### User Guide: Advanced Data Analysis and Visualization Application

This application is for analyzing experimental data using Gaussian Process Regression (GPR), Bayesian Optimization (BO) visualization, violin plots, and statistical analysis. Below is a detailed guide on how to use the application and its technical aspects.

### **1. Required Python Imports**

Before running the application, ensure you have these dependencies installed:

python

import tkinter as tk

from tkinter import filedialog, messagebox, simpledialog, ttk, scrolledtext

from tkinter.simpledialog import Dialog

import numpy as np

import matplotlib

import matplotlib.pyplot as plt

from matplotlib.backends.backend\_tkagg import FigureCanvasTkAgg, NavigationToolbar2Tk

from matplotlib.figure import Figure

from matplotlib.animation import FuncAnimation, FFMpegWriter

from sklearn.gaussian\_process import GaussianProcessRegressor

from sklearn.gaussian\_process.kernels import RBF, ConstantKernel as C, Matern, RationalQuadratic, ExpSineSquared, DotProduct

from sklearn.preprocessing import StandardScaler

from sklearn.decomposition import PCA

import pandas as pd

from matplotlib.colors import LinearSegmentedColormap

import matplotlib.patches as mpatches

import seaborn as sns

import os

import time

import re

import traceback

import io

from scipy import stats

from sklearn.metrics.pairwise import cosine\_similarity

from sklearn.preprocessing import MinMaxScaler

# Use TkAgg backend for interactive plots

matplotlib.use('TkAgg')

### **2. Application Overview**

The application features 5 tabs:

1. **Gaussian Process Regression (GPR)**: For regression modeling
2. **Bayesian Optimization Visualization**: For BO process animation
3. **Data Selection**: For dataset preparation
4. **Violin Plots**: For distribution visualization
5. **Statistics**: For statistical analysis

### **3. Step-by-Step Usage Guide**

#### **Step 1: Load Data**

1. Go to the **Data Selection** tab
2. Click "Browse CSV" to load your experimental data
3. Preview your data in the text area
4. Define up to 4 datasets by clicking "Select Dataset X" buttons
5. Generate a combined "Total" dataset if needed

#### **Step 2: Gaussian Process Regression (GPR)**

1. Select a dataset from the dropdown
2. Choose X and Y columns
3. Select a kernel type (e.g., Matern 5/2)
4. Click "Run GPR" to generate the regression model
5. Use toolbar to zoom/pan, "Reset View" to restore
6. "Save Plot" to export the visualization

#### **Step 3: Bayesian Optimization Visualization**

1. Select a dataset
2. Specify input/output columns (comma-separated)
3. Set weights for output combination (e.g., "1.0,0.5,1.0")
4. Click "Visualize BO" to generate the PCA plot
5. Use "Start Animation" to see BO progression
6. Adjust star offset and animation speed as needed
7. "Save Animation" to export as MP4/GIF  
   (sort of broken, especially with visualization, will have to restart application)

#### **Step 4: Violin Plots**

1. Select variables from the list (Ctrl+click for multiple)
2. Click "Generate Plots" to create distribution visualizations
3. Use "Save Plot" to export the figure (very broken)

#### **Step 5: Statistics**

1. View dataset statistics in the dataset-specific tabs
2. For statistical tests:
   * Select a variable
   * Click "Run Statistical Tests"
   * Results show significant differences between datasets
3. For cosine similarity:
   * Click "Calculate Cosine Similarity"
   * Results show distribution similarity (1.0 = identical)

### **4. Key Technical Features**

#### **Gaussian Process Regression**

* **Kernel Options**: RBF, Matern, RationalQuadratic, Periodic, Linear
* **Data Handling**: Automatic NaN removal and numeric conversion
* **Visualization**: Confidence intervals, interactive navigation
* **Optimization**: Multiple restarts for robust kernel parameter fitting

#### **Bayesian Optimization**

* **PCA Projection**: Visualizes high-dimensional inputs in 2D
* **Score Calculation**: Weighted combination of normalized outputs
* **Animation**: Progressive display of BO iterations
* **Visual Cues**: Gold star marks best point, color gradient shows score quality

#### **Data Processing**

* **Automatic Cleaning**: Handles percentages, commas, and missing values
* **Dataset Management**: Create up to 4 subsets + combined "Total" dataset
* **Numeric Detection**: Smart column type identification

#### **Statistical Analysis**

* **Kruskal-Wallis Test**: Non-parametric ANOVA replacement
* **Post-Hoc Analysis**: Pairwise comparisons with Bonferroni correction
* **Cosine Similarity**: Compares distribution shapes using histogram vectors
* **Violin Plots**: Shows distribution, quartiles, and observation counts

### **5. Troubleshooting Tips**

1. **No Numeric Columns Found**:
   * Ensure your CSV has numeric data
   * Check for consistent formatting (no mixed types)
   * Remove percentage signs if not needed
2. **Animation Not Saving**:
   * Install FFmpeg: **conda install -c conda-forge ffmpeg**
   * For GIFs: **pip install pillow**
3. **GPR Failing**:
   * Try different kernels
   * Increase optimization restarts
   * Ensure Y-column has variation (not constant)
4. **Large Datasets**:
   * Use "Max Points" to limit GPR samples
   * Subset data using the Data Selection tab

### **6. Exporting Results**

* **Plots**: PNG, PDF, SVG formats
* **Animations**: MP4 (video) or GIF formats
* **Statistics**: Copy/paste from results text areas

### **7. Technical Notes**

* **Data Scaling**: Inputs/outputs are automatically standardized
* **PCA Consistency**: Uses same projection for animation frames
* **Memory Management**: Large datasets are downsampled for visualization
* **Error Handling**: Detailed tracebacks provided in error messages