Publication and Researchers:

The first paper I have chosen to study is 'A Cluster Analysis of NBA Players'. There is only one researcher : Dwight Lutz, who is a data scientist for the National Basketball Association (NBA).

Te paper is split into two parts. The first part aims to cluster NBA players based on their on-court attributes. The second part takes those clusters and tries to find what type of players and combination of players are most often found on winning and losing teams. I have personal doubts over the viability of associating the clusters with winning teams (mainly because of the difficulty of the task) and I would want to see it verified against future data before considering it. I will focus on part 1 which has clustering techniques that are new and interesting to me.

Basketball has five traditional positions which can be subclassed into two futher groups the 'bigs' and the 'smalls'. The bigs are centre, power forward, and small forward and these roles are usually assigned in that order with the tallest and strongest at centre. The smalls has two groups the point guard and the shooting guard with the former being the one who passes the ball better.

This publication aims to derive new positions using multivariate cluster analysis by doing away with size and weight and focusing on what people actually do on the basketball court. Or maybe it would be more accurate to say it looks at the how because the paper de-emphasises the effictiveness of the player's shooting and instead looks at the areas from which they took shots.

Dataset:

The clusters were generated from the 2010-11 season on all players who played in at least 30 games and that averaged at least 10 minutes per game. This gives an 329 players and players from the preceding two seasons were placed into these clusters, assuming they met the playing-time thresholds mentioned above, using Fisher's Linear Discriminant to gives us a final total of 924 players used.

Here is a table of the attributes used

Variable	Abbreviation	Formula (if needed)
Games Played (minimum of 30)	GP	-
Minutes played per game (minimum of 10)	Min	-
Percent of made field goals that are assisted	% Ast	$\frac{\text{made field goals that are assisted}}{\text{total made field goals}}$
Assist Ratio	AR	$\frac{\text{Assists} \times 100}{\text{FGA}+(\text{FTA} * .44)+\text{Turnovers}}$
Turnover Ratio	TOR	$\frac{\text{Turnovers} \times 100}{\text{FGA}+(\text{FTA} * .44)+\text{Turnovers}}$
Offensive Rebound Rate	ORR	$\frac{100 \times (\text{Player ORebs} * (\text{Team Min}/5))}{(\text{Player Min} * (\text{Team ORebs}+\text{Opp DRebs}))}$
Defensive Rebound Rate	DRR	$\frac{100 \times (\text{Player DRebs} * (\text{Team Min}/5))}{(\text{Player Min} * (\text{Team DRebs}+\text{Opp ORebs}))}$
Attempted field goals at the rim per 40 minutes	Rim	-
Attempted field goals from 3-9 feet per 40 minutes	3-9	-
Attempted field goals from 10-15 feet per 40 minutes	10-15	-
Attempted field goals from 16-23 feet per 40 minutes	16-23	-
3-point field goals attempted per 40 minutes	3s	-
Steals per 40 minutes	Stls	-
Blocks per 40 minutes	Blks	-

While most attributes are success based the attempted field goals are style based. My conjecture on this is that basketball players are probably thought where there coach wants them to shoot from and if a shot is not available then they should pass. Therefore the location they have attempted from is possibly more telling and almost certainly has less noise than success percentage.

Another thing that can be noted is that all of the variables have strong variability from cluster to cluster and this validates the use of each variable. From the lecture notes good clusters should have low inter-class similarity (and high intra class similarity). If a variable was relatively constant across clusters this would inhibit dissimilarity between the average object in different clusters and the authors should have consider removing that variable.

Findings:

Here are the found clusters and the mean z-score for each statistic:

Cluster (#)	GP	Min	% Ast	AR	TOR	ORR	DRR	Rim	3-9	10-15	16-23	3s	Stls	Blks
Combo Guards (1)	0.38	-0.02	0.61	0.20	-0.47	-1.01	-1.01	-1.32	-0.84	-0.34	0.17	1.05	-0.21	-0.78
Defensive Bigs (2)	-0.41	-0.62	0.23	-0.45	0.88	1.57	0.99	0.82	-0.06	-0.94	-1.34	-1.09	-0.36	1.11
Versatile Swingmen (3)	-0.49	-0.41	0.26	0.09	0.23	-0.08	-0.16	0.15	-0.30	-0.66	-0.66	0.00	0.67	-0.02
Floor Spacers (4)	-0.48	-0.39	0.99	-0.52	-0.82	-0.48	-0.07	-0.89	-0.81	-0.73	-0.25	1.06	-0.37	-0.31
Elite Bigs (5)	0.83	1.04	0.02	-0.47	-0.31	0.86	1.23	0.85	1.93	1.03	0.40	-0.95	-0.20	0.80
Big Bodies (6)	-0.55	-1.06	0.67	-0.60	0.33	0.85	0.52	-0.21	-0.30	-0.03	0.72	-1.03	-0.55	0.22
Active Bigs (7)	-0.08	-0.29	0.29	-0.90	-0.19	1.25	0.98	0.82	0.67	0.65	0.08	-1.08	-0.77	1.29
Ball Handlers (8)	0.04	0.34	-1.53	1.69	0.94	-0.85	-0.95	-0.12	-0.14	0.09	0.15	0.37	0.75	-0.81
Perimeter Scorers (9)	0.04	0.59	-0.56	-0.10	-0.42	-0.60	-0.23	0.16	0.36	1.09	0.90	0.21	-0.03	-0.34
Durable Shooters (10)	1.09	0.38	0.32	-0.21	-0.82	-0.58	-0.53	-0.16	-0.39	-0.28	-0.02	0.88	0.64	-0.52

A good start was that some of the clusters related very closely to traditional basketball positions. For example nearly-all point guards were found in the ball handler section but also some bigs were too. One thing I think may be of interest here is that is that Ball Handlers are -1.53 on %Ast and +1.69 in AR, two of the futhest scores from 0, which would seem to make them less likely to be missclassified.

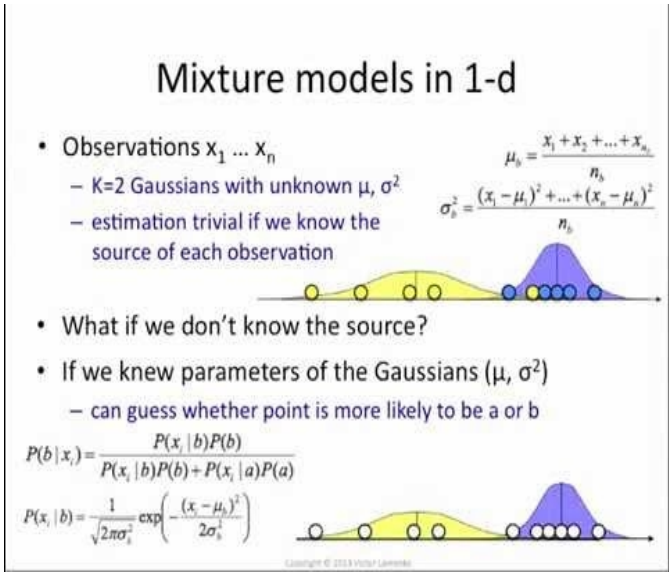
An example of where the method would have advantages over the traditional method of labelling player roles is given by Laker's point guard Steve Blake. Blake had struggled for the Lakers and it would be difficult to see why if one looked at him as playing point guard before joining the Lakers and point guard with the Lakers. However the Lakers played with a unique offence called the Triangle offence. None of their point guards were clustered in the Ball Handlers' cluster. But instead were placed in the Combo Guard cluster. Therefore this method gives a way to identify players who might have played 'out of position' in terms of usage but still played the same traditional basketball rôle.

More succinctly traditional basketball roles are defined relative to teammamtes. Clustering defines roles relative to the league. For a player whose best attribute is passing it is not good enough just to be the player the most on their team. If the strategy for the team spreads passing more evenly amongst players then they are being misused.

Techniques:

Expectation Maximisation algorithm for Gaussian mixture models [1]. There are two ways to do clustering. In our course we have studied hard clustering, such as K-Means and DBScan, where an object either belongs to a cluster or not. In soft clustering each cluster is a model for generating a probability that an object belongs to a cluster.

The algorithm works similar to k-means in that we have to choose our clusters first. Here we do that by assuming their are n normal distributions as sources and . For each point you then assign it to it's most likely cluster (You can see from the image on the right that Bayes' rue is used but since you are assignning to the most likely cluster and all objects will have the same denominator I am not sure why this is neccessary. Unless you can skip it while running the algorithm but need it afterwards to get true probabilities i.e. between 0 and 1).



Now each cluster has a set of objects and the means and variance of these are calculated and a new normal distribution is produced for each cluster.

This process is repeated until we reach convergence.

From Cearbhall's presentation I am now thinking that this might be very useful for taking out people whose defaults were owing to some outside circumstance like a collapse of employment in their industry rather than them personally being the risk. For example you might find that if you look at people in construction with <1% chance of being in a 'yes' default cluster that 90% of them happened between 2008 and 2010 and it might be pertinent to remove that group or those years or some combination of years and players from the dataset.

Mclust function: According to the paper this uses the Bayesian Information Criterion (BIC) to determine the parameters of the model and how many clusters to use. It seems to be the name of an r package that implements EM and BIC introduces a penalty into the model for each parameter used so that each parameter must bring more than a threshold value – as a way to avoid overfitting.

Fisher's Linear Discriminant (FLD) projects high-dimensional data onto a line [2]. It tries to maximise the distance between the means of the classes whilst reducing the variance within classes. In 2-d it uses the Fisher criterion to achieve this [3]:

$$J(w) = \frac{|m_1 - m_2|^2}{s_1^2 + s_2^2}$$

In this paer FLD was used to place the players from the 2008-2009 and
2009-2010 seasons into one of the 10 clusters. Essentially discriminant analysis is another form of distance metric but it, similar to how EM worked, it forms the clusters and then calculates the metric before re-calculating the clusters.

Relation To Work:

In assignment 1 I tried to cluster baseball players and noted that by mixing in a player's future WAR with their other statistics it inhiibited my ability to find good clusters. I also pointed out that, from the lecture notes, there are two use of clustering:

1. To find insight into the distribution of a dataset

2. As an input step into other data mining techniques

The two parts of this paper mirror those steps. First clusters were found and then these clusters were used to find which players and combination of players were found most often on winning and losing teams.

Secondly the techniques used here were not covered in the course. I wasn't even aware of the idea of soft clustering and FLD provides another way to assign objects to clusters that I had not considered.

References:

[1] Expectation Maximisation algorithm for Gaussian mixture models https://www.youtube.com/watch?v=REypj2sy_5U

[2] FLD in higher dimensions http://sebastianraschka.com/Articles/2014_python_lda.html#summarizing-the-lda-approach-in-5-steps

[3] FLD in 2D https://compbio.soe.ucsc.edu/genex/genexTR2html/node12.html