, 1 indicates pure white, and intermediate values correspond to gray with different degrees of hue. We pre-processed the images to cut off empty edges, thus centering the SO particle in the images, and to correct for the varying size and contrast and of the images. The procedure is illustrated in Figure 5.2 and described precisely in Algorithm 6. This pre-processing is meant to be consistently applied to all images, also future ones, and hence is part of the filter constructed here.

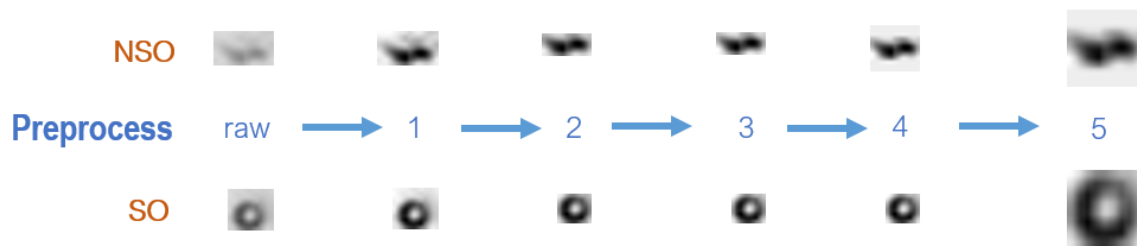


Fig. 5.2 Illustration of the pre-processing procedure for images

As a convention, images of particles of sizes smaller than $5 \mu m$ (as determined by the variable ECD in the metadata of the image) are not processed by any SO filters. Their low resolution prevents

making a confident judgement, also by human experts after close scrutiny. Images of particle of sizes exceeding $5\mu m$ usually have at least 7 rows or columns of pixels.

5.5 Construction of Image-Based (IB) Filter

Our objective is to build an accurate binary filter (classifier) for SO/NSO assignment using full images as inputs. We approach this from the perspective of unsupervised learning, because images of NSO particles, in contrast to SO particles, do not have a clear description and can take infinitely many forms. Therefore, it would be difficult to form a good training set, and a filter based on supervised learning would not be able to generalize its success against NSO stereotypes that are not seen in the training data.

On the other hand, SO particles have fairly predictable 2D projections from any angle, making it possible to use similarity of a new image to images in the training set the basis for classification.

Our filter consists of two sub-filters (or ‘tests’), each of which outputs a binary label SO/NSO. A future image will be labeled as SO if either or both tests label it as SO. The first filter uses data depth to assess similarity of a future image to a set of training SO images. The data depth is computed on features of the images extracted by a principal component analysis. The second filter is more crude and checks a pixel-wise distance of a future image to set of SO stereotypes in a reference book of nonstandard SO images.

We found that the first filter rarely misclassified an image as NSO while it was SO. On the other hand, the second filter was good at recognizing some nonstandard SO images, which were missed by the first filter. Details on how each test was constructed are presented in the following subsections.

5.5.1 IB-Test 1: General SO Pattern Recognition

After pre-processing, an image is represented as a 20×20 matrix, and can be identified with a vector in 400-dimensional space. The first filter is based on measuring the data depth of a future vector in the distribution of a typical vector obtained from an image of a SO particle. Data depth enables ranking of multivariate data in a center-outwards manner, low depth indicating an outlier, and can be used as a measure of similarity. In the present case lower depth means that the future image does not represent a SO particle. See Section 3.3 for a detailed description on depth and its usage in building nonparametric prediction sets.

To reduce noise, the 400-dimensional image vector is first projected on the linear span of the first 20 principal components of the distribution of the SO particles. If the projection of a future image is very different from the original image, then this is a first indication that it does not represent a SO particle (see Figure 5.3). We used the 99.5% percentile of the distribution of the distance between a typical vector obtained from an image of a SO image and its projection as a cut off for future images.

The loadings of the 400-dimensional image vector on the first 20 principal components is a vector in 20-dimensional space. The distribution of a typical vector obtained from an image of a SO image was approximated by a mixture distribution, and data-depth of a future vector is computed relative to all mixture components separately. The maximum data depth over the 30 components is taken as the final similarity measure. The maximum is taken, because the component weights resulting from the training set are arbitrary, while a future image should be labeled as "SO" if it is close to any SO stereotype. As a cut off for classification of a future image as SO, we chose again the 99.5% quantile of the similarity measure in the training data.

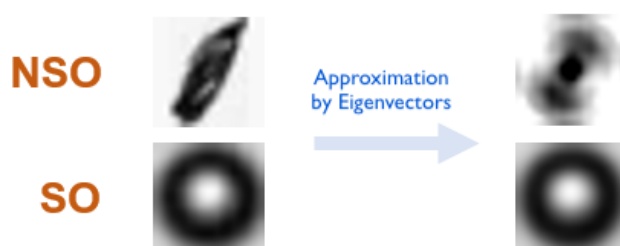


Fig. 5.3 Illustration of Bad (top row) and Good (bottom row) approximation by extracted image features.

We next elaborate on these steps.

The basic training set included 798 SO images, of various sizes. After pre-processing, each image was represented by a 20×20 matrix, with entries between 0 and 1. The training set was augmented by rotating every matrix by 90, 180 and 270 degrees, giving a total of 3192 matrices. Each matrix was vectorized by sequentially concatenating its column vectors. The resulting 3192 vectors were joined as rows in a data matrix X of dimensions 3192×400 .

Principal component analysis (PCA) was performed using the covariance matrix $X^T X$ of the vectorized images, yielding (orthogonal) eigenvectors $v_1, \dots, v_{400}$ (or 'image features'), scaled to unit length ($\|v_i\| = 1$), and corresponding eigenvalues $\lambda_1 \geq \dots \geq \lambda_{400}$. The first 20 eigenvectors, which preserved 98% of the total variation, were retained, and are visualized in Figure 5.4. (In the picture the values in each eigenvector were rescaled to the interval $[0, 1]$, and the vectorization was reverted back to a 20×20 matrix.) The visualization enhances the interpretation of the eigenvectors. For instance, the first eigenvector (top-left of Figure 5.4) captures the spherical shape of the droplet and the inner white spot, the second and the third eigenvectors (the second and the third in the first row of Figure 5.4) encapsulate the different camera lighting angles on the droplet, and the fourth and fifth features represent a diagonal shift of the inner white spot. The decision to retain the first 20 eigenvectors depended, besides on explained variance, on these graphical evaluations. The 21st and onward eigenvectors showed features that are not desired for a typical SO image (e.g. vertical or horizontal elongation of a sphere).

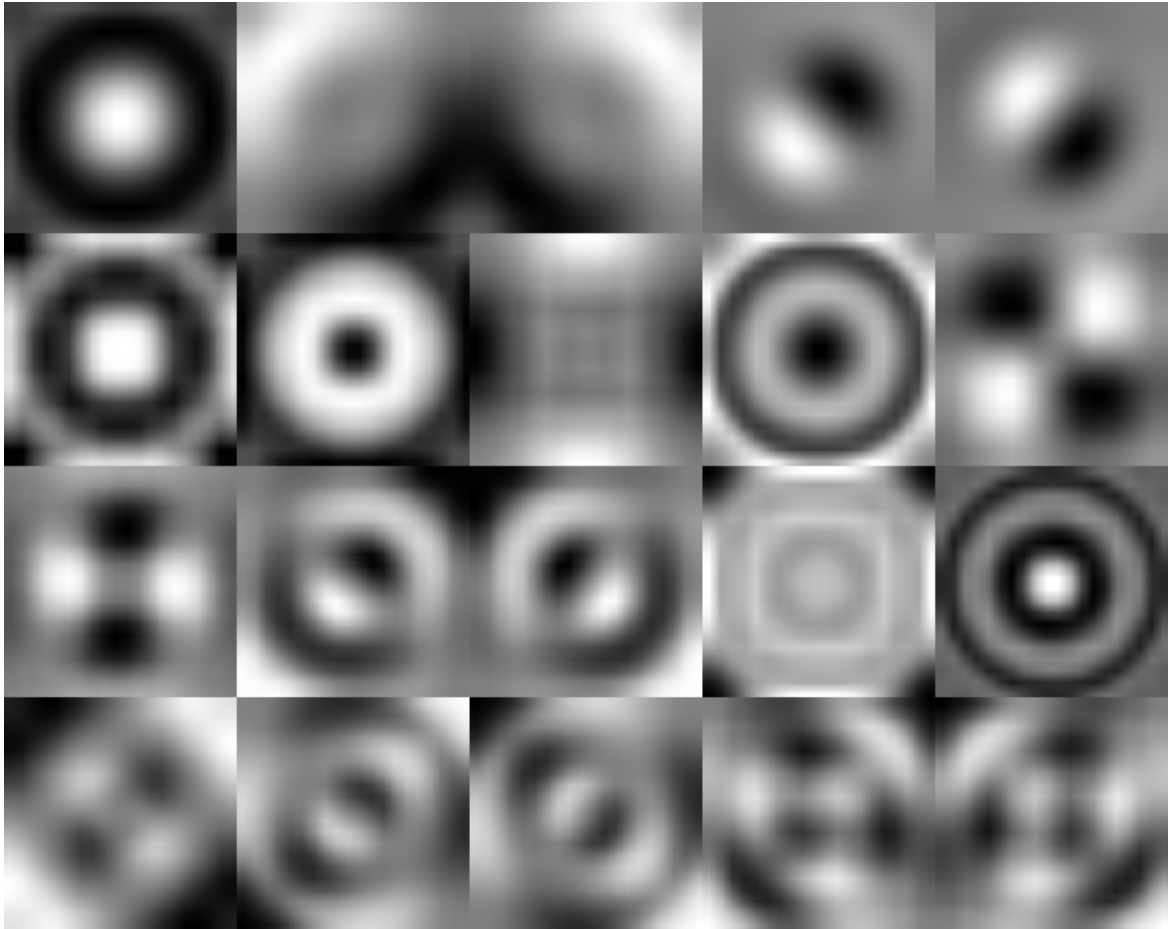


Fig. 5.4 First 20 (rescaled) eigenvectors of the covariance matrix of the 3192 training data (798 pre-processed 20×20 images with 3 augmentations for each via rotation). The first 5 eigenvectors are in the first row from left to right, the second 5 in the second row, and so on.

For $V_k = [v_1, \dots, v_k]$ the $400 \times k$ matrix with as columns the first k eigenvectors, the loading matrix of the original data can be computed as $\Lambda_k = XV_k$. (We used $k = 20$.) The projection of a training image X_j . (1×400 vector, j^{th} row of X) on the first k eigenvectors is given by $\sum_{i=1}^k (X_j \cdot v_i) v_i$, which leads to the matrix $X^* = \Lambda_k V_k^T$ for the projections of all training images relative to the standard basis. The squared Euclidean distanced between the original training images and their projections (the squared distances between the rows of X and X^*) ranged from 0.58 to 2.04, with 99.5% sample percentile approximately equal to 1.4. The latter sample percentile was chosen as the upper threshold for assessing whether a future image is well approximated by its projection on the eigenvectors (and hence similar to SO), or not.

We clustered the loadings (the rows of Λ_{20}) of the training data using the K-Means algorithm, with $K = 30$ clusters. The latter value was based on evaluation of the average patterns across the clusters, with a focus on assessing their variability and avoiding duplication. Intuitively, increasing the number of clusters (i.e. choosing larger K) would enhance the ability to recognize a future SO image, but lower the cardinalities of the clusters and increase the chances of duplicated stereotypes. In the end we determined that 30 clusters are appropriate in this case, leading to cluster sizes between 44 and 163. Trimmed means of the images of the 30 clusters are shown in Figure 5.5.

For each of the 30 clusters, the mean vector and covariance matrix of the loadings of the training data were calculated, and next the 0.5% percentile of Mahalanobis depth was derived via simulation. The details are as follows. Let $c_j \in \{1, \dots, 30\}$ be the member indicators of the training data ($j = 1, \dots, 3192$), and $B_l = \{j: c_j = l, \forall j\}$ be the indices of the clusters. For each cluster, we simulated 5000 data points from the multivariate Gaussian distribution with parameters the sample mean and sample covariance matrix of the vectors $\{\Lambda_j\}_{j \in B_l}$, and then calculated 0.5% percentile of the Mahalanobis data depth of these simulated data relative to $\{\Lambda_j\}_{j \in B_l}$. Taking 0.5% percentile as the lower-sided threshold gives a 99.5% prediction set per cluster, since larger depth means the data point is closer to the center. Across the 30 clusters these 0.5% percentiles range from 0.023 to 0.025, which motivated a lower threshold of 0.02 for the maximum depth over the 30 clusters of a future image.

The projection and loadings of a future image, and its data depth to the clusters, can be computed similarly. The filter will consider the image to be SO if the square Euclidean distance between the image and its projection is small enough and its depth relative to at least one of the clusters is large enough. The two thresholds were set at 1.4 and 0.02, respectively, both reflecting a 99.5 % prediction set based on the training data.

In summary, the first part of our filter is as follows.

IB-Filter: Test 1. *A future image is pre-processed in the manner described in Section 5.4 and vectorized to a 1×400 vector y . Its loadings to the extracted 20 features V_{20} are calculated as $\eta_y = yV_{20}$. The projected image is calculated as $\eta_y V_{20}^T$ and its square distance to the original image is $d_{PC}(y) = (y - \eta_y V_{20}^T)(y - \eta_y V_{20}^T)^T$. The Mahalanobis depths relative to the clusters $l = 1, \dots, 30$*

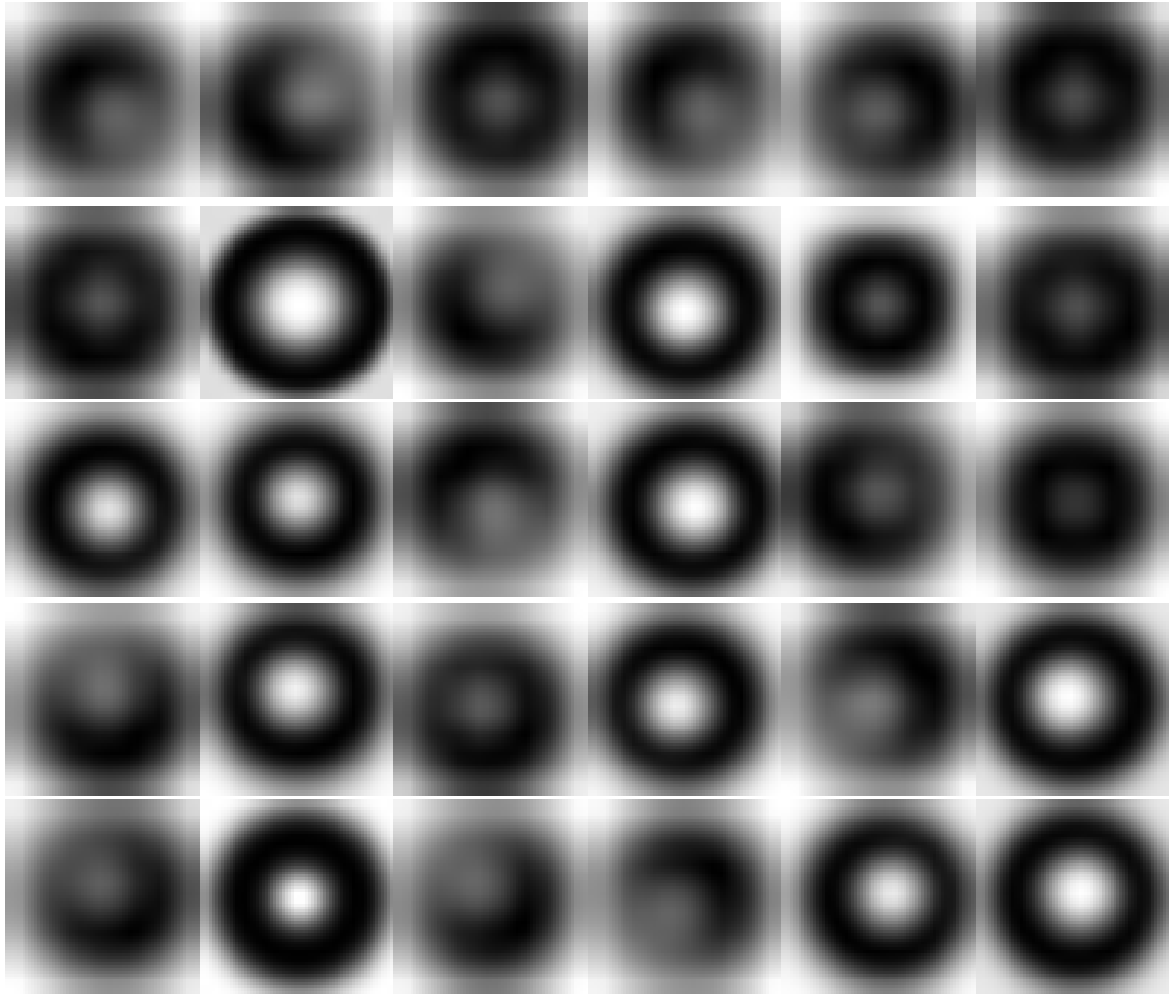


Fig. 5.5 Trimmed means of the 30 clusters obtained via 30-Means applied to the loading matrix. Trimmed means were computed as pixel-wise averages, after discarding the 10% most extreme values.

are $MD_{y,l} = MD(y, \{\Lambda_j\}_{j \in B_l})$, giving the maximum depth $MD_y = \max(MD_{y,1}, \dots, MD_{y,30})$ as the final similarity measure. The future image is considered to be a SO particle if $d_{PC}(y) < 1.4$ and $MD_y > 0.02$.

5.5.2 IB-Test 2: Reference books of SO stereotypes

The second part of the filter is to check similarity of a future image against a reference book of SO stereotypes. The rationale of this test is our expectation that the first part of the filter might neglect some important nonstandard (unpopular) SO patterns (e.g. non-spherical white spot for a modestly large SO image, or a small SO image with gently breached sphere hull).

This test is performed separately on ‘large’ and ‘small’ images, small images being defined as images that originally had fewer than 9 rows or columns. The reason is that small-size SO images tend to show more idiosyncratic patterns than modest or large ones (discussed at the start of Section 5.4), and sizing them up to 20×20 would introduce profound distortion. Instead we resized smaller images to 9×9 , by the same pre-processing procedure discussed in algorithm 6.

Thus two reference books (or ‘galleries’) of SO images were formed, one of 68 images of 20×20 pixels and a second one of 33 images of 9×9 pixels (see Appendix of the chapter). A future image is pre-processed to either 20×20 or 9×9 , based on whether its number of rows and columns is greater than 10 or smaller than 9, and its pixel-wise distance is computed to all images in the corresponding reference book. If the minimum distance is smaller than a threshold, the future image is classified as SO. The distance employed was the L_1 -distance between the pixel values, and the threshold was set to the 99% and 90% percentiles of the distances of the images in the training datasets, for larger and smaller images, respectively.

The reference books were prepared from training datasets, as follows. The training dataset included the same matrix of dimensions 3192×400 containing as rows the vectorized 20×20 images, used to construct the first part of the filter, and in addition a 2060×81 matrix, in which each row is a vectorized 9×9 image. The second matrix was prepared from the 419 ‘small’ images in the original training sample of 798 images, resized to 9×9 pixels by algorithm 6, and augmented to a set of 1676 images by rotating each image over 90, 180 and 270 degrees. The data matrices were reduced in dimension by PCA and next clustered to 30 clusters by the K -means algorithm applied to the rows of the loading matrix, following the procedure described in Section 5.5.1. For small pictures we retained 22 principal components, explaining 99% of the variation in the loadings (see Figure 5.6 for a graphical representation of the eigenvectors) and also 30 clusters. The 30 clusters means were placed in initial reference books, separately for large and small images, which were next enlarged as follows. We computed the L_1 -distance

$$d(x, y) = \sum_{i=1}^h |x_i - y_i|,$$

between every training image (row of the data matrix) and every element of the initial reference book (cluster average), where $h = 400$ or $h = 81$ for large and small images, respectively, and selected the 10 % images that were most deviant from the reference book, i.e. had the largest minimum distance. These images were added to the initial reference books. Following visual inspection the latter sets of images were next reduced to sets of 68 and 33 images. Visual inspection focused on the picture clearly being a SO particle, and, for instance, excluded pictures with a severe breach of the spherical shape of the SO particle.

Having thus formed reference books $R^{(20)}$ and $R^{(9)}$ of large and small images (matrices of dimensions 68×400 and 33×81), we computed the minimum L_1 -distance of all training images to the corresponding reference book, and choose the 99% and 90% percentiles of these minima (38 and 6, respectively) as cut offs for the L_1 -distance of a future distance. Choosing 90% percentile as the cut off for small images is due to our finding of large portion of unsatisfactory SO pattern during the visual inspection for its final reference book.

In summary, the second part of the filter is as follows.

IB-Filter: Test 2. *A future image of dimensions $a \times b$ is pre-processed to 20×20 pixels if $\max(a, b) > 9$, and to 9×9 otherwise, in the manner described in Section 5.4, and vectorized to a 1×400 or 1×81 vector y . The L_1 -distance of y to every row of the reference book $R^{(h)}$ is computed, where $h = 20$ or $h = 9$ for large and small images, respectively. The distance $d_{REF}(y, R^{(h)})$ of y to the reference book $R^{(h)}$ is defined as the minimum of these distances. If $d_{REF}(y, R^{(h)}) < c$, the image is labeled as SO, otherwise NSO. Here $c = 38$ when $h = 20$ and $c = 6$ when $h = 9$.*

5.6 Performance Evaluation

Three groups of data were used to evaluate the performance of the benchmark filters and our proposal, categorized as: (I) clearly SO, (II) unclear SO, (III) clearly NSO. Table 5.1 summarizes these data. These three groups correspond to common situations, which a specialist faces in practice when manually labeling an image as SO or NSO. For each group, the metadata were used for the benchmark filters, and the raw images were the input to our image-based filter. The majority of test images are small to modest-size (almost all particles are $< 20\mu m$), with around 30% of the images falling into the small category (number of rows and columns ≤ 9). This is useful for testing purposes, since the main challenge will be in recognizing SO images of lower resolution. Simple cases of NSO particles, e.g. very elongated images, were also excluded from the datasets. The performance of the filters is expected to be significantly better when applied to real sample measurements.

The filters were compared on their false negative rate (i.e. error in labeling SO as NSO) for test sets (I) and (II), and on their false positive rate (i.e. error in labeling NSO as SO) for test set (III). As SO particles are not immunogenic, while some types of NSO particles could cause immunogenic

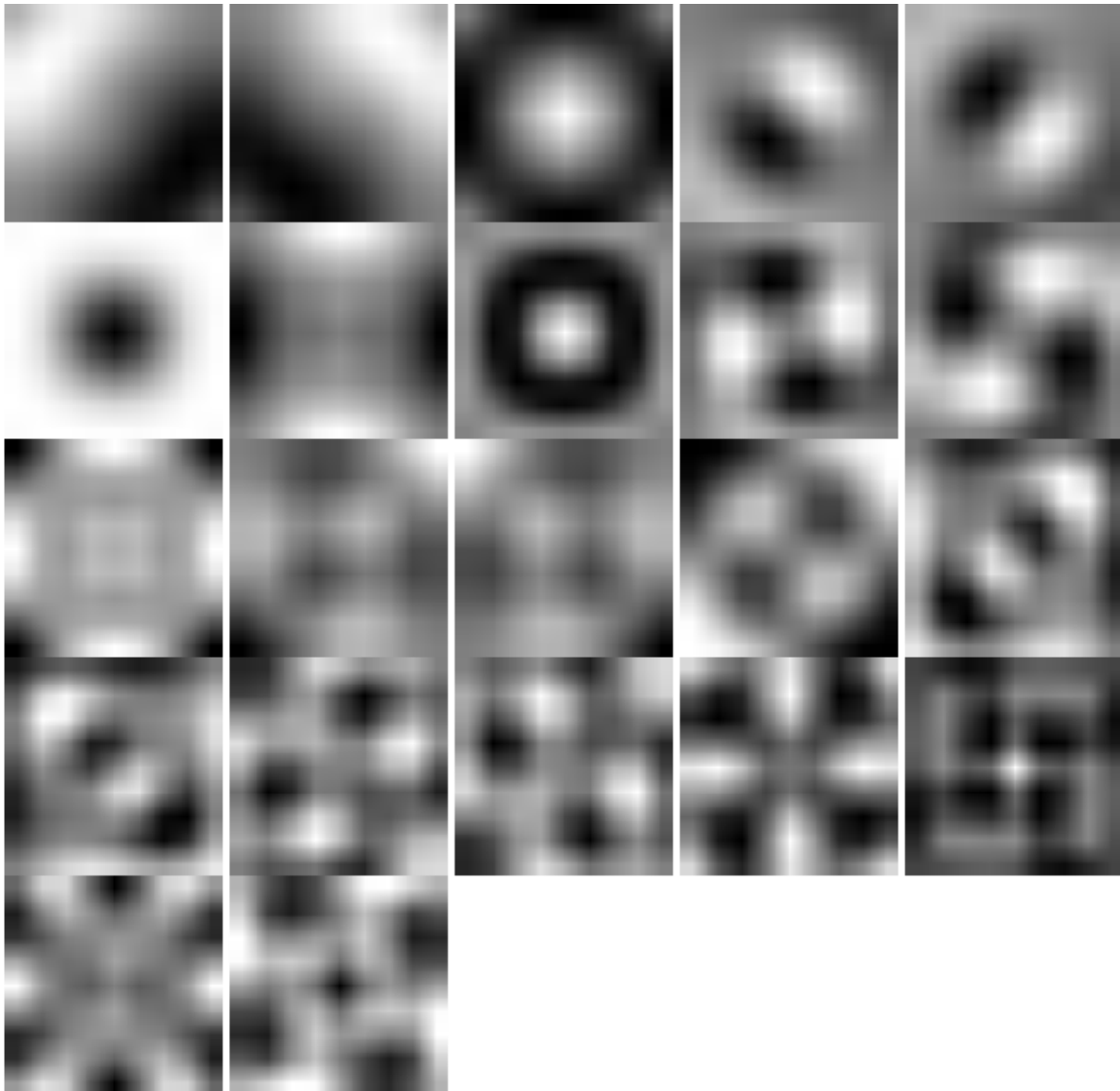


Fig. 5.6 First 22 (rescaled) eigenvectors of the covariance matrix of 1676×81 training data (419 pre-processed 9×9 images with 3 augmentations for each via rotation). The first 5 eigenvectors are in the first row from left to right and the sequence follows.

Table 5.1 Summary of Testing datasets

	Category	Selection Criteria	Small Image		Not-Small Image	
			#	ECD (99% perc.)	#	ECD (99% perc.)
(I)	Clearly SO	Recognize easily	233	7.375	700	16.375
(II)	Unclear SO	Moderately certain	537	7.125	373	8.625
(III)	Clearly NSO	Recognize easily	263	6.47	1361	19.375

concerns for bio-pharmaceuticals, the false positive rates could be considered as the more important criterion. The results are summarized in Tables 5.2–5.4, and visualized in Figure 5.7.

Table 5.2 Misclassification rates (False Negative) in group (I): Clearly SO

	Number	Benchmark AR	Benchmark SDR	Benchmark RF	IB-Filter (Test 1)	IB-Filter (Test 2)	IB-Filter (full)
Small	233	0.013	0.021	0.652	0.021	0.266	0.009
Not-Small	700	0.163	0.051	0.491	0.099	0.023	0.017
Total	933	0.125	0.044	0.532	0.079	0.084	0.015

Table 5.3 Misclassification rates (False Negative) in group (II): Unclear SO

	Number	Benchmark AR	Benchmark SDR	Benchmark RF	IB-Filter (Test 1)	IB-Filter (Test 2)	IB-Filter (full)
Small	537	0.084	0.324	0.873	0.235	0.378	0.212
Not-Small	373	0.496	0.603	1.000	0.464	0.249	0.206
Total	910	0.253	0.253	0.925	0.329	0.325	0.210

Table 5.4 Misclassification rates (False Positive) in group (III): Clearly NSO

	Number	Benchmark AR	Benchmark SDR	Benchmark RF	IB-Filter (Test 1)	IB-Filter (Test 2)	IB-Filter (full)
Small	263	0.517	0.319	0.015	0.137	0.076	0.163
Not-Small	1361	0.248	0.282	0.324	0.012	0.071	0.071
Total	1624	0.291	0.288	0.274	0.032	0.071	0.086

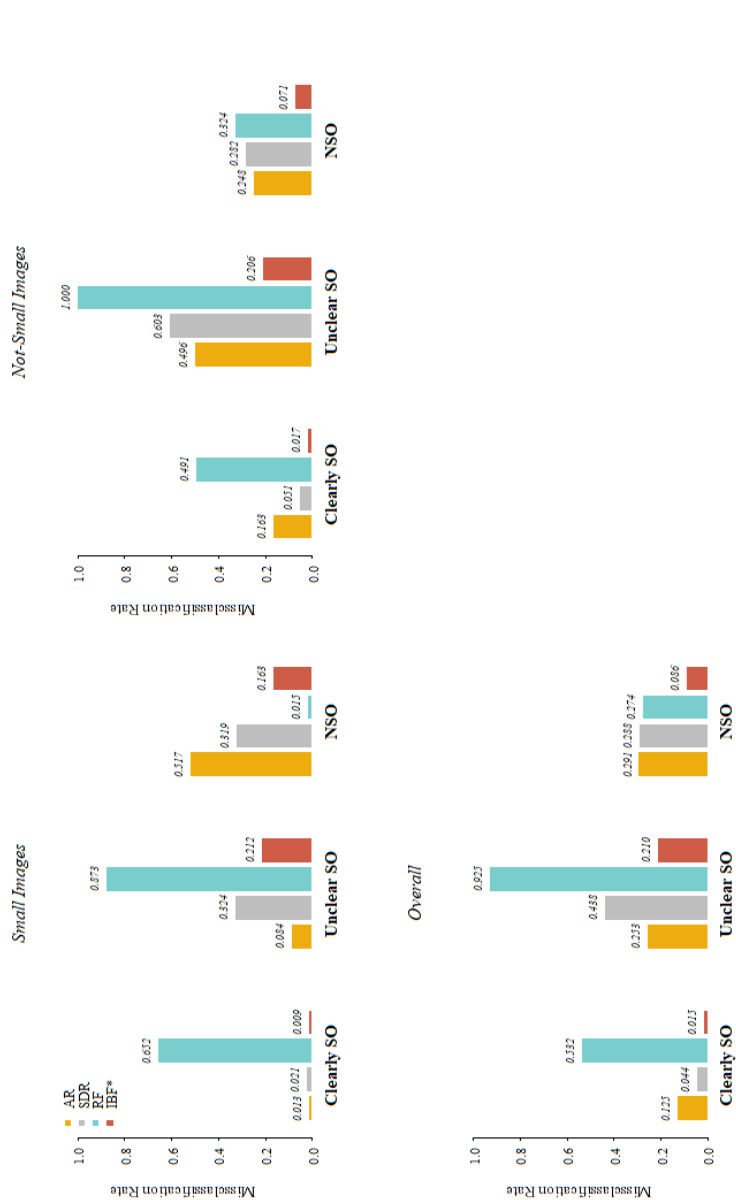


Fig. 5.7 Misclassification rates of the three benchmark filters (AR, SDR, RF) and our IB-filter for the three testing set (Clearly SO, Unclear SO and Clearly NSO)

5.7 Conclusion

As shown in the previous section, our IB-filter keeps a good balance in trading off the false positive and false negative rates. It outperforms the benchmark filters in almost all cases, the exceptions being that it gives the second lowest false negative rate for small unclear SO images (trailing AR) and the second lowest false positive rate for small NSO images (trailing RF). AR achieves the lowest error rate in labeling small unclear SO at the cost of having the highest error rate in labeling small NSO images. RF achieves its excellent false positive rate for small images by labeling the majority of testing set as NSO, which leads to very high error rates for classifying small clearly SO images (65.2%) and unclear SO images (87.3%).

Although it is not the focus of this chapter, we further studied the highly unsatisfactory performance of RF, which may be surprising, as the error on the training set was only 0.4%, while the tree sizes had been properly tuned by cross validation to prevent overfitting. We suspected that the key issue was that the NSO images in the RF training set were obtained from different materials than the NSO material in the testing set, giving little overlap between the NSO images in training and test sets. We prepared another training set for RF, consisting of the original 2316 SO samples and a new set of 1317 NSO samples, obtained from measurements from similar material as used for the NSO samples in the test set. The maximum number of terminal nodes of the trees in the random forest was determined at 11, by a similar cross-validation procedure as stated in the caption of Figure 5.1. The performance of the two RF models (training the original and the new dataset) is visualized in Figure 5.8. It can be observed that the error rate in labeling small NSO images increases, but the overall false positive and false negative rates decrease. This is a warning to practitioners that an RF-based SO filter will need to be trained per drug product. Even with this improvement, the performance with respect to the false positive rate comes close to our IB-filter, but still falls behind with respect to the false negative rates.

5.8 Discussion

This study confirms the potential of image-based filters based on the data depth of a future image in the distribution of SO images. In particular, our proposed filter makes a good trade-off between false positive and false negative rates. The advantages over the benchmark approaches come from two parts: (1) metadata has limited representation of image features that are important to distinguish SO and NSO particles, particularly for small particles, and (2) the benchmark filters were mostly built in a supervised manner.

We elaborate on the second point. The patterns of NSO images are arbitrary, as they are random 2D projections of free-shape 3D objects. A formulation of the binary classification problem SO/NSO as a supervised learning task is therefore ill-posed. A training dataset will necessarily differ from

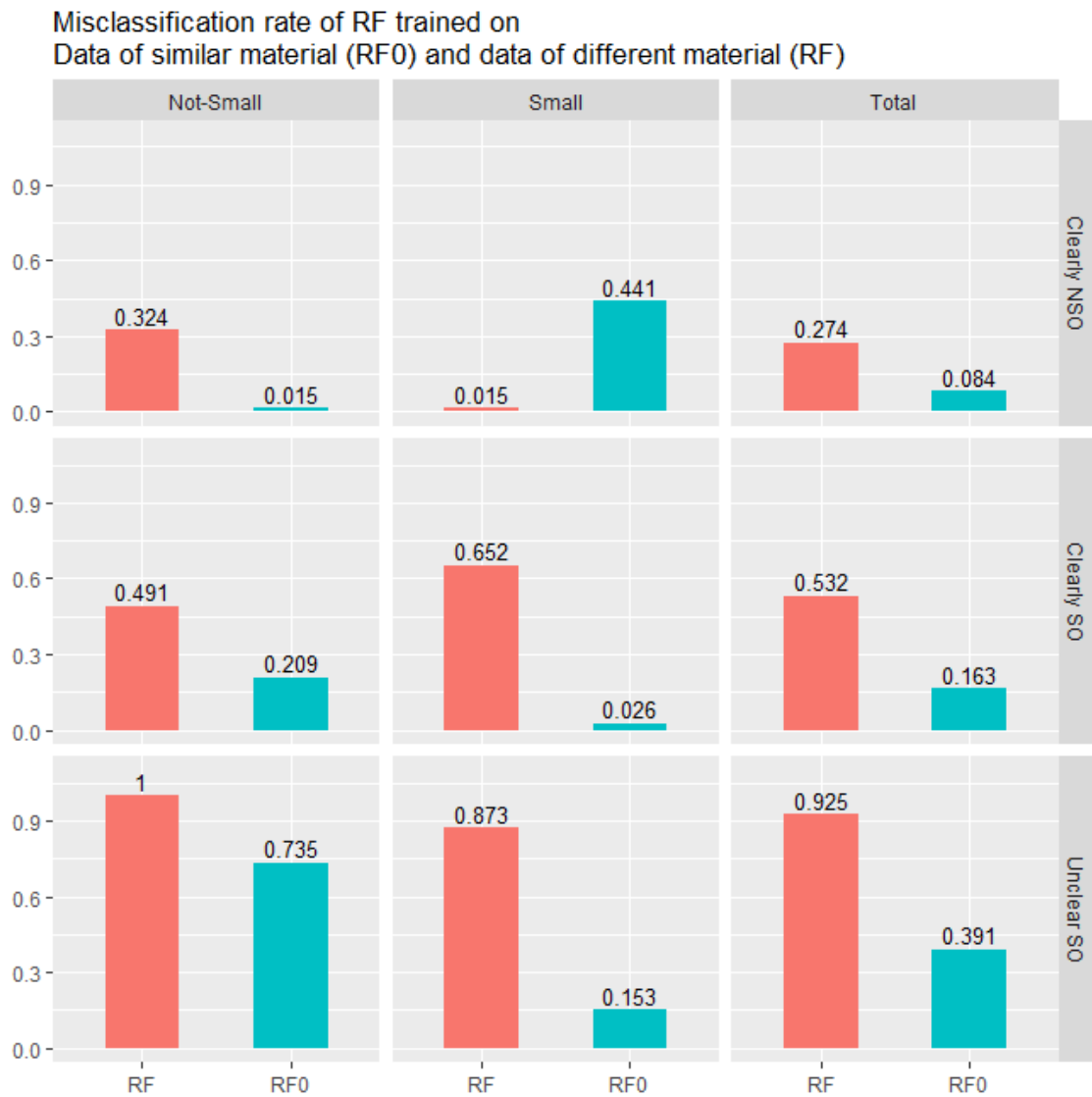


Fig. 5.8 Misclassification rates of the two RF models across the three testing sets. The original RF model (RF, red bars) was trained on NSO images of a material different from the NSO testing set. The second model (RF0, blue bars) was trained on NSO images from a similar material as the NSO testing set.

future data in essential features, and the usefulness of a supervised filter will be limited to a restricted domain.

By choosing to train a filter based only on SO images, we accept that there will be some patterns of NSO images that overlap with the distribution of SO images, but this is acceptable if the primary focus is on recognizing particles that give strong evidence that they are SO. The overlapping patterns introduce an inevitable trade-off between false positives and false negatives, which is shared by any image-based or metadata-based filters. We learnt that this challenge is more serious for small particles, since small SO images have less regular patterns, e.g. the spherical hull may be breached by the light reflection, or the white spot may all but disappear.

Another minor merit of pursuing an image-based filter is its applicability across different versions of MFI systems. For instance, our image-based filter is theoretically applicable to images in gray scale, regardless of the type of flow image system it originates from. In contrast, metadata from different versions of MFI system may summarize the images under different definitions, even if the summarized features are labeled with the same names.

The following points can be improved in future work.

- While pre-processing is necessary to standardize input images of arbitrary size for analysis, the current algorithm for padding and resizing can be improved with respect to distortion of the original patterns.
- Although this study is adequate to demonstrate the value of an image-based filter, more careful selection of a larger set of SO training images could easily improve performance. Furthermore, instead of our distinction between small and particles (number of rows and columns smaller than 10), a systematic study may be carried out to determine appropriate bin sizes.
- The settings of the first part of the filter (Test 1) in our image-based filter may be further optimized. Instead of using the maximum depth across 30 clusters, one may seek the average of the highest two depth values, or measure another degree of agreement with several key clusters. The intuition is that, if the testing image has a standard SO pattern, it is likely to get high depth values in more than one cluster.
- One may develop a better algorithm to extract individual images from the intact images, in order to minimize the chance of presence of multiple particles in one image, or truncation of a particle.

5.9 Appendix 1: Gallery on variations of SO/NSO images

This Appendix shows the gallery of selected SO and NSO particles: examples of Clearly SO images in Figure 5.9, Unclear SO images in Figure 5.10, and Clearly NSO images in Figure 5.11.

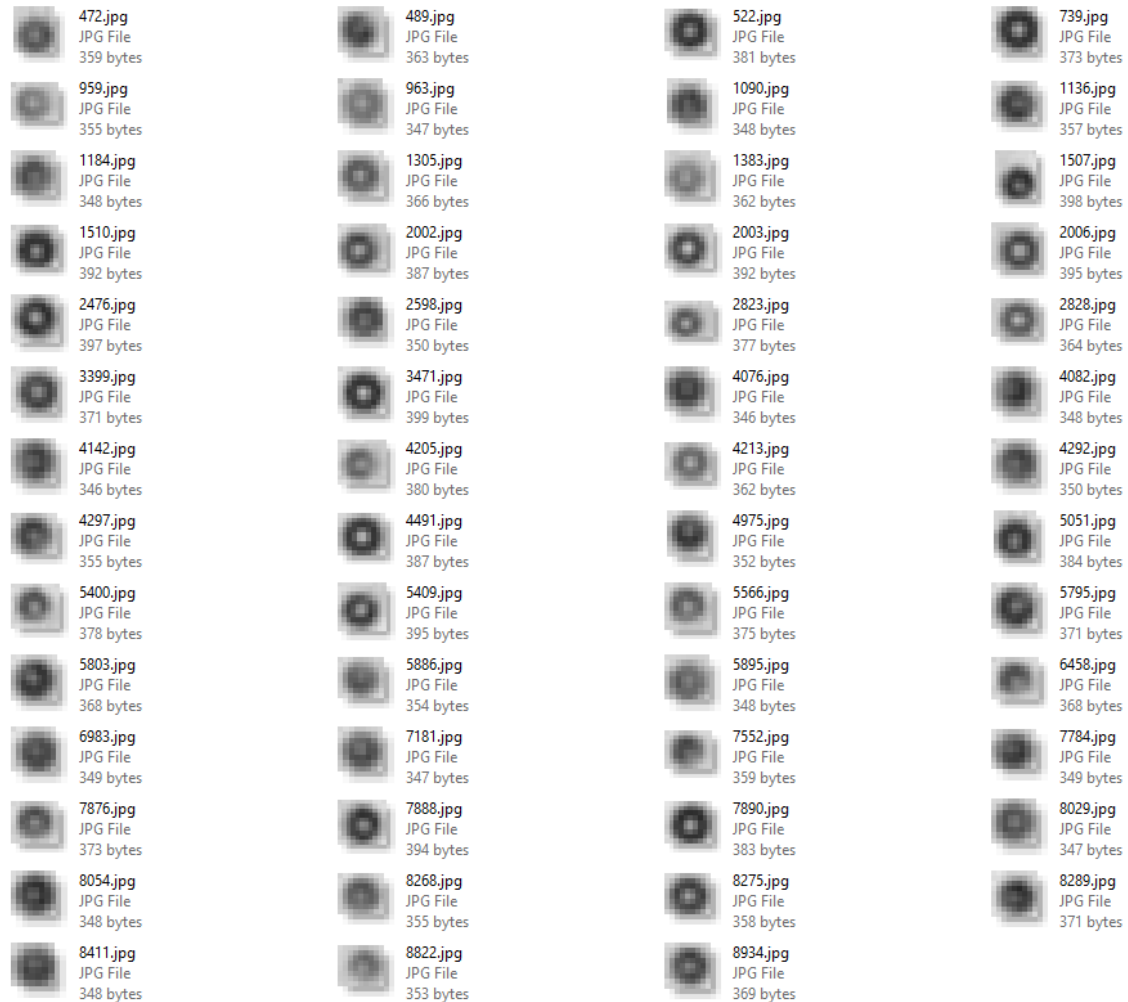


Fig. 5.9 Examples of Clearly silicon oil (SO) particles that can be easily recognized.

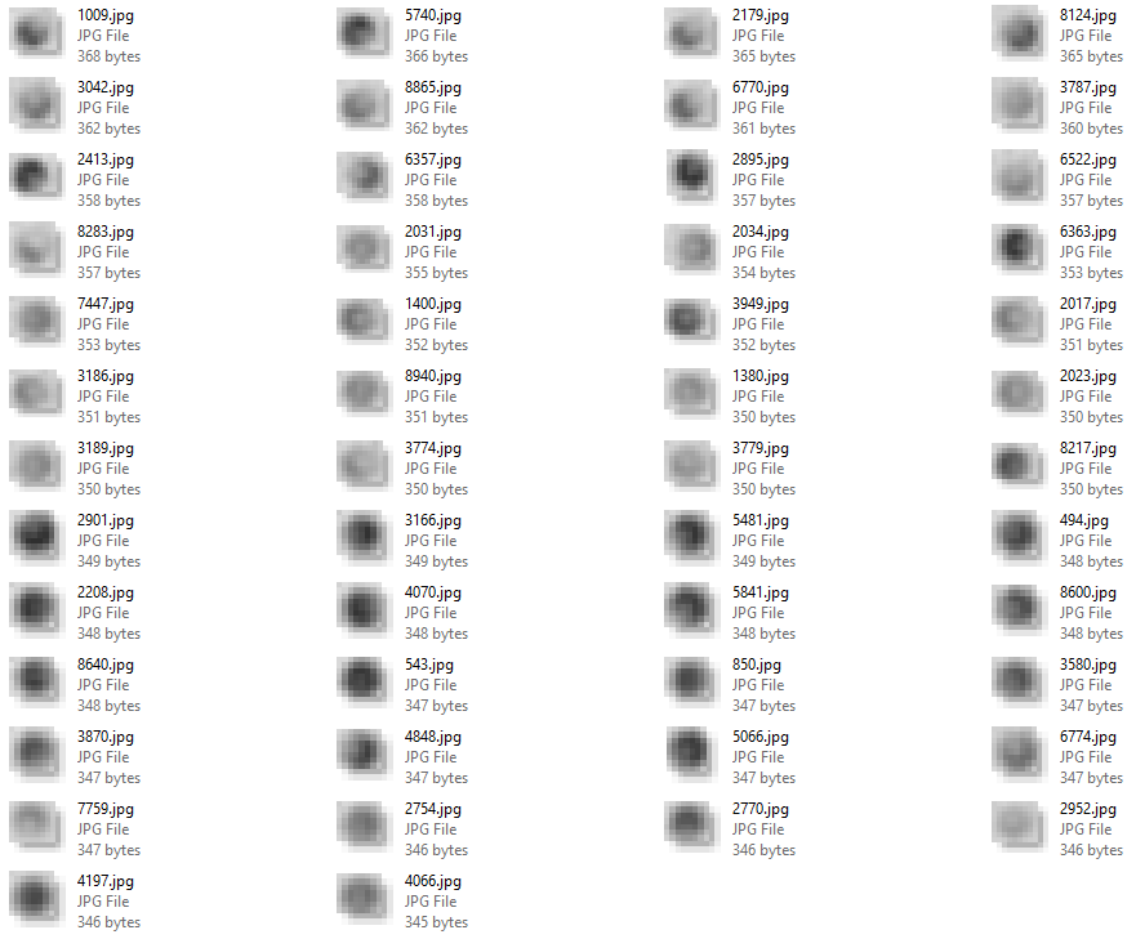


Fig. 5.10 Examples of Unclear silicon oil (SO) particles that experts are moderately certain that they are SO based on images.

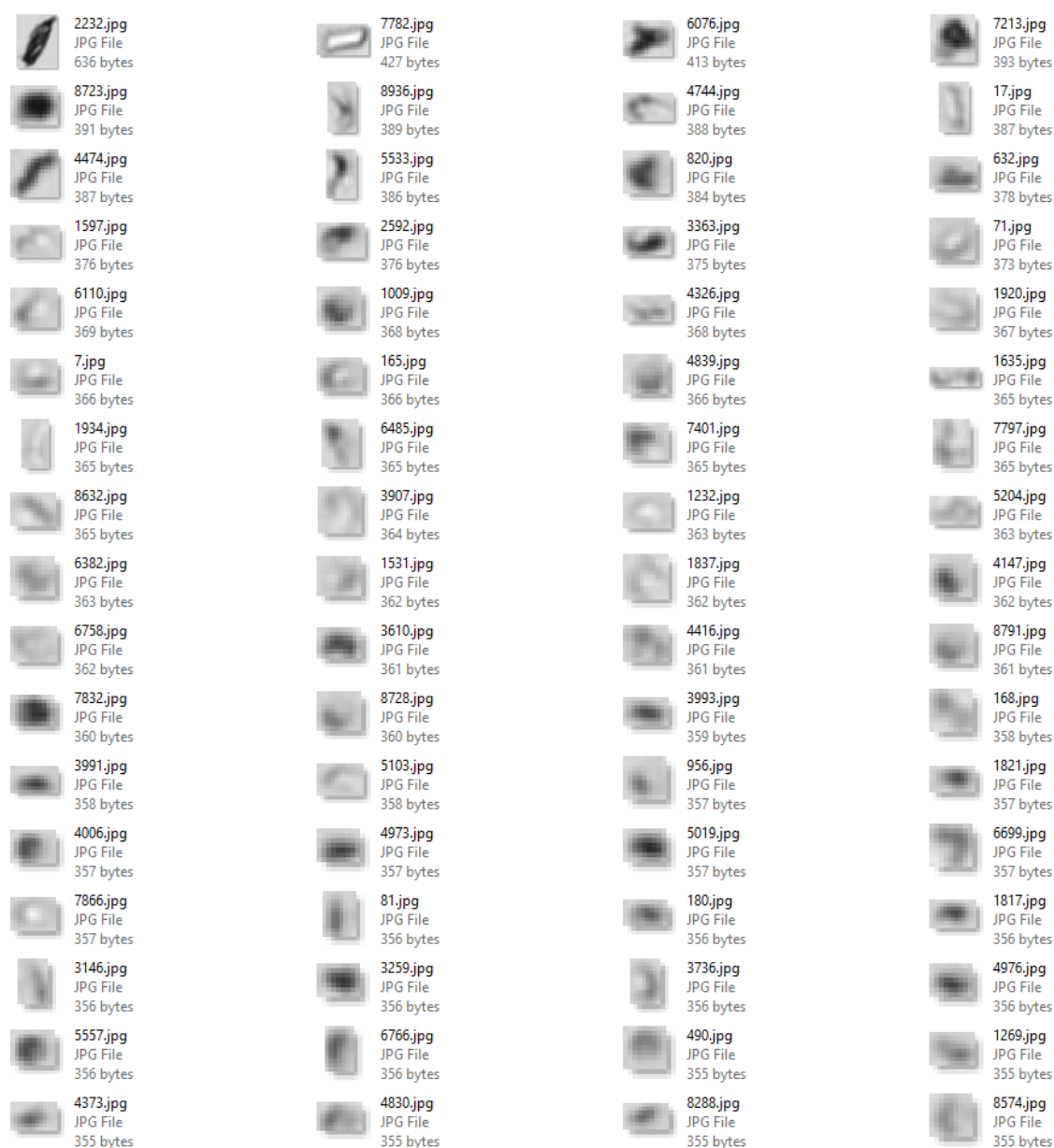


Fig. 5.11 Examples of Clearly non-silicon oil (NSO) particles that can be easily recognized.

5.10 Appendix 2: Gallery of the Reference book in Test 2 of IB-Filter

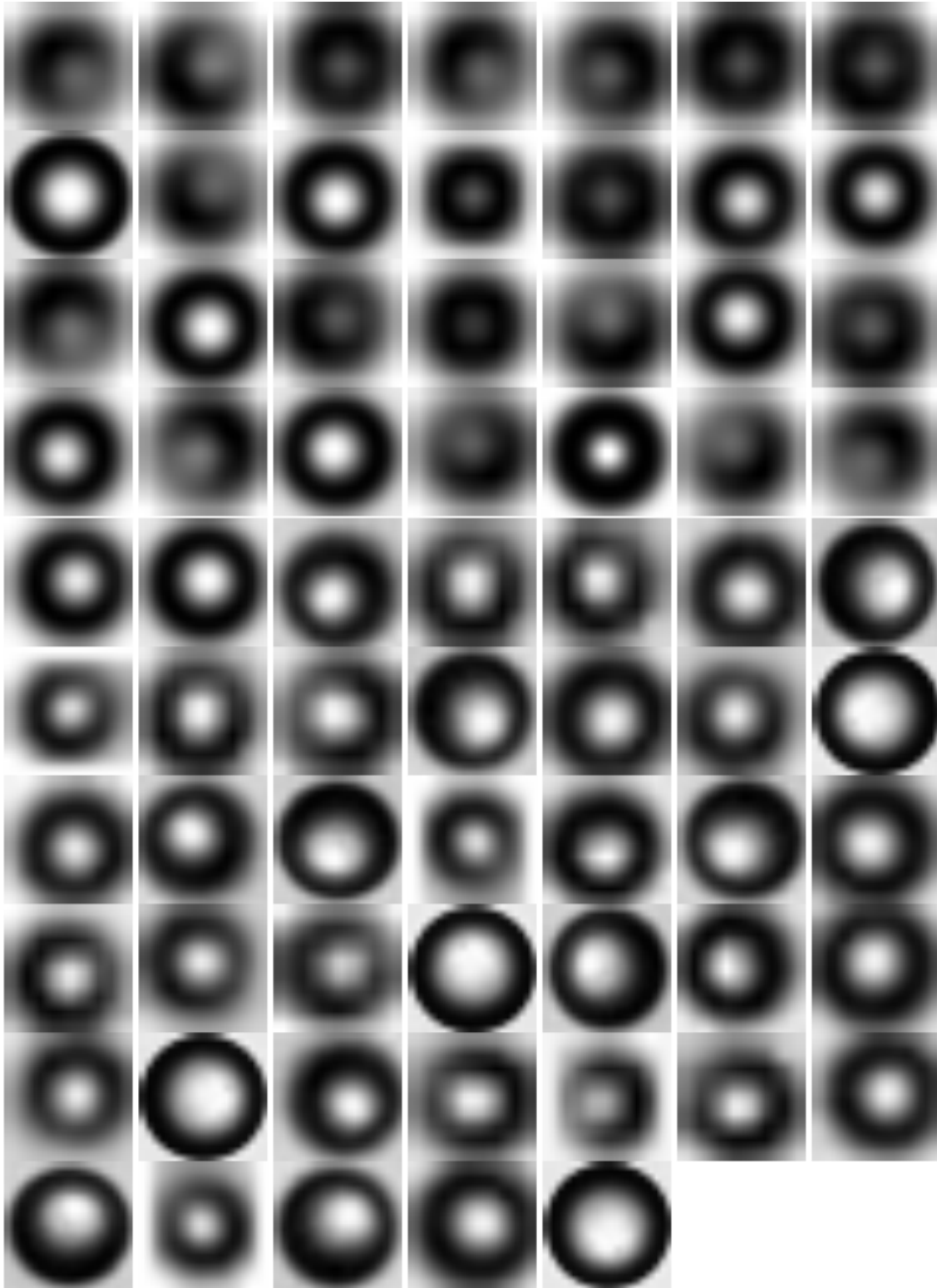


Fig. 5.12 Reference book of SO images of size 20×20 .

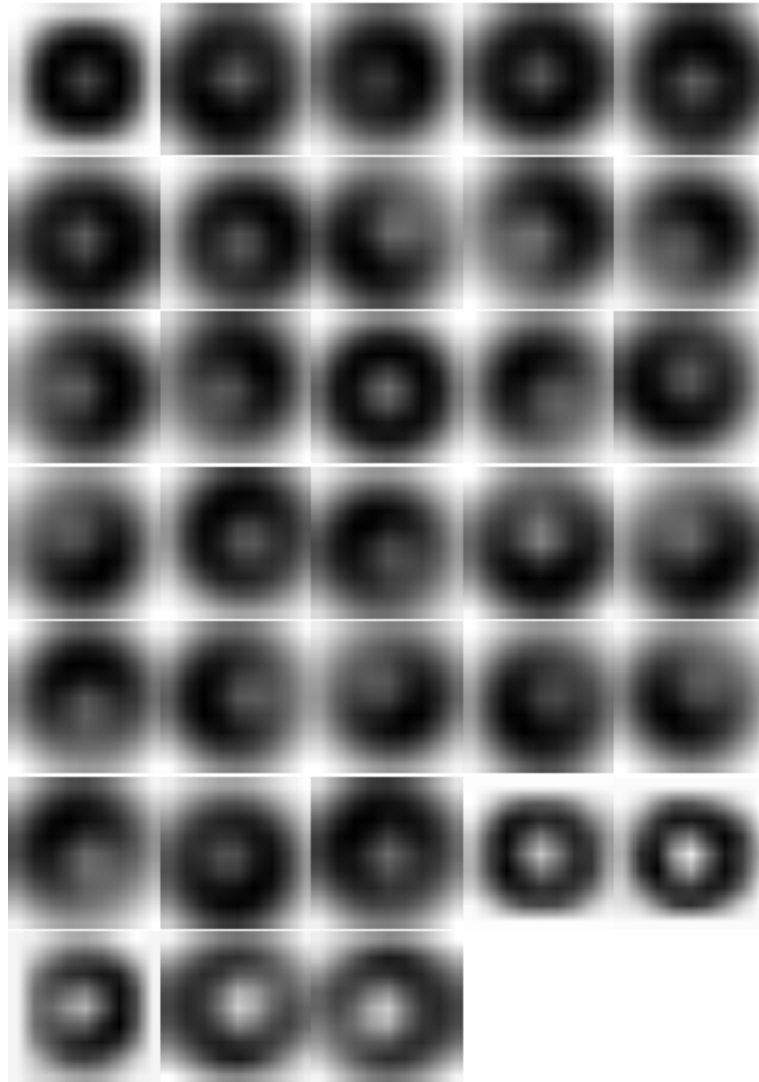


Fig. 5.13 Reference book of SO images of size 9×9 .

5.11 Appendix 3: Gallery of example images from the testing dataset.

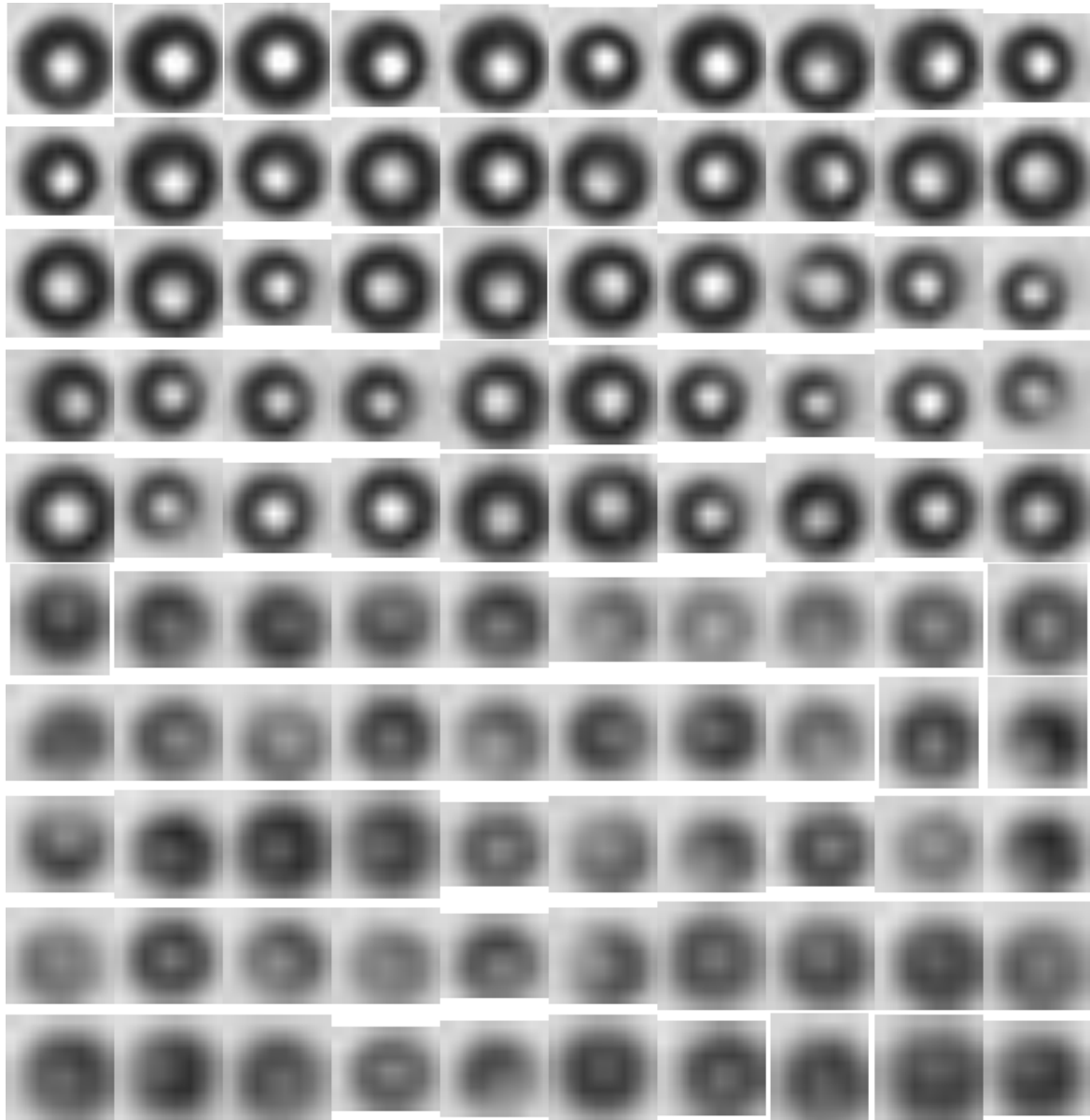


Fig. 5.14 Examples of Clearly Oil images in the testing set.

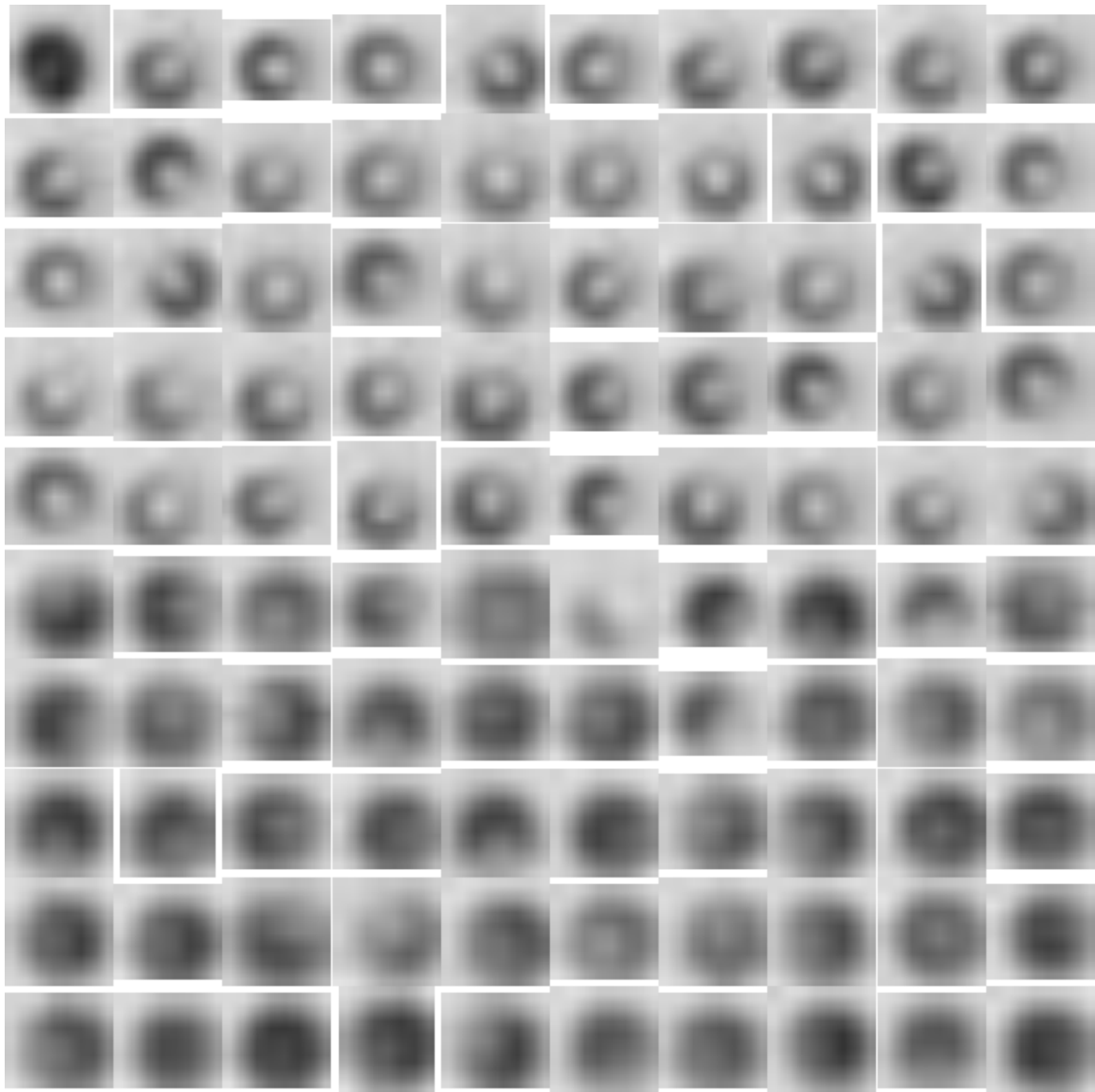


Fig. 5.15 Examples of Unclear Oil images in the testing set.

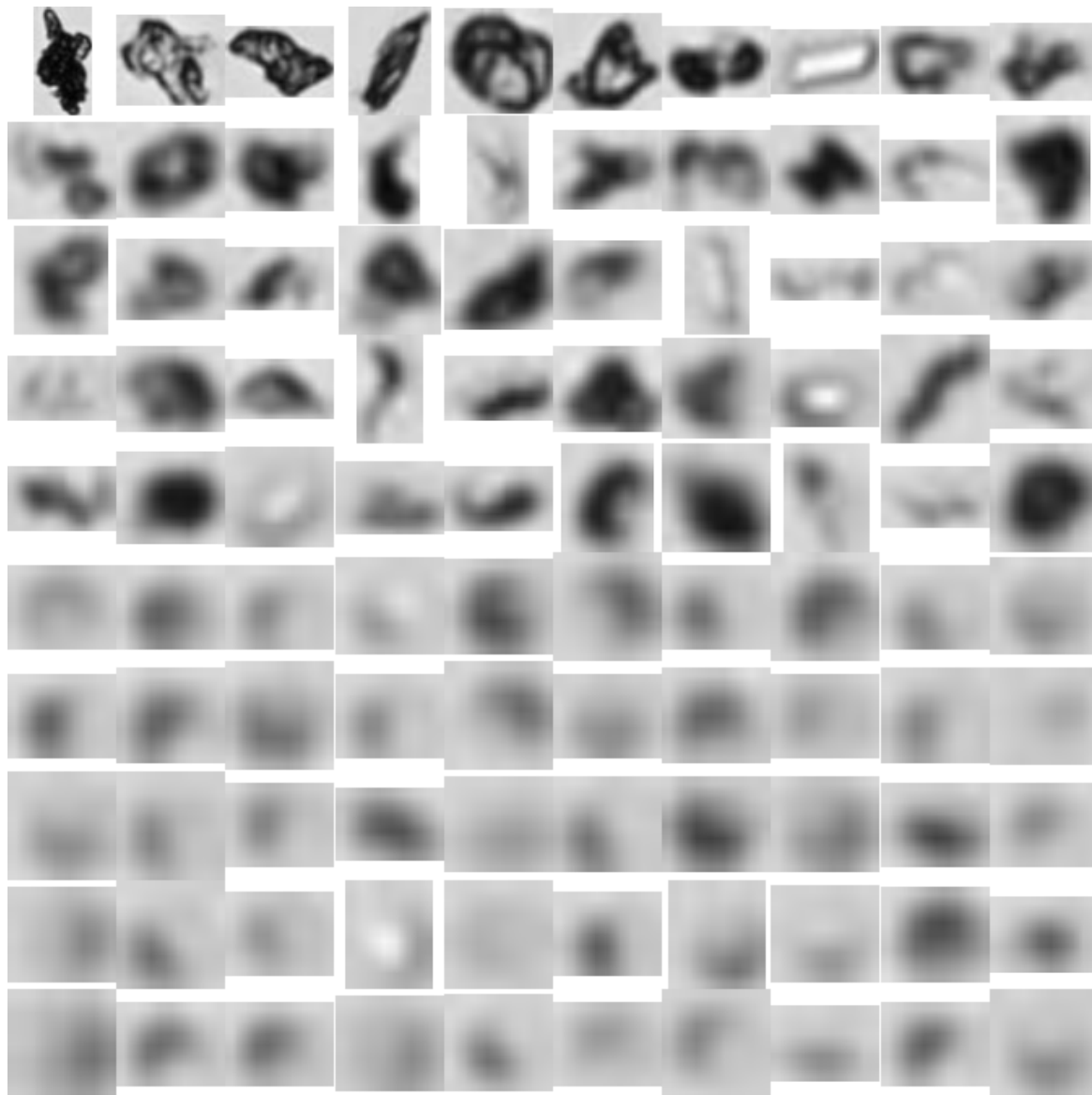


Fig. 5.16 Examples of Clearly NSO images in the testing set.

References

- [1] Aitchison, J. (1964). Two papers on the comparison of bayesian and frequentist approaches to statistical problems of prediction: Bayesian tolerance regions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):161–175.
- [2] Aitchison, J. (1966). Expected-cover and linear-utility tolerance intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):57–62.
- [3] Allen, P. V., Dukes, G. R., and Gerger, M. E. (1991). Determination of release limits: a general methodology. *Pharmaceutical research*, 8(9):1210–1213.
- [4] Blackwell, D., MacQueen, J. B., et al. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, 1(2):353–355.
- [5] Blei, D. M., Jordan, M. I., et al. (2006). Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–144.
- [6] Browne, W. J. and Draper, D. (2006). A comparison of bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, 1(3):473–514.
- [7] Cardot, H., Goga, C., Lardin, P., et al. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic journal of statistics*, 7:562–596.
- [8] Cardot, H. and Josserand, E. (2011). Horvitz–thompson estimators for functional data: Asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98(1):107–118.
- [9] Chakraborty, A., Chaudhuri, P., et al. (2014). The spatial distribution in infinite dimensional spaces and related quantiles and depths. *The Annals of Statistics*, 42(3):1203–1231.
- [10] Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, 91(434):862–872.
- [11] Claeskens, G., Hubert, M., Slaets, L., and Vakili, K. (2014). Multivariate functional halfspace depth. *Journal of the American Statistical Association*, 109(505):411–423.

- [12] Cuevas, A., Febrero, M., and Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3):481–496.
- [13] Dahl, D. B. (2005). Sequentially-allocated merge-split sampler for conjugate and nonconjugate dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 11.
- [14] Easterling, R. G. and Weeks, D. L. (1970). An accuracy criterion for bayesian tolerance intervals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 32(2):236–240.
- [15] Food, Administration, D., et al. (2013). Immunogenicity assessment for therapeutic protein products. *FDA: Rockville, MD* (<http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM338856.pdf>).
- [16] Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.
- [17] Gelman, A., Van Dyk, D. A., Huang, Z., and Boscardin, J. W. (2008). Using redundant parameterizations to fit hierarchical models. *Journal of Computational and Graphical Statistics*, 17(1):95–122.
- [18] Ghosal, S. and Van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press.
- [19] Gijbels, I. and Nagy, S. (2015). Consistency of non-integrated depths for functional data. *Journal of Multivariate Analysis*, 140:259–282.
- [20] Gijbels, I., Nagy, S., et al. (2017). On a general definition of depth for functional data. *Statistical Science*, 32(4):630–639.
- [21] Hamada, M. (2002). Bayesian tolerance interval control limits for attributes. *Quality and Reliability Engineering International*, 18(1):45–52.
- [22] Hamada, M., Johnson, V., Moore, L. M., and Wendelberger, J. (2004). Bayesian prediction intervals and their relationship to tolerance intervals. *Technometrics*, 46(4):452–459.
- [23] Ishwaran, H. and James, L. F. (2011). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*.
- [24] Jain, S. and Neal, R. M. (2012). A split-merge markov chain monte carlo procedure for the dirichlet process mixture model. *Journal of Computational and Graphical Statistics*.
- [25] Jain, S., Neal, R. M., et al. (2007). Splitting and merging components of a nonconjugate dirichlet process mixture model. *Bayesian Analysis*, 2(3):445–472.
- [26] Kalli, M., Griffin, J. E., and Walker, S. G. (2011). Slice sampling mixture models. *Statistics and computing*, 21(1):93–105.

- [27] Katki, H. A., Engels, E. A., and Rosenberg, P. S. (2005). Assessing uncertainty in reference intervals via tolerance intervals: application to a mixed model describing hiv infection. *Statistics in Medicine*, 24(20):3185–3198.
- [28] Krishnamoorthy, K. and Mathew, T. (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation*. Wiley Series in Probability and Statistics. Wiley.
- [29] Lei, J., Robins, J., and Wasserman, L. (2013). Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287.
- [30] Li, J., Liu, R. Y., et al. (2008). Multivariate spacings based on data depth: I. construction of nonparametric multivariate tolerance regions. *The Annals of Statistics*, 36(3):1299–1323.
- [31] Liu, R. Y., Parelius, J. M., Singh, K., et al. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by liu and singh). *The annals of statistics*, 27(3):783–858.
- [32] López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.
- [33] López-Pintado, S. and Romo, J. (2011). A half-region depth for functional data. *Computational Statistics & Data Analysis*, 55(4):1679–1695.
- [34] MacEachern, S. N. and Müller, P. (1998). Estimating mixture of dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238.
- [35] Miller, R. W. (1989). Parametric empirical bayes tolerance intervals. *Technometrics*, 31(4):449–459.
- [36] Mukerjee, R. and Reid, N. (2001). Second-order probability matching priors for a parametric function with application to bayesian tolerance limits. *Biometrika*, 88(2):587–592.
- [37] Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.
- [38] Nieto-Reyes, A., Battey, H., et al. (2016). A topologically valid definition of depth for functional data. *Statistical Science*, 31(1):61–79.
- [39] Organization, W. H. (2017). 5.7: Tests for particulate contamination. *The International Pharmacopoeia, seventh edition*, (Seventh Edition).
- [40] Papaspiliopoulos, O. and Roberts, G. O. (2008). Retrospective markov chain monte carlo methods for dirichlet process hierarchical models. *Biometrika*, 95(1):169–186.
- [41] Pathmanathan, D., Mukerjee, R., and Ong, S. H. (2014). Two-sided bayesian and frequentist tolerance intervals: general asymptotic results with applications. *Statistics*, 48(3):524–538.

- [42] Pathmanathan, D. and Ong, S. H. (2014). A monte carlo simulation study of two-sided tolerance intervals in balanced one-way random effects model for non-normal errors. *Journal of Statistical Computation and Simulation*, 84(11):2329–2344.
- [43] Pitman, J., Yor, M., et al. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900.
- [44] Pokotylo, O., Mozharovskiy, P., and Dyckerhoff, R. (2016). Depth and depth-based classification with r-package ddalpha. *arXiv preprint arXiv:1608.04109*.
- [45] Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer.
- [46] Saggu, M., Patel, A. R., and Koulis, T. (2017). A random forest approach for counting silicone oil droplets and protein particles in antibody formulations using flow microscopy. *Pharmaceutical research*, 34(2):479–491.
- [47] Saggu, M., Patel, A. R., and Koulis, T. (2018). S-factor optimization for mfi using lumetics link. *Breckenridge, CO in Protein Aggregation and Immunogenicity Conference*.
- [48] Serfling, R. and Wijesuriya, U. (2017). Depth-based nonparametric description of functional data, with emphasis on use of spatial depth. *Computational Statistics & Data Analysis*, 105:24–45.
- [49] Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.
- [50] Sguera, C., Galeano, P., and Lillo, R. (2014). Spatial depth-based classification for functional data. *Test*, 23(4):725–750.
- [51] Sharma, G. and Mathew, T. (2012). One-sided and two-sided tolerance intervals in general mixed and random effects models using small-sample asymptotics. *Journal of the American Statistical Association*, 107(497):258–267.
- [52] Singh, S. K., Afonina, N., Awwad, M., Bechtold-Peters, K., Blue, J. T., Chou, D., Cromwell, M., Krause, H.-J., Mahler, H.-C., Meyer, B. K., et al. (2010). An industry perspective on the monitoring of subvisible particles as a quality attribute for protein therapeutics. *Journal of pharmaceutical sciences*, 99(8):3302–3321.
- [53] Strehl, R., Rombach-Riegraf, V., Diez, M., Egodage, K., Bluemel, M., Jeschke, M., and Koulov, A. V. (2012). Discrimination between silicone oil droplets and protein aggregates in biopharmaceuticals: a novel multiparametric image filter for sub-visible particles in microflow imaging analysis. *Pharmaceutical research*, 29(2):594–602.
- [54] Tsiligkaridis, T. and Forsythe, K. (2015). Adaptive low-complexity sequential inference for dirichlet process mixture models. In *Advances in Neural Information Processing Systems*, pages 28–36.

- [55] Turco, S. (1994). *Sterile Dosage Forms: Their Preparation and Clinical Application*. Lea & Febiger.
- [56] Van der Merwe, A. and Hugo, J. (2007). Bayesian tolerance intervals for the balanced two-factor nested random effects model. *TEST*, 16(3):598–612.
- [57] Van der Merwe, A., Pertorius, A. L., and Meyer, J. H. (2006). Bayesian tolerance intervals for the unbalanced one-way random effects model. *Journal of Quality Technology*, 38(3):280–293.
- [58] van der Vaart, A., van der Vaart, A., van der Vaart, A., and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.
- [59] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.
- [60] Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*. New York: Springer.
- [61] Wald, A. (1943). An extension of wilks' method for setting tolerance limits. *The Annals of Mathematical Statistics*, 14(1):45–55.
- [62] Walker, S. G. (2007). Sampling the dirichlet mixture model with slices. *Communications in Statistics—Simulation and Computation*®, 36(1):45–54.
- [63] Wang, L. and Dunson, D. B. (2011). Fast bayesian inference in dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216.
- [64] Wei, G. C. (1998). Simple methods for determination of the release limits for drug products. *Journal of biopharmaceutical statistics*, 8(1):103–114.
- [65] Weinbuch, D., Zölls, S., Wiggenhorn, M., Friess, W., Winter, G., Jiskoot, W., and Hawe, A. (2013). Micro–flow imaging and resonant mass measurement (archimedes)–complementary methods to quantitatively differentiate protein particles and silicone oil droplets. *Journal of pharmaceutical sciences*, 102(7):2152–2165.
- [66] Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, 12(1):91–96.
- [67] Wolfinger, R. D. (1998). Tolerance intervals for variance component models using bayesian simulation. *Journal of quality technology*, 30(1):18–32.
- [68] Young, D. S., Gordon, C. M., Zhu, S., and Olin, B. D. (2016). Sample size determination strategies for normal tolerance intervals using historical data. *Quality Engineering*, 28(3):337–351.
- [69] Young, D. S. and Mathew, T. (2014). Improved nonparametric tolerance intervals based on interpolated and extrapolated order statistics. *Journal of Nonparametric Statistics*, 26(3):415–432.

-
- [70] Zölls, S., Weinbuch, D., Wiggenghorn, M., Winter, G., Friess, W., Jiskoot, W., and Hawe, A. (2013). Flow imaging microscopy for protein particle analysis—a comparative evaluation of four different analytical instruments. *The AAPS journal*, 15(4):1200–1211.
- [71] Zuo, Y. et al. (2003). Projection-based depth functions and associated medians. *The Annals of Statistics*, 31(5):1460–1490.
- [72] Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of statistics*, pages 461–482.

Summary

This thesis studies how to construct prediction sets for different types of data and models.

In Chapter 1, two common types of Bayesian prediction sets (also known as *tolerance region*) are introduced in comparison to their frequentist counterparts. The $(1 - \delta)$ -expectation type aims to achieve a nominal coverage on average, while the (δ, β) -type targets to cover at least $1 - \delta$ mass with a prescribed confidence $1 - \beta$. We focus on the prediction set (i.e. interval) for a univariate variable in this chapter, where the two-sided prediction interval needs particular attention. Assume that we derive a desired two-sided interval by expanding symmetrically around a center point. We provide both analytical lemmas and numerical examples to show that it might not lead to the shortest interval by choosing the posterior mean as the center point in different models. Specifics depend on the model and the prior. Discussion starts with i.i.d. sample from a normal distribution to a linear mixed model. We give conditions under which for large samples a Bayesian tolerance region also has frequentist validity, and study this for smaller samples in a simulation study. A computational strategy for computing a Bayesian two-sided tolerance interval for a Gaussian future variable is provided and compared with some benchmarks, and this is applied in details to the case of possibly unbalanced linear mixed models. The motivation to consider a mixed-effect model comes from the context of quality control in the pharmaceutical industry, where data usually include several replicates per batch and hence are not i.i.d.

In Chapter 2, we discuss how to derive a conditional one-sided prediction limit in a linear random-coefficients model, motivated by a common task in stability risk control. Given empirical stability data, the shelf life (SL) and the stability specification, the objective is to determine a limit for a measurement at release (i.e. time zero) that incorporates stability risk. We discuss the two practically most relevant formulations of "incorporate stability risk". The first one is to ensure that the calculated limit at release and the stability specification will define the riskiness of a future batch consistently, provided it has similar stability behavior as the empirical batches. The second one is to set a control limit at release that can screen out future batches that are at high risk of failing the stability specification at the end of their shelf life. This second formulation is natural for quality assurance and popular in practice, but in our study turns out to be stringent. We provide insight when this control strategy is too ambitious and impractical to be implemented. We also introduce a heat-map visualization that enables practitioners to determine when the limits calculated in the second formulation are feasible, and sheds light on

directions to improve in the other case. Bayesian procedures for inference under both formulations are discussed. Particularly, two possible estimators under the second formulation are provided, and their asymptotic properties are discussed in the annex. In addition, the current industrial benchmark is reviewed and contrasted to our two formulations. All discussed approaches are illustrated on two example datasets, representing two commonly-encountered scenarios.

In Chapter 3, we study how to construct the two types of Bayesian prediction set defined in Chapter 1 in a nonparametric setting. Past and future observations are assumed to be sampled from an unknown distribution F , for which we assume a Dirichlet Process mixture (DPM) of Gaussian distributions as the prior. Details about the DPM prior and numerical procedures for its posterior inference are provided in the main text and the annexes. For the sake of computational scalability and efficiency for large sample size in the multi-dimensional setting, we study the SUGS algorithm from [63] as an alternative to a Gibbs sampler. To alleviate the dependency of SUGS on the sequence of input data, i.e. one of its key drawbacks, we also discuss how to select the final output from the algorithm among many randomly-permuted sequences of the input data. We confirm later in the simulation that SUGS is fast and can provide good approximated solution to $(1 - \delta)$ -expectation type prediction set. However, its pseudo posterior sample tends to under-represent the variation in the true posterior, and hence it falls short to meet the nominal confidence for (δ, β) -type. In this chapter, we also briefly introduce the classical nonparametric prediction sets for univariate observations, data depth methodology and its application in building nonparametric prediction sets. We include them as the baseline methods in the numerical evaluation via simulations. Data depths rank data points in a center-outwards manner, where a higher depth value indicates that a point is more "central" relative to the sample or a distribution. The depth paradigm provides a way to form prediction sets without particular shape assumptions. It is particularly useful in the multivariate case, since ranking of more than one-dimensional data points is not straightforward otherwise. We also adopt it as a key part of our Bayesian solution.

In Chapter 4, we study how to utilize depth methodology for functional data to (i) detect outliers, (ii) construct prediction sets, and (iii) perform two-sample testing. An overview of functional data depth (FDD) is provided at the start, where two new depth measures are proposed. These two depths are later shown to be particularly useful in outlier detection when the population is multi-modal (e.g. a mixture of Gaussian processes with distinct mean functions), which commonly occurs in industrial practice. The sensitivity of different FDDs to different types of outliers is illustrated via a simulated dataset. We also show the potential advantage of centering the curves before calculating FDDs. For topic (ii), we review two existing approaches to build a prediction set for functional data, and propose an alternative. The calculation procedure of our proposed approach is detailed via an example dataset. Then a simulation study confirms that the nominal coverage can be achieved on average via this procedure for a modest sample size. For topic (iii), we presents three different approaches to perform two-sample testing with functional data: a rank test from literature, our proposed procedures with permutation and bootstrap samples. Given a widely-known dataset, the results of these three

approaches is compared and discussed. Through this exercise we advocate mindful interpretation of the testing outcome where visual aid seems particularly valuable for functional data. Depending on the chosen test statistics, obvious “local” difference in visualized patterns may be masked by similarity elsewhere when the support space (of observed functions) is wide.

In Chapter 5, we presents an application that uses multivariate data depth to build prediction sets for image recognition. The objective is to provide a binary classification of future images as either a silicon oil (SO) or a non-silicon oil (NSO) particle. SO particles have a clear profile, while NSO particles may have many, also unknown forms, Therefore a classifier based on supervised learning would not be able to generalize its success against NSO stereotypes that are not seen in the training data. We demonstrate this with empirical evidence in the example of Random Forest (one of the industrial benchmarks). Our approach is to label a future image based on its similarity to the training set of SO particle images. A working procedure to pre-process the input images with different resolutions and sizes is proposed. The performance of our image-based filter is compared with three industrial benchmarks on a test set with 1843 SO images and 1624 NSO images. It stands out from the competitors with respect to both false positive (label SO as NSO) and false negative rate (label NSO as SO) overall, particularly when the input images are not small (i.e. larger than 9 by 9 pixels).

Samenvatting

In dit proefschrift wordt onderzocht hoe voorspellingsgebieden kunnen worden gemaakt voor verschillende soorten gegevens en modellen.

In Hoofdstuk 1 worden twee veelvoorkomende typen Bayesiaanse voorspellingsgebieden (beter bekend als *tolerantieregios*) geïntroduceerd in vergelijking met hun frequentistische tegenhangers. Het $(1 - \delta)$ -verwachtingstype streeft naar gemiddeld een nominale dekking, terwijl het (δ, β) -type zich richt op het dekken van ten minste $1 - \delta$ massa met een voorgeschreven vertrouwen $1 - \beta$. We richten ons in dit hoofdstuk op een voorspellingsinterval voor een univariate variabele, waarbij het tweezijdige voorspellingsinterval bijzondere aandacht vereist. Stel dat we een gewenst tweezijdig interval afleiden door symmetrisch rond een middelpunt uit te breiden. We bieden zowel analytische lemma's als numerieke voorbeelden om aan te tonen dat het misschien niet tot het kortste interval leidt door in verschillende modellen het a-posteriori gemiddelde als middelpunt te kiezen. Precieze keuzen zijn afhankelijk van het model en de prior. De discussie begint met een i.i.d. steekproef uit een normale verdeling afgeleid uit een lineair mixed model. We geven condities waaronder voor grote steekproeven een Bayesiaans tolerantiegebied ook frequentistische validiteit heeft en bestuderen dit voor kleinere steekproeven in een simulatiestudie. Een rekenstrategie wordt gegeven voor het berekenen van een Bayesiaans tweezijdig tolerantie-interval voor een Gaussische toekomstige variabele en vergeleken met enkele benchmarks. Dit wordt in detail toegepast op het geval van mogelijk ongebalanceerde lineaire mixed modellen. De motivatie om een mixed-effect-model te overwegen, komt voort uit de context van kwaliteitscontrole in de farmaceutische industrie, waar de gegevens meestal meerdere replicaties per batch bevatten en dus niet i.i.d. zijn.

In Hoofdstuk 2 bespreken we hoe we een voorwaardelijke eenzijdige voorspellingslimiet kunnen afleiden in een lineair random-coëfficiënten model, gemotiveerd door een veelvoorkomen de taak in stabiliteitsrisicobeheersing. Gegeven empirische stabiliteitsgegevens, de houdbaarheid (SL) en de stabiliteitsspecificatie, is het doel om een limiet te bepalen voor een meting bij vrijgave (d.w.z. tijdstip nul) die het stabiliteitsrisico omvat. We bespreken de twee praktisch meest relevante formuleringen van "geïncorporeerd stabiliteitsrisico". De eerste is ervoor te zorgen dat de berekende limiet bij vrijgave en de stabiliteitsspecificatie het risico van een toekomstige batch consistent zullen definiëren, op voorwaarde dat deze vergelijkbaar stabiliteitsgedrag vertoont als de empirische batches. De tweede is om een controlelimiet in te stellen bij vrijgave die toekomstige batches kan uitsluiten die een hoog

risico lopen om niet aan de stabiliteitsspecificatie te voldoen aan het einde van hun houdbaarheid. Deze tweede formulering is natuurlijk voor kwaliteitsborging en populair in de praktijk, maar blijkt in ons onderzoek stringent te zijn. We geven inzicht wanneer deze controlestrategie te ambitieus en onpraktisch is om geïmplementeerd te worden. We introduceren ook een warmtekaartvisualisatie waarmee gebruikers kunnen bepalen wanneer de limieten die in de tweede formulering zijn berekend haalbaar zijn, en die licht werpen op richtingen die in het andere geval kunnen worden verbeterd. Bayesiaanse procedures voor inferentie onder beide formuleringen worden besproken. In het bijzonder worden twee mogelijke schatters onder de tweede formulering gegeven en hun asymptotische eigenschappen besproken in de bijlage. Bovendien wordt de huidige industriële benchmark herzien en vergeleken met onze twee formuleringen. Alle besproken benaderingen worden geïllustreerd aan de hand van twee voorbeelddatasets, die twee veelvoorkomende scenario's vertegenwoordigen.

In Hoofdstuk 3 bestuderen we hoe we de twee typen Bayesiaanse voorspellingsgebieden, gedefinieerd in Hoofdstuk 1, kunnen construeren in een niet-parametrische setting. Waarnemingen uit het verleden en de toekomst worden verondersteld te zijn bemonsterd uit een onbekende distributie F , waarvoor we een Dirichlet Process Mixture (DPM) van Gauss-distributies als prior aannemen. Details over de prior en numerieke DPM-procedures voor de a-posteriori inferentie zijn te vinden in de hoofdttekst en de bijlagen. Omwille van computationele schaalbaarheid en efficiëntie voor een grote steekproefomvang in de multidimensionale setting, bestuderen we het SUGS-algoritme van [63] als alternatief voor een Gibbs-sampler. Om de afhankelijkheid van SUGS van de reeks invoergegevens, d.w.z. een van de belangrijkste nadelen ervan, te verminderen, bespreken we ook hoe de uiteindelijke uitvoer van het algoritme kan worden geselecteerd uit vele willekeurig gepermuteerde reeksen van de invoergegevens. We bevestigen later in de simulatie dat SUGS snel is en een goede geschatte oplossing kan bieden voor de $(1 - \delta)$ -verwachtingstype voorspellingsset. Zijn pseudo-posterieure steekproef neigt er echter naar de variatie in de werkelijke a-posteriori verdeling te onderschatten en daarom schiet hij tekort om te voldoen aan het nominale vertrouwen voor (δ, β) -type. In dit hoofdstuk geven we een korte introductie van de klassieke niet-parametrische voorspellingsgebieden voor univariate waarnemingen, datadiepte-methodologie en de toepassing ervan bij het bouwen van niet-parametrische voorspellingsgebieden. We nemen ze als basismethode op in de numerieke evaluatie via simulaties. Datadiepte rangschikt datapunten van uit het midden naar buiten, waarbij een hogere dieptewaarde aangeeft dat een punt meer “centraal” is ten opzichte van een steekproef of een distributie. Dit diepteparadigma biedt een manier om voorspellingsgebieden te vormen zonder een bepaalde vormveronderstelling. Het is met name handig in het multivariate geval, aangezien de rangschikking van meer dan eendimensionale gegevenspunten niet eenvoudig is. We nemen het ook over als een belangrijk onderdeel van onze Bayesiaanse oplossing.

In Hoofdstuk 4 bestuderen we hoe we dieptemethodologie kunnen gebruiken voor functionele gegevens om (i) uitschieters te detecteren, (ii) voorspellingsgebieden te construeren en (iii) twee steekproeven toetsen uit te voeren. Bij de start wordt een overzicht gegeven van de functionele datadiepte (FDD), waar twee nieuwe dieptematen worden voorgesteld. Deze twee diepten blijken

later bijzonder nuttig te zijn bij het opsporen van uitschieters wanneer de populatie multimodaal is (bijvoorbeeld een mengsel van Gaussische processen met verschillende gemiddelde functies), wat vaak voorkomt in de industriële praktijk. De gevoeligheid van verschillende FDD's voor verschillende soorten uitschieters wordt geïllustreerd aan de hand van een gesimuleerde dataset. We laten ook het potentiële voordeel zien van het centreren van de curven voordat FDD's worden berekend. Voor onderwerp (ii) bekijken we twee bestaande benaderingen om een voorspellingsgebied voor functionele gegevens te bouwen en stellen we een alternatief voor. De berekeningsprocedure van onze voorgestelde aanpak wordt gedetailleerd via een voorbeelddataset. Vervolgens bevestigt een simulatiestudie dat de nominale dekking gemiddeld via deze procedure kan worden bereikt voor een bescheiden steekproefomvang. Voor onderwerp (iii) presenteren we drie verschillende benaderingen om toetsen met twee steekproeven met functionele gegevens uit te voeren: een rangtoets uit de literatuur, onze voorgestelde procedures met permutatie en bootstrap-steekproeven. Gegeven een algemeen bekende dataset, worden de resultaten van deze drie benaderingen vergeleken en besproken. Door middel van deze oefening pleiten we voor een bewuste interpretatie van de testresultaten waar visuele hulp bijzonder waardevol lijkt voor functionele gegevens. Afhankelijk van de gekozen toetsingsgrootheden, kunnen duidelijke “lokale” verschillen in gevisualiseerde patronen worden gemaskeerd door gelijkenis elders wanneer het domein (van waargenomen functies) breed is.

In Hoofdstuk 5 presenteren we een toepassing die gebruik maakt van multivariate datadiepte om voorspellingsgebieden voor beeldherkenning te bouwen. Het doel is om een binaire classificatie van toekomstige foto's te bieden als een siliciumolie (SO) of een niet-siliciumolie (NSO) deeltje. SO-deeltjes hebben een duidelijk profiel, terwijl NSO-deeltjes veel, ook onbekende vormen kunnen hebben. Daarom zou een supervised classificatie op basis van begeleid leren zijn succes niet kunnen generaliseren tegen NSO-stereotypen die niet in de trainingsgegevens zijn gezien. We demonstreren dit met empirisch bewijs in het voorbeeld van gebruik van Random Forest (één van de industriële benchmarks). Onze aanpak is om een nieuwe foto te labelen op basis van de gelijkenis met de trainingsset van SO-deeltjesafbeeldingen. Er wordt een werkprocedure voorgesteld om de invoerbeelden met verschillende resoluties en formaten voor te bewerken. De prestaties van ons op afbeeldingen gebaseerde filter worden vergeleken met drie industriële benchmarks op een testset met 1843 SO-afbeeldingen en 1624 NSO-afbeeldingen. Het onderscheidt zich van de concurrenten met betrekking tot zowel fout-positief (label SO als NSO) als fout-negatief percentage (label NSO als SO) in het algemeen, vooral wanneer de invoerafbeeldingen niet klein zijn (d.w.z. groter dan 9 bij 9 pixels).

Acknowledgment

First and Foremost, I want to express my most heartfelt gratitude to my supervisor Prof. Aad van der Vaart. Thank you Aad, for the teaching, mentoring, inspiration, patience, flexibility and many other things in the past seven years of my doctorate journey. Without your kind support and making yourself so easily accessible, it won't be possible for me to complete this thesis along side a full-time job. I consider myself extremely lucky to be your student since the project of my Master thesis. Other than academic knowledge, I have learnt more valuable things from you in life that has already helped me through many bumps and kept things in perspective during the career pursue outside of academia.

Sincere thanks to all professors and colleagues in Leiden that have taught, helped and inspired me in the course, especially to Prof. Jacqueline Meulman, who I have the privilege to be connected already since the very start of my journey outside of China. Thanks, Jacqueline, for many many helps since the master program, they have tremendous meaning to me.

I would like to thank my mother for the continuous "gentle" nudging, passing on loads and loads of cheer that keeps me going. I guess I pick up the tenacity and optimism from her. Also big thanks to my aunts Yanyue Luo, Yanxing Luo, Yanli Luo, Lewen Zhuang, uncle Shiyu Chen, and cousins Rong-Rong Zhuang, Lin Li for their support and cheers, remote but very powerful.

I would also like to thank all my dear friends and colleagues along the way for the encouragement, discussion, collaboration, anecdote and food sharing, laughter and accompanying in general. Especially to Baobao, Beibei, Xinru, Pieter from Leiden; Marieke and Zheng Guan from CHDR; Erik, Pieta, Michiel, Henk, Thanh, Abu, Cathelijne from MSD Netherlands; Ting Li, Lin Zhang, Mian Deng, Jianxun Ji, Shengbin He from my hometown; Liliana, Ivan from Basel; Mark, Antonio, Paolo, Robert, Doruk, Katjusa, Nico from Novartis Quality; Björn, Mingle, John from Novartis BTD; Tony, Frank, Mark from Novartis GDD; Andreas D. from Athens and Bern; Big thanks, your accompanying makes this seven years seemly shorter.

Special thank-you to my past and current managers, Marieke de Kam, Erik Talens and Mark Schweitzer, I have learnt a great deal about distinct styles of pleasant leadership from each of you. Allegedly reality is overpopulated with "wrong-kind" managers, yet somehow I am lucky enough to always work with the good ones so far. Who want to do the math here? Hope the charm continues throughout my whole career.

Curriculum Vitae

Xiangyi Chen was born on October 13, 1987, in a southern coastal city of China, called Xiamen. He is known also as Gregory Chen in the industry or X. Gregory Chen in publication,

He obtained a Bachelor degree in 2010 from the School of Economics in Xiamen University, specializing in Statistics and Econometrics. Starting 2010 September, he studied in a research master program on Statistical Science for the Life and Behavioral Sciences in Mathematical Institute (MI), Leiden University, and obtained a degree of MSc. in Mathematics at 2012 August. Since 2013 May, he started as an external doctorate candidate in MI, Leiden University under the supervision of Prof. dr. A. W. van der Vaart. For the dissertation, he conducted researches on how to construct prediction sets via parametric and nonparametric Bayesian paradigm for different types of data and models, with applications in pharmaceutical industry.

Next to his academia journey, Gregory has already around 10 years working experience in the pharmaceutical industry, held roles of increasing responsibilities.

- During 2010 December to 2012 August, he started as a part-time assistant statistician, in a team of statisticians and pharmacometricians in Centre for Human Drug Research (CHDR) in Leiden, where he worked on several statistical projects in clinical phase I and IB.
- From 2012 November, he started a full-time position as a non-clinical statistician with MSD (d.b.a. Merck&Co. in USA and Canada) in Oss, the Netherlands, and promoted to a senior role in 2014 August. During the five years there, he worked in a team of statisticians, called Center for Mathematical Sciences, mainly to provide statistical consultancy and training to support regulatory filing or requests, development of laboratory analytical methods, in-vivo/in-vitro assays and manufacture processes, stability assessment and other issues for European sites, mostly for biologics and some exposure with small-molecule products.
- On 2017 September, he joined Novartis in its headquarter and moved to Basel, Switzerland, taking on a global role in quality division with strong strategic and later on leadership components, reporting to Global Head of Analytical Science and Technology and Global Head of QC. Among other responsibilities, he is the global lead of statistical affairs in laboratory testing and quality control across all product types and legacies, particularly for the commercial phase; lead or co-lead cross-divisional effort in authoring several global policies (e.g. standard operation

procedure) and own a sub-process in the quality system of lab operation; correspond with industrial associations on statistics-related regulations and inquiries; develop and maintain a pipeline of web-based tools (via R Shiny) for interactive and automated analysis, reporting and visualization for GMP and non-GMP purposes; deliver validation and deployment procedures for Shiny Apps that meets regulatory standard of computerized system verification (CSV) for GxP usage, via forging and leading cross-divisional collaboration; oversee best statistical practices and ensure adequate statistical capacity at manufacturing sites across the globe, via managing a network of regional representatives, mentoring a small group of technical associates, and offering training. He is also an proactive advocate for data-science-driven innovations in and outside of Novartis, providing counsel to senior business leaders on key digital initiatives or lighthouse projects, making proof-of-concept demo, verifying AI vendors, etc. He was the company representative in the digital work-stream in EFPIA during 2018-2019.

- Starting from 2021 March, he will re-join MSD in Zurich, Switzerland to practice statistics in the clinical sector, in the team of clinical safety statistics.