CO902
Probabilistic and statistical inference

Lab 9 – Course Review

Tom Nichols
Department of Statistics &
Warwick Manufacturing Group

t.e.nichols@warwick.ac.uk

Lab 9: Last Lab (not a Lab)

- Lab 8 Redux
  - Review/restate key observations
- Written Reports - Feedback
- Oral presentations – Feedback
- Course Review
- Practice Exam Exercises

Lab 8 Redux
"Classification" with GMM

- Once a GMM is fit, each observations can be assigned to the class that is most likely to have generated it
  - Precisely, it is the class that maximizes the posterior probability of class k given x…

$$P(Z = k|X = x) \propto p(x|Z = k)\, p(Z = k) = \mathcal{N}(x|\mu_k, \Sigma_k)\pi_k$$

  - That is, it is *not* the class k that minimizes the Mahalanobis distance between x & μ_k!
  - It is the class that maximizes $\mathcal{N}(x|\mu_k, \Sigma_k)\pi_k$
    - The joint likelihood of X & latent class variable Z

“Classification” with GMM

- Probability of Y given X (up to a constant)

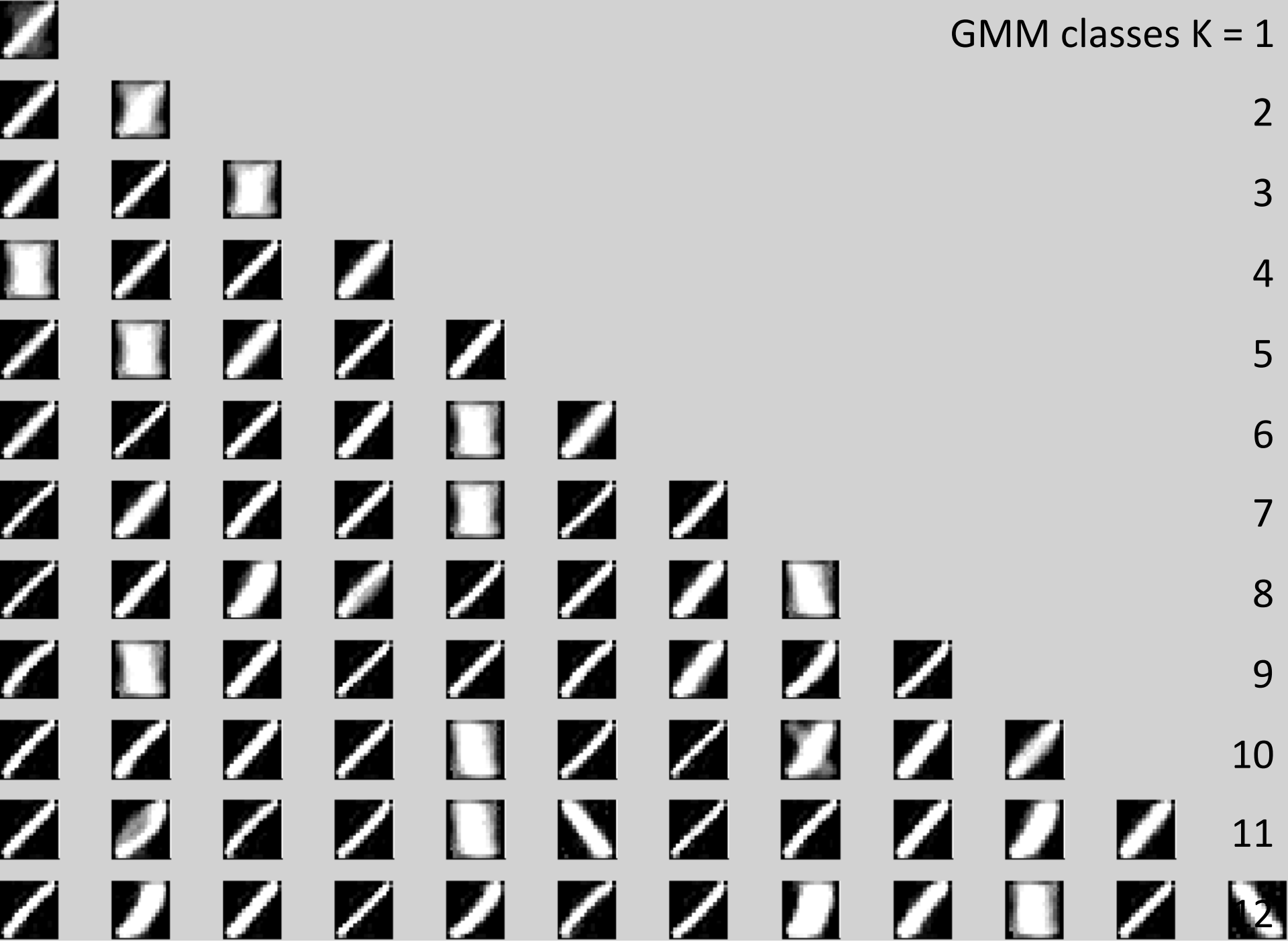
$$\mathcal{N}(x|\mu_k, \Sigma_k)\pi_k =$$

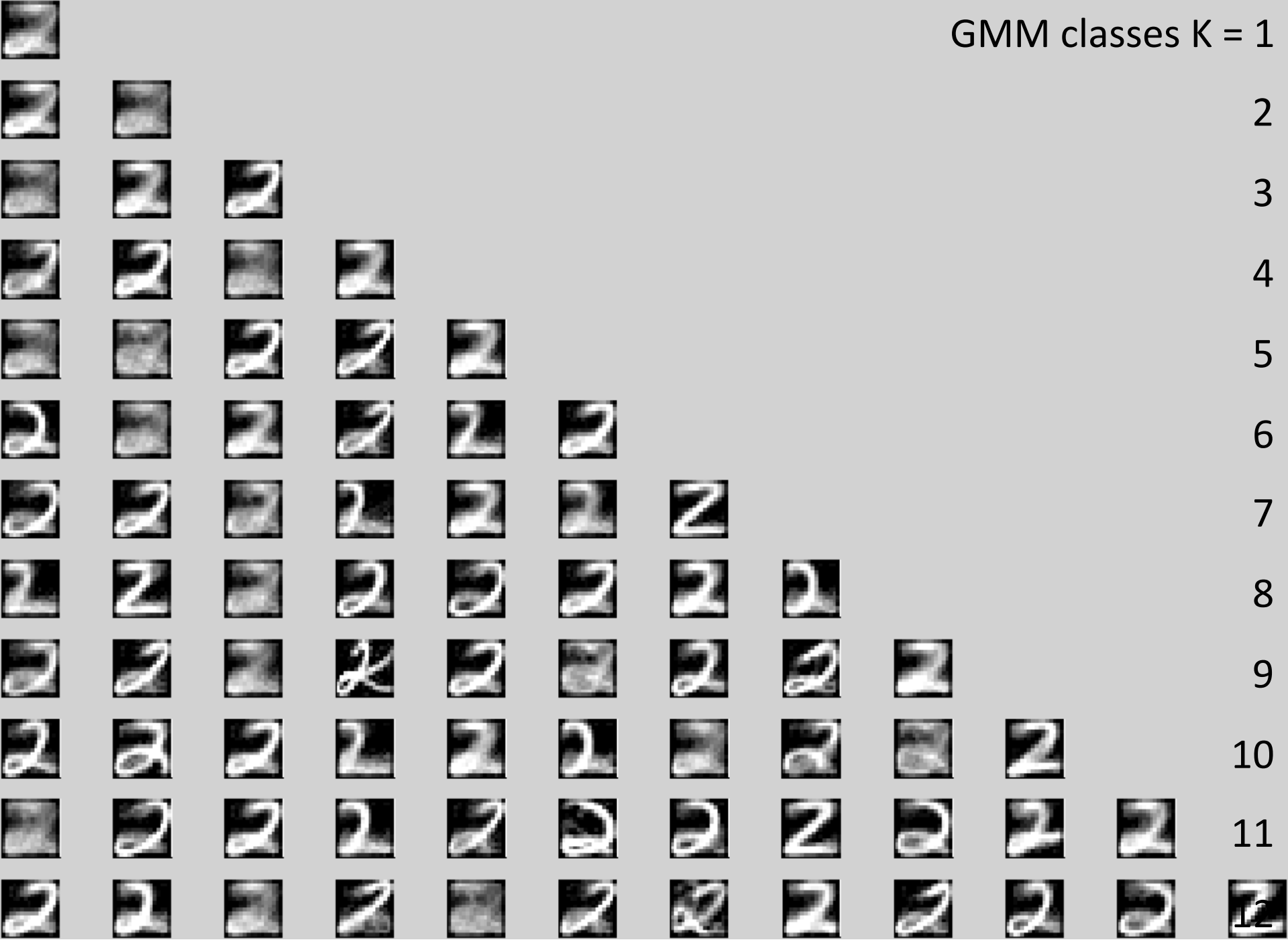
$$(2\pi)^{-d/2}|\Sigma_k|^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_k)^\top\Sigma_k^{-1}(x-\mu_k)\right) \quad \pi_k$$

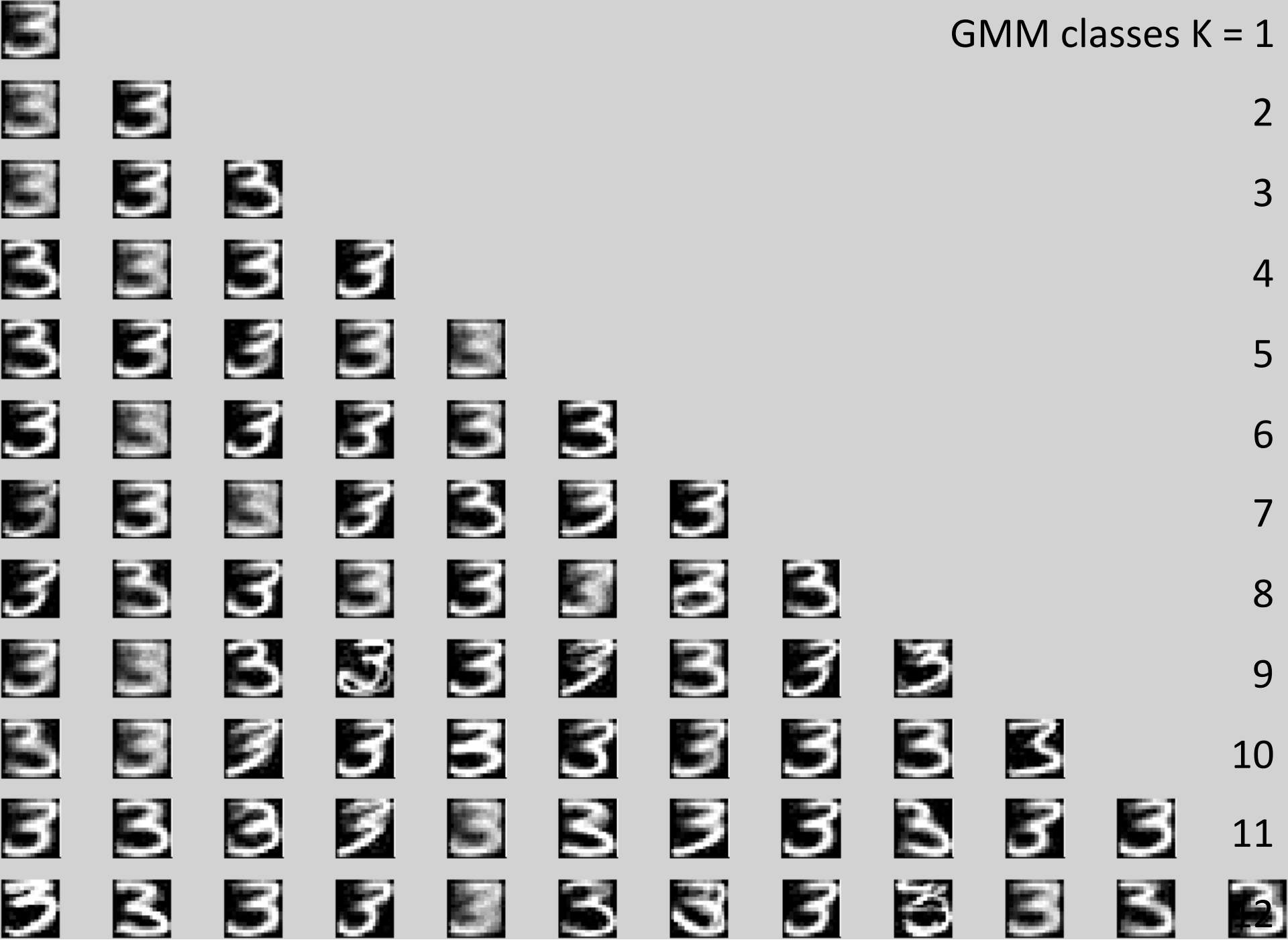
- Mahalanobis distance

$$(x-\mu_k)^\top\Sigma_k^{-1}(x-\mu_k)$$

- Using only Mahalanobis ignores
  - Scaling information in $|\Sigma_k|$, &
  - Prior weight π_k, i.e. prevalence of group k

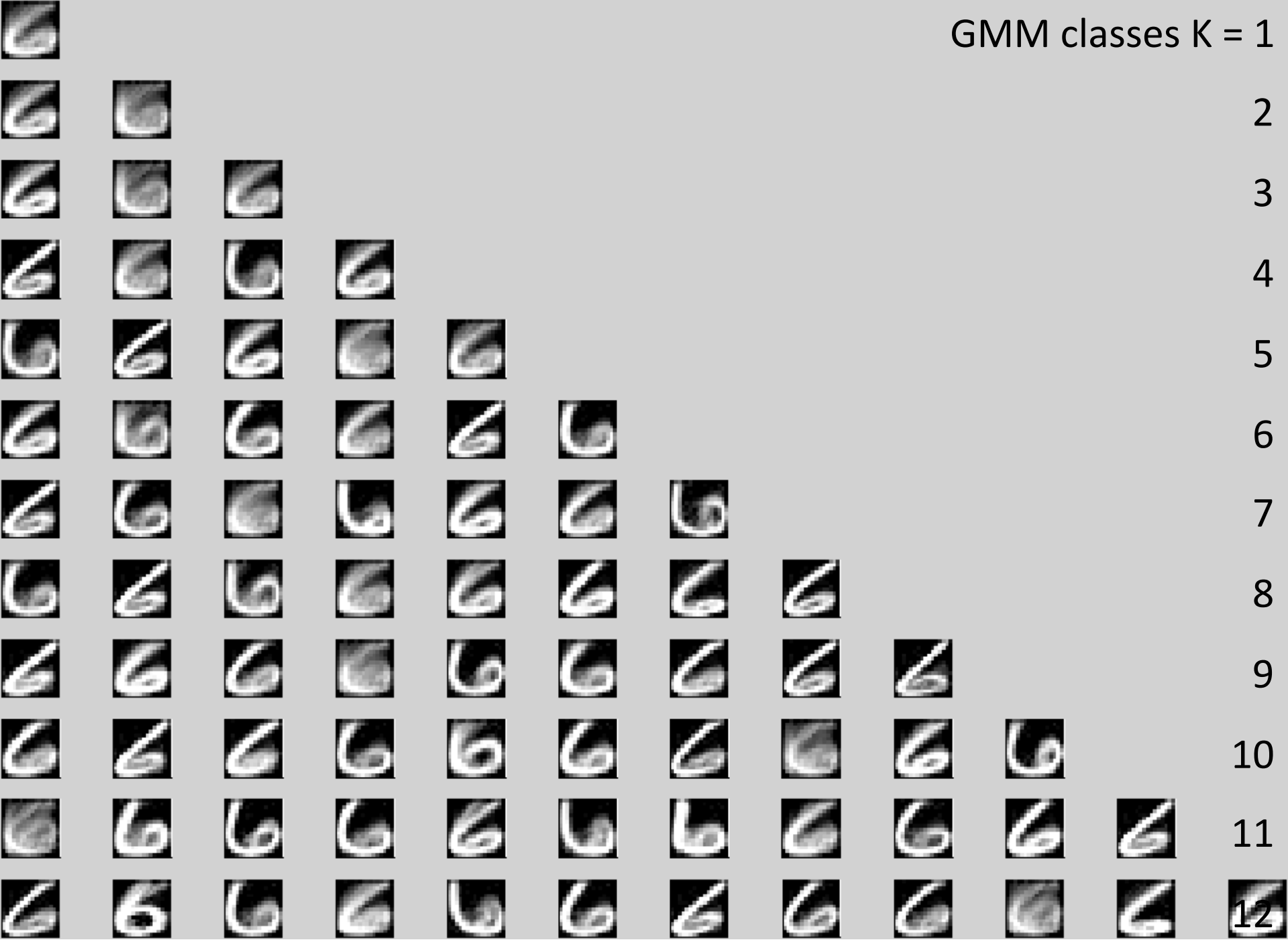

GMM classes K = 1
2
3
4
5
6
7
8
9
10
11
12

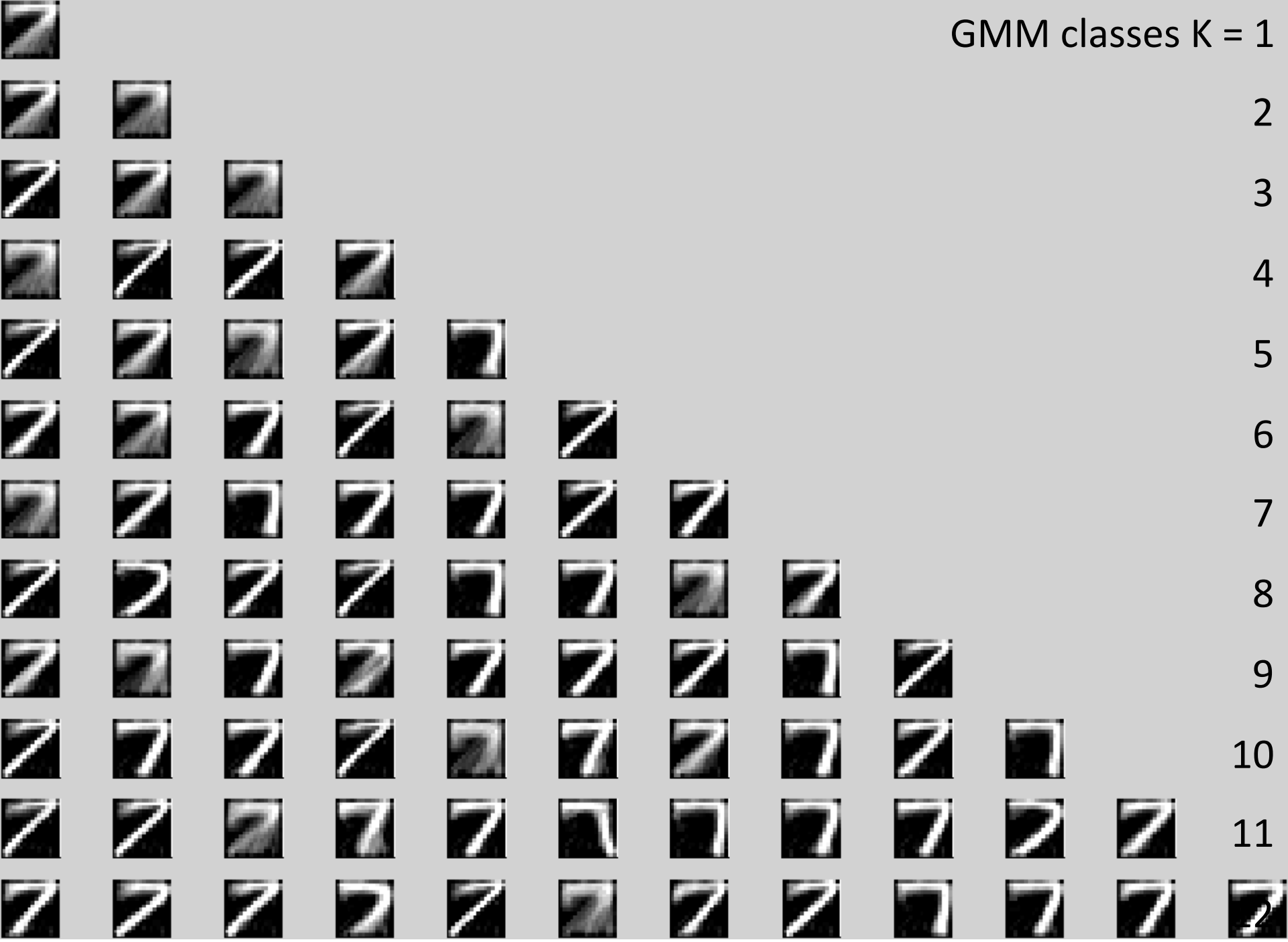

GMM classes K = 1
2
3
4
5
6
7
8
9
10
11
12

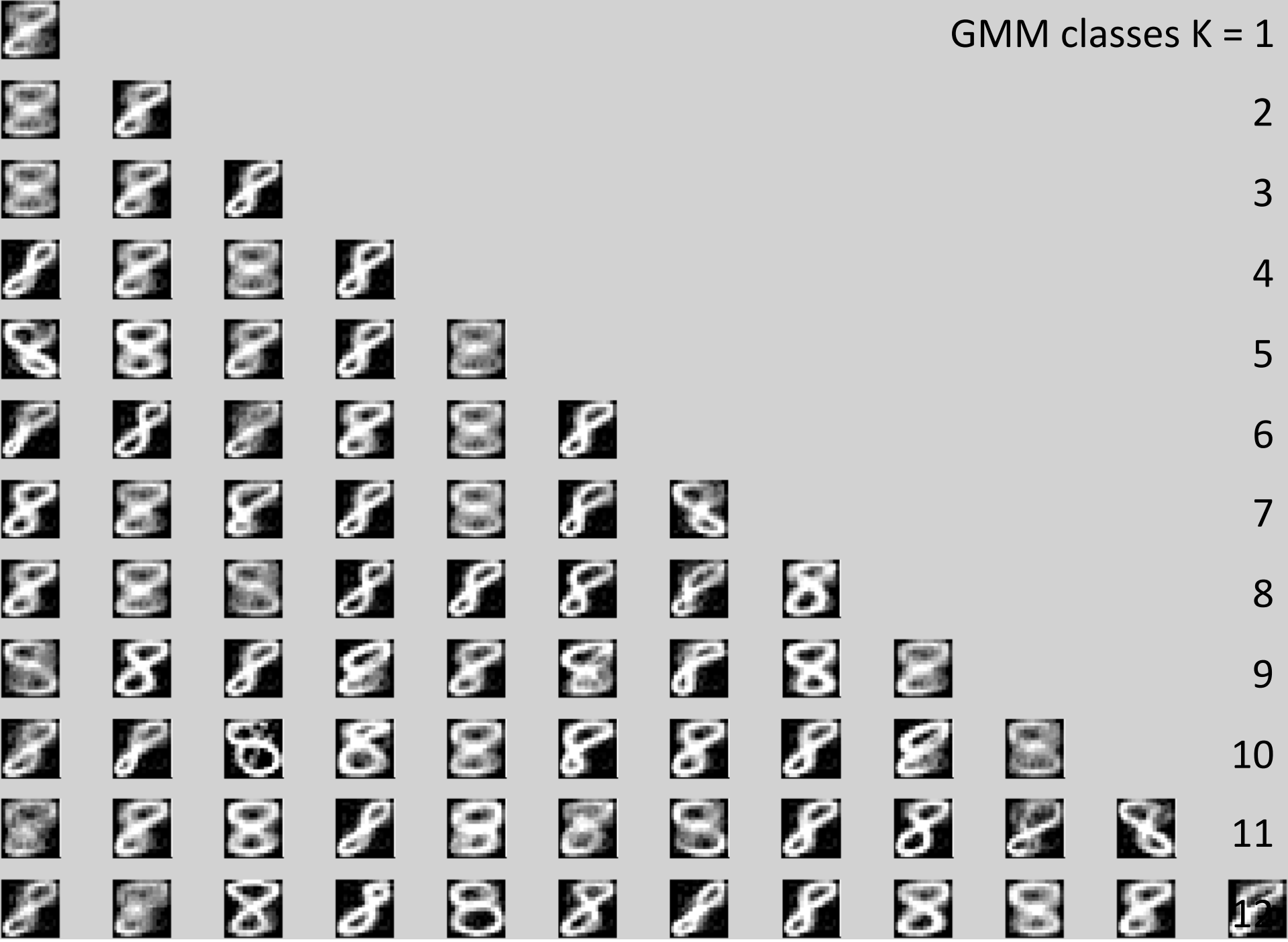

GMM classes K = 1
2
3
4
5
6
7
8
9
10
11
12

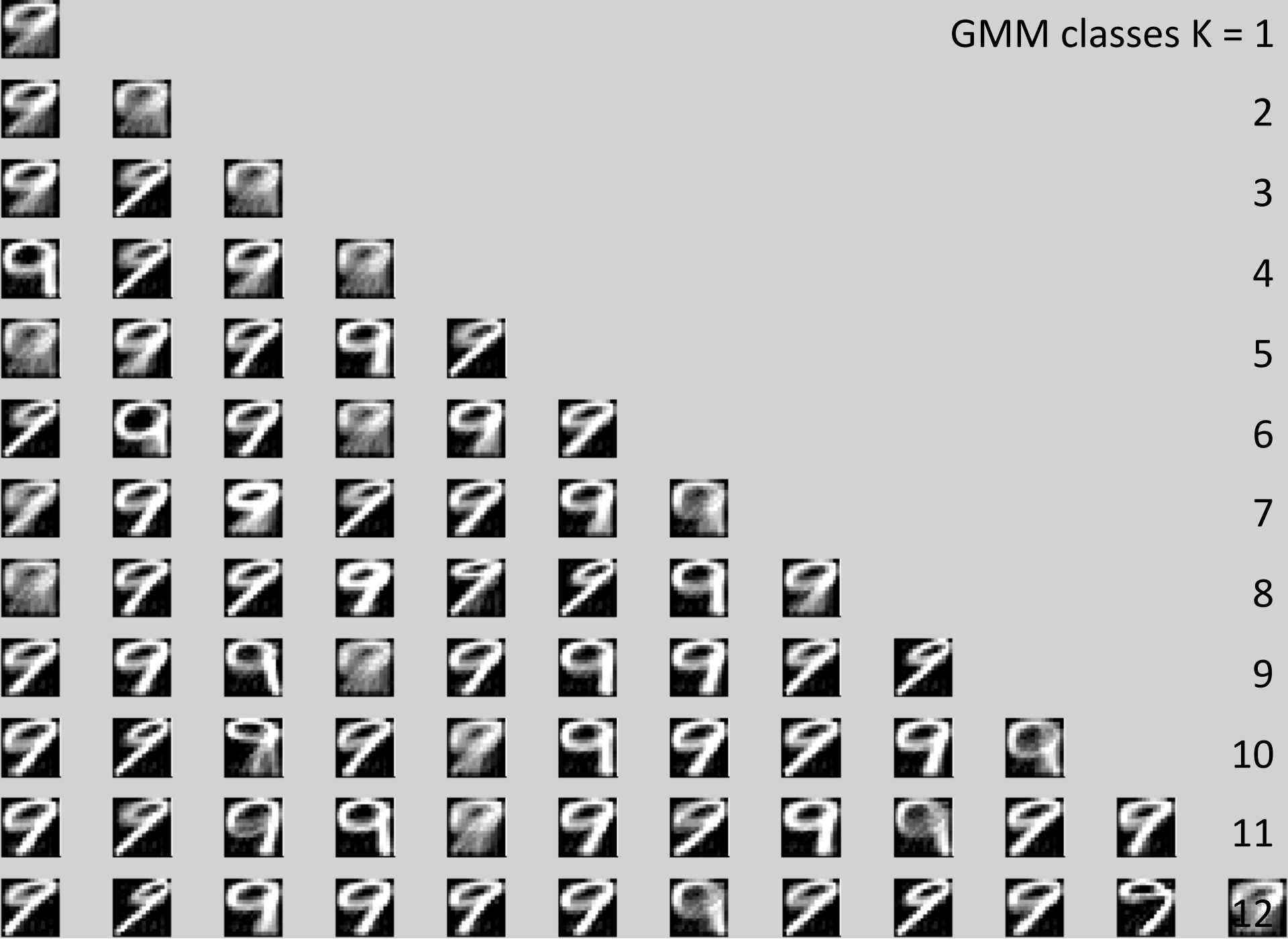

GMM classes K = 1
2
3
4
5
6
7
8
9
10
11
15

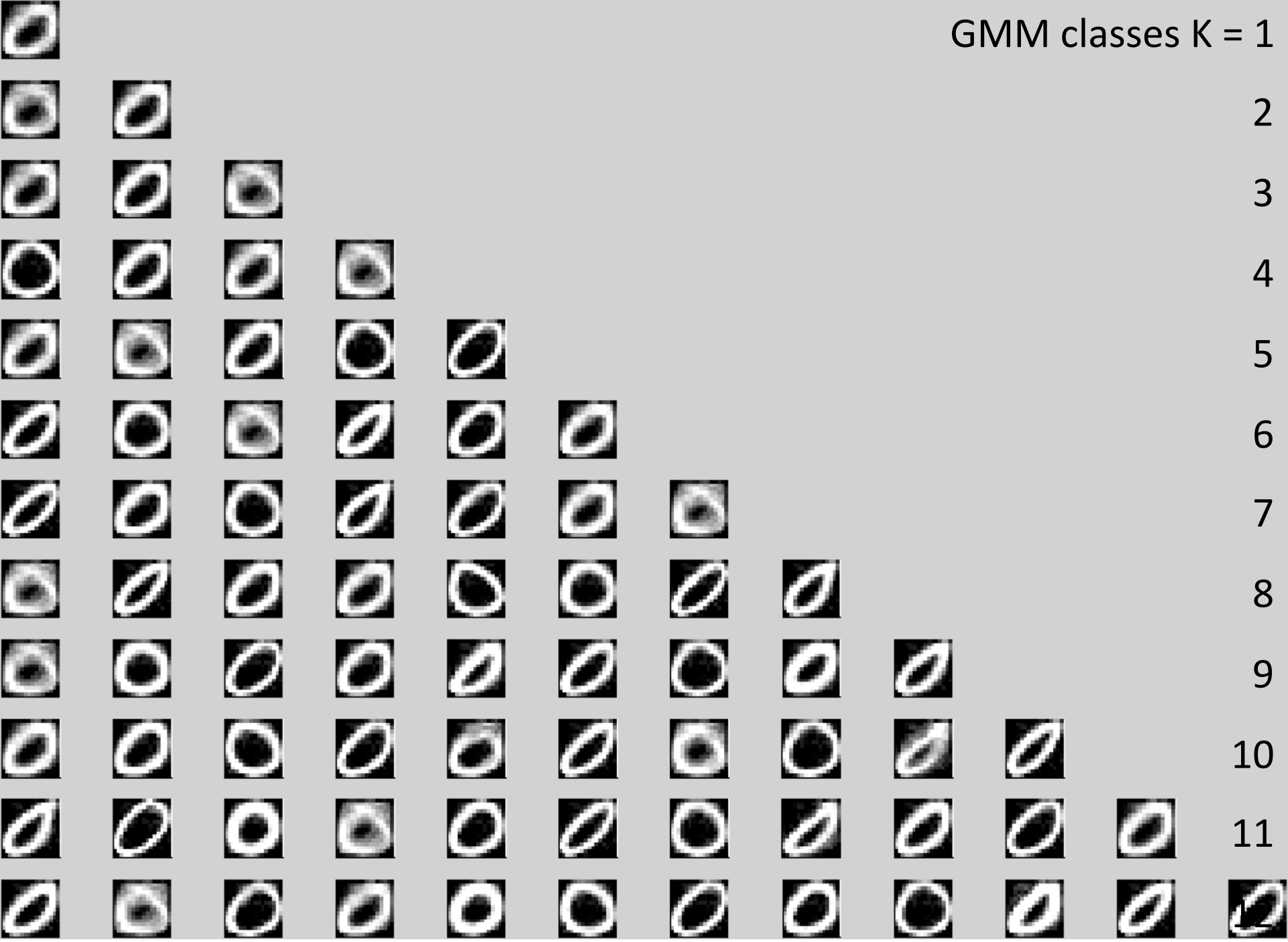

GMM classes K = 1
2
3
4
5
6
7
8
9
10
11
12


GMM classes K = 1
2
3
4
5
6
7
8
9
10
11
12

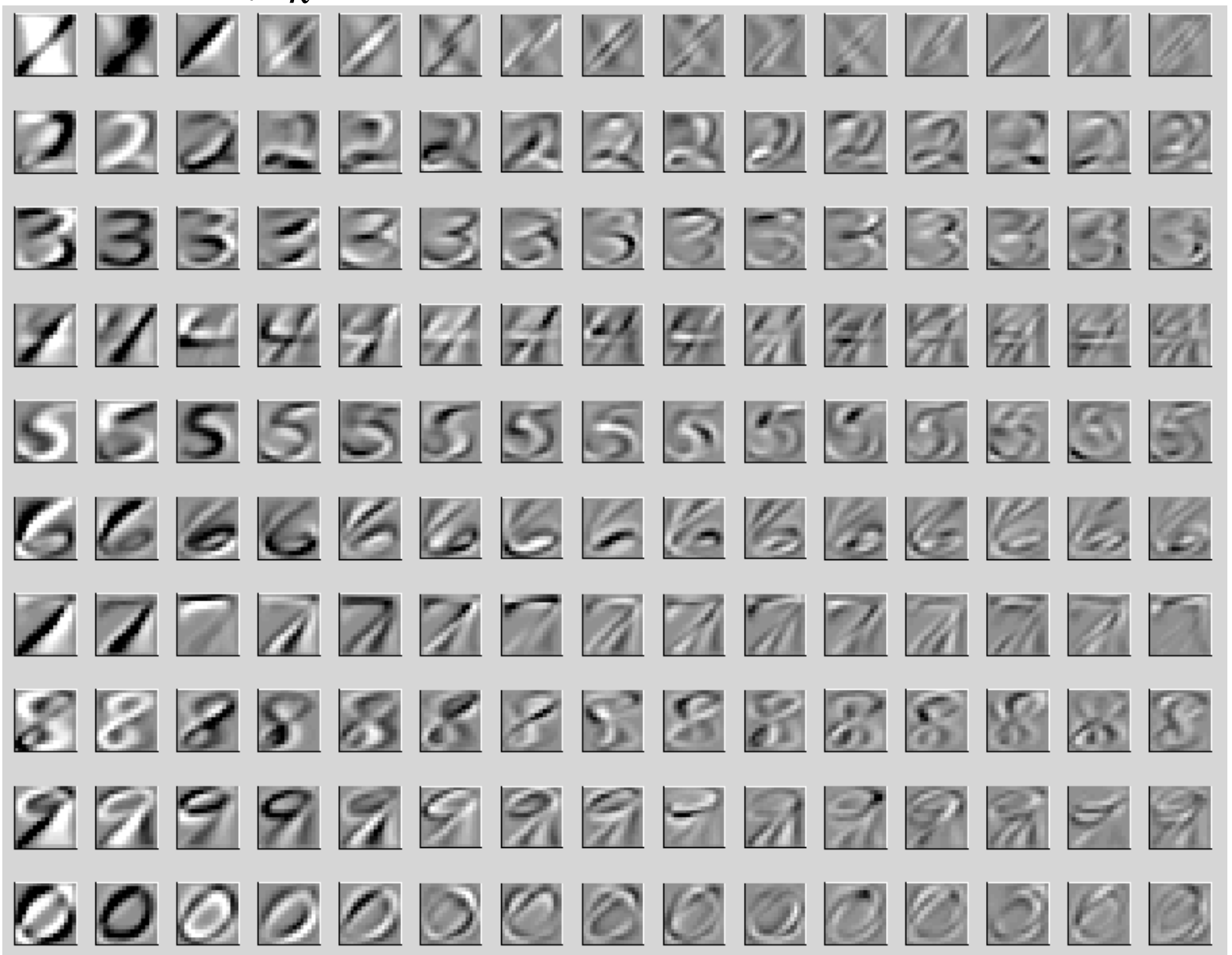

GMM classes K = 1
2
3
4
5
6
7
8
9
10
11
2


GMM classes K = 1
2
3
4
5
6
7
8
9
10
11
12


GMM classes K = 1
2
3
4
5
6
7
8
9
10
11
12


GMM classes K = 1
2
3
4
5
6
7
8
9
10
11
12

Compare to eigenvectors from PCA… GMM $\boldsymbol{\mu}_k$'s much more interpretable!



Written Reports

- Overall, really excellent work
  - Polished, adhered to structure of scientific paper
- Result
  - Everyone got the same "answer"
  - But varying degrees of variable selection
    - (Not required, but didn't hurt)
- Problems/Shortcomings
  - Used $|\theta_{i0} - \theta_{i1}|$ without motivating/explaining
  - Examined θ_{i0} or θ_{i1} alone, not as difference; or only one direction, e.g. $\theta_{i0} > \theta_{i1}$
  - Judgment of "OK classification" of 70%, when dumb "Always successful" classifier would give 73.8%!
    - Exploring False Positive / False Negative balance addresses this some
  - Some imprecise language on "iid"
    - "X_{ij} is iid"… over samples? variables?
      - Not over variables! Not identical!

Oral Reports

- Again, really nice work
  - Polished presentations
  - Betrayed comprehension of the paper
  - Conveyed understanding of the work
- Shortcomings
  - Only a few directly addressed “shortcomings”
    - Criticism of the work is vital
  - Some “slide reading”, not facing audience

Vivas

- Tuesday 8th March
- 20 minutes each
  - Schedule emailed by Jen; room TBA
- Verbal/whiteboard
- Expect 3-4 questions
  - Some "Define this…" or "Explain difference between…"
  - Some "Write down the model and assumptions for…"
  - Some derivations, "Show…"
  - More questions possible
  - Level of rigor used in class derivations
- Dr. Stefan Grosskinsky – 2nd marker
- "Code of honour"
  - No discussions with other class members!

Course Review: Basics (1)

- Random variables
- Independence
- Joint distributions
- Sum rule (Law of total probability)
- Product rule
- Bayes rule
- Expectation, Variance
- PMF's, PDF's and CDF's
- Parameterized Distributions
  - Bernoulli, Normal, Multivariate Normal, Beta

Course Review: Basics (2)

- Likelihood (vs. a PMF/PDF)
- Maximum Likelihood Estimation
- Dependent Random Variables
  - Conditional Distributions
  - Binary Markov Chain
- Properties of Estimators
  - Bias, Variance, Mean Squared Error, Consistency
- Bayesian Inference
  - Posterior \propto Likelihood × Prior
  - Conjugacy
- Maximum a Posteriori Estimation

Course Review: Supervised Learning (1)

- Discrete response (Bernoulli or Multinomial)
  - Discrete data (Bernoulli)
  - Continuous data (Gaussian)
- Class conditional distributions
- Classification by maximizing posterior of Y|X
  - Informally & via Decision Theory
- Naïve Bayes Classification
- Classification Accuracy
  - In-sample vs. Out-of-sample accuracy
- Cross Validation
  - LOOCV vs k-fold to address dangers of overfitting

Course Review: Supervised Learning (2)

- Continuous response (Gaussian), Continuous data (Gaussian)
  - "Discriminant Analysis"
    - Linear (LDA) vs Quadratic (QDA) decision boundary (why!?)
- Regression
  - From algebraic or Gaussian likelihood POV
- Ridge Regression
  - To address over-fitting, d>n problem
  - From "Penalized" optimization or Bayesian POV

Course Review: Unsupervised Learning (1)

- Dimension Reduction
  - PCA, as motivated by two criterion
  - PCA vs SVD
- Clustering
  - PCA
  - k-means
  - Gaussian Mixture Models
    - Intuition on EM

Example Exam Questions (1)

- Let $X_1, X_2, \ldots, X_n$ be *iid* samples of a Bernoulli with mean θ. Find the MLE of θ. Find the mean and the variance of the MLE of θ. Is it biased? Is it consistent?
- Now suppose we have a prior for θ, specifically a Beta(α,β). $p(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$
  Find the MAP estimator of θ.
  Is it biased? Is it consistent?
- Explain the findings on bias and consistency in intuitive terms

Example Exam Questions (2)

- Write down the model for a 2-state, time-invariant, 1st order Markov Chain. How many free parameters are there?
- Now, modify the model to make it a 2nd order Markov chain. How many free parameters are there?

Example Exam Questions (3)

- In the context of supervised learning with continuous data (X) and continuous response (Y), describe the "Normal Equations", and show how they are derived.
- I have n=50 measurements and d=70 predictor variables. How do I estimate $\mathbf{w}$?
  - (Follow up)

Example Exam Questions (4)

- What is the difference between Ridge Regression and Ordinary Least Squares (OLS) Regression? Write the criterion that each optimizes, and the estimator of w for each.
- What are the advantages of Ridge Regression over OLS? What are the disadvantages?

Example Exam Questions (5)

- What do the acronyms PCA and SVD stand for? What is the difference?
- Show how one is used to accomplish the other.

Example Exam Questions (5)

- For data matrix X ($d \times n$), consider transforming the centered data matrix,

$$\mathbf{Y} = \mathbf{U}^\top (\mathbf{X} - \bar{\mathbf{X}})$$

  Show that these derived variables are orthogonal.