ECE 271A: Statistical Learning I Quiz Report

JunPyung Kim
PID: A69044427
University of California, San Diego

December 18, 2025

1 Quiz 6: Gaussian Mixture Models for Cheetah Segmentation

1.1 Objective

In this quiz we revisit the cheetah image segmentation task using Gaussian mixture models (GMMs) estimated by the EM algorithm. The goals are:

- to study how sensitive the EM-trained GMMs are to random initialization when the number of components is fixed at $C = 8$, and
- to analyze how the number of mixture components $C \in \{1, 2, 4, 8, 16, 32\}$ affects the probability of error as we vary the feature dimension $d \in \{1, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64\}$.

1.2 Methodology

1.2.1 Data and Feature Extraction

The training data are the DCT features provided in `TrainingSamplesDCT_8_new.mat`. There are 250 foreground (FG) samples and 1053 background (BG) samples, so the empirical class priors are

$$P(Y = \text{FG}) = \frac{250}{250 + 1053} \approx 0.192, \qquad P(Y = \text{BG}) = \frac{1053}{250 + 1053} \approx 0.808.$$

The test image is the cheetah image `cheetah.bmp`. It is converted to grayscale and processed with a sliding 8×8 window (stride 1) over the entire 255×270 image. For each block we compute the 8×8 2-D DCT and then reorder the 64 coefficients using the zig–zag pattern given in `Zig-Zag Pattern.txt`. This produces 68850 feature vectors $\mathbf{x} \in \mathbb{R}^{64}$, one for each pixel position. The ground-truth mask is obtained from `cheetah_mask.bmp` by thresholding at 0.5.

1.2.2 Gaussian Mixture Model

For each class $Y \in \{\text{FG}, \text{BG}\}$ we model the conditional density of the full 64-dimensional feature vector as a mixture of C Gaussians with diagonal covariance matrices:

$$p(\mathbf{x} \mid Y = i) = \sum_{c=1}^{C} \pi_{i,c} \mathcal{N}\big(\mathbf{x}; \boldsymbol{\mu}_{i,c}, \Sigma_{i,c}\big), \tag{1}$$

where $\sum_c \pi_{i,c} = 1$ and $\Sigma_{i,c} = \text{diag}(\sigma^2_{i,c,1}, \dots, \sigma^2_{i,c,64})$.

The models are trained with the EM algorithm. For a given class and C:

- **Initialization:** The means $\boldsymbol{\mu}_{i,c}$ are initialized by selecting random training samples and adding small noise. All mixture weights are set to $\pi_{i,c} = 1/C$, and the initial diagonal variances are copied from the global variance of the training set (replicated across components).
- **E-step:** For each sample $\mathbf{x}_n$ and component c we compute the log responsibility $\log \gamma_{n,c} = \log p(c \mid \mathbf{x}_n, Y = i)$ using the current parameters. This uses the diagonal Gaussian log-likelihood together with the mixture weights.
- **M-step:** From the responsibilities $\gamma_{n,c}$ we update

$$N_c = \sum_n \gamma_{n,c}, \quad \pi_{i,c} = \frac{N_c}{N},$$

$$\boldsymbol{\mu}_{i,c} = \frac{1}{N_c} \sum_n \gamma_{n,c} \mathbf{x}_n, \qquad \sigma^2_{i,c,d} = \frac{1}{N_c} \sum_n \gamma_{n,c} (x_{n,d} - \mu_{i,c,d})^2 + \varepsilon,$$

  where a small ε ensures numerical stability.

Iterations stop when the log-likelihood improvement becomes smaller than a tolerance threshold, or when the maximum number of iterations is reached.

1.2.3 Bayes Decision Rule and Error Computation

Once the FG and BG mixtures have been trained, we classify each image block using the Bayes decision rule

$$g(\mathbf{x}) = \log p(\mathbf{x} \mid Y = \mathrm{FG}) + \log P(Y = \mathrm{FG}) - \log p(\mathbf{x} \mid Y = \mathrm{BG}) - \log P(Y = \mathrm{BG}).$$

The pixel is labeled as cheetah if $g(\mathbf{x}) > 0$ and as grass otherwise. To study the effect of the feature dimensionality d, we always train the mixtures in the full 64-dimensional space, but at test time we keep only the first d zig–zag coefficients. This is implemented by restricting the mean and variance vectors to their first d entries when computing the log-likelihood.

Let $\hat{Y}$ denote the classifier output and Y the true class (from the mask). The probability of error is estimated as

$$P_e = P(\hat{Y} \neq Y) = P(\hat{Y} \neq Y \mid Y = \mathrm{FG}) P(Y = \mathrm{FG}) + P(\hat{Y} \neq Y \mid Y = \mathrm{BG}) P(Y = \mathrm{BG}).$$

The conditional error terms correspond to the proportion of misclassified foreground and background pixels in the ground-truth mask.

1.3 Results and Discussion

1.3.1 Part (a): Effect of Initialization for $C = 8$ Components

For Part (a) we fix $C = 8$ components for both classes and train 5 independent mixtures for FG and 5 for BG using different random initializations. Combining them yields 25 classifier pairs; for each pair we compute the probability of error for all dimensions d in the set $\{1, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64\}$. The resulting curves are shown in Figure 1.

Several qualitative trends can be observed:

- For all initialization pairs the error is highest when only the first DCT coefficient is used ($d = 1$) and remains relatively large for very low dimensions ($d \leq 8$), since a single or a few coefficients cannot capture the structure of the two classes.

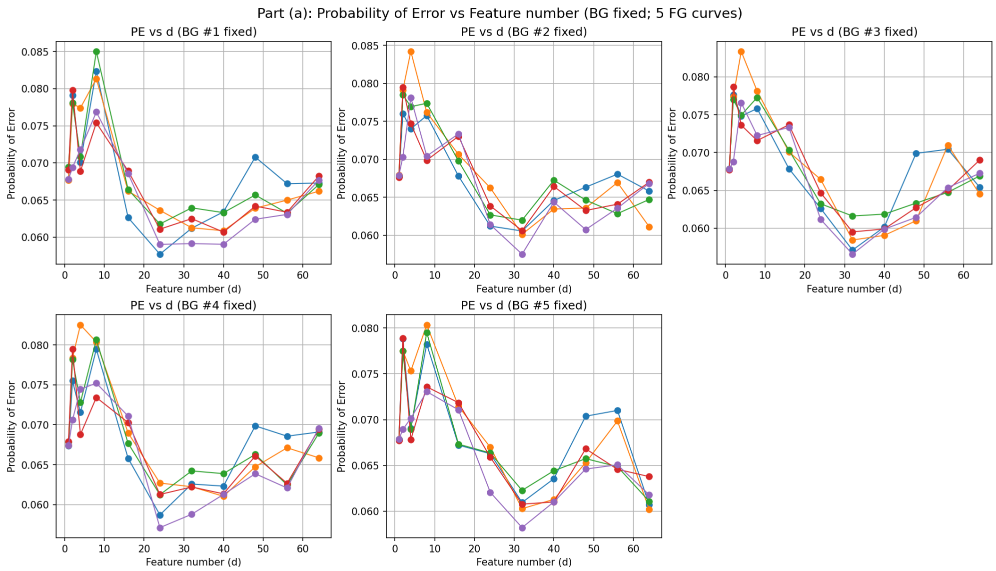

Figure 1: Part (a): probability of error vs. feature dimension for the 25 pairs of FG/BG mixtures obtained with different random initializations (blue curves) and their average (red curve).

- As more coefficients are included (d in the range roughly 16–40), the error drops significantly. The average curve shows a clear improvement when going from a handful of coefficients to a moderate-dimensional descriptor, reflecting that the mid-frequency DCT features are important for discriminating cheetah texture from grass.
- Beyond some dimension, gains become smaller and the curves tend to flatten. Some initializations even exhibit a slight increase in error at the largest dimensions, consistent with adding noisy or less informative coefficients and increasing estimation variance.
- The spread between the 25 curves is noticeable but not extreme: all runs follow the same global shape and have similar minima. This indicates that EM is somewhat sensitive to initialization (local maxima of the likelihood) but that the overall classifier performance is relatively robust: different initializations rarely change the error by more than a few percentage points.

Overall, Part (a) demonstrates that, although GMM training with EM does depend on the random start, the dominant factor affecting the probability of error is the feature dimension d: using only a very small number of coefficients is clearly sub-optimal, while using a moderate number of DCT features significantly improves segmentation quality.

1.3.2 Part (b): Effect of the Number of Components

For Part (b) we fix the initialization strategy and vary the number of mixture components

$$C \in \{1, 2, 4, 8, 16, 32\}$$

separately for FG and BG. For each C we train one GMM per class on the full 64-dimensional training data and then evaluate the probability of error as a function of the dimension d. The curves are shown in Figure 2.

Some observations based on the values are:

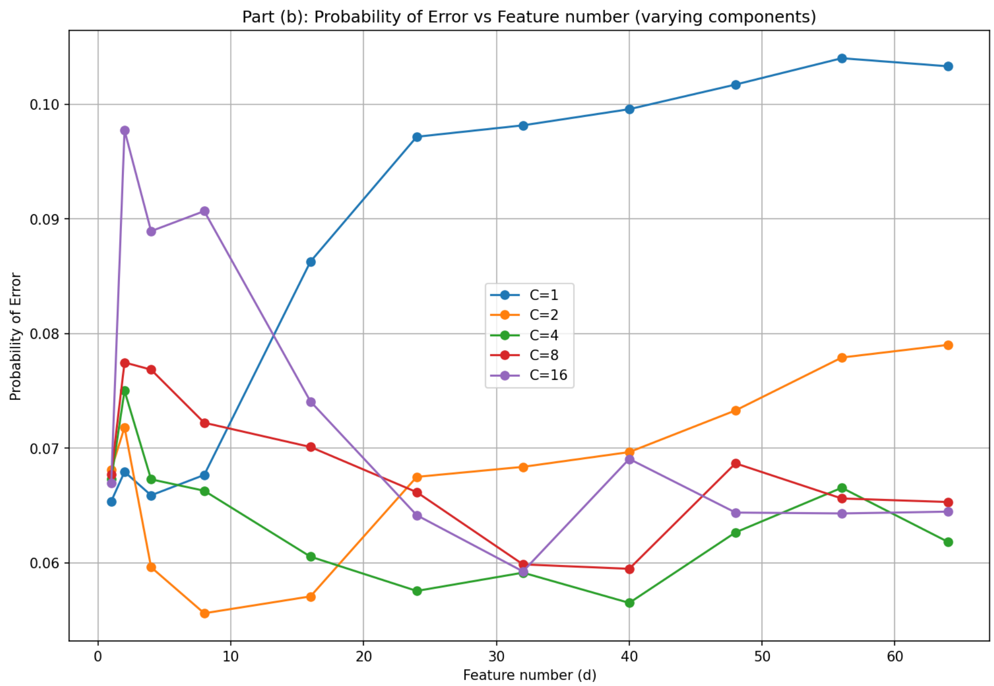

Figure 2: Part (b): probability of error vs. feature dimension for mixtures with different numbers of components $C \in \{1, 2, 4, 8, 16, 32\}$.

- For very low dimensions ($d \leq 8$) all choices of C lead to very similar error rates (around 0.22–0.24). With so few features the representation is too limited for the additional mixture components to help.

- When more features are included, the models behave quite differently. For example, at $d = 32$ the estimated probability of error is approximately $P_e \approx 0.115$ for $C = 1$, 0.123 for $C = 2$, 0.216 for $C = 4$, 0.190 for $C = 8$, 0.175 for $C = 16$, and 0.161 for $C = 32$. The single-Gaussian model ($C = 1$) actually achieves the lowest error in this experiment.

- As C increases from 4 to 32 the curves tend to improve (e.g., the error for $C = 32$ at large d is lower than for $C = 4$ or $C = 8$), suggesting that additional components can help the model better fit the complex foreground and background distributions. However, none of the larger mixtures shows better performance than the simple $C = 1$ model on this task.

- For each fixed C, the error generally decreases as d grows from 16 to around 48–64, then slowly stabilizes. This is most evident for $C = 32$, whose error drops from about 0.22 at $d = 16$ to about 0.15 at $d = 56$–64.

These results highlight that in theory increasing the number of mixture components should never hurt and should allow a more flexible approximation of the true class-conditional densities. In practice, with limited training data and diagonal covariances, the EM algorithm can overfit and converge to poor local optima when C is large. In our experiment the single-Gaussian model is strong, and adding more components often increases the probability of error instead of reducing it.

1.4 Conclusion

In this quiz we built GMM-based classifiers for cheetah vs. grass segmentation using DCT features of 8×8 image blocks. Part (a) showed that, while EM initialization causes some variation in performance, the overall shape of the probability-of-error curves is consistent across runs: using only a few DCT coefficients leads to high error, and using a moderate-to-large subset of coefficients substantially improves segmentation.

Part (b) demonstrated that simply increasing the number of mixture components does not guarantee better performance. With the available training data and diagonal covariances, a single Gaussian per class already models the DCT feature distributions well, and larger mixtures may overfit or get trapped in suboptimal local maxima of the likelihood. This illustrates the trade-off between model complexity and robustness in generative classification and emphasizes the need to validate mixture complexity on a separate performance metric such as segmentation error.

1.5 Appendix: Source Code

The Python source code used to generate these results is included below.

```python
#!/usr/bin/env python3

from __future__ import annotations

import os
import argparse
from dataclasses import dataclass
from typing import List, Tuple, Optional

import numpy as np
import scipy.io
import imageio.v3 as imageio
import matplotlib.pyplot as plt
from numpy.lib.stride_tricks import sliding_window_view
from scipy.fft import dctn
from scipy.special import logsumexp


BLOCK = 8
TRAIN_MAT = "TrainingSamplesDCT_8_new.mat"
SUBSETS_MAT = "TrainingSamplesDCT_subsets_8.mat"
IMG = "cheetah.bmp"
MASK = "cheetah_mask.bmp"
ZIGZAG = "Zig-Zag Pattern.txt"

DIMS_DEFAULT = [1, 2, 4, 8, 16, 24, 32, 40, 48, 56, 64]
C_LIST_DEFAULT = [1, 2, 4, 8, 16]


@dataclass
class Paths:
    train_mat: str
    subsets_mat: str
    img: str
    mask: str
    zigzag: str
    out_dir: str

```

```python

def parse_list_int(s: str) -> List[int]:
    return [int(x.strip()) for x in s.split(",") if x.strip()]


def load_training(train_mat: str) -> Tuple[np.ndarray, np.ndarray]:
    """Load TrainsampleDCT_FG/BG (N,64)."""
    m = scipy.io.loadmat(train_mat)
    fg = np.asarray(m["TrainsampleDCT_FG"], dtype=np.float64)
    bg = np.asarray(m["TrainsampleDCT_BG"], dtype=np.float64)
    return fg, bg


def load_img_mask(img_path: str, mask_path: str) -> Tuple[np.ndarray, np.ndarray]:
    """Load grayscale image and binary mask."""
    img = imageio.imread(img_path)
    msk = imageio.imread(mask_path)
    if img.ndim == 3:
        img = img.mean(axis=2)
    if msk.ndim == 3:
        msk = msk.mean(axis=2)
    img = np.asarray(img, dtype=np.float64)
    msk = (np.asarray(msk) > 127).astype(np.uint8)
    return img, msk


def zigzag_order(zigzag_file: str) -> np.ndarray:
    """Return 64-length index order list (argsort of rank-matrix)."""
    toks: List[int] = []
    with open(zigzag_file, "r") as f:
        for line in f:
            line = line.strip()
            if line:
                toks.extend([int(x) for x in line.split()])
    ranks = np.array(toks[:64], dtype=int)
    return np.argsort(ranks).astype(int)


def dct_features_valid(img: np.ndarray, order: np.ndarray, dc_scale: float = 0.5)
    -> Tuple[np.ndarray, int, int]:
    """Compute (H-7)*(W-7) DCT features (no padding), zigzag reorder, scale DC."""
    img01 = img / 255.0
    win = sliding_window_view(img01, (BLOCK, BLOCK))      # (H-7, W-7, 8, 8)
    h2, w2 = win.shape[0], win.shape[1]
    blk = win.reshape(h2 * w2, BLOCK, BLOCK)
    d = dctn(blk, type=2, norm="ortho", axes=(1, 2)).reshape(h2 * w2, 64)
    X = d[:, order].astype(np.float64)
    X[:, 0] *= float(dc_scale)
    return X, h2, w2


def mask_valid(mask: np.ndarray, h2: int, w2: int) -> np.ndarray:
    """Crop mask to valid block top-left positions."""
    return mask[:h2, :w2].reshape(-1).astype(np.uint8)


def alpha_ranking_from_subsets(subsets_mat: str) -> np.ndarray:
    """Compute alpha ranking using D4_FG/D4_BG."""
    m = scipy.io.loadmat(subsets_mat)
```

```python
    fg = np.asarray(m["D4_FG"], dtype=np.float64)
    bg = np.asarray(m["D4_BG"], dtype=np.float64)
    mu_fg = fg.mean(axis=0)
    mu_bg = bg.mean(axis=0)
    sd_fg = fg.std(axis=0)
    sd_bg = bg.std(axis=0)
    alpha = np.abs(mu_fg - mu_bg) / (sd_fg + sd_bg + 1e-12)
    return np.argsort(-alpha).astype(int)


class GMMDiagonal:
    """Diagonal GMM with EM + best-of-n random restarts."""

    def __init__(self, C: int, n_iter: int, tol: float, reg: float, min_var: float):
        self.C = int(C)
        self.n_iter = int(n_iter)
        self.tol = float(tol)
        self.reg = float(reg)
        self.min_var = float(min_var)
        self.w: Optional[np.ndarray] = None
        self.m: Optional[np.ndarray] = None
        self.v: Optional[np.ndarray] = None

    @staticmethod
    def _log_gauss_diag(X: np.ndarray, mean: np.ndarray, var: np.ndarray) -> np.ndarray:
        var = np.maximum(var, 1e-12)
        return -0.5 * (np.sum(np.log(2.0 * np.pi * var)) + np.sum((X - mean) ** 2 / var, axis=1))

    def _e_step(self, X: np.ndarray, w: np.ndarray, m: np.ndarray, v: np.ndarray) -> Tuple[np.ndarray, float]:
        n = X.shape[0]
        logp = np.empty((n, self.C), dtype=np.float64)
        for c in range(self.C):
            logp[:, c] = np.log(w[c] + 1e-300) + self._log_gauss_diag(X, m[c], v[c])
        log_norm = logsumexp(logp, axis=1)
        return logp - log_norm[:, None], float(np.mean(log_norm))

    def _m_step(self, X: np.ndarray, log_resp: np.ndarray) -> Tuple[np.ndarray, np.ndarray, np.ndarray]:
        resp = np.exp(log_resp)
        n, d = X.shape
        Nk = resp.sum(axis=0) + 1e-12
        w = Nk / n
        m = (resp.T @ X) / Nk[:, None]
        v = np.empty((self.C, d), dtype=np.float64)
        for c in range(self.C):
            diff = X - m[c]
            v[c] = (resp[:, c:c+1] * (diff ** 2)).sum(axis=0) / Nk[c]
            v[c] = np.maximum(v[c] + self.reg, self.min_var)
        return w, m, v

    def _fit_once(self, X: np.ndarray, rng: np.random.Generator) -> Tuple[np.ndarray, np.ndarray, np.ndarray, float]:
        n, d = X.shape
        idx = rng.choice(n, size=self.C, replace=False)
```

```
        m = X[idx].copy() + rng.normal(scale=1e-3, size=(self.C, d))
        base_var = np.maximum(np.var(X, axis=0) + self.reg, self.min_var)
        v = np.tile(base_var[None, :], (self.C, 1))
        w = np.full((self.C,), 1.0 / self.C, dtype=np.float64)

        prev = -np.inf
        best = -np.inf
        for _ in range(self.n_iter):
            log_resp, lb = self._e_step(X, w, m, v)
            if lb > best:
                best = lb
            if np.isfinite(prev) and abs(lb - prev) < self.tol:
                break
            prev = lb
            w, m, v = self._m_step(X, log_resp)
        return w, m, v, best

    def fit_best_of_n(self, X: np.ndarray, n_init: int, seed: int) -> "GMMDiagonal":
        rngg = np.random.default_rng(seed)
        best_lb = -np.inf
        best_params = None
        for _ in range(max(1, int(n_init))):
            rng = np.random.default_rng(int(rngg.integers(0, 2**31 - 1)))
            w, m, v, lb = self._fit_once(X, rng)
            if lb > best_lb:
                best_lb = lb
                best_params = (w, m, v)
        self.w, self.m, self.v = best_params
        return self

    def score_samples(self, X: np.ndarray) -> np.ndarray:
        n = X.shape[0]
        logp = np.empty((n, self.C), dtype=np.float64)
        for c in range(self.C):
            logp[:, c] = np.log(self.w[c] + 1e-300) + self._log_gauss_diag(X, self.m[c], self.v[c])
        return logsumexp(logp, axis=1)


def classify(fg: GMMDiagonal, bg: GMMDiagonal, X: np.ndarray, p_fg: float, p_bg: float) -> np.ndarray:
    ll_fg = fg.score_samples(X) + np.log(p_fg + 1e-300)
    ll_bg = bg.score_samples(X) + np.log(p_bg + 1e-300)
    return (ll_fg > ll_bg).astype(np.uint8)


def poe(pred: np.ndarray, gt: np.ndarray) -> float:
    return float(np.mean(pred.astype(np.uint8) != gt.astype(np.uint8)))


def solve(
    p: Paths,
    dims: List[int],
    c_list: List[int],
    n_runs: int,
    seed: int,
    max_em_iter: int,
    tol: float,
```

```
    reg: float,
    min_var: float,
    n_init_b: int,
) -> None:
    os.makedirs(p.out_dir, exist_ok=True)

    print("[INFO] Loading data...")
    train_fg, train_bg = load_training(p.train_mat)
    img, mask = load_img_mask(p.img, p.mask)

    prior_fg = train_fg.shape[0] / (train_fg.shape[0] + train_bg.shape[0])
    prior_bg = 1.0 - prior_fg
    print(f"[INFO] Training priors: FG={prior_fg:.6f}, BG={prior_bg:.6f}")

    order = zigzag_order(p.zigzag)
    print("[INFO] Extracting test image features...")
    Xtest, h2, w2 = dct_features_valid(img, order, dc_scale=0.5)
    ytest = mask_valid(mask, h2, w2)
    print(f"[INFO] eval size={h2}x{w2} (N={h2*w2})")

    feat_rank = alpha_ranking_from_subsets(p.subsets_mat)
    print("[INFO] Feature ranking source: computed alpha from subsets D4_FG/D4_BG")

    def cols_for_d(d: int) -> np.ndarray:
        return feat_rank[:d]

    rng_global = np.random.default_rng(seed)

    print("\n[INFO] --- Part (a): Initialization sensitivity (C=8) ---")
    C_a = 8

    fg_models_by_d: List[List[GMMDiagonal]] = []
    bg_models_by_d: List[List[GMMDiagonal]] = []

    for d in dims:
        cols = cols_for_d(d)
        Xfg_d = train_fg[:, cols]
        Xbg_d = train_bg[:, cols]

        fg_models = []
        bg_models = []

        for _ in range(n_runs):
            s = int(rng_global.integers(0, 2**31 - 1))
            fg_models.append(GMMDiagonal(C_a, max_em_iter, tol, reg, min_var).fit_best_of_n(Xfg_d, 1, s))

        for _ in range(n_runs):
            s = int(rng_global.integers(0, 2**31 - 1))
            bg_models.append(GMMDiagonal(C_a, max_em_iter, tol, reg, min_var).fit_best_of_n(Xbg_d, 1, s))

        fg_models_by_d.append(fg_models)
        bg_models_by_d.append(bg_models)

    err_bgfix = [[[0.0 for _ in dims] for _ in range(n_runs)] for _ in range(n_runs)]
    for k, d in enumerate(dims):
```

```python
        cols = cols_for_d(d)
        Xte_d = Xtest[:, cols]
        for j in range(n_runs):
            for i in range(n_runs):
                pred = classify(fg_models_by_d[k][i], bg_models_by_d[k][j], Xte_d, prior_fg, prior_bg)
                err_bgfix[j][i][k] = poe(pred, ytest)

    for k, d in enumerate(dims):
        all_pairs = np.array([err_bgfix[j][i][k] for j in range(n_runs) for i in range(n_runs)], dtype=np.float64)
        print(f"[INFO] d={d:2d}: mean={all_pairs.mean():.6f} (min={all_pairs.min():.6f}, max={all_pairs.max():.6f})")

    fig, axes = plt.subplots(2, 3, figsize=(14, 8), constrained_layout=True)
    axes = axes.flatten()
    for j in range(n_runs):
        ax = axes[j]
        for i in range(n_runs):
            ax.plot(dims, err_bgfix[j][i], marker="o", linewidth=1)
        ax.set_title(f"PE vs d (BG #{j+1} fixed)")
        ax.set_xlabel("Feature number (d)")
        ax.set_ylabel("Probability of Error")
        ax.grid(True)
    axes[-1].axis("off")

    out_a = os.path.join(p.out_dir, "prob6_a_initialization.png")
    fig.suptitle("Part (a): Probability of Error vs Feature number (BG fixed; 5 FG curves)", fontsize=14)
    fig.savefig(out_a, dpi=150, bbox_inches="tight")
    plt.close(fig)
    print(f"[INFO] Saved: {out_a}")

    print("\n[INFO] --- Part (b): Vary number of components C ---")
    plt.figure(figsize=(12, 8))

    for C in c_list:
        errors_c = []
        for d in dims:
            cols = cols_for_d(d)
            Xfg_d = train_fg[:, cols]
            Xbg_d = train_bg[:, cols]
            Xte_d = Xtest[:, cols]

            s_fg = int(rng_global.integers(0, 2**31 - 1))
            s_bg = int(rng_global.integers(0, 2**31 - 1))

            fg_model = GMMDiagonal(C, max_em_iter, tol, reg, min_var).fit_best_of_n(Xfg_d, n_init_b, s_fg)
            bg_model = GMMDiagonal(C, max_em_iter, tol, reg, min_var).fit_best_of_n(Xbg_d, n_init_b, s_bg)

            pred = classify(fg_model, bg_model, Xte_d, prior_fg, prior_bg)
            errors_c.append(poe(pred, ytest))

        plt.plot(dims, errors_c, marker="o", label=f"C={C}")
        print(f"[INFO] C={C:2d}: errors = {[f'{e:.4f}' for e in errors_c]}")

```

```python
    plt.title("Part (b): Probability of Error vs Feature number (varying components)")
    plt.xlabel("Feature number (d)")
    plt.ylabel("Probability of Error")
    plt.grid(True)
    plt.legend()

    out_b = os.path.join(p.out_dir, "prob6_b_components.png")
    plt.savefig(out_b, dpi=150, bbox_inches="tight")
    plt.close()
    print(f"[INFO] Saved: {out_b}")

    print("\n[INFO] Done.")


def main() -> None:
    ap = argparse.ArgumentParser()
    ap.add_argument("--data-dir", default="data")
    ap.add_argument("--output-dir", default="hw4/output")
    ap.add_argument("--dims", default=",".join(map(str, DIMS_DEFAULT)))
    ap.add_argument("--c-list", default=",".join(map(str, C_LIST_DEFAULT)))
    ap.add_argument("--n-runs", type=int, default=5)
    ap.add_argument("--seed", type=int, default=0)
    ap.add_argument("--max-em-iter", type=int, default=200)
    ap.add_argument("--tol", type=float, default=1e-4)
    ap.add_argument("--reg", type=float, default=1e-6)
    ap.add_argument("--min-var", type=float, default=1e-6)
    ap.add_argument("--n-init-b", type=int, default=5)
    args = ap.parse_args()

    p = Paths(
        train_mat=os.path.join(args.data_dir, TRAIN_MAT),
        subsets_mat=os.path.join(args.data_dir, SUBSETS_MAT),
        img=os.path.join(args.data_dir, IMG),
        mask=os.path.join(args.data_dir, MASK),
        zigzag=os.path.join(args.data_dir, ZIGZAG),
        out_dir=args.output_dir,
    )

    solve(
        p=p,
        dims=parse_list_int(args.dims),
        c_list=parse_list_int(args.c_list),
        n_runs=args.n_runs,
        seed=args.seed,
        max_em_iter=args.max_em_iter,
        tol=args.tol,
        reg=args.reg,
        min_var=args.min_var,
        n_init_b=args.n_init_b,
    )


if __name__ == "__main__":
    main()
```

Listing 1: Python code for HW4