# Machine Learning Project

Hammerer Lena, Ibele Luisa, Janez Isabel, Romer Judith, Steinwender Hanna

2023-09-08

# Gliederung

- Motivation und Zielsetzung
- Vorgehen und Methodik
- Versuchsaufbau
- Bewertungsmatrix
- Umsetzung und Auswertung
- Schlussbetrachtung

Hanna Steinwender

# Motivation und Zielsetzung

# Motivation

- Entwicklung und Bewertung von Optimierungsalgorithmen in realen Anwendungen
- Herausforderung traditioneller Bewertung mit statischen Testdatensätzen
- Notwendigkeit von Testfunktionen für Algorithmen
- Schwierigkeit bei der Suche nach passenden Testfunktionen
- Begrenzte Verfügbarkeit von Ground-Truth-Funktionen
- Modelle mit realen Daten als Ersatz für Ground-Truth-Funktionen erstellen

# Zielsetzung

- Entwicklung von Testfunktionen
- Approximation von Ground-Truth-Funktionen
- Erfüllung von Anforderungen: Difficulty, Diversity, Flexibility, Relevance, Evaluation Cost, Non-Smoothing

**Forschungsfragen**

1. Ist der Einsatz eines Variational Autoencoder (VAE) als Datenerhebungstrategie sinnvoll?
2. Ist ein Deep Neural Network (DNN) basierend auf der erstellten Bewertungsmatrix geeignet zur Erzeugung der Testfunktion?
3. Ist der Einsatz eines DNNs auf Basis der Kriterien aus der Zielsetzung geeignet?

Hanna Steinwender

# Vorgehen und Methodik

# Vorgehen und Methodik

- Methodik: Empirische Untersuchung mit Experimenten


- Sampling aus Groundtruth

- Modelltraining
- Generalisierungstests
- Funktionsgenerierung

- Visuelle Vergleichsanalyse
- Evaluation

- Auswertung Matrix

Lena Hammerer

# Versuchsaufbau

# Ground-Truth

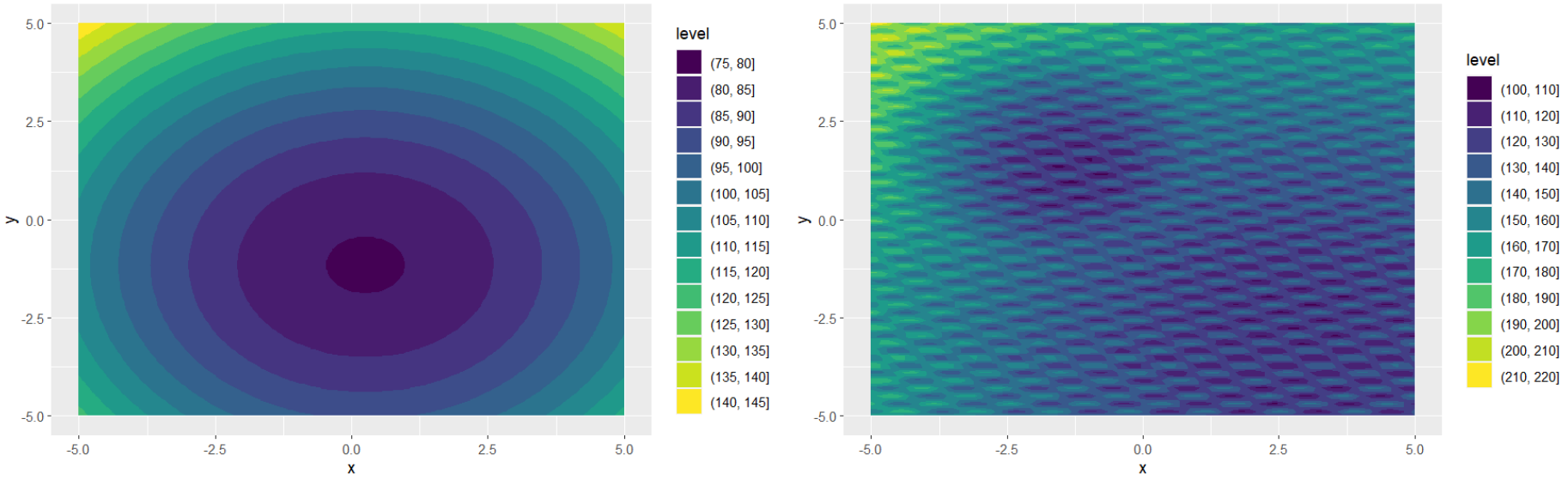
- *Wahre / korrekte* Vorhersage für ein gegebenes Problem
- Referenzpunkt, um die Leistung von ML-Modellen zu bewerten und ihre Qualität sicherzustellen
- Probleme in der Praxis:
  - (Noch) nicht verfügbar
  - Teuer in der Auswertung, eingeschränkte Experimentiermöglichkeiten
  - Vertraulich (Veröffentlichung nicht möglich)

Hansen, N., Auger, A., Mersmann, O., Tušar, T. and Brockhoff, D., 2016. COCO: A platform for comparing continuous optimizers in a black-box setting.

ArXiv e-prints. arXiv preprint arXiv:1603.08785, 172.

# Ground-Truth

- **f1 Sphere Function**
  - Einfachstes kontinuierliches Domänensuchproblem, da unimodal und hochsymmetrisch
- **f24 Lunacek bi-Rastrigin Function**
  - Hoch multimodal und trügerisch für evolutionäre Algorithmen mit großer Populationsgröße



# Parameter der Datenerhebung

Ground-Truth-Funktion f1 und f24 mit jeweils 2 oder 3 Dimensionen

```
numBbobf <- 1/24
dim <- 2/3
```

25 oder 600 Datenpunkten mit Random/Grid Sampling oder LHS

```
dataGenerationMethod <- "lhs", "random", "grid"
numDataPoints <- 25/600
```

Split von Trainings- und Evaluationsdaten für die Modellierung

```
trainTestSplit <- 0.8
```

# Datenerhebungstrategie

## Random Sampling

- Zufällige Werte für x und y generieren
- Reproduzierbar durch Random Seed
- Verteilung entspricht ggf. nicht den tatsächlichen Daten
- Erzeugung innerhalb von fixen Grenzen

## Grid Sampling

- Vordefinierte Menge von Parameterkombinationen
- Reproduzierbar, da jeder Punkt im Parameterraum einmal besucht
- Fluch der Dimensionen
- Besser geeignet für diskrete Parameter

# Datenerhebungstrategie

## Latin Hypercube Sampling

- Kombination der RS und GS Ansätze
- Parameterraum in gleich wahrscheinliche Intervalle oder Bins entlang jeder Dimension unterteilt
- Aus jedem Bin wird ein Zufallswert ausgewählt
- Gleichmäßige Verteilung von Stichprobenpunkte über den Parameterraum
- Repräsentativität wird erhöht

McKay, M.D., Beckman, R.J. and Conover, W.J., 2000. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code. Technometrics, 42(1), pp.55-61.

# Rechenressourcen

- CPU: Intel(R) Core(TM) i5-10300H CPU @ 2.5 GHz
- GPU: *nicht verfügbar*
- Memory: 16,0 GB RAM
- Storage: 0.5 TB SSD

# Modellaufbau DNN

## Deep Neural Network für Regression

- Manuelle Architektursuche unter Berücksichtigung limitierter Rechenressourcen

| | | |
|---|---|---|
| **Input Layer** | layer_dense(units=128, input_shape=2) | Leaky ReLU |
| **Hidden Layer** | layer_dense(units=32) | Leaky ReLU |
| **Hidden Layer** | layer_dense(units=128) | Leaky ReLU |
| **Dropout** | layer_dropout(rate=0.001) | - |
| **Hidden Layer** | layer_dense(units=64) | Leaky ReLU |
| **Output Layer** | layer_dense(units=1) | Linear |

# Modellaufbau DNN

## Mean Squared Logarithmic Error

- Besonders nützlich wenn Zielgrößen stark variieren

$$\text{MSE Log Error} = \frac{1}{N}\sum_{i=1}^{N}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2$$

## Adam Optimizer

- SGD-Methode, die auf der adaptiven Schätzung von Momenten erster und zweiter Ordnung beruht
- Gut geeignet für Probleme mit großer Anzahl von Daten/Parametern

Kingma, D.P. and Ba, J., 2014. Adam: A Method for Stochastic Optimization. arXiv preprint arXiv:1412.6980.

# Optimierung mit Differential Evolution

- Metaheuristik zur globalen Optimierung von Problemen
- Nutzt Population von Kandidatenlösungen, Mutation, Rekombination und Selektion, um schrittweise bessere Lösungen zu finden

**Populationsgröße**: Anzahl gleichzeitig betrachteter Lösungen

- Klein => schnell, aber ggf. vorzeitige Kovergenz zu lokalen Optima
- Groß => ggf. bessere Lösungen, aber mehr Rechenleistung

```
popSize = 4          # Vergleich zu gradientenbasierten Verfahren
popSize = 10\*dim    # Default
popSize = 20\*dim

funEval <- 200/400
```

Storn, R. and Price, K., 1997. Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization, 11, pp.341-359.

Luisa Ibele

# Bewertungsmatrix

# Grundlegendes Bewertungsschema

- Grundlegende Parameter der einzelnen Versuchsdurchläufe:

```
numDataPoints <- 25/600
dataGenerationMethod <- "lhs", "random", "grid"
numBbobf <- 1/24
dim <- 2/3
```

- Bewertung der Modelle und Optimierung anhand verschiedener Kriterien

# Grundlegendes Bewertungsschema

- Bewertungsskala von 1 bis 5 Punkten mit:
  - 1: sehr schlecht, schwach
  - 2: schlecht, unterhalb des Erwarteten
  - 3: durschnittlich, akzeptabel
  - 4: gut, über dem Erwarteten
  - 5: sehr gut, ausgezeichnet

# Modellbewertung

| Gewichtung | 0,1 | 0,3 | 0,4 | 0,2 | |
|---|---|---|---|---|---|
| Methode | Loss Value | Loss Function Verlauf | Generated Function vs. Original Function | Performance | Ergebnis |
| Grid Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |
| Random Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |
| Latin Hypercube Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |

# Gewichtung

- Durchschnittlicher Trainingsloss - 10%
- (Visual) Loss Function Verlauf - 30%
- (Visual) Generated Function vs. Original Function - 40%
- Performance - 20%

# Optimierungsbewertung

| | Gewichtung | 0,3 | 0,2 | 0,2 | 0,3 | |
|---|---|---|---|---|---|---|
| Popsize | Methode | Optimum Wert | Performance | error/ evaluations | y/ evalutions | Ergebnis |
| 4 | Grid Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |
| 4 | Random Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |
| 4 | Latin Hypercube Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |
| Popsize | Methode | Optimum Wert | Performance | error/ evaluations | y/ evalutions | Ergebnis |
| 10*dim | Grid Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |
| 10*dim | Random Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |
| 10*dim | Latin Hypercube Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |
| Popsize | Methode | Optimum Wert | Performance | error/ evaluations | y/ evalutions | Ergebnis |
| 20*dim | Grid Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |
| 20*dim | Random Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |
| 20*dim | Latin Hypercube Sampling | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | Beschreibung \1-5 | \1-5 |

# Gewichtung

- Optimum Wert - 30%
- Performance - 20%
- Error / Evaluations - 20%
- Y / Evaluations - 30%

# Finale Bewertungsmatrix

| Gewichtung | | 0,3 | 0,7 | | | |
|---|---|---|---|---|---|---|
| Versuch | Modellentwicklung | | Optimierung | Gesamt | | |
| Random_4_f1 | | | | 0 | | |
| Random_10xdim_f1 | | | | 0 | | |
| Random_20xdim_f1 | | | | 0 | | |
| Grid_4_f1 | | | | 0 | | |
| Grid_10xdim_f1 | | | | 0 | | |
| Grid_20xdim_f1 | | | | 0 | | |
| LHS_4_f1 | | | | 0 | | |
| LHS_10xdim_f1 | | | | 0 | | f1_best |
| LHS_20xdim_f1 | | | | 0 | | f1_worst |
| | | | | | | |
| Random_4_f24 | | | | 0 | | |
| Random_10xdim_f24 | | | | 0 | | |
| Random_20xdim_f24 | | | | 0 | | |
| Grid_4_f24 | | | | 0 | | |
| Grid_10xdim_f24 | | | | 0 | | |
| Grid_20xdim_f24 | | | | 0 | | |
| LHS_4_f24 | | | | 0 | | |
| LHS_10xdim_f24 | | | | 0 | | f24_best |
| LHS_20xdim_f24 | | | | 0 | | f24_worst |

# Gesamtbewertung

- Gewichtung:
  - Modellbewertung - 30%
  - Optimierungsbewertung - 70%

- Systematische Analyse und Bewertung der Versuche im Rahmen der Modell- und Optimierungsbewertung
- Fundierte Beurteilung der Modellleistung, Optimierungsalgorithmen und Versuche anhand der gewählten Bewertungskriterien
- Berücksichtigung qualitativer und quantitativer Aspekte
- Bewertung des gesamten Versuchs möglich

# Versuchsbewertung

Bewertung der einzelnen Versuche anhand von:

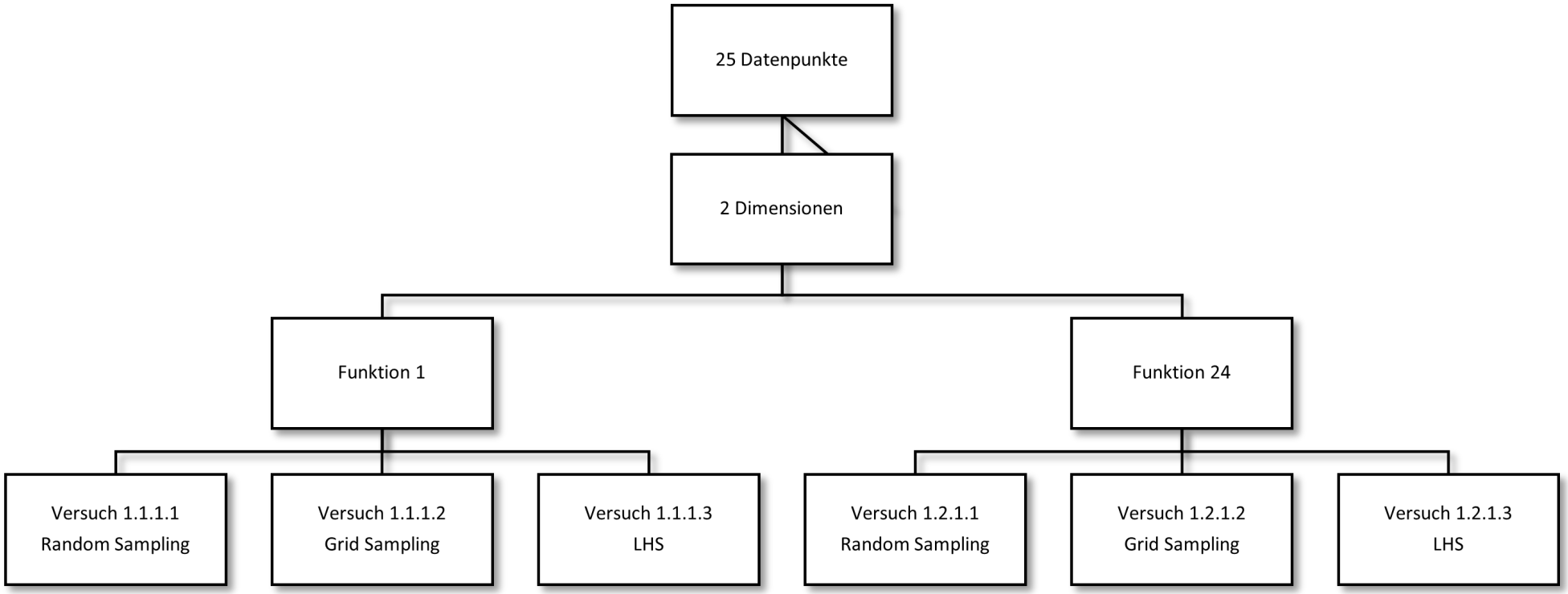
- Difficulty
- Diversity
- Flexibility
- Relevance
- Evaluation cost
- Non-Smoothing

Zaefferer, M., Fischbach, A., Naujoks, B. and Bartz-Beielstein, T., 2017, July. Simulation-based Test Functions for Optimization Algorithms. In Proceedings of the genetic and evolutionary computation conference (pp. 905-912).

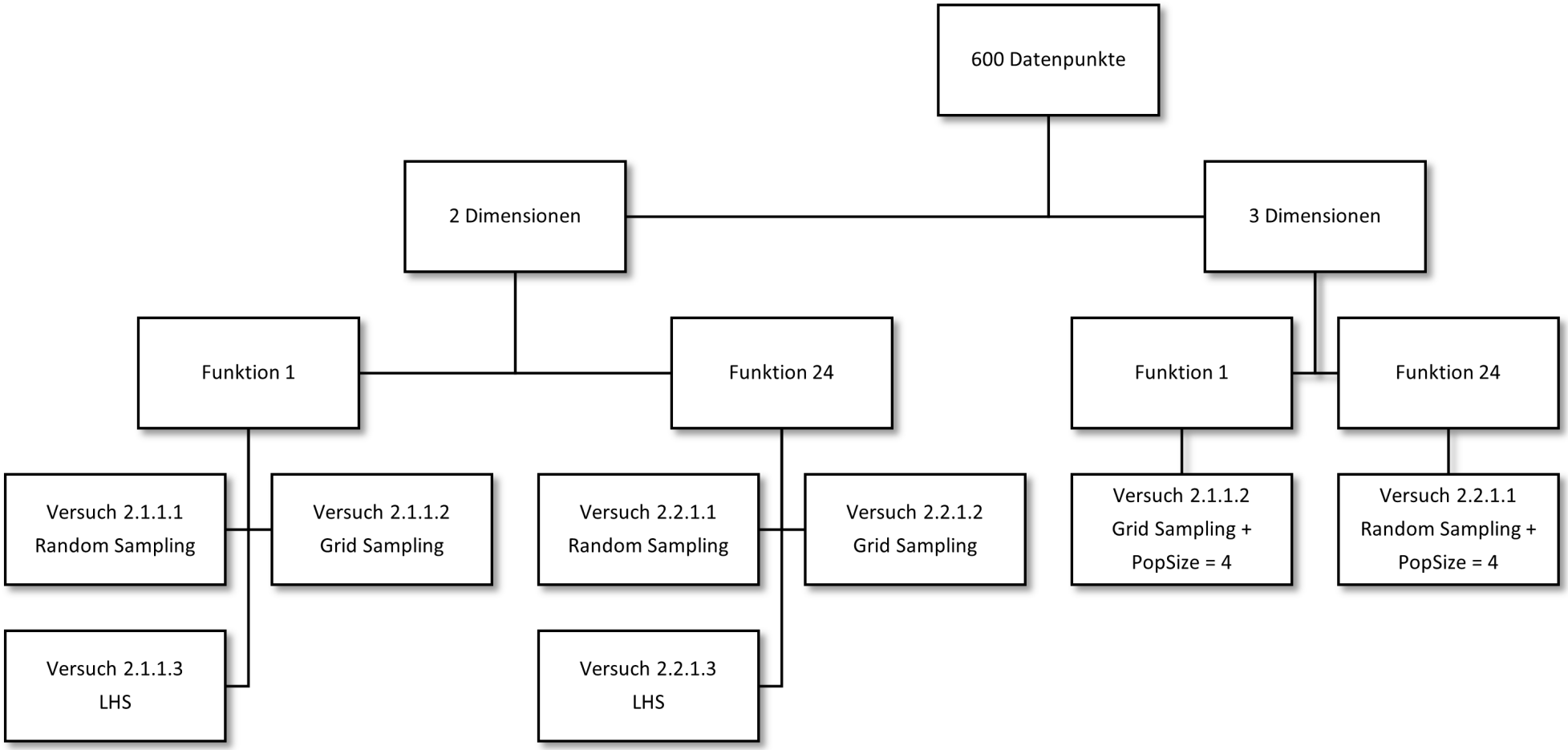
Luisa Ibele

# Umsetzung

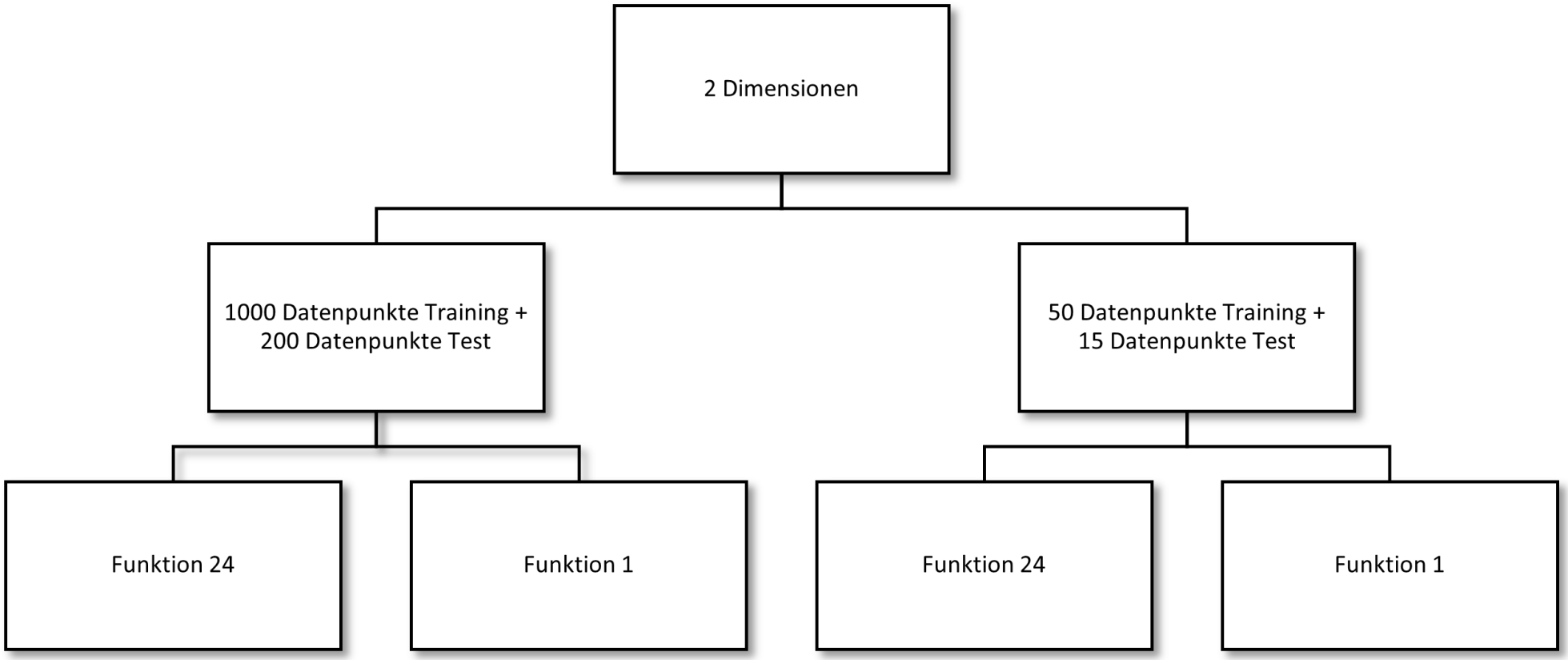
# Übersicht 1. Experiment



# Übersicht 2. Experiment



# Übersicht 3. Experiment

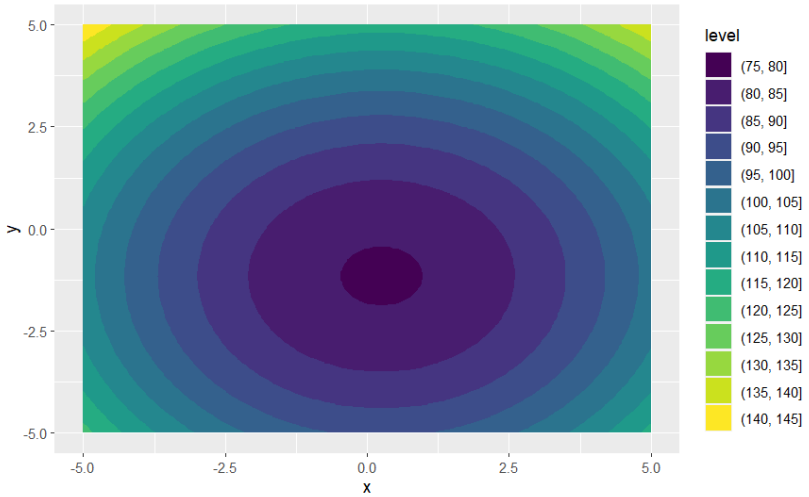


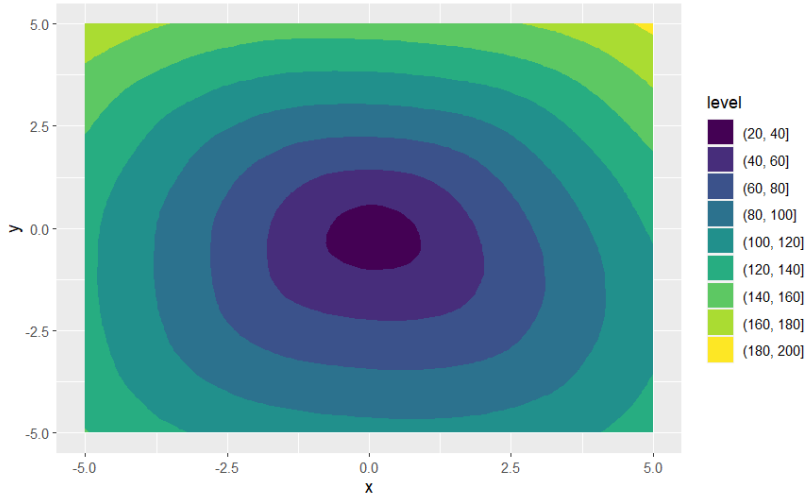
# Auswertung

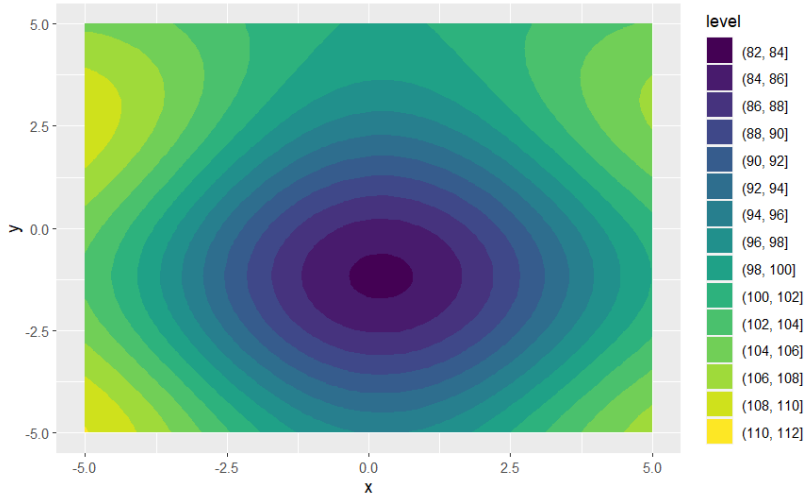
Judith Romer

# Eperiment 1

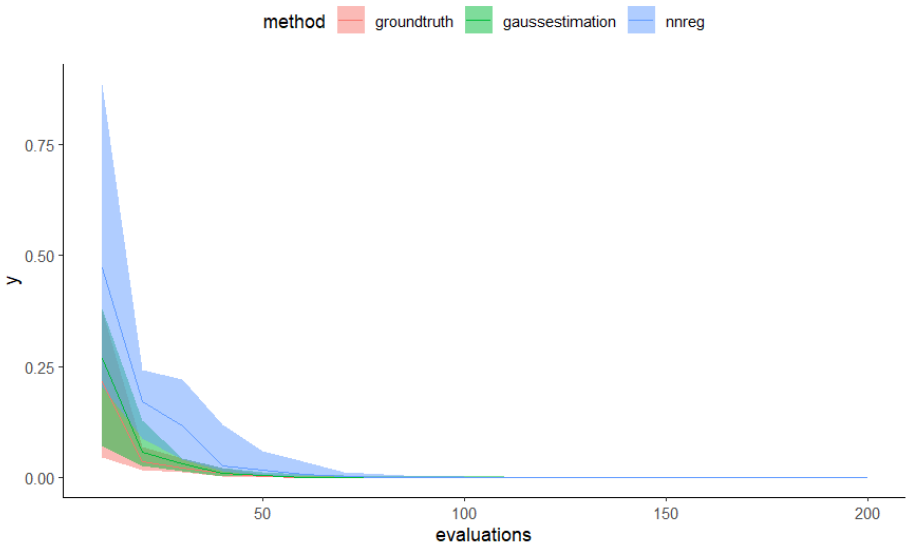
# Versuch 1.1.1.2 - Bildvergleich

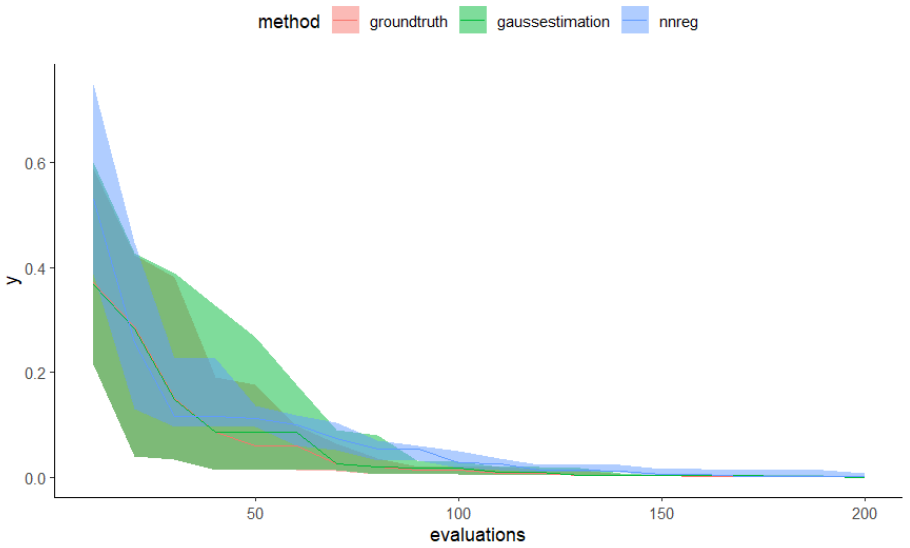


# Versuch 1.1.1.2 - Y Wert / Evaluation



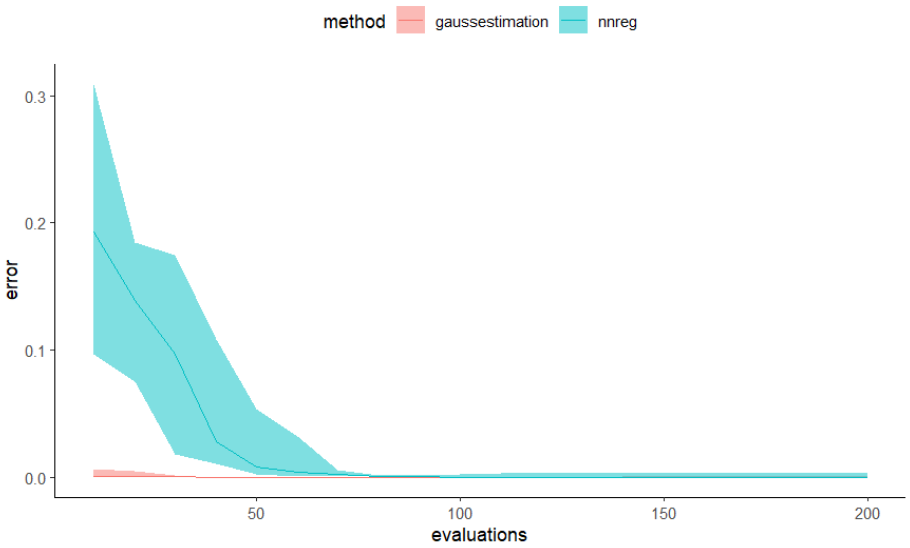

Population Size = 4

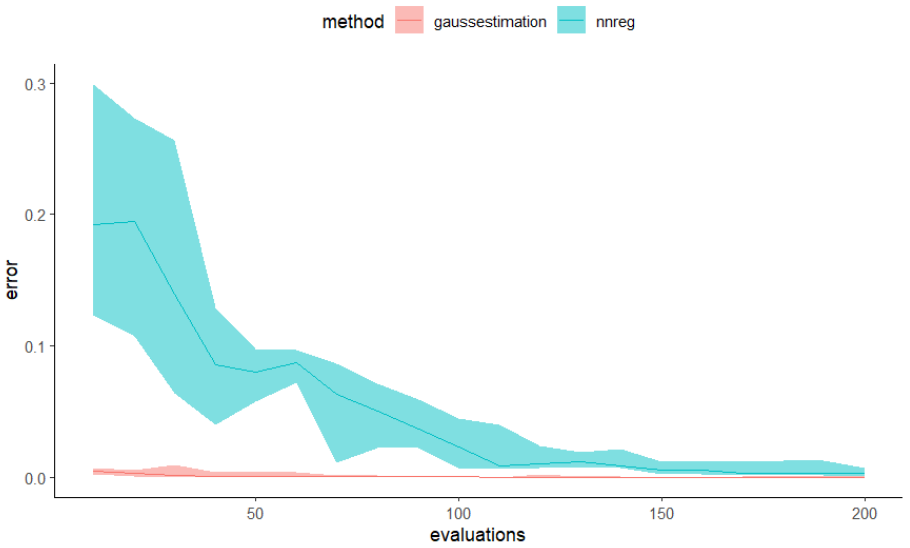
- Groundtruth (GT) und Gaußsche Prozessmodelle (GPM): gute Optimierung
- DNN: mehr Evaluationen

Population Size = 10 * Dimension

- GT und GPM: im Median gut
- DNN: Optimierung deutlich schlechter

# Versuch 1.1.1.2 - Fehler / Evaluation




Population Size = 4

- GT und GPM: Fehler von Anfang an sehr niedrig
- DNN: Fehler am Anfang sehr hoch, jedoch zuverlässige Verringerung

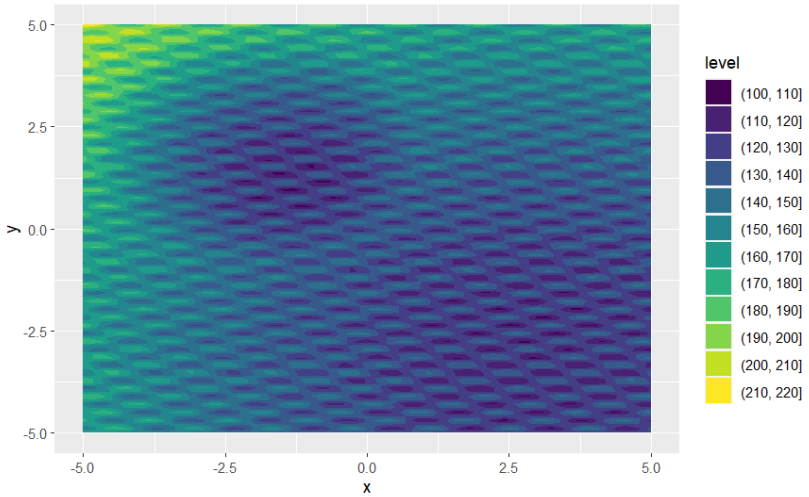
Population Size = 10 * Dimension

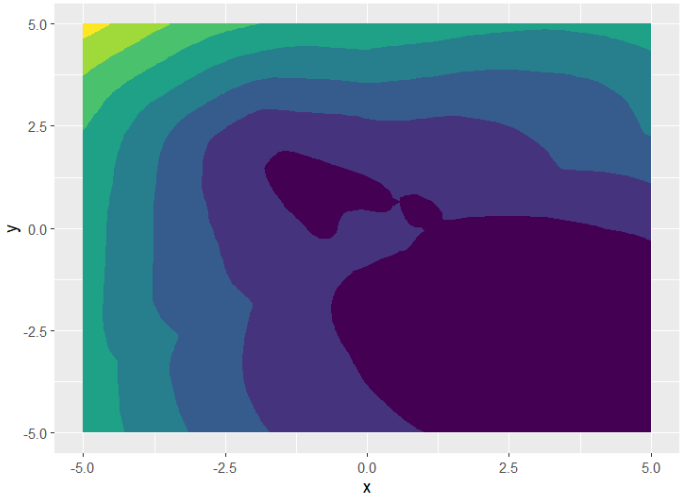
- GT und GPM: Fehler von Anfang an sehr niedrig
- DNN: selbst nach 200 Evaluationen noch schlechter als GPM

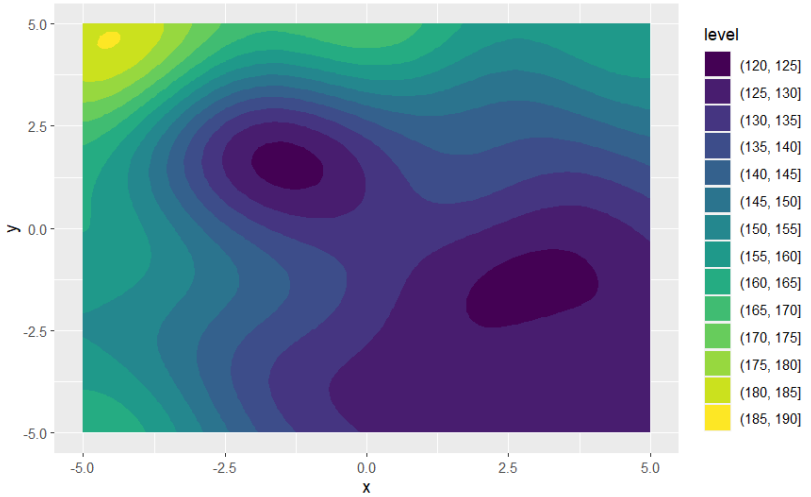
Judith Romer

# Experiment 2

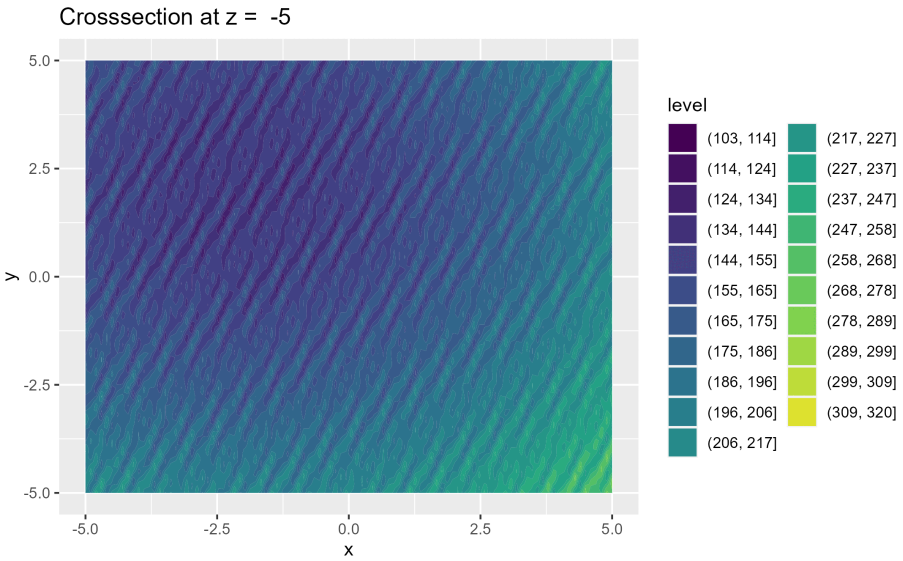
# Versuch 2.2.1.1 - Bildvergleich 2D

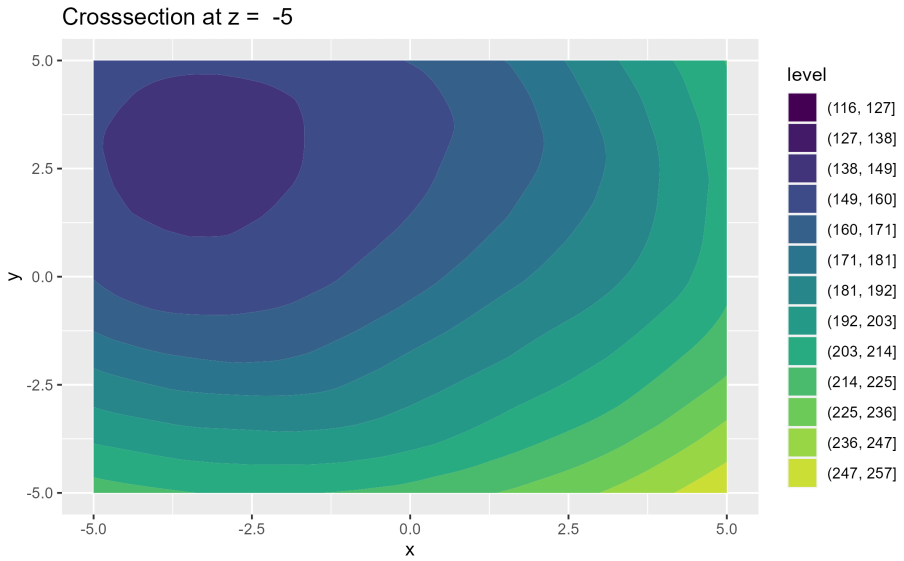


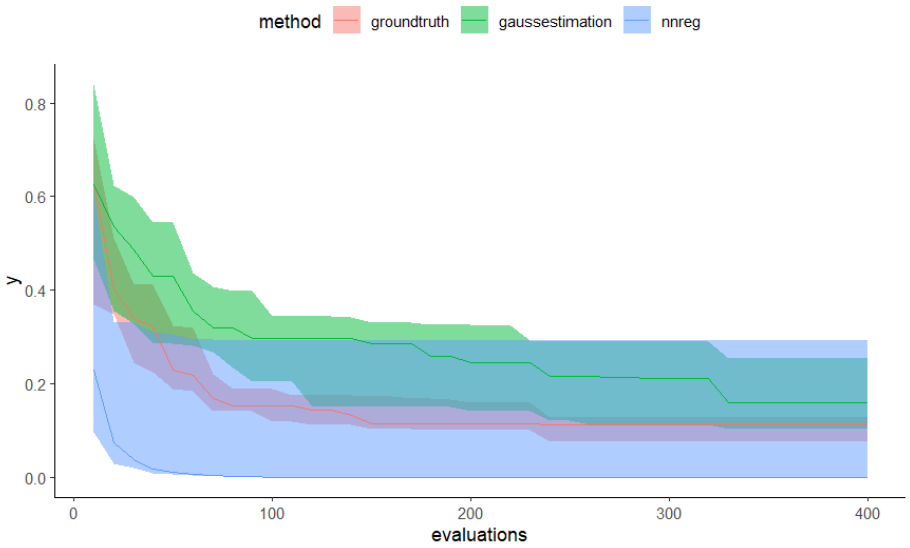



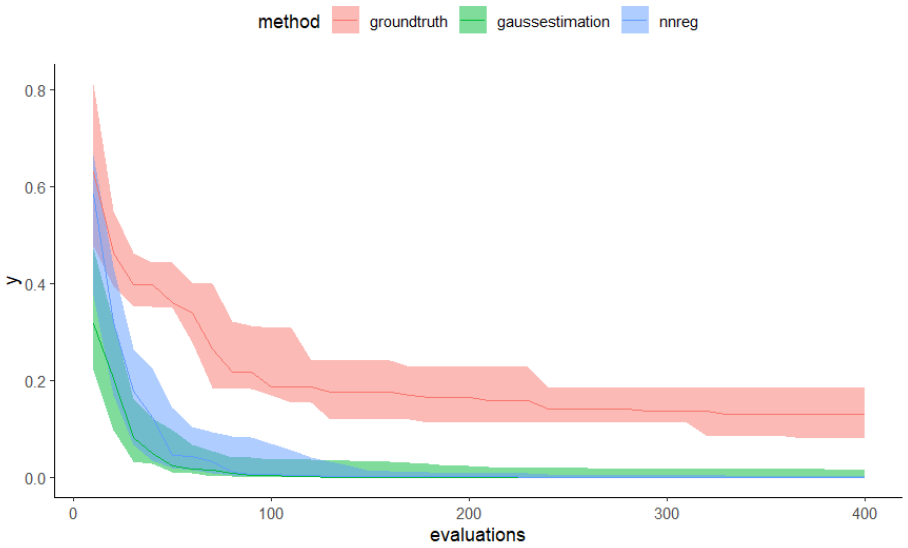
# Versuch 2.2.2.1 - Bildvergleich 3D



# Versuch 2.2.1.1 & 2.2.2.1 - Y Wert / Evaluation



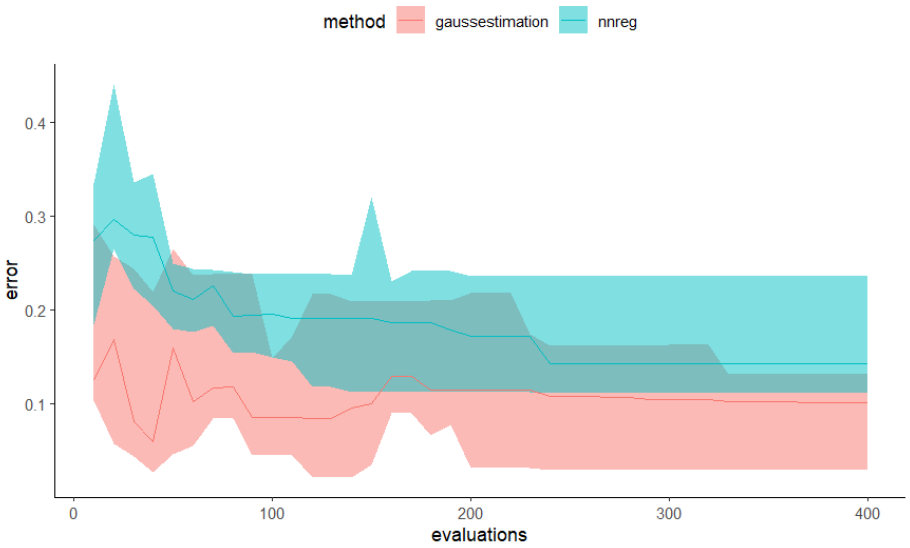

## 2 Dimensionen

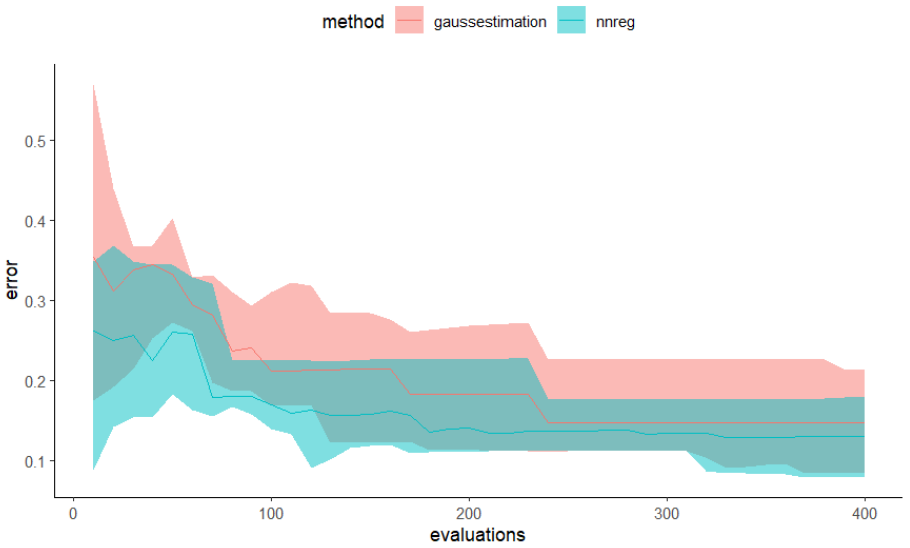
- GT und GP: Minimum nach 400 Evalutationen nicht gefunden
- DNN: im Median in Ordnung, aber Durchläufe in denen nicht optimiert wird

## 3 Dimensionen

- GT: Minimum nach 400 Evaluationen nicht gefunden
- GPM und DNN: ähnlich gut

# Versuch 2.2.1.1 & 2.2.2.1 - Fehler / Evaluation




2 Dimensionen

- GPM: in manchen Durchläufen teils nache 0
- DNN: deutlich höherer Fehler

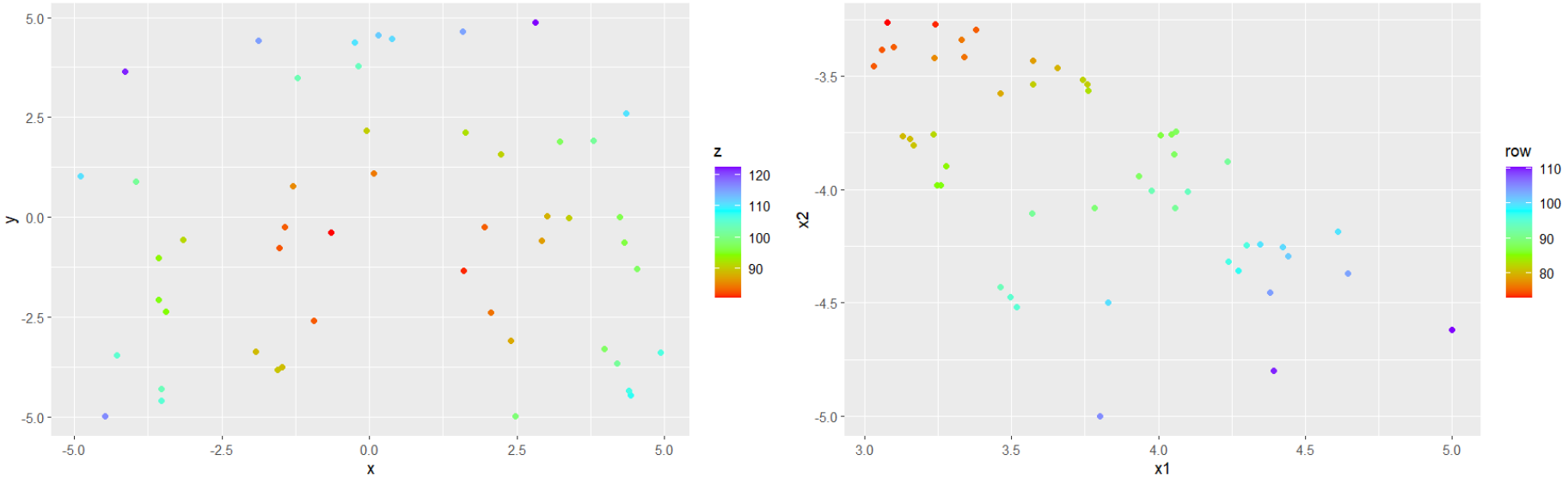
3 Dimensionen

- GPM und DNN: deutlich höherer Fehler

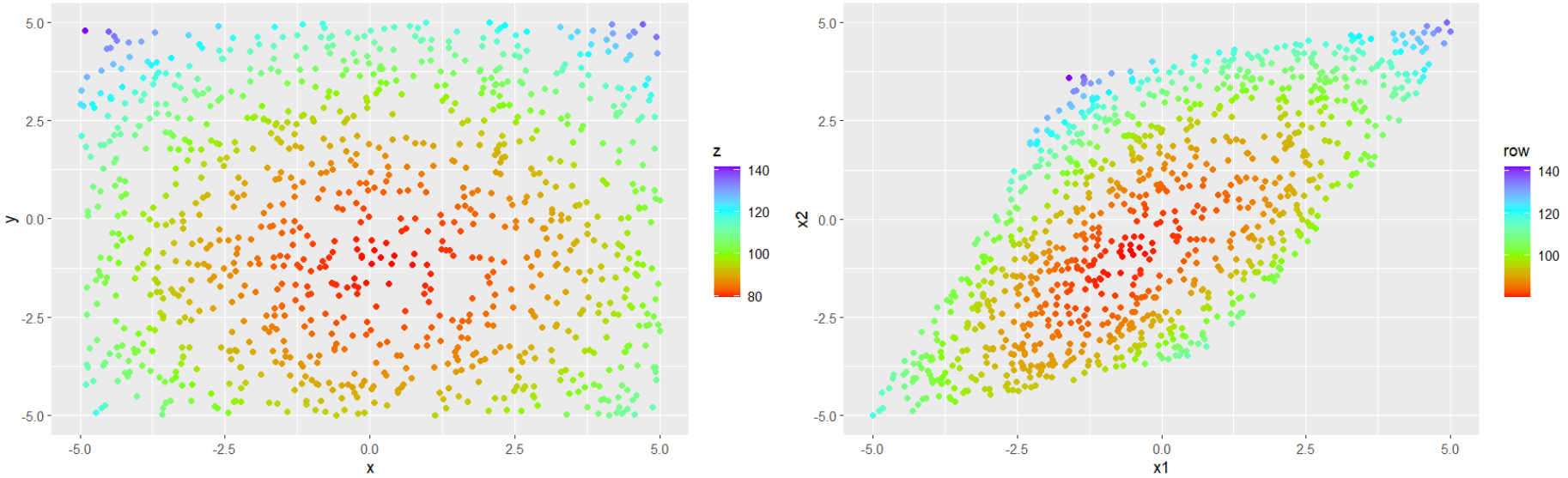
Isabel Janez

# Experiment 3

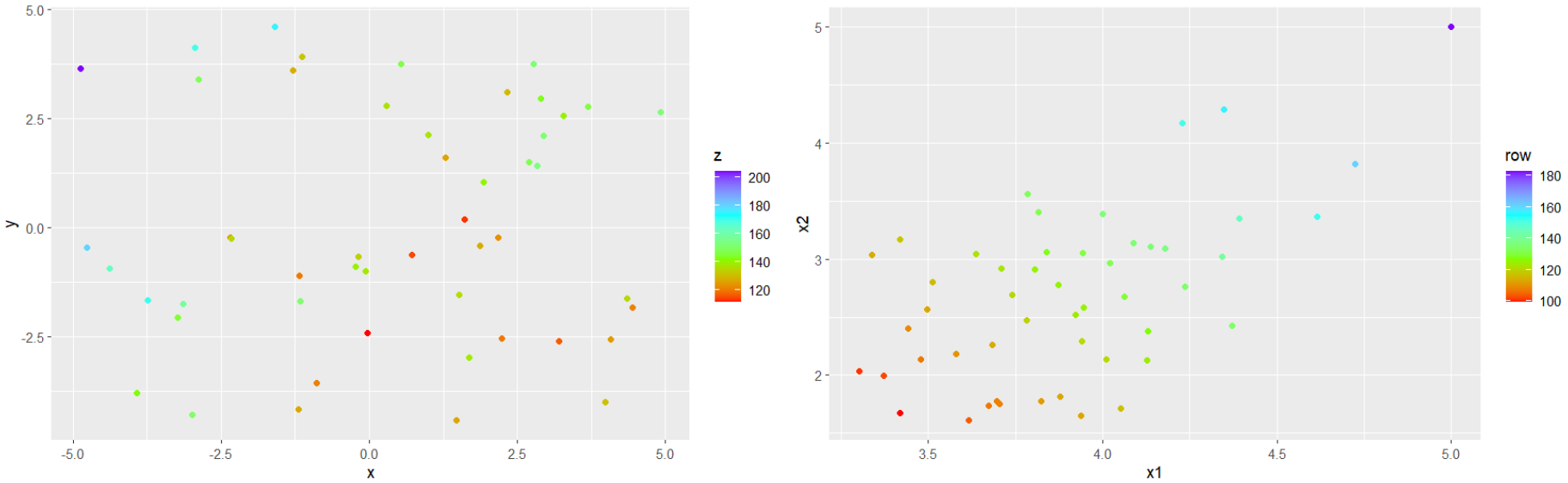
# VAE mit 50 Datenpunkten Funktion 1



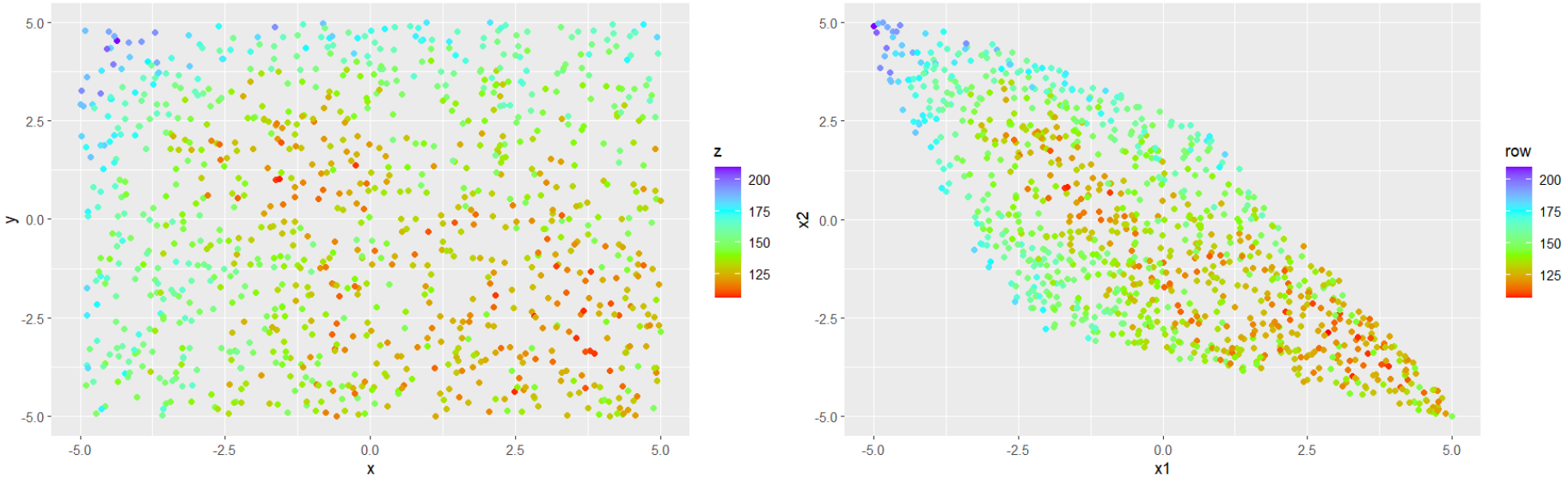
# VAE mit 1000 Datenpunkten Funktion 1



# VAE mit 50 Datenpunkten Funktion 24



# VAE mit 1000 Datenpunkten Funktion 24



Isabel Janez

# Beantwortung der Forschungsfragen

# Ist der Einsatz eines VAE als Datenerhebungstrategie sinnvoll?

- Versuch wurde abgebrochen → sehr schlechte Ergebnisse vor allem bei 50 Datenpunkten
- Nicht in der Lage die Verteilung über gesamten Raum zu lernen
- Datenpunkteskala passt nur bedingt
- Repräsentiert die ‘Realität’ nicht ausreichend
- Dimensionsreduktion führt zu Problemen, dadurch verzerrte / gestauchte Darstellung

# Ist der Einsatz eines VAE als Datenerhebungstrategie sinnvoll?

Für das vorliegende Problem konnte der VAE nicht zur Datenerhebung eingesetzt werden

Weiteres Vorgehen:

- Modellarchitektur überarbeiten
- Generative Adversial Network (GAN)
- Conditional Variational Autoencoder (CVAE)

# Ist ein DNN basierend auf der erstellten Bewertungsmatrix geeignet zur Erzeugung der Testfunktion?

- Threshold: 4 / 5
- DNN = 4 → gleich gut wie GPM
- DNN > 4 → besser als GPM

- Experiment 1: 2,13 / 5
- Experiment 2: 3 / 5

Im vorliegenden Fall: DNN liegt unter Threshold, somit nicht geeignet

# Ist der Einsatz eines DNNs auf Basis der Kriterien aus der Zielsetzung geeignet?

Threshold: 3 / 5

**[4|5] Difficulty**

**[3|5] Diversity**

**[3|5] Flexibilty**

**[1|5] Relevance**

**[1|5] Evaluation cost**

**[1|5] Non-Smoothing**

---

→ DNN: 2,2 / 5

Im vorliegenden Fall: DNN liegt unter Threshold, somit nicht geeignet

Isabel Janez

# Schlussbetrachtung

# Fazit

- Gaußsche Prozessmodelle im vorliegenden Anwendungsfall besser geeignet
- Anmerkung: es wurde nur die Estimation verglichen; kein Vergleich von Simulation
- Tradeoff zwischen Rechenleistung und Komplexität des DNN

# Ausblick

- Abbildung der lokalen Struktur mit dem neuronalen Netz durch Erstellung neuer Loss-Funktion, die nicht nur einzelne Punkte optimiert, sondern mehrere Punkte betrachtet
- Vergleich von Gaußscher Prozessmodell Simulation anstelle von Estimation

# Vielen Dank für die Aufmerksamkeit! Noch Fragen?

# Prüfungsleistung Aufteilung

- Hanna Steinwender: Folie 1 bis Folie 7
- Lena Hammerer: Folie 8 bis Folie 17
- Luisa Ibele: Folie 18 bis Folie 31
- Judith Romer: Folie 32 bis Folie 41
- Isabel Janez: Folie 42 bis 55