LECTURE 5 TERM 2:

MSIN0097

UCL SCHOOL OF MANAGEMENT

PREDICTIVE ANALYTICS

A P MOORE

Individual Coursework assignment has been extended by one week to **Friday 5th March 2021** at **10:00 am**

USING OTHER PEOPLE'S CODE

UCL
SCHOOL OF
MANAGEMENT

When you copy code from stackoverflow and GitHub

https://twitter.com/woj_zaremba/status/1357432527540035585

MACHINE LEARNING JARGON

UCL SCHOOL OF MANAGEMENT

- Model
- Interpolating / Extrapolating
- Data Bias
- Noise / Outliers
- Learning algorithm
- Inference algorithm
- Supervised learning
- Unsupervised learning
- Classification
- Regression
- Clustering
- Decomposition
- Parameters
- Optimisation
- Training data
- Testing data
- Error metric
- Linear model
- Parametric model
- Model variance
- Model bias
- Model generalization
- Overfitting
- Goodness-of-fit
- Hyper-parameters
- Failure modes
- Confusion matrix
- True Positive
- False Negative
- Partition
- Margin
- Data density
- Hidden parameter
- High dimensional space
- Low dimensional space
- Separable data
- Manifold / Decision surface
- Hyper cube / volume / plane

A - B - C- D

ALGORITHMIC APPROACHES

UCL SCHOOL OF MANAGEMENT

A. ClAssification

Model requirements

Classifiation

$x \in [-\infty, \infty]$

$y \in \{0, N\}$

B. Regression

Model requirements

Regression

$x \in [-\infty, \infty]$

$y \in [-\infty, \infty]$

Supervised

C. Clustering

Model requirements

Clustering

$x \in [-\infty, \infty]$

$y \in \{0, N\}$

Hidden variables

Density estimation

D. Decomposition

Model requirements

Decomposition

$x \in [-\infty, \infty]$

$y \in [-\infty, \infty]$

Subspaces

Manifolds

Unsupervised

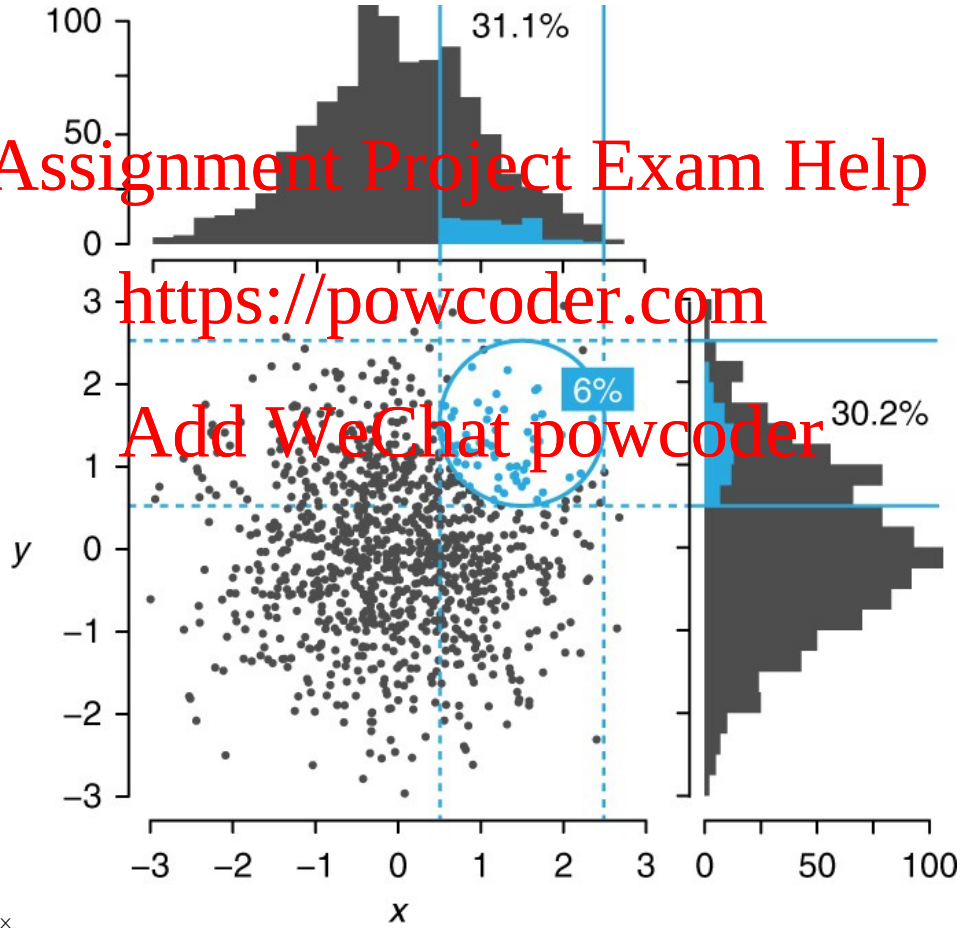
QUESTIONS

UCL
SCHOOL OF
MANAGEMENT

— How would I know if my data will be benefitted from a transformation to a higher or lower dimensional space?

CURSE OF DIMENSIONALITY

UCL SCHOOL OF MANAGEMENT



https://www.nature.com/articles/s41592-018-0019-x

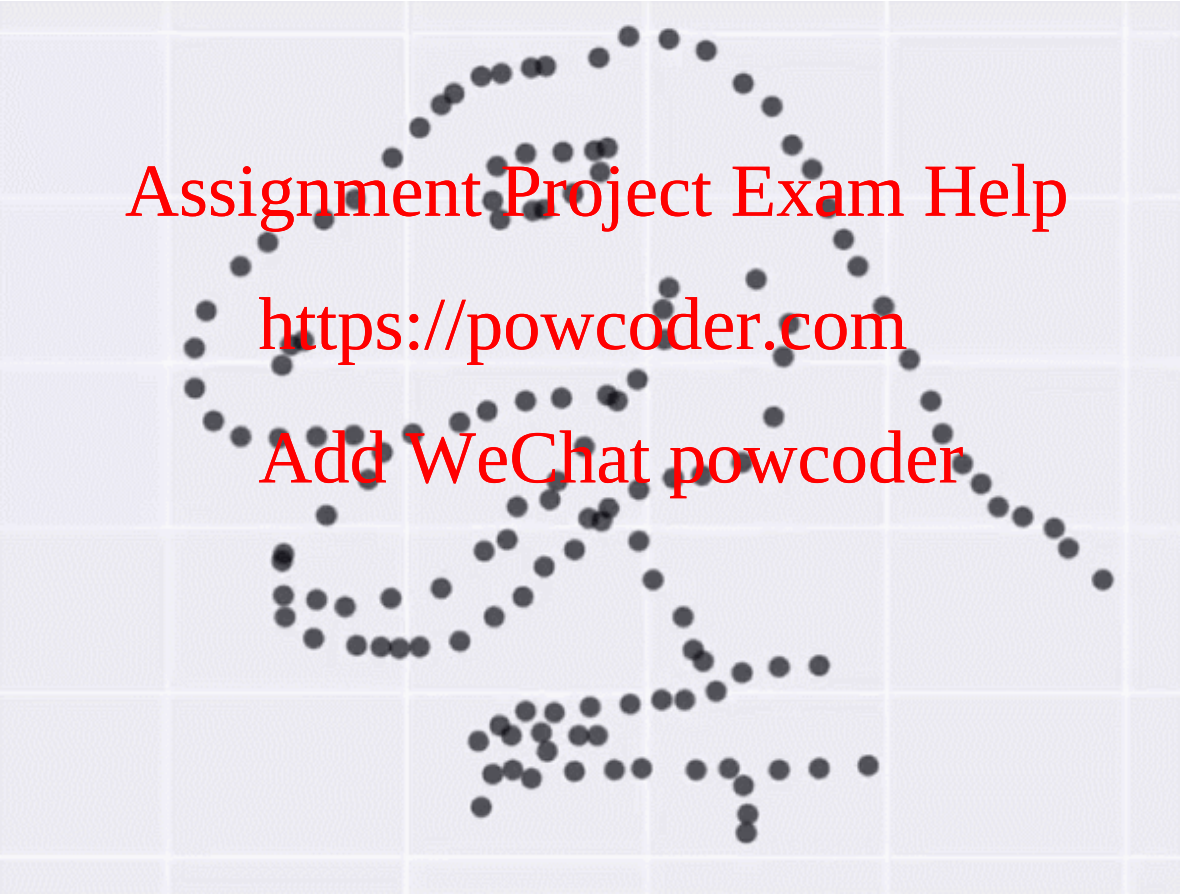
QUESTIONS

UCL
SCHOOL OF
MANAGEMENT

— Would I always have to visualize the data at a 2D or 3D level to visually understand if the data can be better separable? (but then this would defeat the idea of going a higher dimensional space which can't be visualized).

SUMMARY STATISTICS

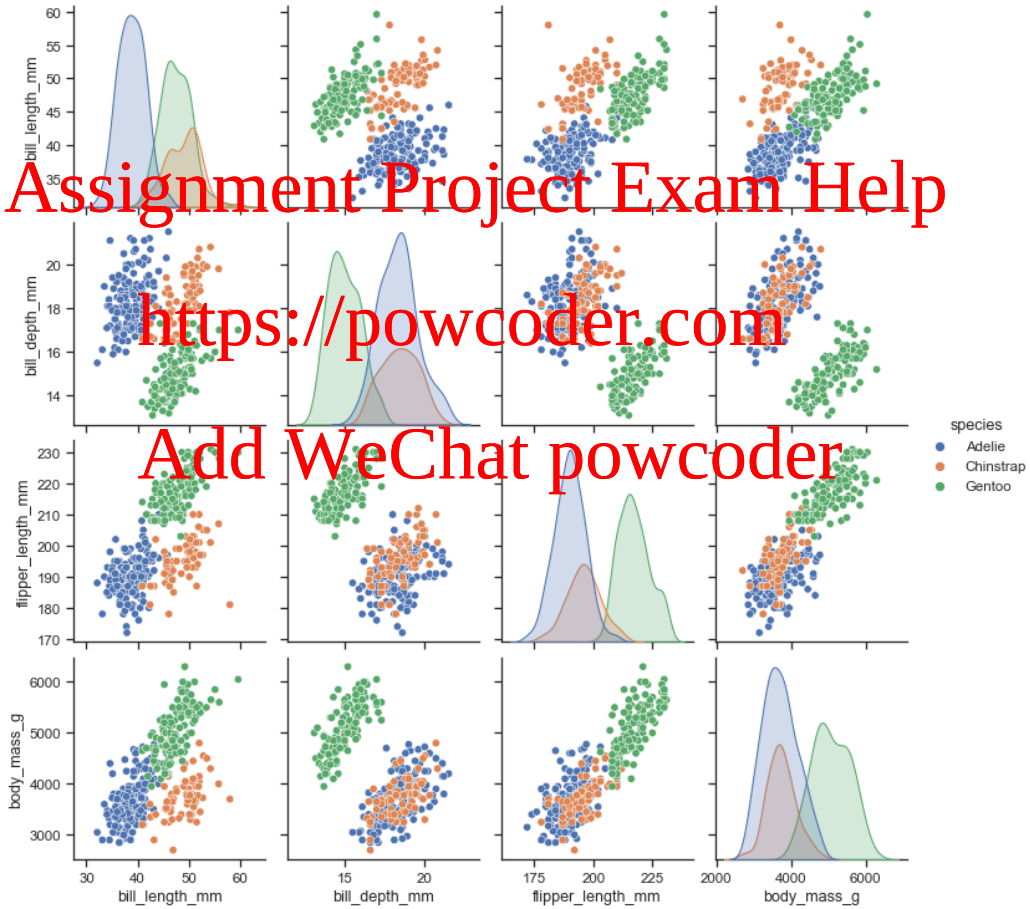
UCL
SCHOOL OF
MANAGEMENT



X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

SUMMARY STATISTICS

UCL
SCHOOL OF
MANAGEMENT



https://seaborn.pydata.org/examples/scatterplot_matrix.html

QUESTIONS?

UCL
SCHOOL OF
MANAGEMENT

— Should I have to go all the way through modelling (e.g. classification) and evaluate a metric such as the Gini coefficient and then go back to comparing different Gini scores from (addition of) extra dimensions?

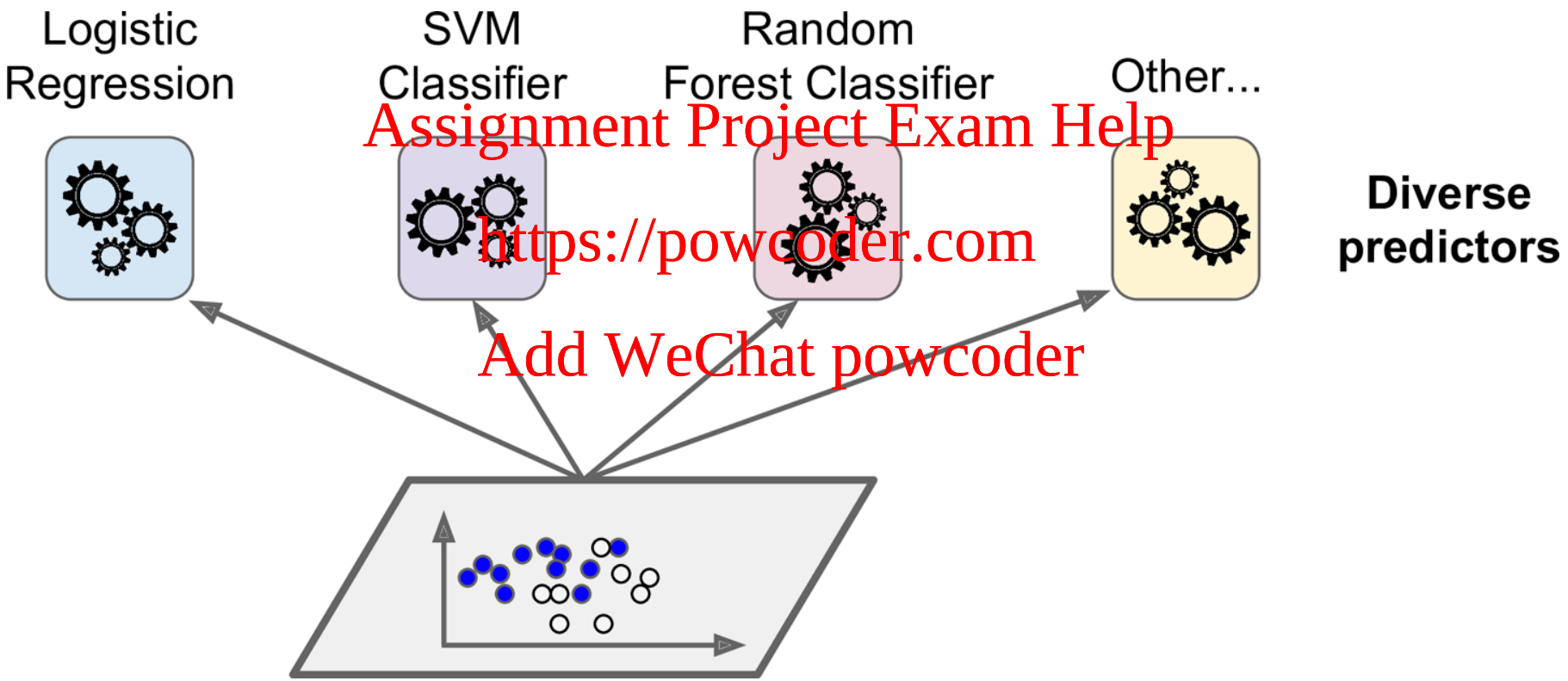
QUESTIONS?

UCL
SCHOOL OF
MANAGEMENT

— I understand that it might be better to go up a dimension in certain cases and other cases it will be better to go lower a dimension?

MULTIPLE MODELS

UCL
SCHOOL OF
MANAGEMENT

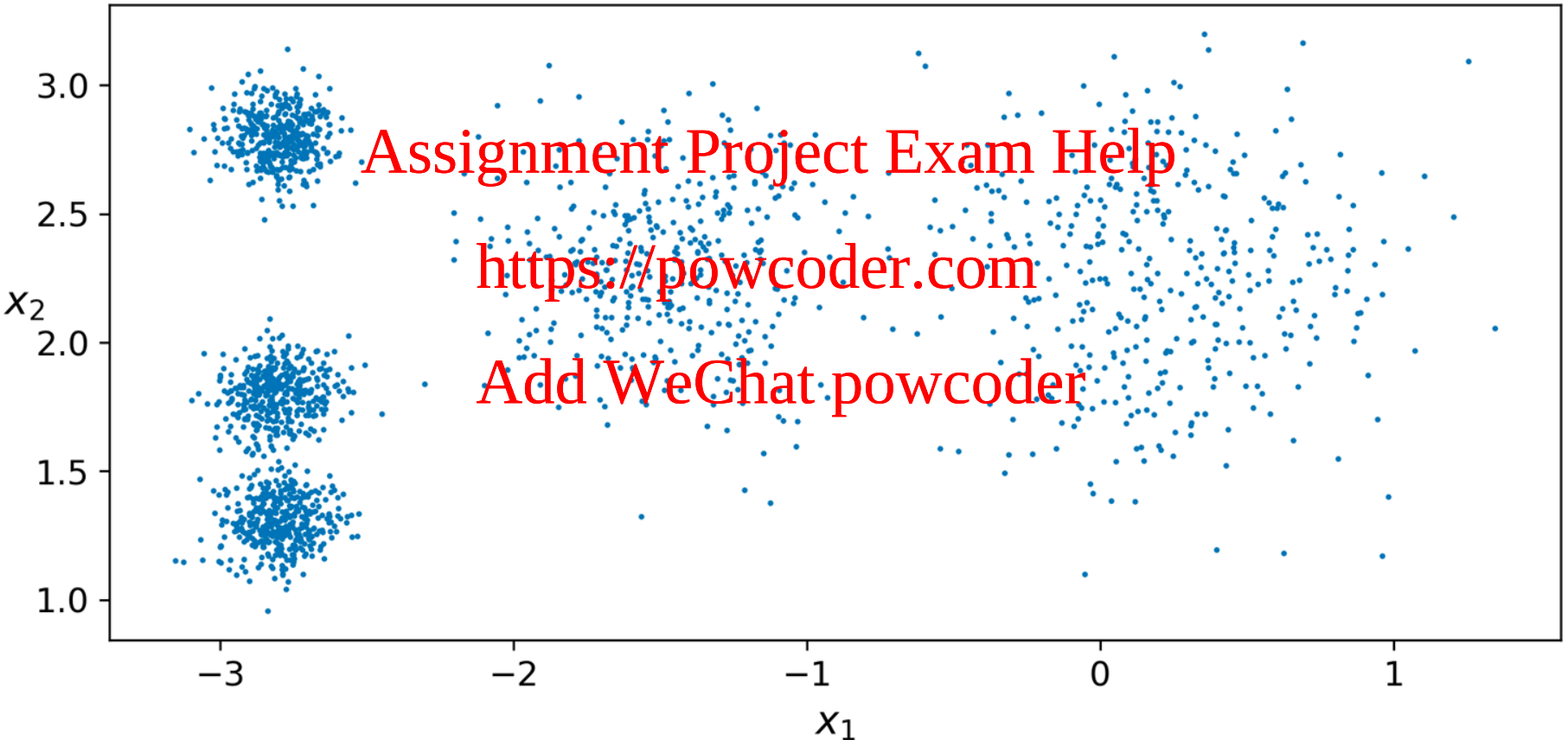


K-means

K-MEANS

LLOYD–FORGY ALGORITHM

UCL SCHOOL OF MANAGEMENT



K-MEANS

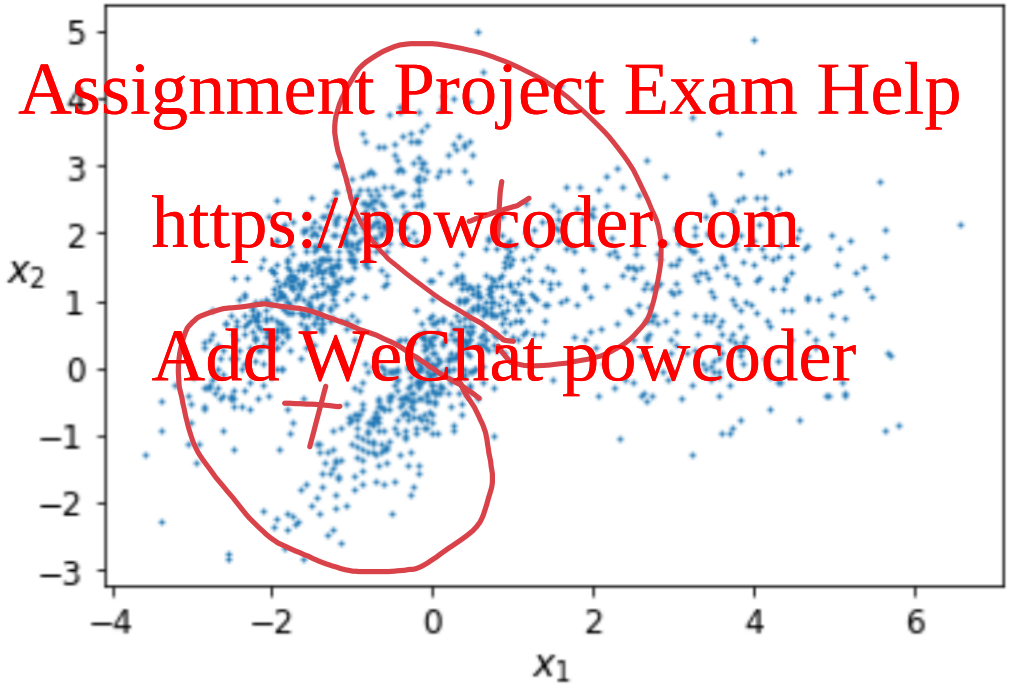
UCL SCHOOL OF MANAGEMENT

— Advantages

— Disadvantages

ELLIPSOIDAL DISTRIBUTED DATA

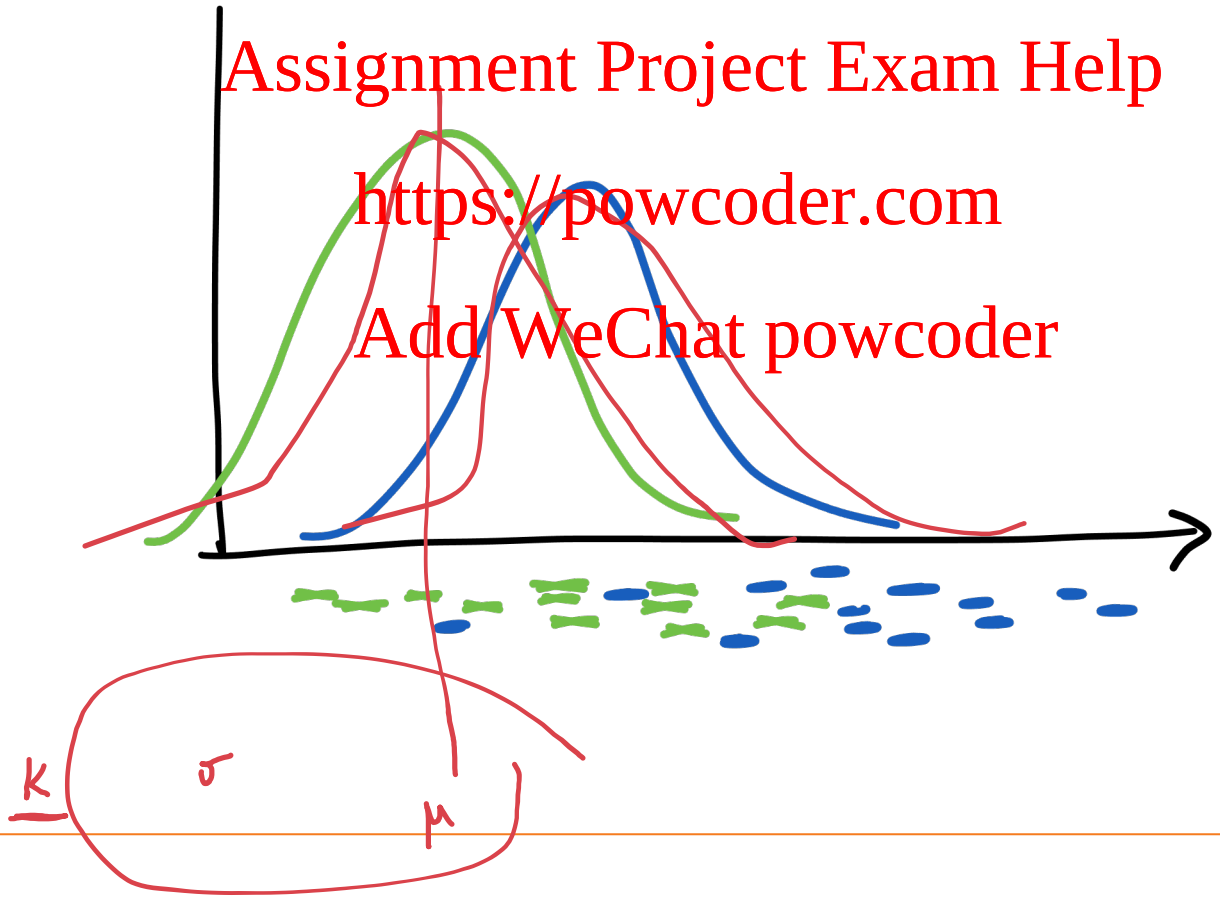
UCL SCHOOL OF MANAGEMENT



Gaussian mixtures

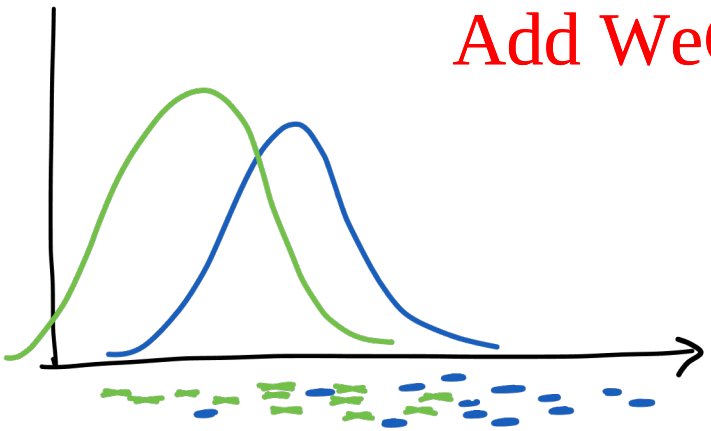
PARTITIONAL

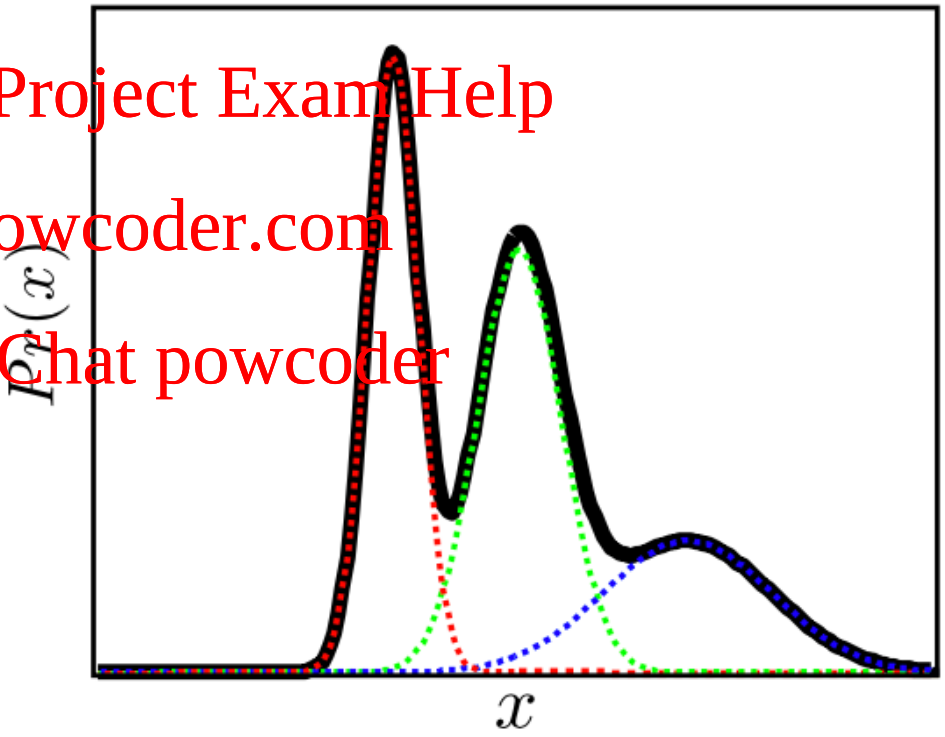
UCL
SCHOOL OF
MANAGEMENT



MIXTURE OF GAUSSIANS (1D)

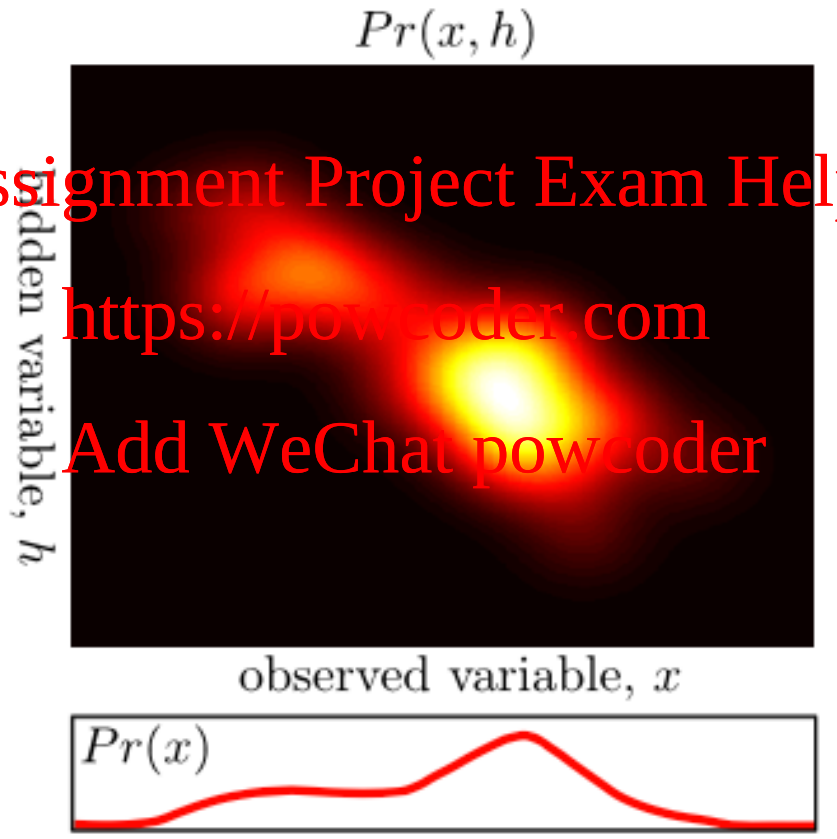
UCL
SCHOOL OF
MANAGEMENT





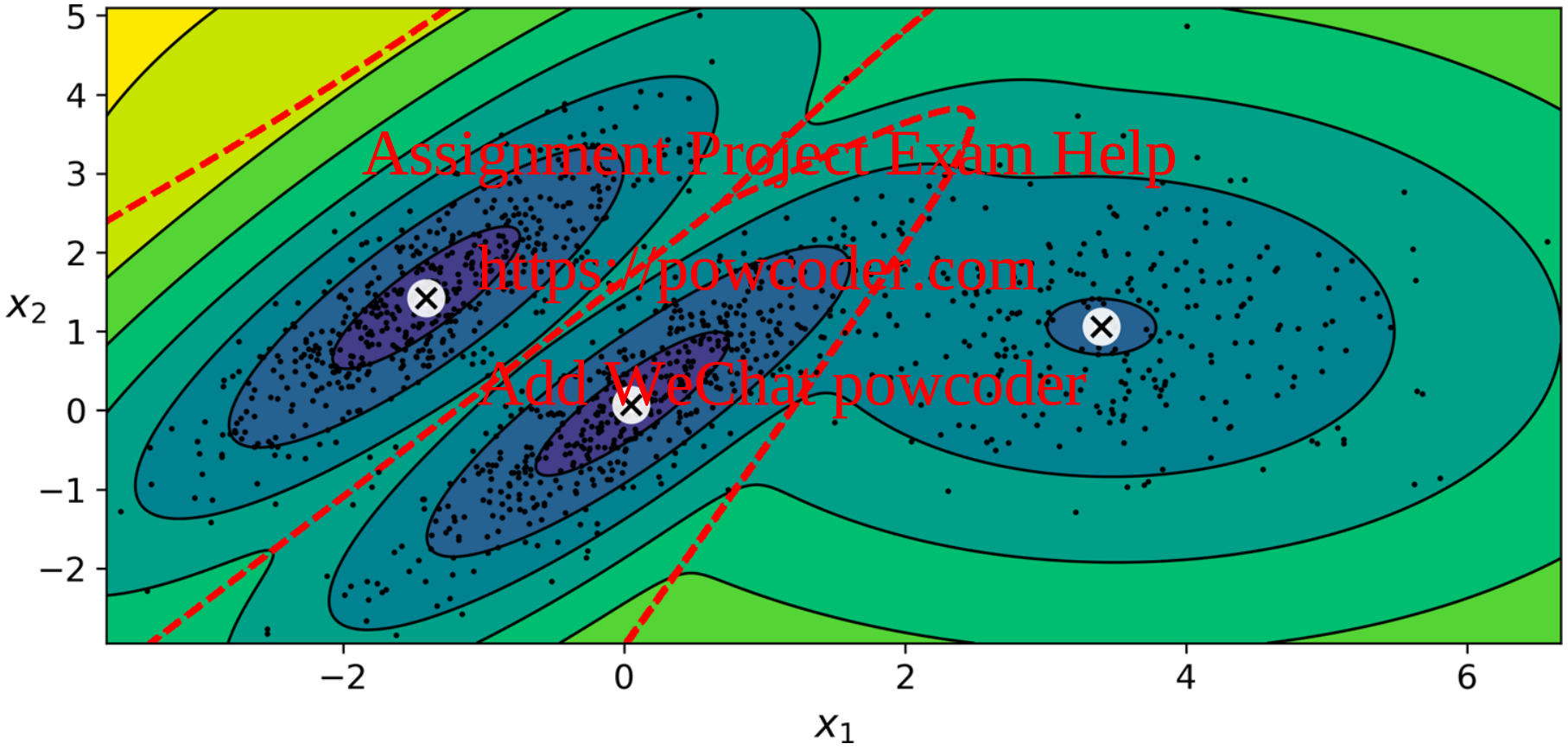
HIDDEN (LATENT) VARIABLES

UCL SCHOOL OF MANAGEMENT



MIXTURE OF GAUSSIANS (2D)

UCL SCHOOL OF MANAGEMENT



GRAPHICAL MODELS

GAUSSIAN MIXTURES

UCL SCHOOL OF MANAGEMENT

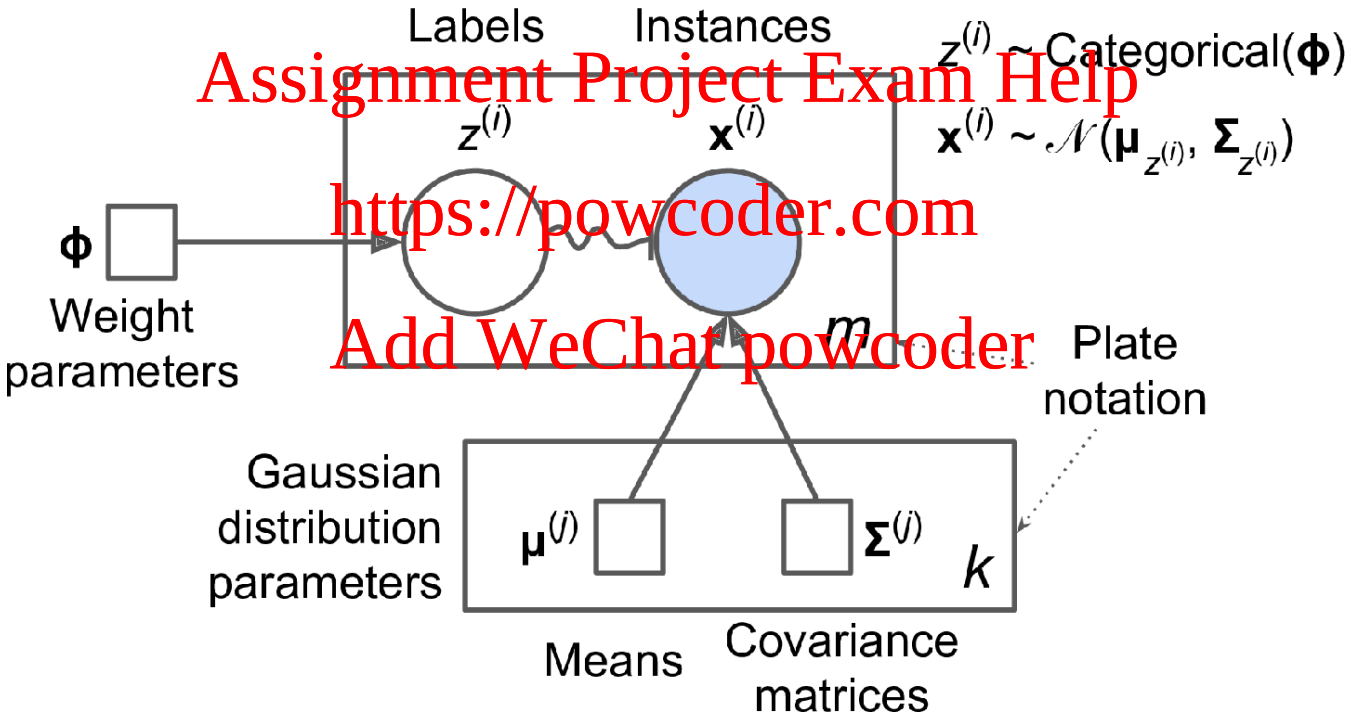
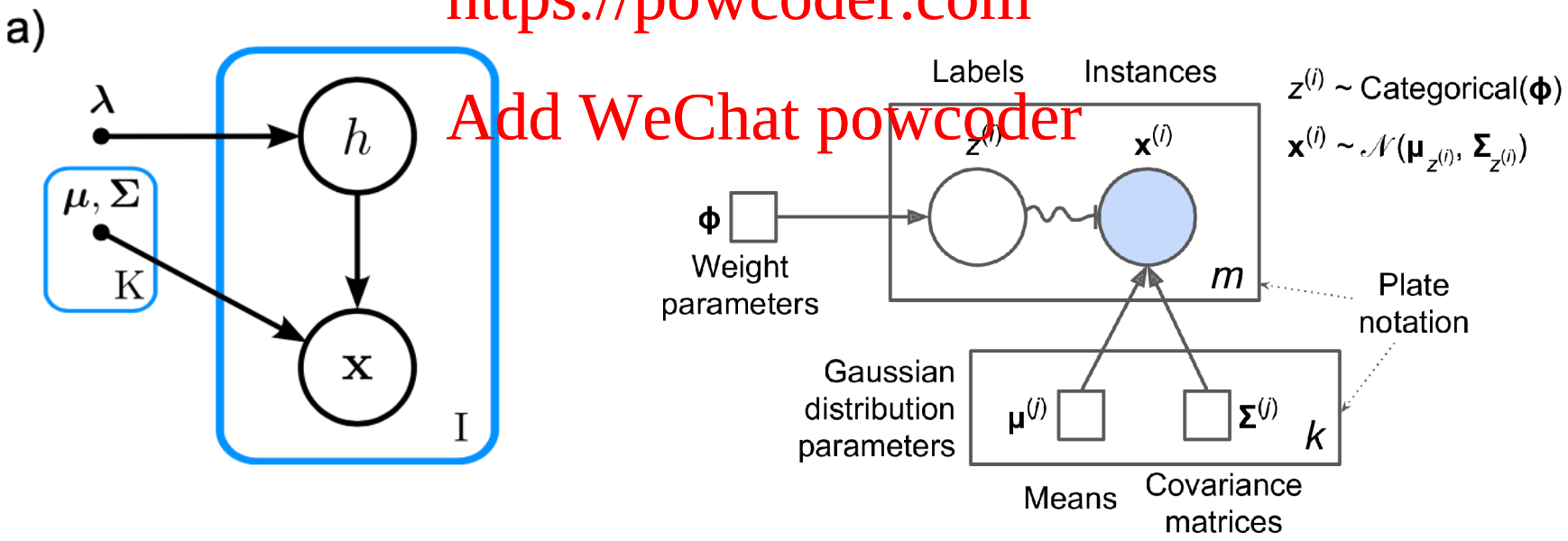


PLATE NOTATION

UCL SCHOOL OF MANAGEMENT

- *including its parameters (squares, solid circles, bullet)*
- *random variables (circles)*
- *conditional dependencies (solid arrows)*



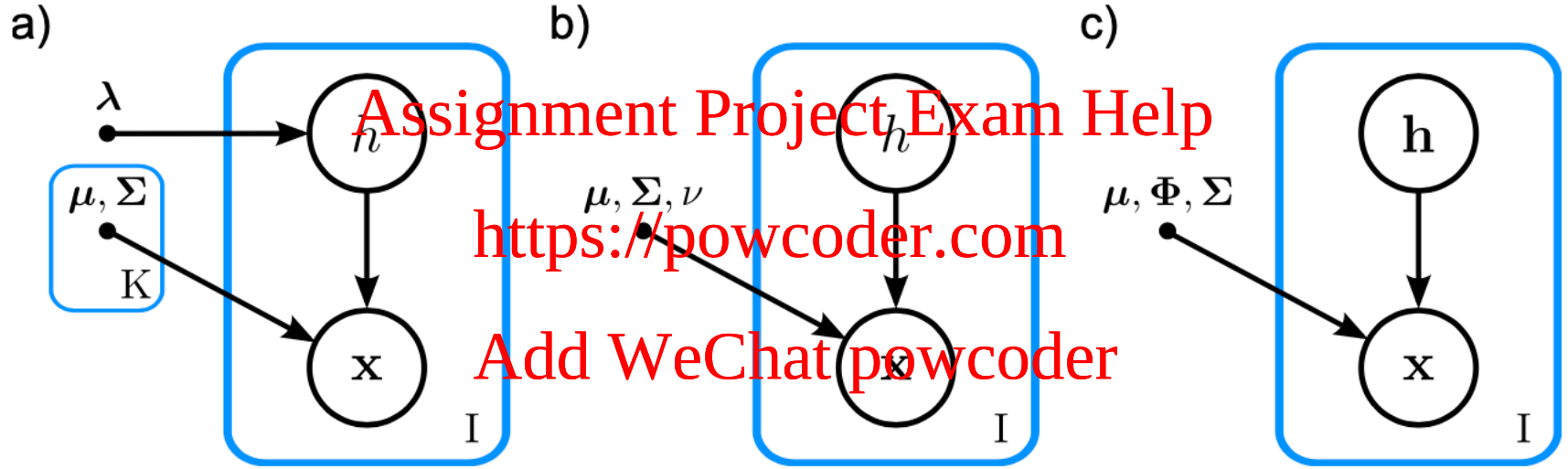
FAMILIES OF MODELS

UCL SCHOOL OF MANAGEMENT

Gaussian mixture

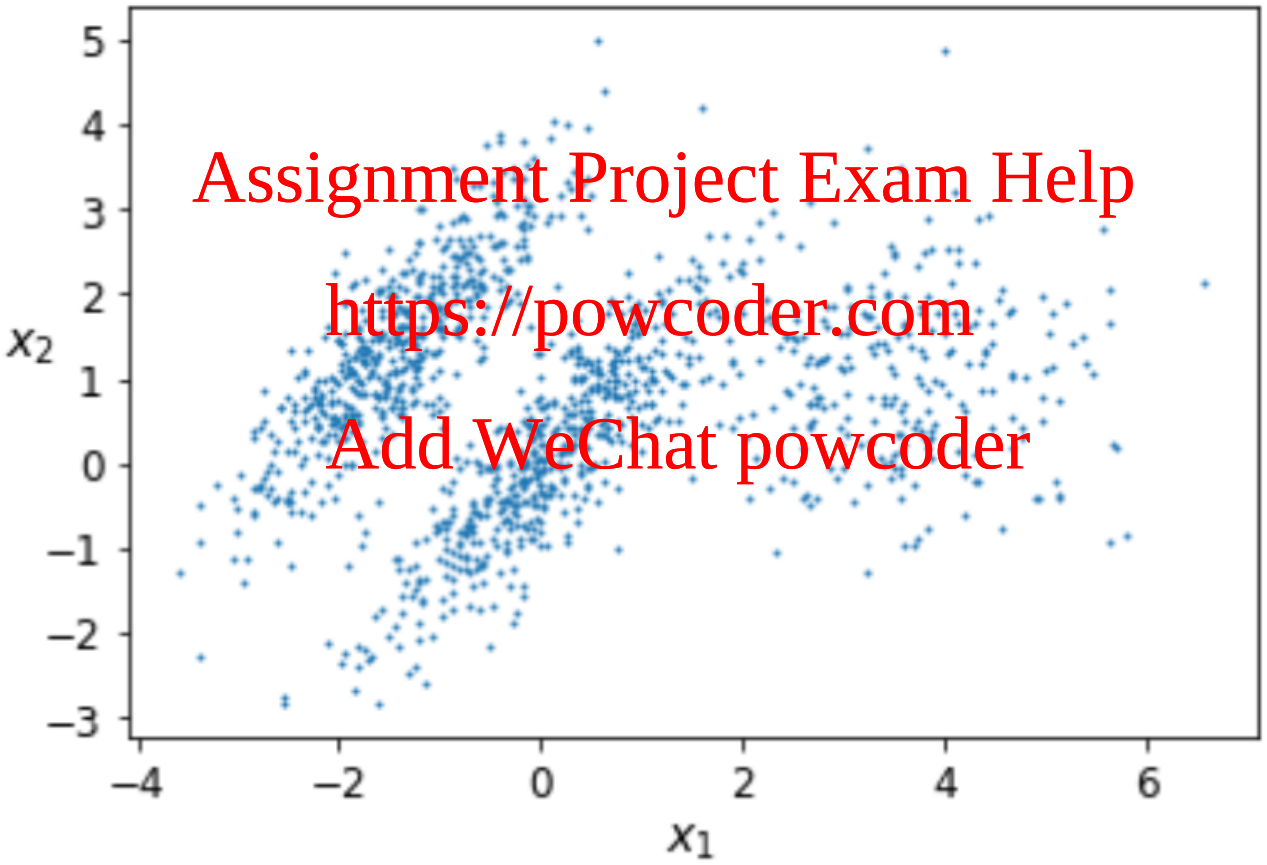
T-distribution mixture

Factor Analysis



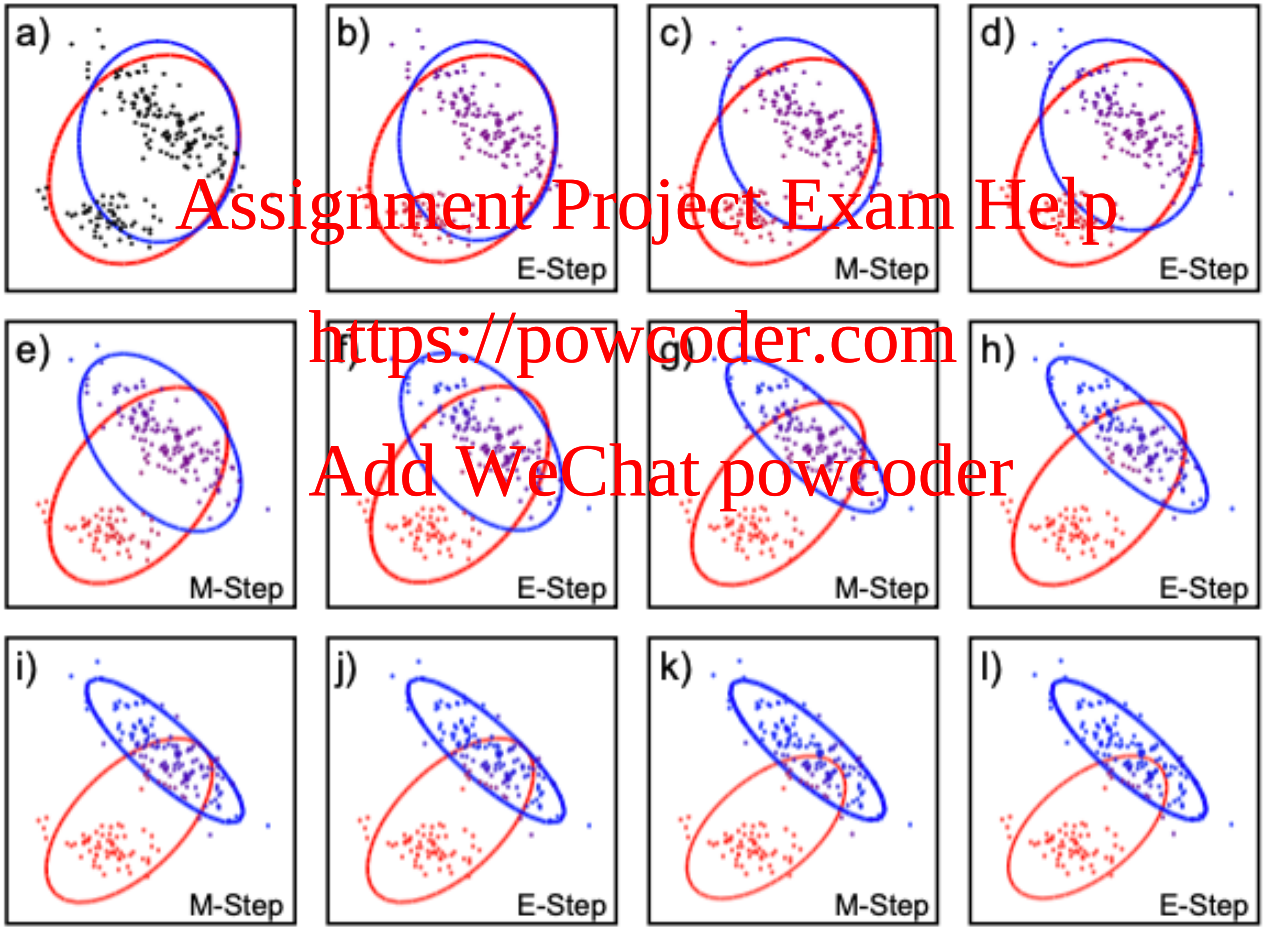
TWO STEP – EM ALGORITHM

UCL SCHOOL OF MANAGEMENT



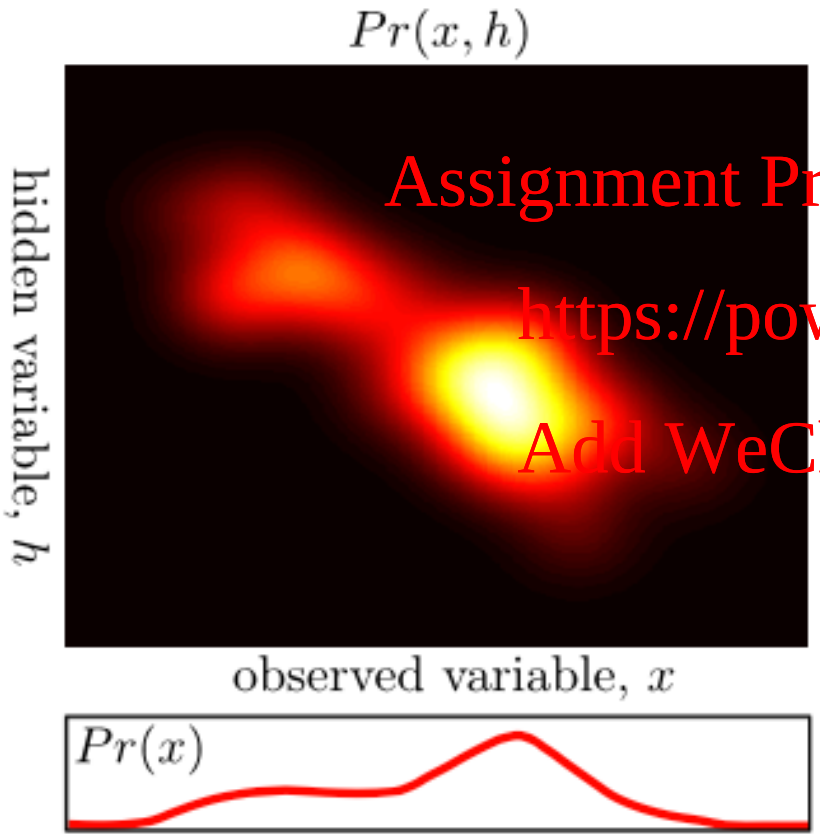
EM ALGORITHM

UCL
SCHOOL OF
MANAGEMENT



EXPECTATION MAXIMIZATION

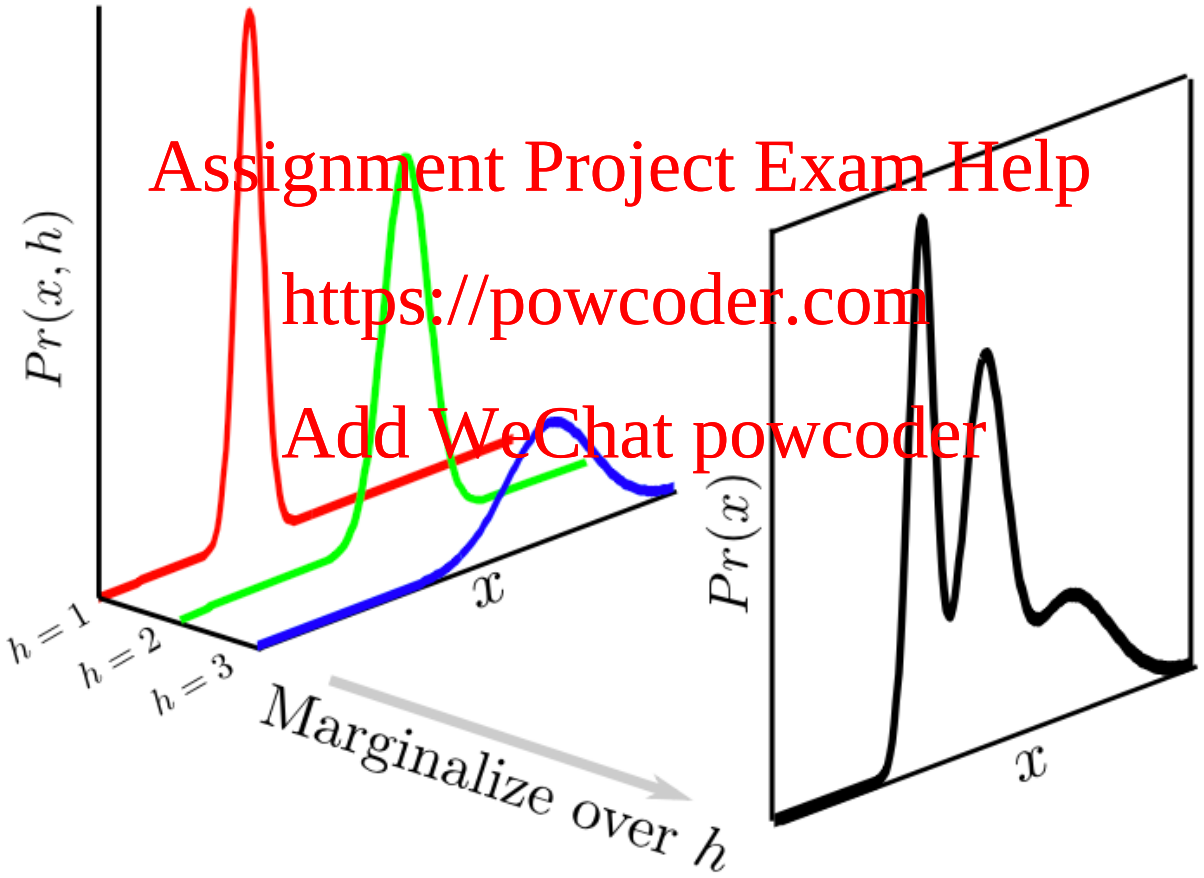
UCL SCHOOL OF MANAGEMENT



$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \log \left[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta}) \, d\mathbf{h}_i \right] \right]$$

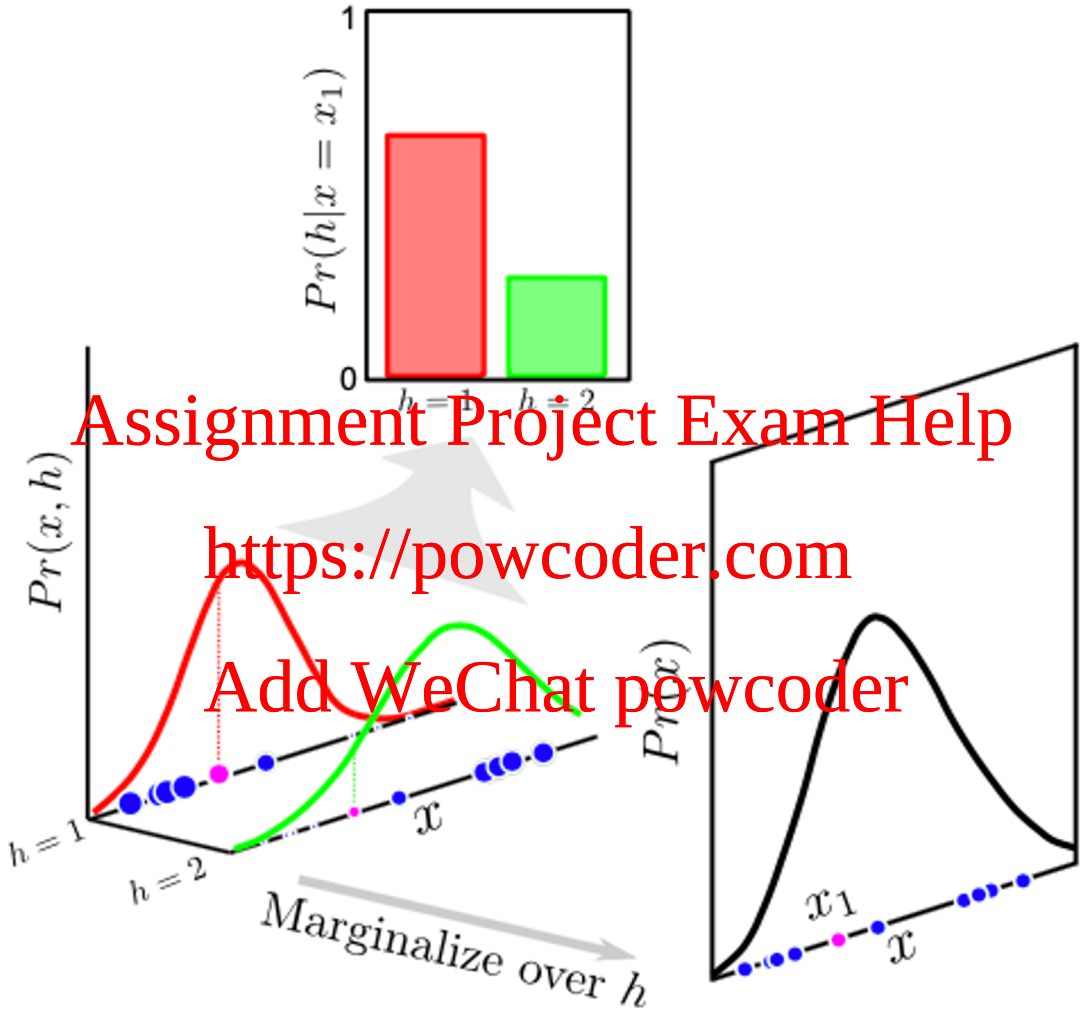
MIXTURE OF GAUSSIANS AS MARGINALIZATION

UCL SCHOOL OF MANAGEMENT



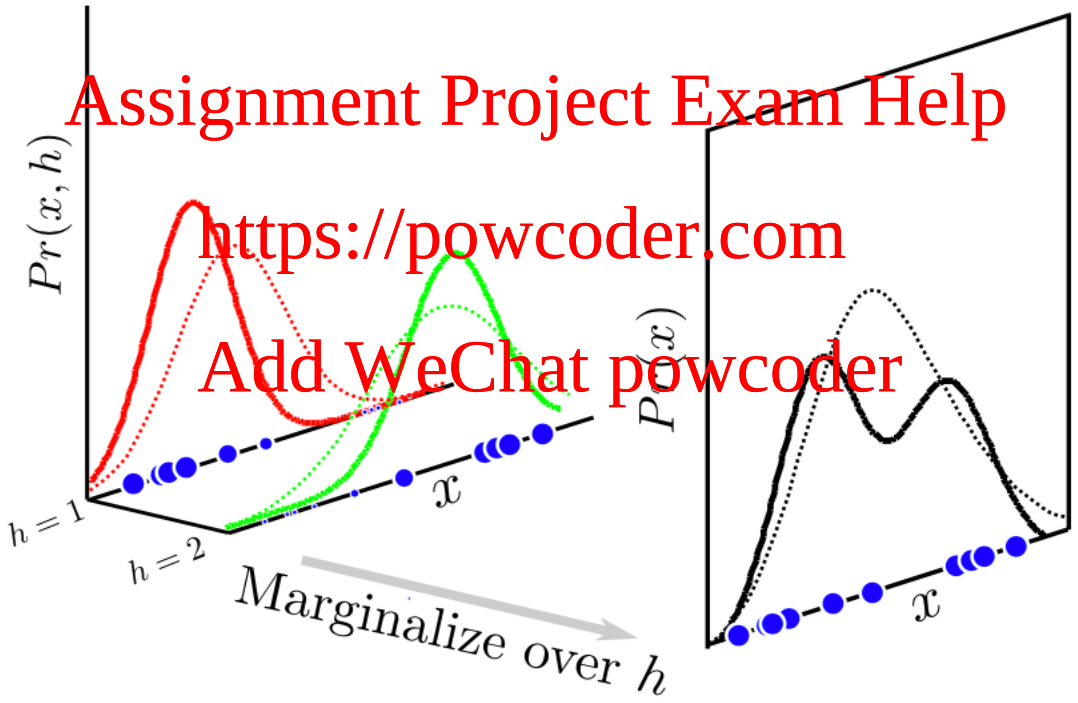
E-STEP

UCL SCHOOL OF MANAGEMENT



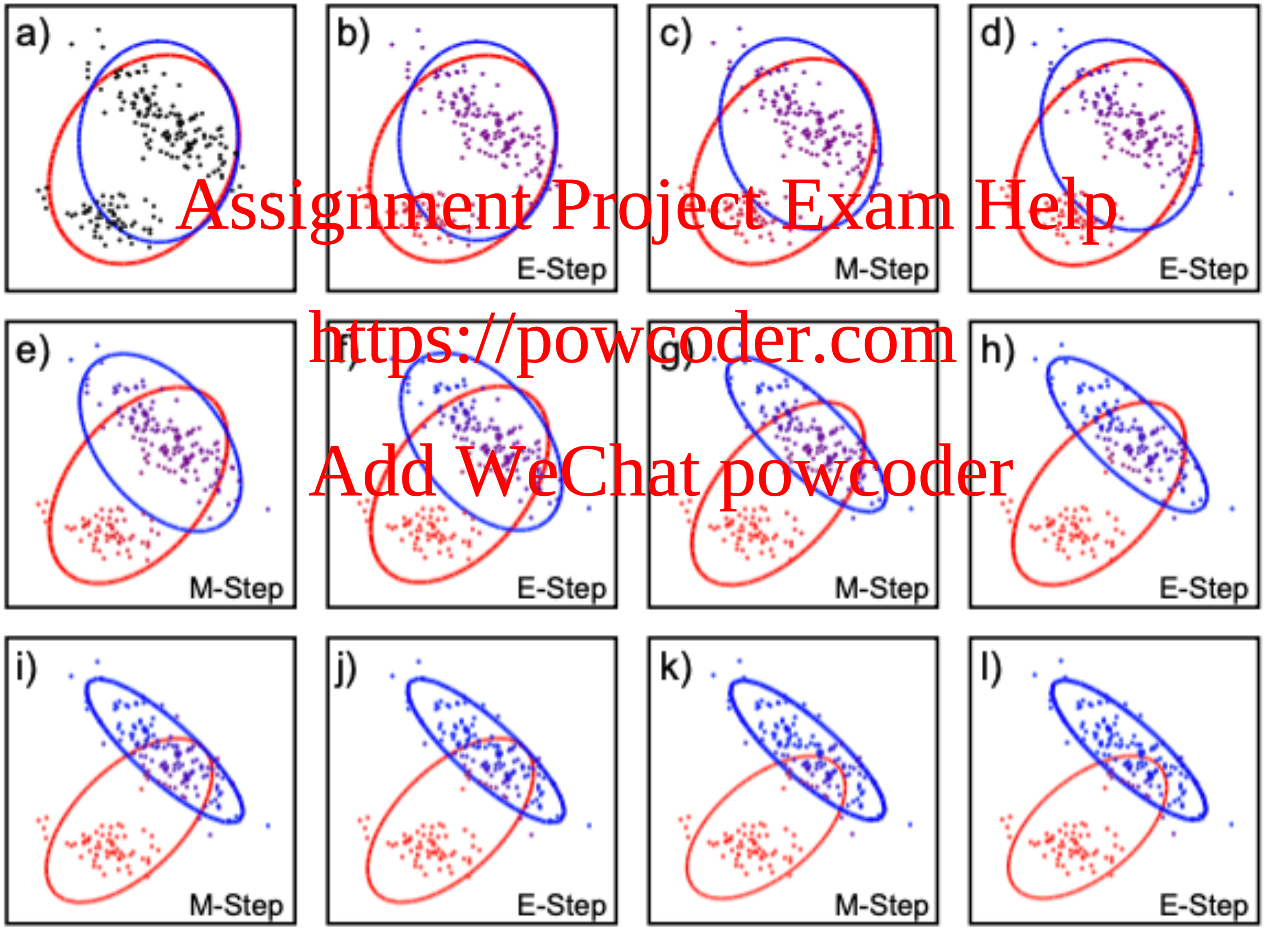
M-STEP

UCL
SCHOOL OF
MANAGEMENT



EM ALGORITHM

UCL
SCHOOL OF
MANAGEMENT



EXPECTATION MAXIMIZATION

UCL SCHOOL OF MANAGEMENT

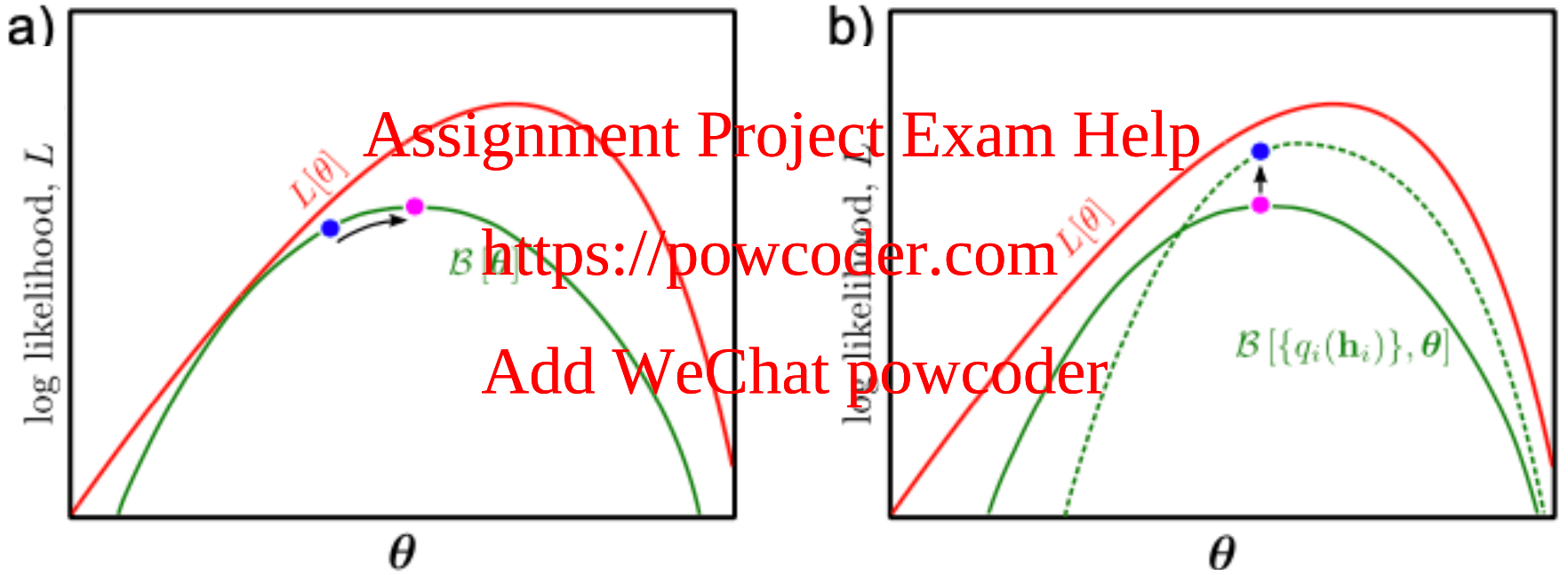
$$p(\mathbf{z}|\mathbf{X}) = \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}} = \frac{p(\mathbf{X}|\mathbf{z})\, p(\mathbf{z})}{p(\mathbf{X})}$$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\sum_{i=1}^{I} \log \left[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})\, d\mathbf{h}_i \right] \right]$$

$$\begin{aligned} \mathcal{B}\left[\{q_i(\mathbf{h}_i)\}, \boldsymbol{\theta}\right] &= \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log \left[\frac{Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)} \right] d\mathbf{h}_i \\ &\leq \sum_{i=1}^{I} \log \left[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})\, d\mathbf{h}_i \right]. \end{aligned}$$

MANIPULATING THE LOWER BOUND

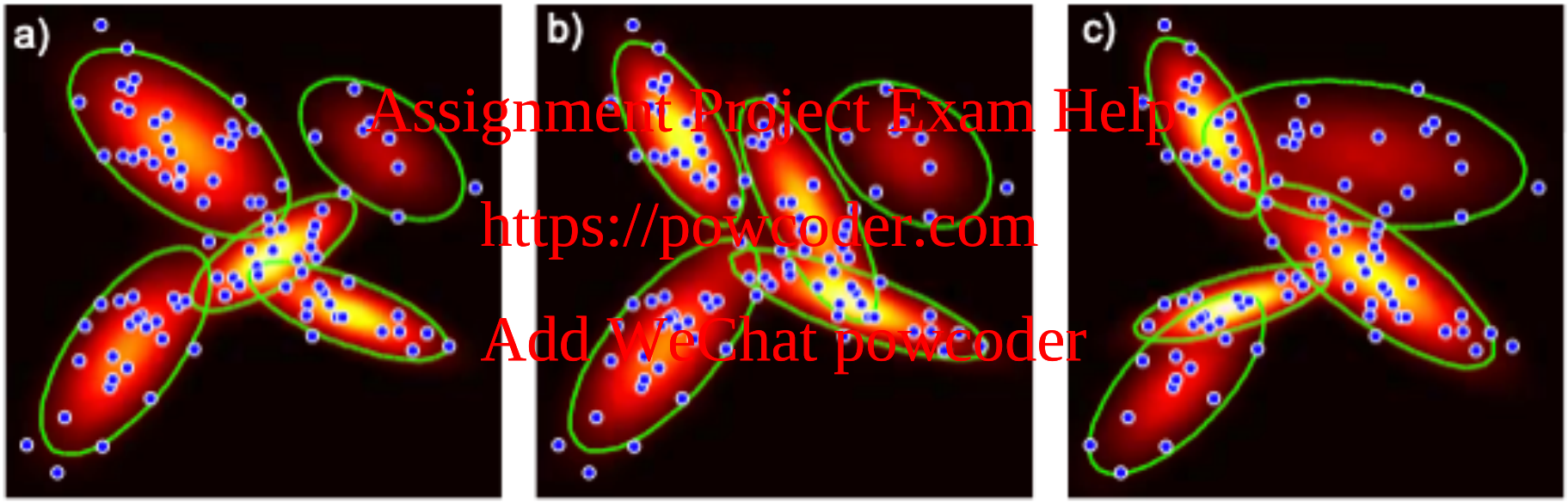
UCL SCHOOL OF MANAGEMENT



$$\begin{aligned}\mathcal{B}[\{q_i(\mathbf{h}_i)\}, \boldsymbol{\theta}] &= \sum_{i=1}^{I} \int q_i(\mathbf{h}_i) \log\left[\frac{Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})}{q_i(\mathbf{h}_i)}\right] d\mathbf{h}_i \\ &\leq \sum_{i=1}^{I} \log\left[\int Pr(\mathbf{x}_i, \mathbf{h}_i|\boldsymbol{\theta})\, d\mathbf{h}_i\right].\end{aligned}$$

LOCAL MAXIMA

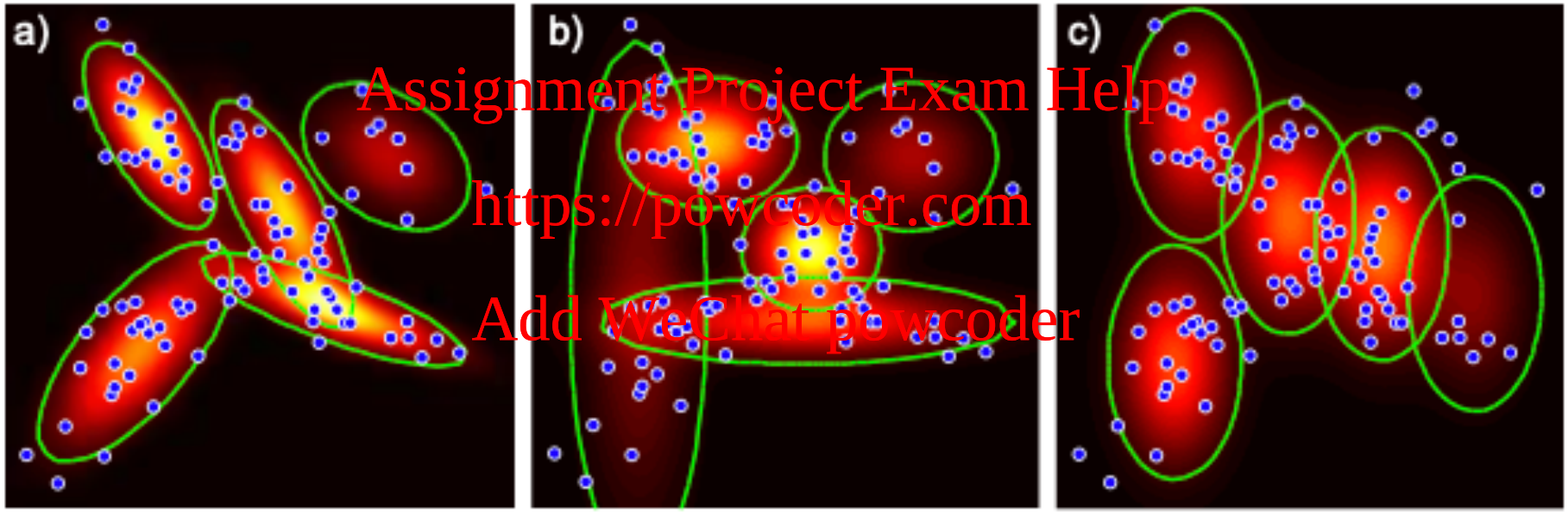
UCL SCHOOL OF MANAGEMENT



Repeated fitting of mixture of Gaussians model with different starting points results in different models as the fit converges to different local maxima.
Log likelihoods are a) 98.76 b) 96.97 c) 94.35, respectively, indicating that (a) is the best fit.

COVARIANCE COMPONENTS

UCL
SCHOOL OF
MANAGEMENT



a) Full covariances.
b) Diagonal covariances.
c) Identical diagonal covariances.

LEARNING GMM

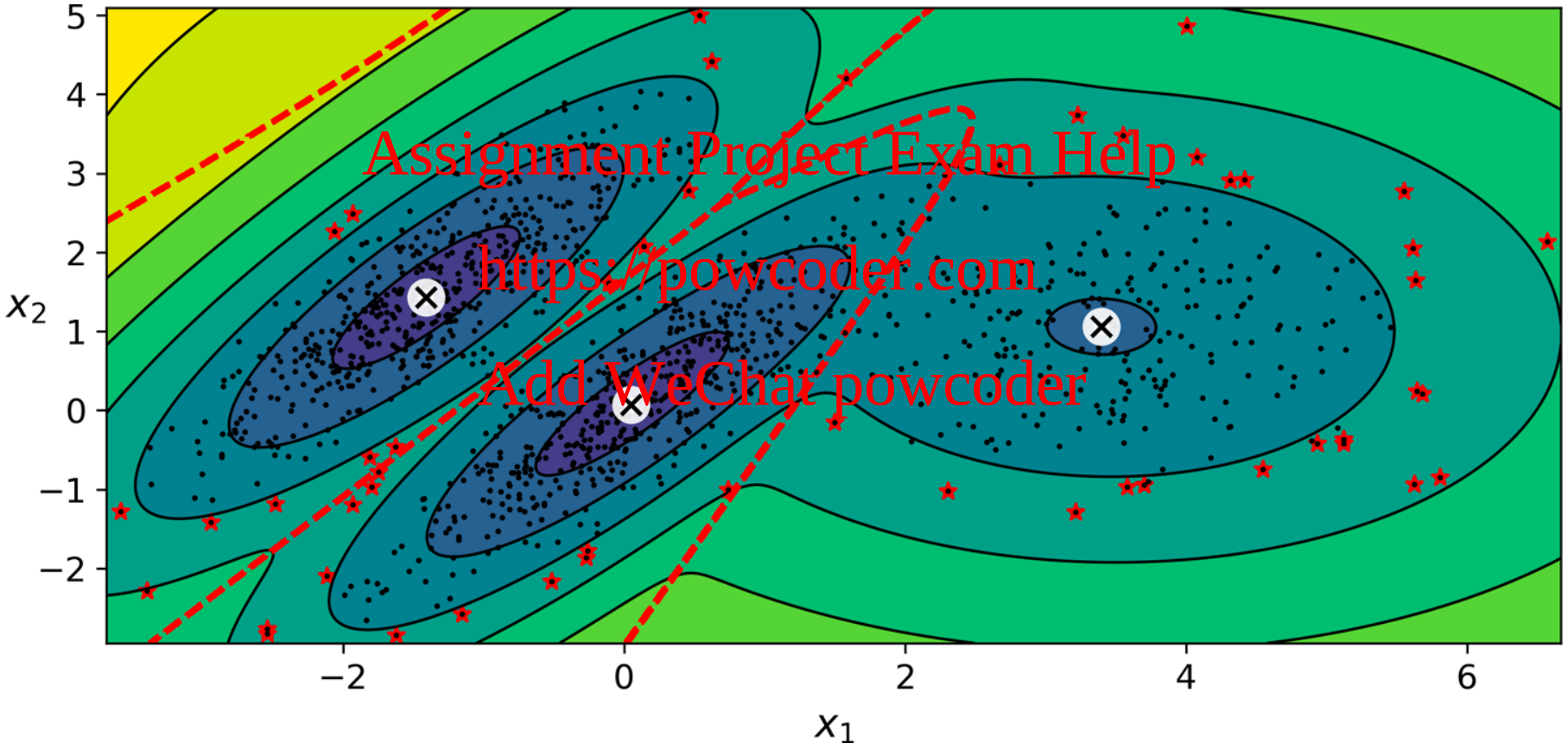
PSEUDO CODE

UCL SCHOOL OF MANAGEMENT

Input : Training data $\{\mathbf{x}_i\}_{i=1}^{I}$, number of clusters K
Output: ML estimates of parameters $\boldsymbol{\theta} = \{\lambda_{1\ldots K}, \boldsymbol{\mu}_{1\ldots K}, \boldsymbol{\Sigma}_{1\ldots K}\}$
begin
  Initialize $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ [a]
  repeat
    // Expectation Step
    for i=1 **to** I **do**
      for k=1 **to** K **do**
        $l_{ik} = \lambda_k \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]$ // numerator of Bayes' rule
      end
      // Compute posterior (responsibilities) by normalizing
      for k=1 **to** K **do**
        $r_{ik} = l_{ik}/(\sum_{k=1}^{K} l_{ik})$
      end
    end
    // Maximization Step [b]
    for k=1 **to** K **do**
      $\lambda_k^{[t+1]} = (\sum_{i=1}^{I} r_{ik})/(\sum_{k=1}^{K}\sum_{i=1}^{I} r_{ik})$
      $\boldsymbol{\mu}_k^{[t+1]} = (\sum_{i=1}^{I} r_{ik}\mathbf{x}_i)/(\sum_{i=1}^{I} r_{ik})$
      $\boldsymbol{\Sigma}_k^{[t+1]} = (\sum_{i=1}^{I} r_{ik}(\mathbf{x}_i - \boldsymbol{\mu}_k^{[t+1]})(\mathbf{x}_i - \boldsymbol{\mu}_k^{[t+1]})^T)/(\sum_{i=1}^{I} r_{ik})$.
    end
    // Compute Data Log Likelihood and EM Bound
    $L = \sum_{i=1}^{I} \log\left[\sum_{k=1}^{K} \lambda_k \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]\right]$
    $B = \sum_{i=1}^{I}\sum_{k=1}^{K} r_{ik} \log[\lambda_k \text{Norm}_{\mathbf{x}_i}[\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k]/r_{ik}]$
  until *No further improvement in* L
end

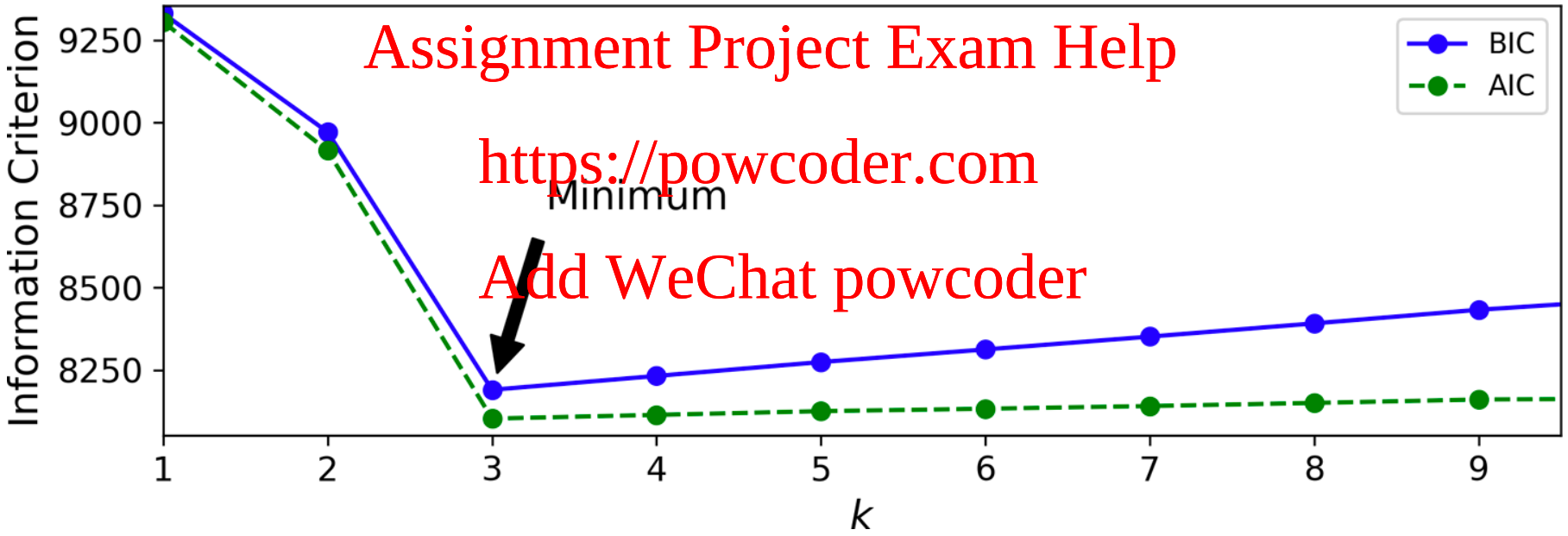
ANOMALY DETECTION

UCL
SCHOOL OF
MANAGEMENT



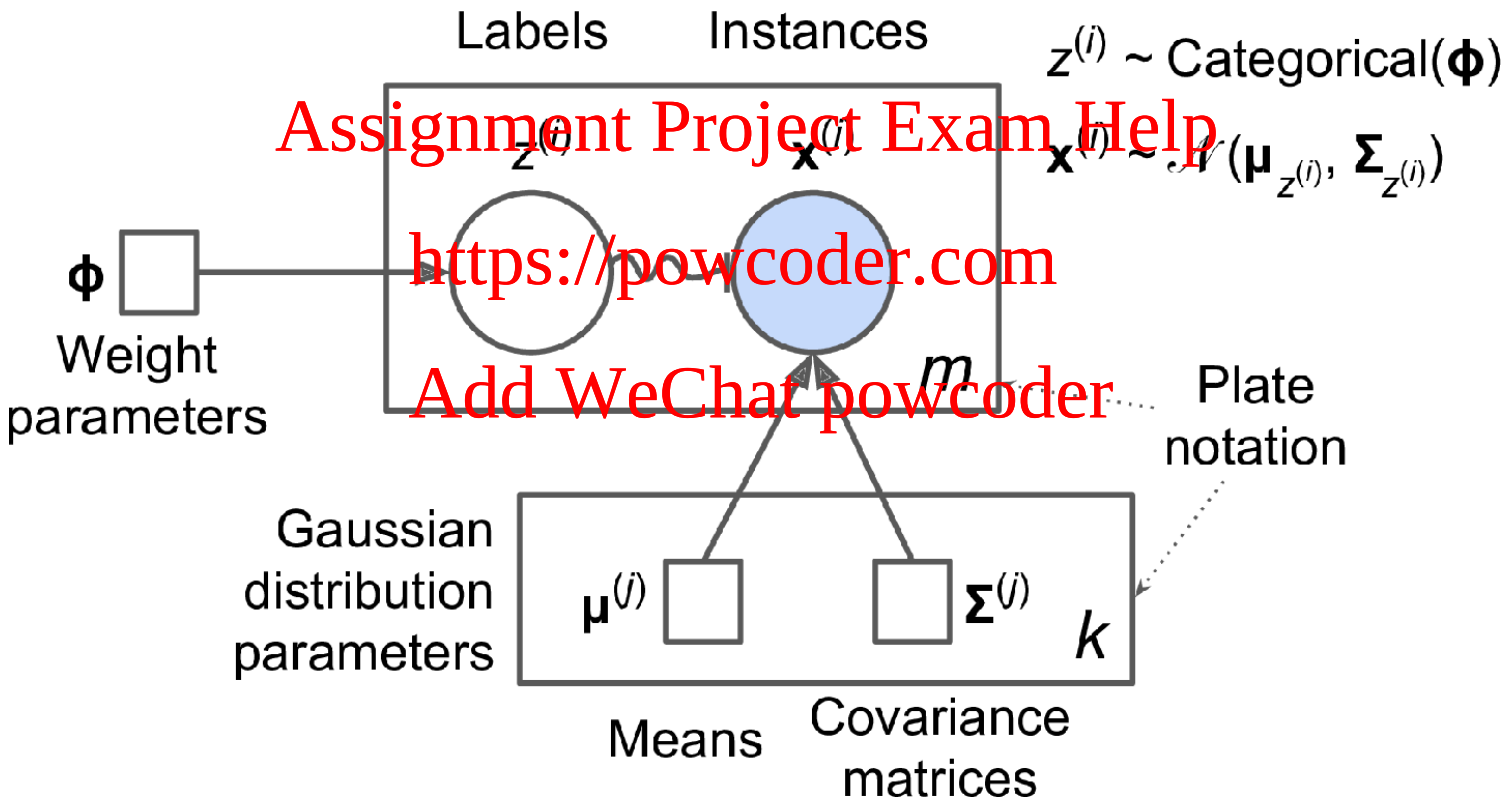
BIC AND AIC

UCL
SCHOOL OF
MANAGEMENT



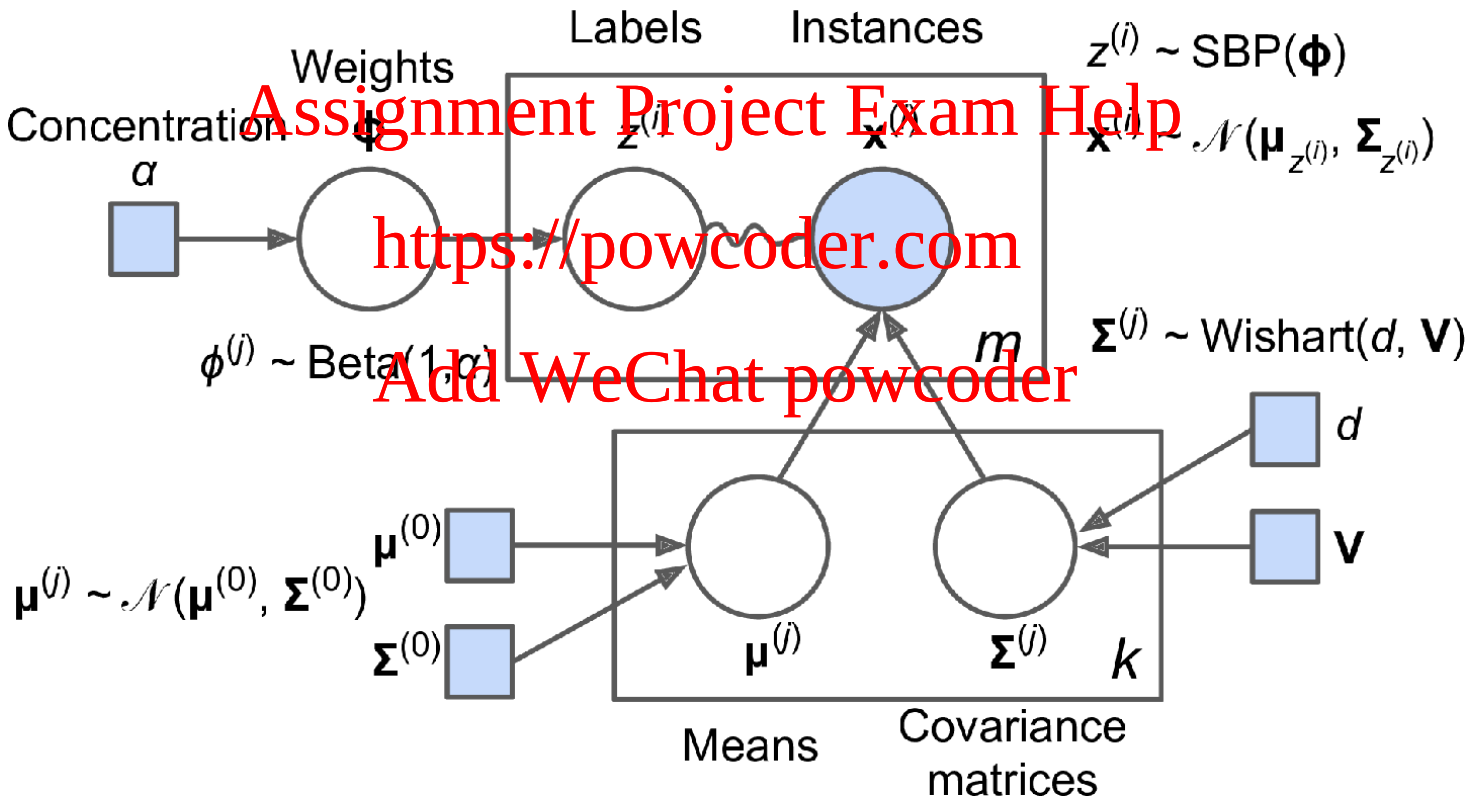
GAUSSIAN MIXTURES

UCL SCHOOL OF MANAGEMENT



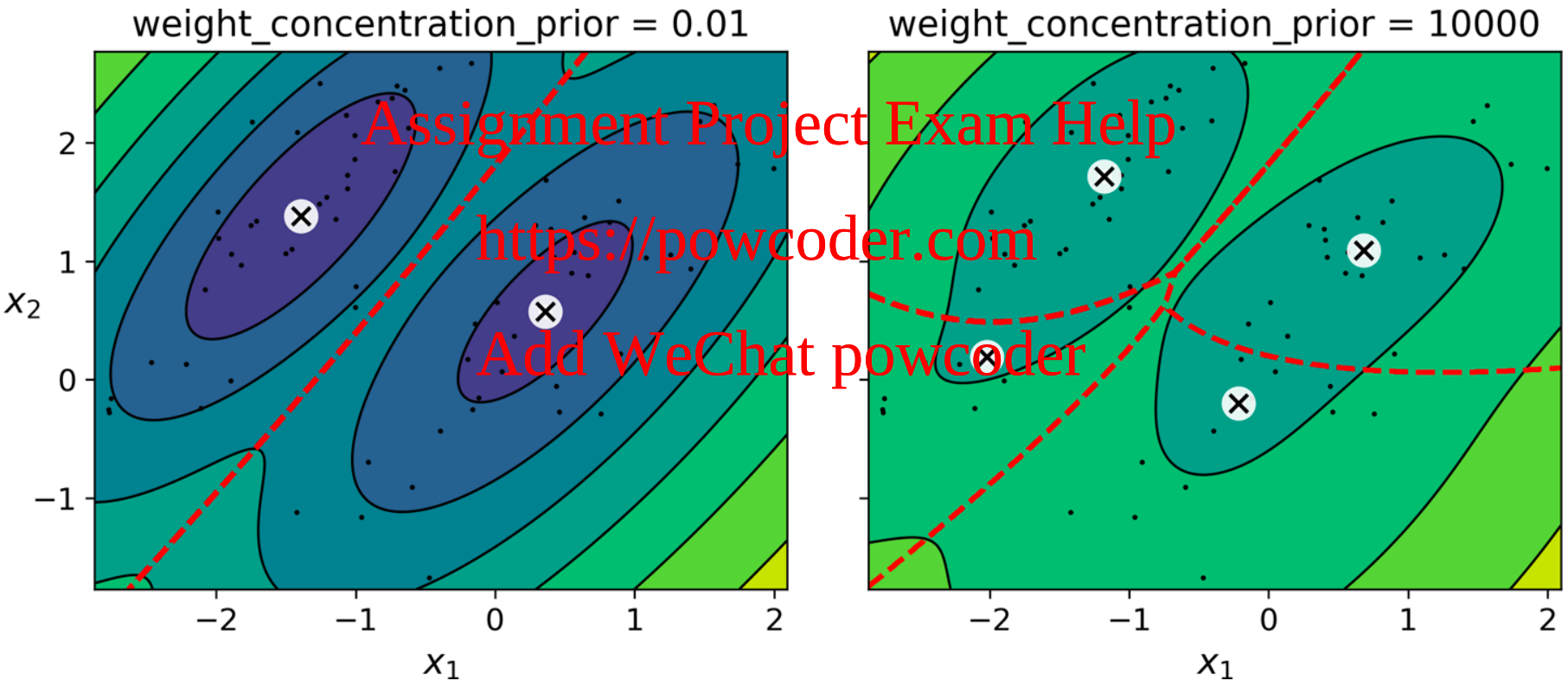
BAYESIAN GMMS

UCL SCHOOL OF MANAGEMENT



CONCENTRATION PRIORS

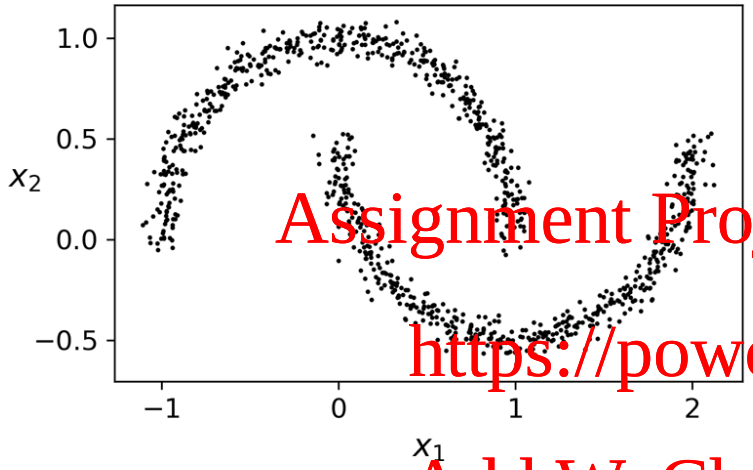
UCL SCHOOL OF MANAGEMENT

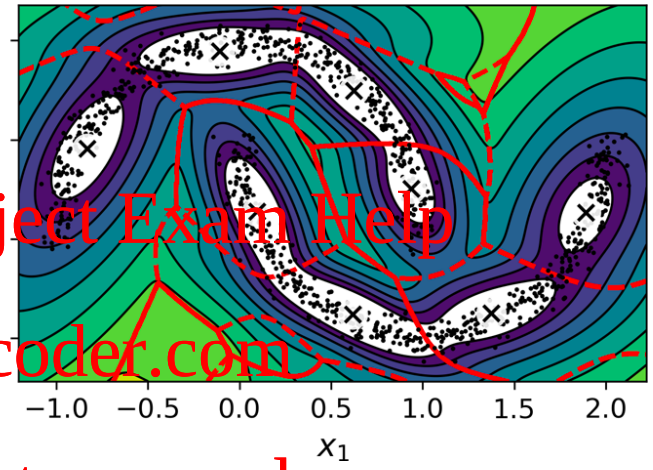


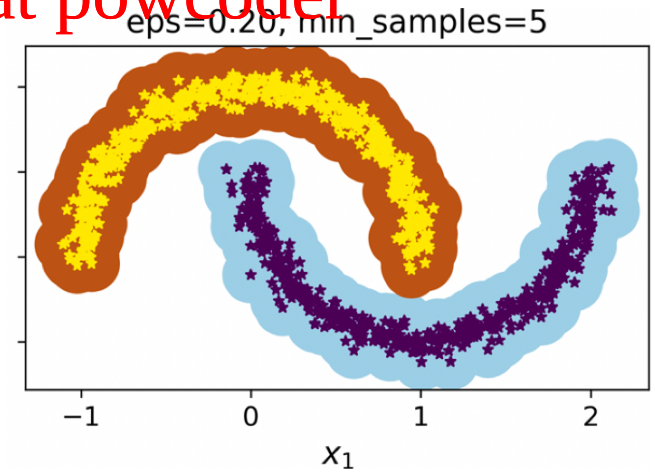
The more data we have, however, the less the priors matter. In fact, to plot diagrams with such large differences, you must use very strong priors and little data.

TWO MOONS DATA

UCL SCHOOL OF MANAGEMENT

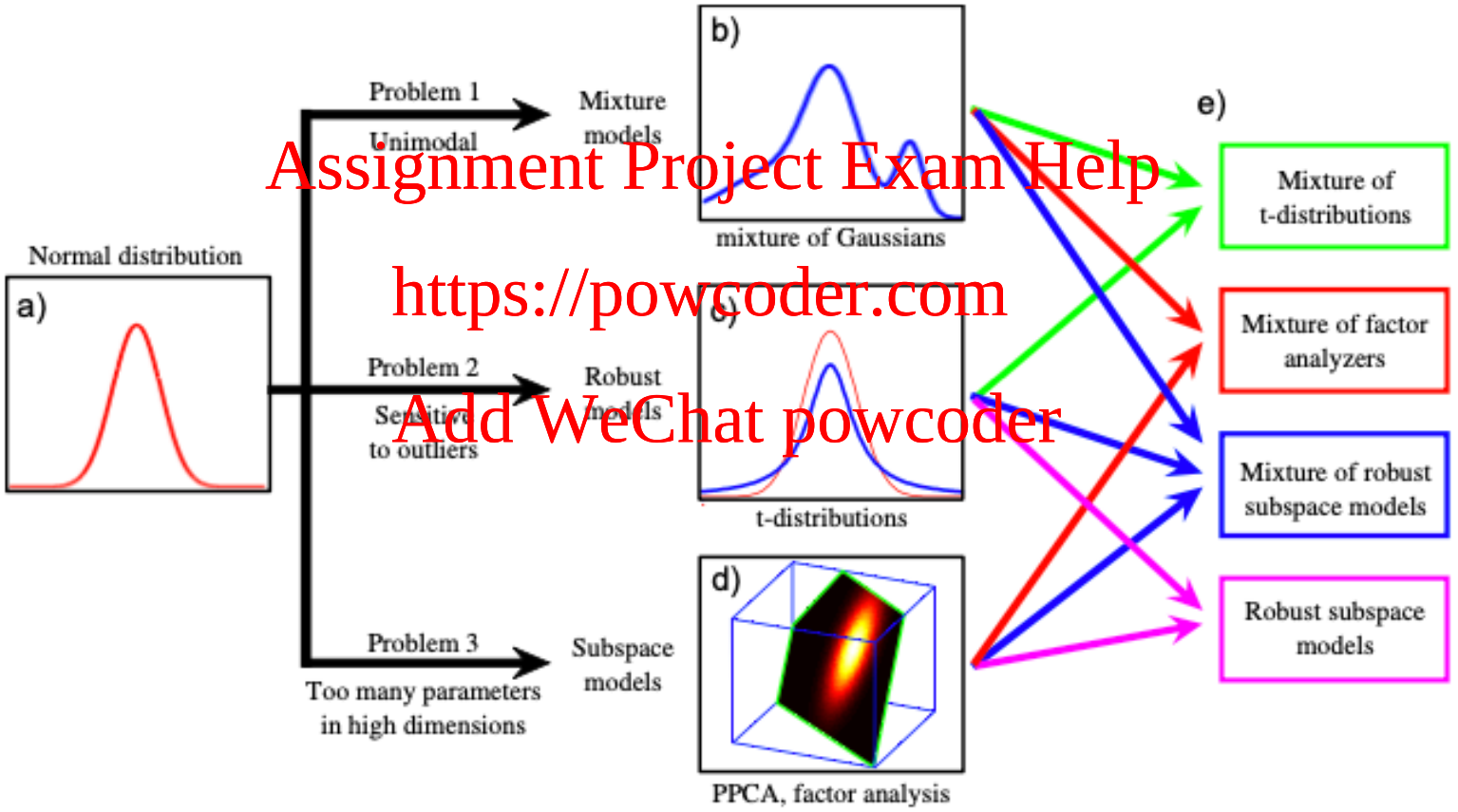






PROBLEMS WITH MULTI-VARIATE NORMAL DENSITY

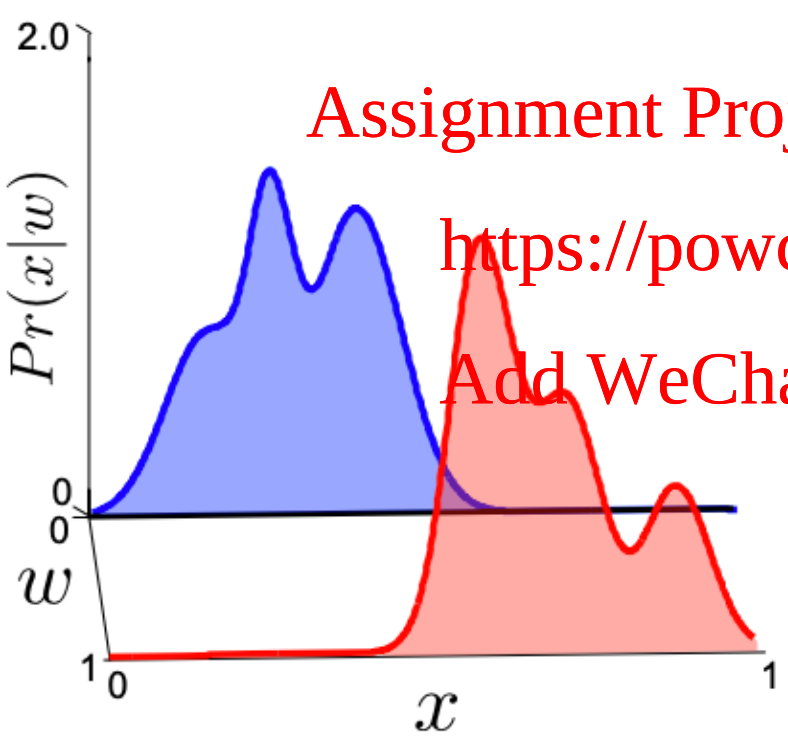
UCL SCHOOL OF MANAGEMENT

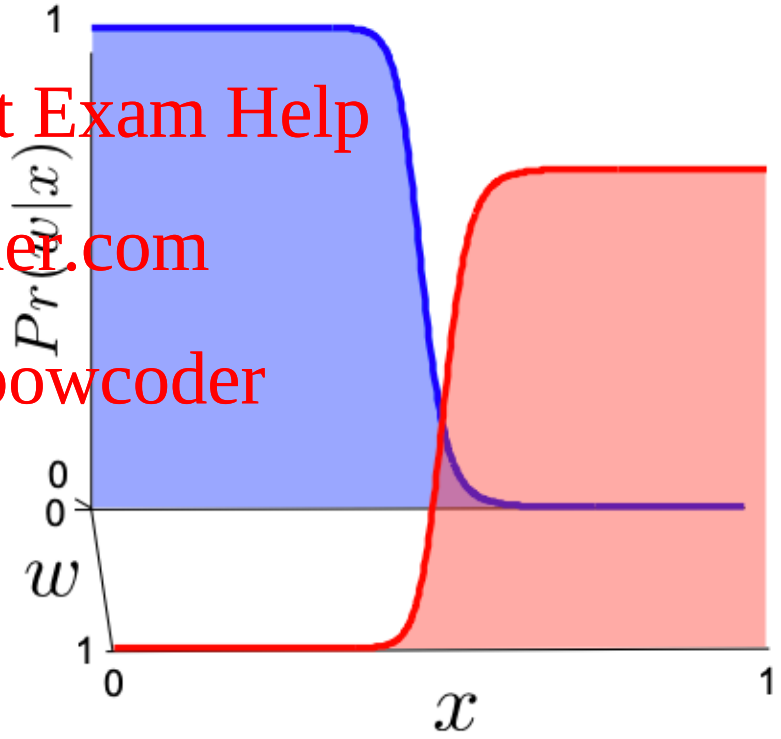


Types of models

GENERATIVE VS DISCRIMINATIVE

UCL SCHOOL OF MANAGEMENT

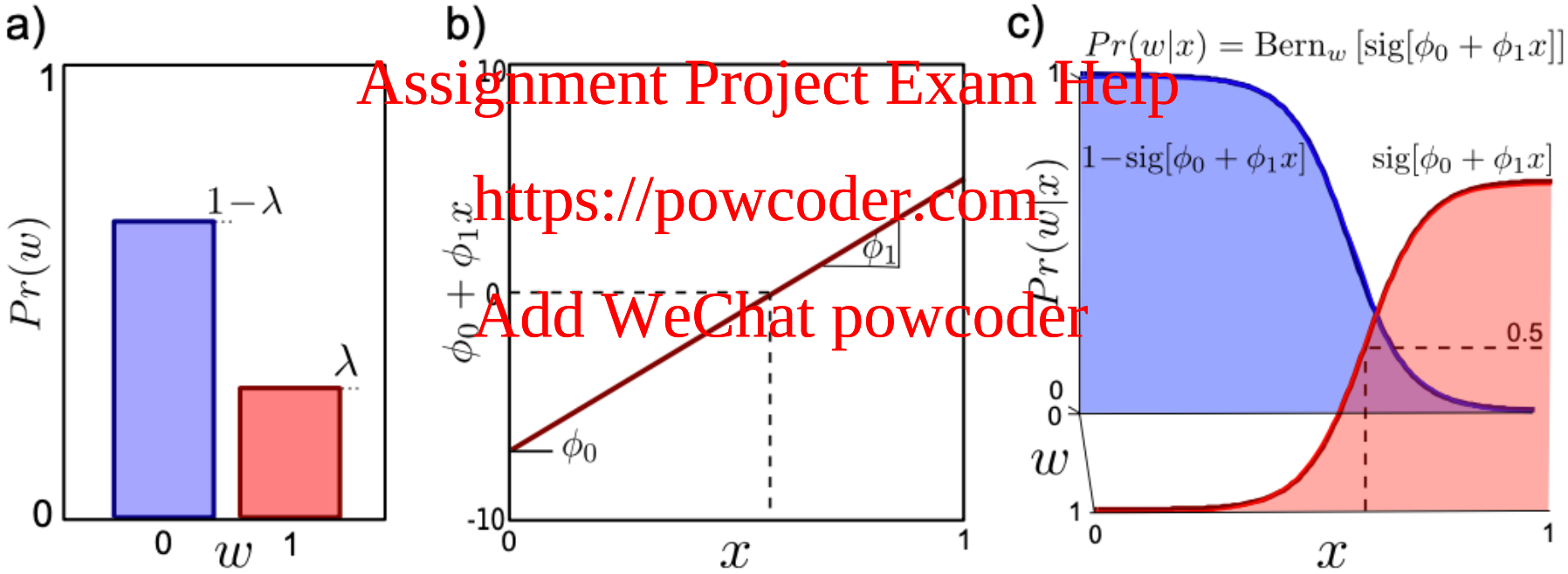




CLASSIFICATION (DISCRIMINATIVE)

LOGISTIC REGRESSION REVISITED
MODEL CONTINGENCY OF THE WORLD ON DATA

UCL SCHOOL OF MANAGEMENT



World state:
Bernoulli distribution

Linear model

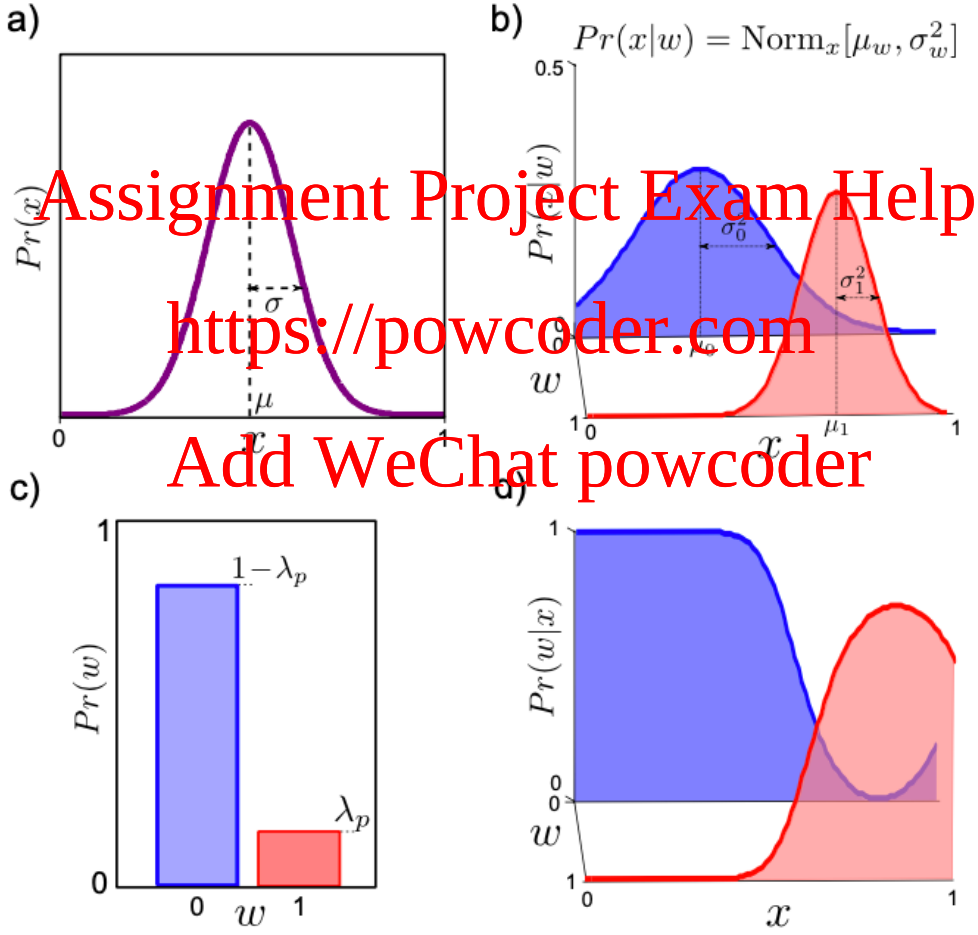
Probability /
Decision surface

CLASSIFICATION (GENERATIVE)

GAUSSIAN MIXTURE

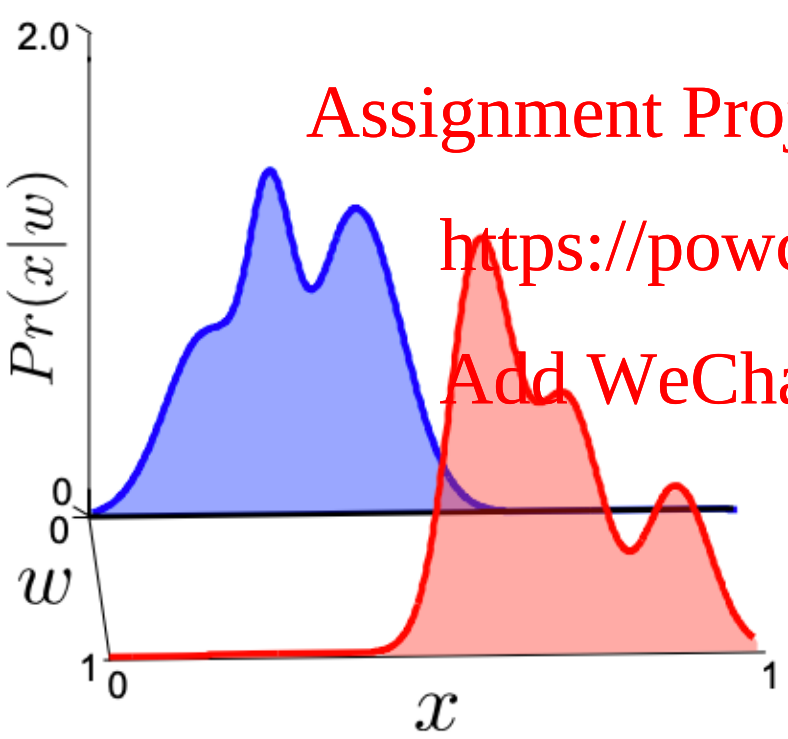
MODEL CONTINGENCY OF DATA ON THE WORLD

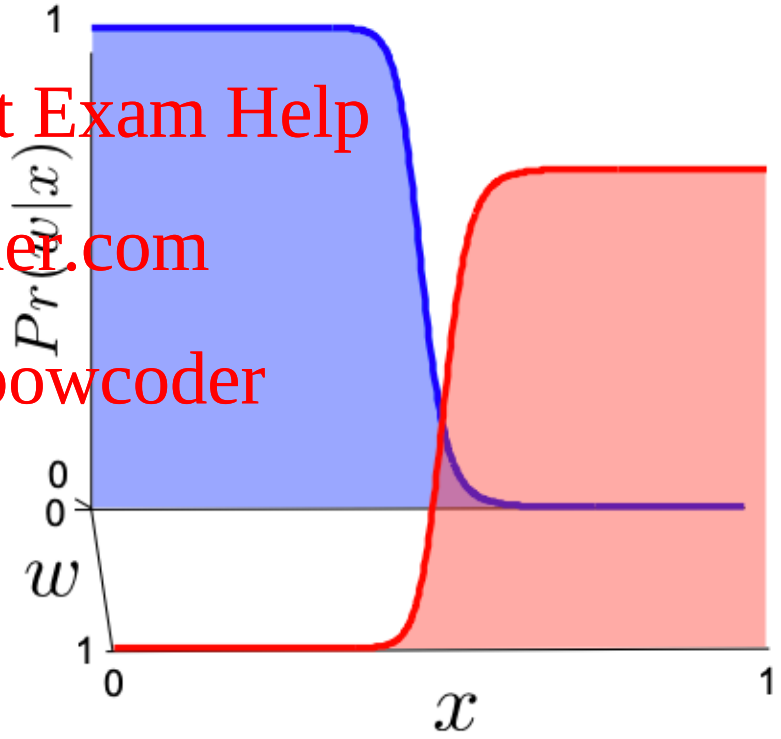
UCL SCHOOL OF MANAGEMENT



WHAT SORT OF MODEL SHOULD WE USE?

UCL SCHOOL OF MANAGEMENT





WHAT SORT OF MODEL SHOULD WE USE?

TL;DR NO DEFINITIVE ANSWER

UCL SCHOOL OF MANAGEMENT

- Inference is generally simpler with discriminative models.
- Generative models calculate this probability via Bayes' rule, and sometimes this requires a computationally expensive algorithm.
- Generative models might waste modelling power.
  The data are generally of much higher dimension than the world, and modelling it is costly. Moreover, there may be many aspects of the data which do not influence the state;
- Using discriminative approaches, it is harder to exploit this knowledge: essentially we have to re-learn these phenomena from the data.
- Sometimes parts of the training or test data vector **x** may be missing. Here, generative models are preferred.
- It is harder to impose prior knowledge in a principled way in discriminative models.

SUMMARY OF APPROACHES

UCL SCHOOL OF MANAGEMENT

	Model $Pr(w\|x)$	Model $Pr(x\|w)$
Regression $x \in [-\infty, \infty], w \in [-\infty, \infty]$	Linear regression	Linear regression
Classification $x \in [-\infty, \infty], w \in \{0, 1\}$	Logistic regression	Probability density function

BEST PRACTICE…

UCL SCHOOL OF MANAGEMENT

Julius Adebayo
@julius_adebayo

Reading through open source code for a nice paper, found this gem :)

```
# This controls the behavior of BatchNormalization layers.
# For some reason I don't fully understand, even though we are
# not training these models where this flag is used,
# it must be set to True or training does not converge.
TRAINING_KWARG_FOR_SECOND_MODEL = True
```

8:54 AM · Feb 8, 2020 · Twitter Web App

7 Retweets **61** Likes

Sean Augenstein @seanAugenstein · 23h
Replying to @julius_adebayo
Hi, I'm one of the authors of this paper! (and the #tensorflowfederated #tff GAN code). Thanks for appreciating this and not finding it infuriating :-) Credit to my esteemed colleague Brendan McMahan for the exact wording of the comment. We have some scars from debugging this.

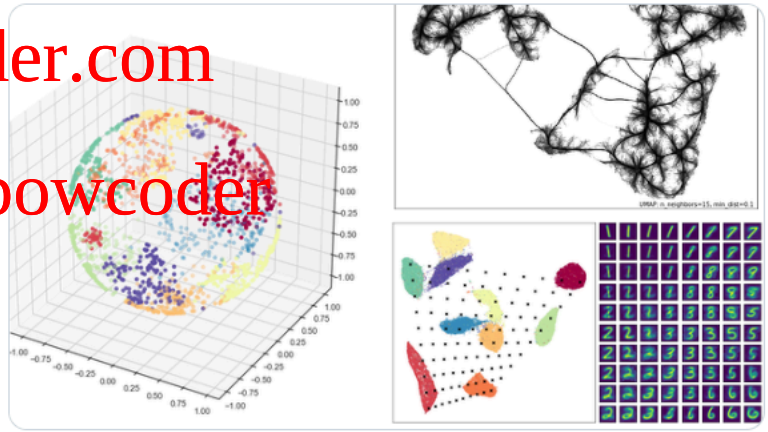
1 4

Sean Augenstein @seanAugenstein · 23h
I'll be @ICLR2020 presenting this work, if you're attending I'd love to chat about using generative models with #differentialprivacy and #federatedlearning to gain insights into decentralized data in a privacy-preserving way (for debugging ML models and other use cases).

1 2

Leland McInnes
@leland_mcinnes

The first release candidate for UMAP 0.4 is out providing lots of new features, including performance improvements, embedding to different manifolds, inverse transform, and plotting tools.



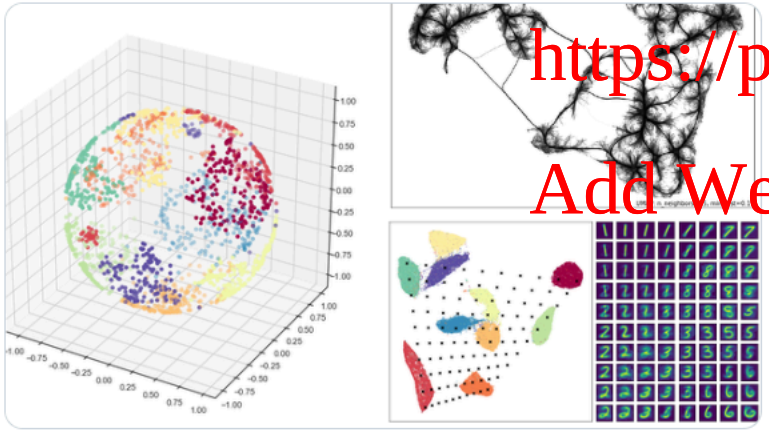
5:35 PM · Feb 9, 2020 · Twitter Web App

BEST PRACTICE…

UCL SCHOOL OF MANAGEMENT

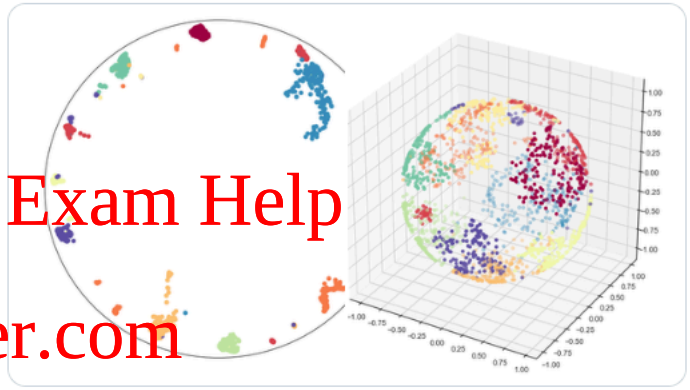
Leland McInnes
@leland_mcinnes

The first release candidate for UMAP 0.4 is out providing lots of new features, including performance improvements, embedding to different manifolds, inverse transform, and plotting tools.



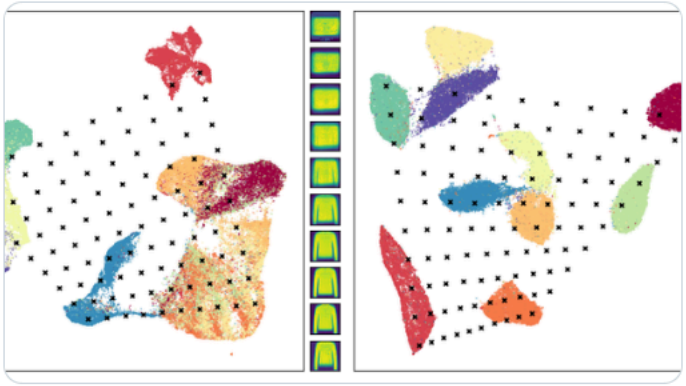
5:35 PM · Feb 9, 2020 · Twitter Web App

Leland McInnes @leland_mcinnes · 3h
Replying to @leland_mcinnes
UMAP 0.4 supports embedding to non-Euclidean manifolds, including spheres, Poncare disks, and more. umap-learn.readthedocs.io/en/latest/embe...



2 3 26

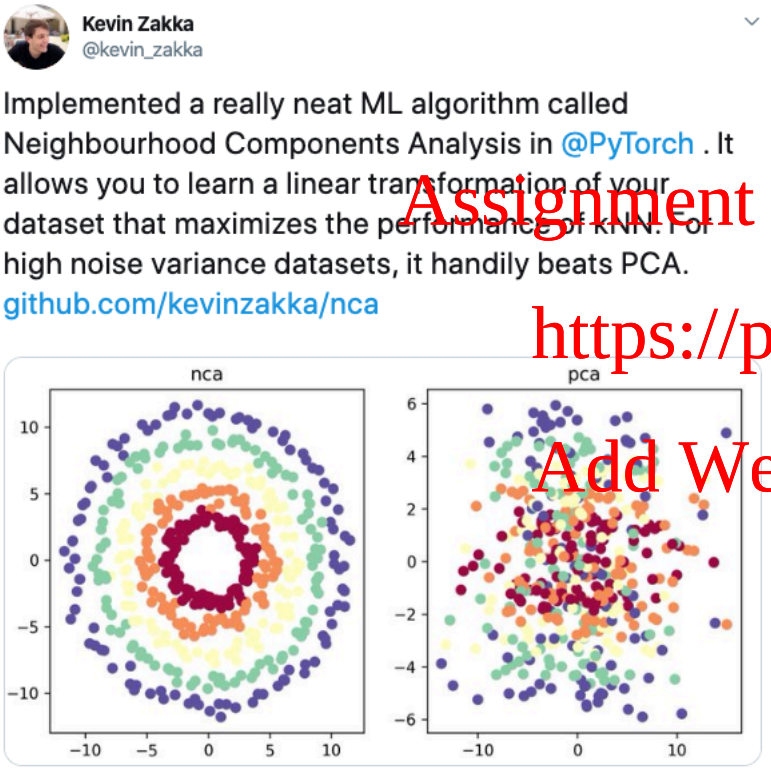
Leland McInnes @leland_mcinnes · 3h
Support for an "inverse transform" has been added to UMAP 0.4, providing the ability to generate a high dimensional representation of a point in the embedding space. umap-learn.readthedocs.io/en/latest/inve...

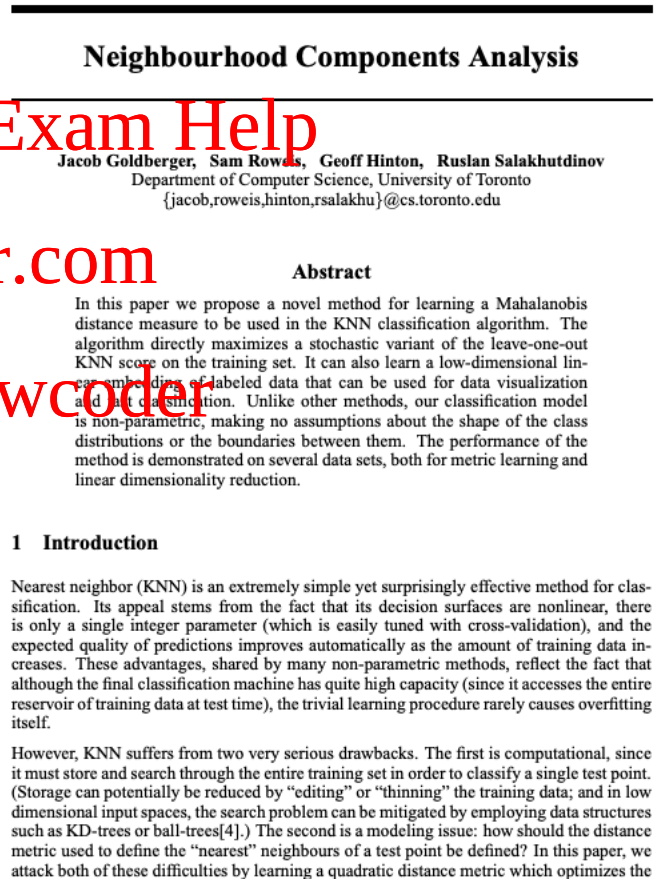


1 4 19

BEST PRACTICE…

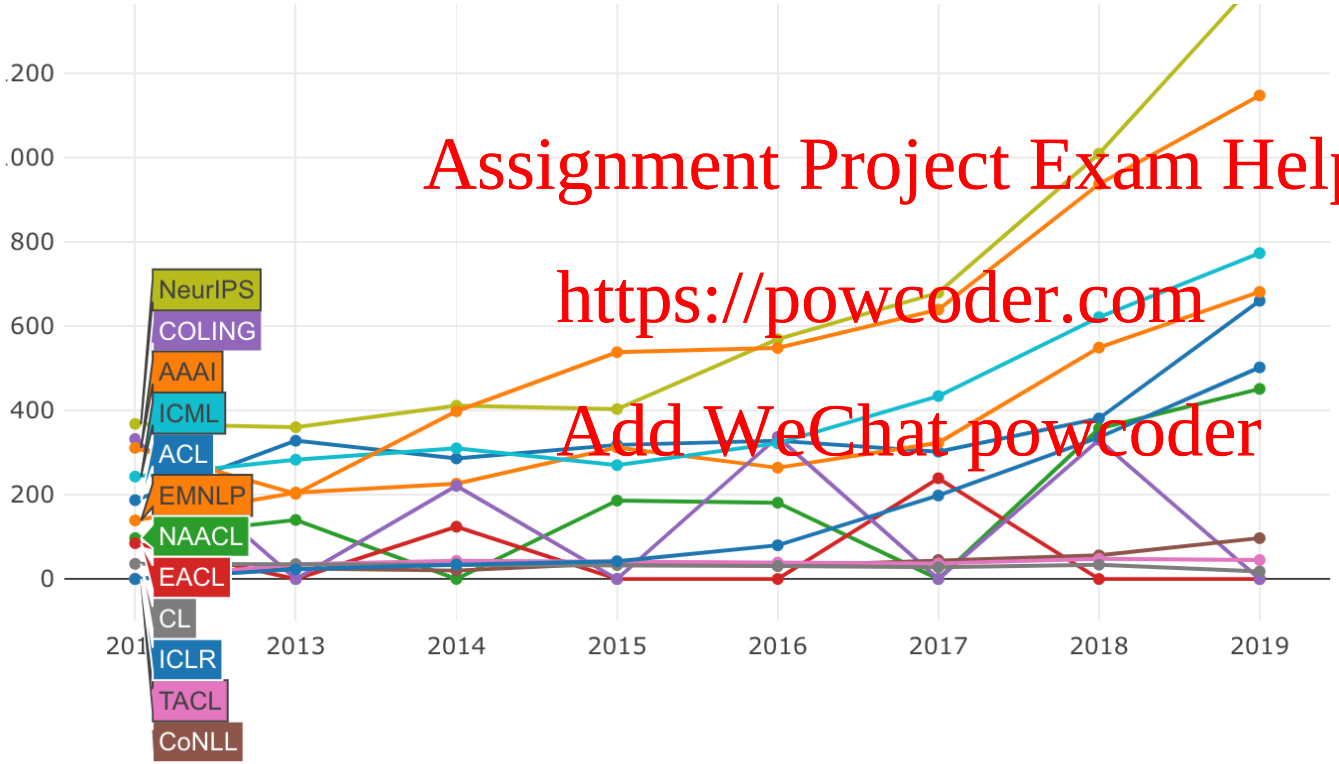
UCL SCHOOL OF MANAGEMENT





BEST PRACTICE…

UCL SCHOOL OF MANAGEMENT

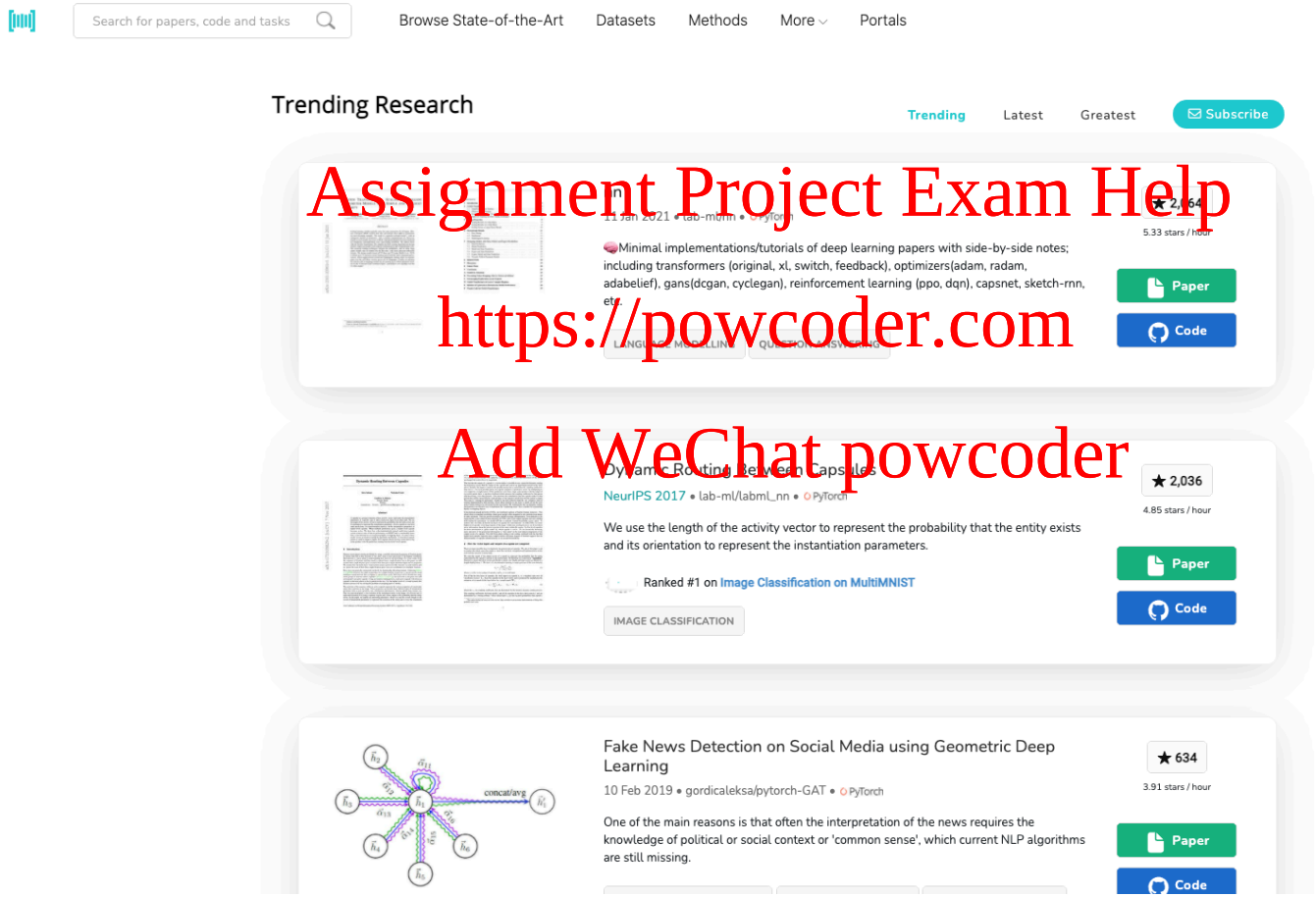


Percentage of papers mentioning GitHub (indicating that the code is made available):
ACL 70%,
EMNLP 69%,
NAACL 68%
ICLR 56%,
NeurIPS 46%,
ICML 45%,
AAAI 31%.
It seems the NLP papers are releasing their code much more freely.

Source: https://www.marekrei.com/blog/ml-and-nlp-publications-in-2019/

PAPERS WITH CODE

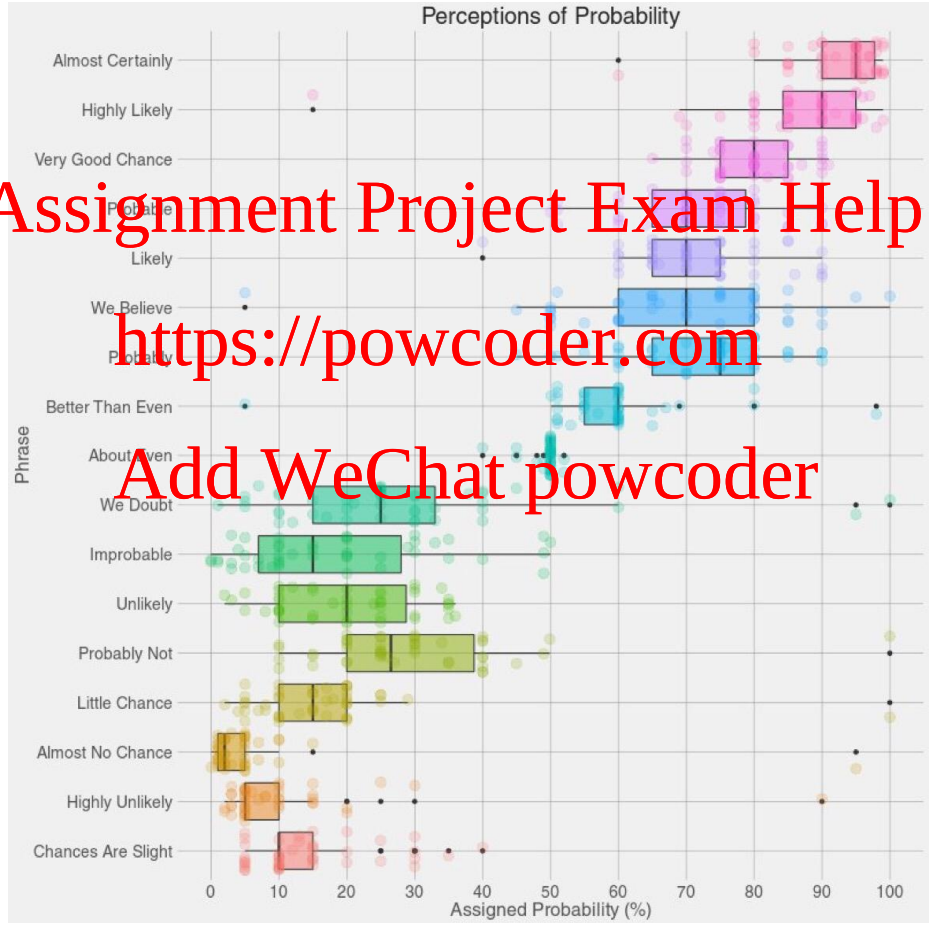
UCL SCHOOL OF MANAGEMENT



https://paperswithcode.com/

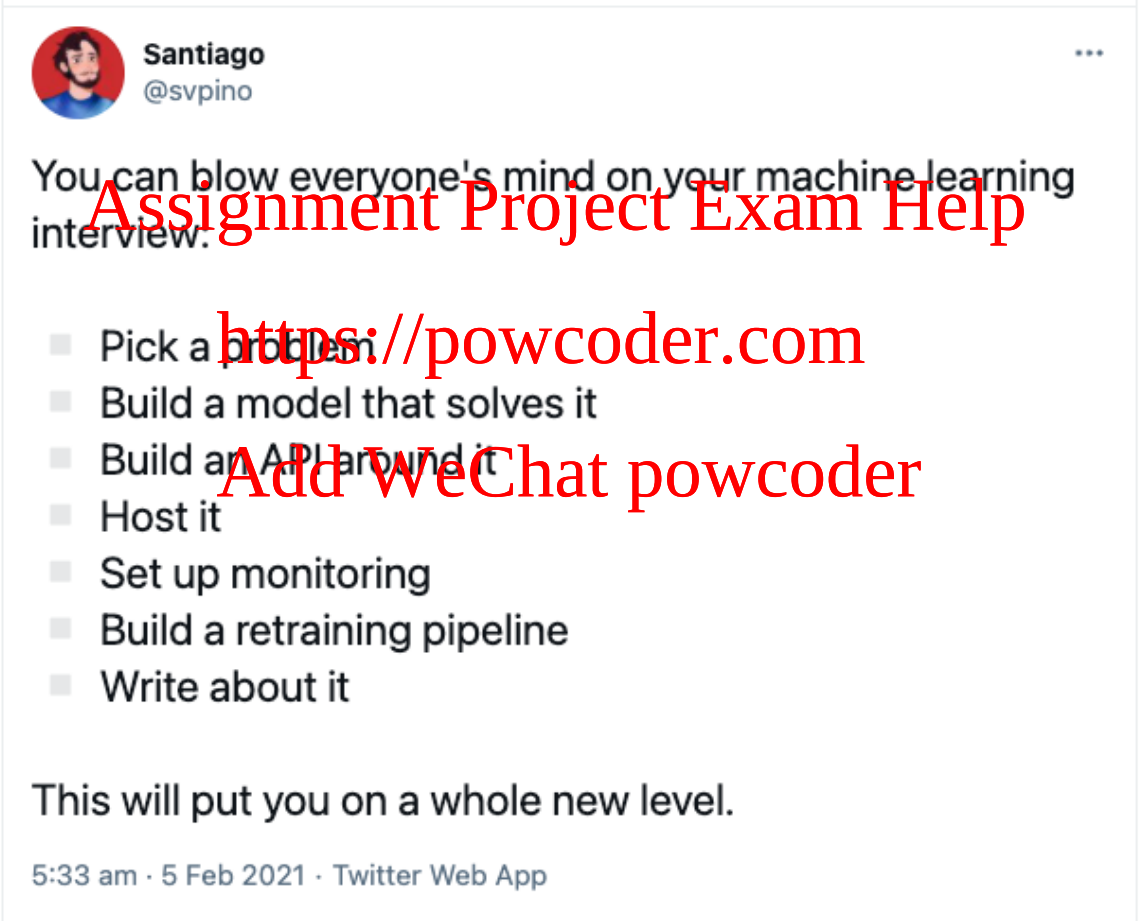
PERCEPTIONS OF PROBABILITY

UCL SCHOOL OF MANAGEMENT



DEPLOYMENT

UCL
SCHOOL OF
MANAGEMENT



Santiago
@svpino

You can blow everyone's mind on your machine learning interview:

- Pick a problem
- Build a model that solves it
- Build an API around it
- Host it
- Set up monitoring
- Build a retraining pipeline
- Write about it

This will put you on a whole new level.

5:33 am · 5 Feb 2021 · Twitter Web App

@SOCIAL

UCL
SCHOOL OF
MANAGEMENT

@chipro
@random_forests
@zacharylipton
@yudapearl
@svpino
@jackclarkSF

TEACHING TEAM

UCL SCHOOL OF MANAGEMENT

Dr Alastair Moore
Senior Teaching Fellow
a.p.moore@ucl.ac.uk
@latticecut

Kamil Tylinski
Teaching Assistant
kamil.tylinski.16@ucl.ac.uk

Jiangbo Shangguan
Teaching Assistant
j.shangguan.17@ucl.ac.uk

Individual Coursework workshop

to **Thursday 11th Feb 2021** at **12:00 am**

LECTURE 3 TERM 2:

MSIN0097

UCL SCHOOL OF MANAGEMENT

PREDICTIVE ANALYTICS

A P MOORE