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 "geincorporeerd 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 hoofdtekst 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).