Developing Multivariate Extreme Value Methods to Explore NFL Prospect Data

New England Symposium on Statistics in Sports

Brook T. Russell and Sydney Newman
CU School of Mathematical and Statistical Sciences

Introductions

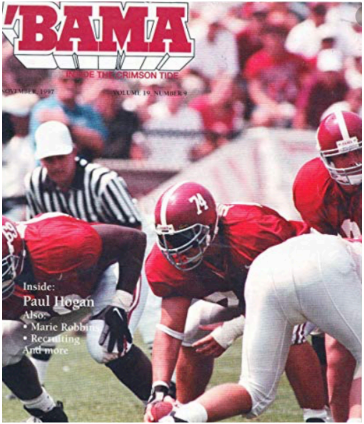
My Background:

- Associate Professor of Statistics at Clemson University's SMSS

Collaborators:

- Sydney Newman: CU SMSS Graduate student
- Paul Hogan: CU Senior Assistant Football Strength and Conditioning Coach

Meet Paul



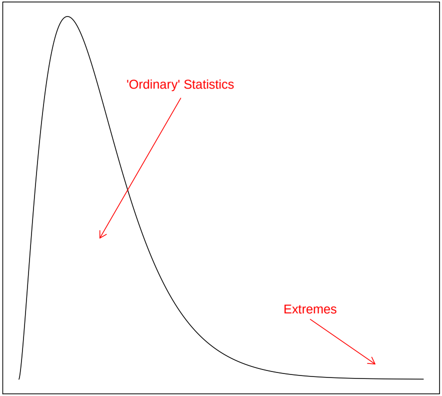
Today's Talk

Outline:

- Multivariate extremes in the context of sports
  - Univariate vs. multivariate extremes
  - Asymptotic dependence
- Case Study 1
  - Characterizing MV extremes: NFL Combine analysis
- Case Study 2
  - Optimize tail dependence: Sydney's current work

What is Extremes?

- In extremes we are only interested in the far upper tail
- We're (almost) always data poor!



Univariate Extremes

Aims and Procedures:

- Model the far upper tail
- Keep best observation in each large block of observations: *Block Maxima*
- Keep all observations larger than some large threshold: *Peaks Over Threshold*

Numerous Examples of Univariate Extremes in Sports:

- Track and Field (Einmahl and Magnus, 2008; Einmahl and Smeets, 2011; Henriques-Rodrigues et al., 2011; Grycmann et al., 2015; Noubary, 2010)
- Swimming (Gomes and Henriques-Rodrigues, 2019; Spearing et al., 2021)
- Horseracing (Pieramati et al., 2011)

Multivariate Extremes

Aims and Procedures:

- Characterize dependence in far upper joint tail
- One approach: keep all observations larger than some large threshold based on chosen norm

Fewer Examples of Multivariate Extremes in Sports:

- Swimming (Adam and Tawn, 2012)
- NFL Combine (Russell and Hogan, 2018)

Dependence vs Asymptotic Dependence

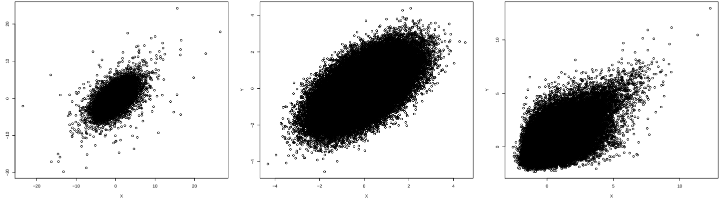
- Standard correlation metrics can be effective in quantifying association in the center of a bivariate distribution
  - Pearson's
  - Spearman's
  - Kendall's
- Two random variables with common marginals are asymptotically (tail) dependent if

$$P(X_1 \text{ is extreme } | X_2 \text{ is extreme}) > 0$$

- Termed asymptotically independent otherwise
- Standard correlation metrics typically perform poorly in terms of modeling asymptotic dependence
- Bivariate Gaussian random vector with cor < 1: asymptotically independent (Sibuya, 1960)

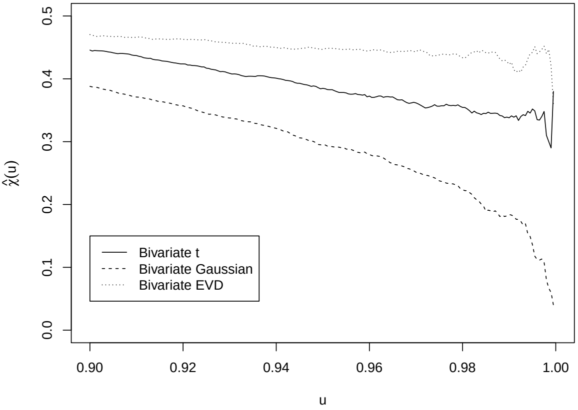
Dependence vs Asymptotic Dependence

- Simulate three data sets (10^5 realizations each)
  1. Bivariate t
  2. Bivariate Gaussian
  3. Bivariate extreme value distribution (EVD)
- Sample correlations: 0.5899, 0.5967, and 0.5766 (L to R)



Dependence vs Asymptotic Dependence

Estimate $P(X_1 \text{ is extreme } | X_2 \text{ is extreme})$ via empirical proportion



Case Study 1

- Explore multivariate dependence in NFL Combine event performances
- Consider dependence in bulk of data and in multivariate tail

JQAS Paper:

Russell, B. Hogan, P. (2018). Analyzing dependence matrices to investigate relationships between national football league combine event performances. Journal of Quantitative Analysis in Sports, 14(4), 201-212. https://doi.org/10.1515/jqas-2017-0086

NFL Combine Physical Tests

1. the *bench press* (BP), where players bench press a barbell weighing 225 lbs. (102.1 kg) for as many repetitions as possible;
2. the *vertical jump* (VJ), where players jump as high as possible from a standing position;
3. the *broad jump* (BJ), where players jump as far a possible horizontally from a standing position;
4. the *sprint* (40YD), where the players is timed while running 40 yds. (36.6 m) as quickly as possible;
5. the *shuttle drill* (SD), where the players run a total of 20 yds. (18.3 m) while changing direction two times; and
6. the *three cone drill* (3CD), where players run around three cones placed 5 yds. (4.6 m) apart and placed in the shape of a "L".

Objectives of this Work

- Great interest in player performances in these events – they can greatly influence how highly a prospect is drafted (or if a prospect is even drafted at all)
- There have been recent calls to rethink the makeup of the physical tests included at the Combine (Robbins and Goodale, 2012)
- Goals of this work:
    1. Characterize the dependence between prospect performances (relative to player position), for both typical and elite level Combine performers.
    2. Use this information to make recommendations regarding how Combine events might be modified
- Predicting NFL performance is not a goal in this work!!

Comparison to Similar Work

- Robbins (2012) has a similar aim, and uses NFL Combine data from the years 2005-2009 to perform a correlation analysis for Combine events
- We supplement Robbins (2012) in three key ways:
  - We take player position into account
  - We consider dependence in the tail as well as dependence in the center of the distribution
  - We consider decompositions of pairwise dependence matrices

Pairwise Dependence Matrices (PDM)

- Consider d dimensional random vector
- Characterizing MV dependence is challenging (even in relatively low dimensions)
- One solution – PDM
- PDM: Symmetric square positive semidefinite matrix whose off diagonal elements give a measure of pairwise association between the two variables corresponding to that row/column
  - Covariance matrices
  - Correlation matrices
  - Tail pairwise dependence matrices (TPDM)

Covariance/Correlation Matrices

Covariance Matrices

- Diagonal elements – variances
- Off-diagonal – covariances
- MV Gaussian framework: completely characterizes dependence structure
- Non-Gaussian case: may still provide useful information regarding pairwise dependence in bulk of data
- Simple to calculate empirical analogue

Correlation Matrices:

- Diagonal elements – all 1
- Off-diagonal – pairwise correlations
- Ignores variable scale
- MV Gaussian case: 0 correlation $\iff$ independence

Eigendecomposition

- Eigendecomposition of covariance/correlation matrices often useful
- Eigendecomposition of standardized variables (centered and scaled) equivalent to eigendecomposition of correlation matrix
- Linked to principal components analysis (PCA)

Tail Pairwise Dependence Matrices

- Assume d dimensional random vector is regularly varying
- TPDM – entry in the ith row and jth column is an extremal dependence measure (EDM) (Cooley and Thibaud, 2019)
- $0 < \text{EDM} < 1$: describes bivariate asymptotic dependence (Larsson and Resnick, 2012)
- Estimate TPDM via sample analogue (Cooley and Thibaud, 2019)
- Eigendecomposition of this matrix may also be useful

Description of Combine Data

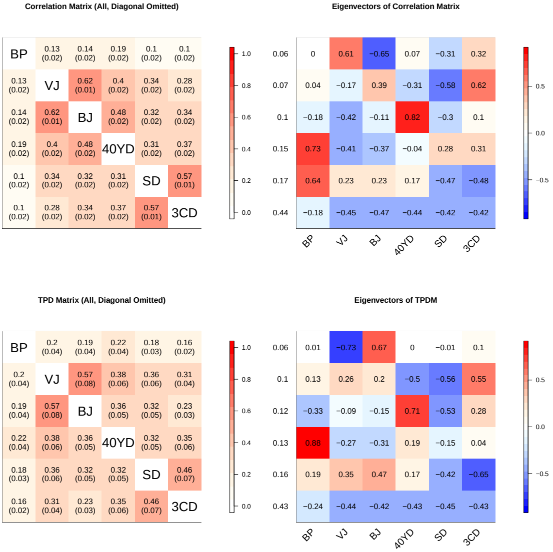
- Data for years 1999-2016[1]
- 2,647 complete multivariate observations (center and scaled)
- For BP, VJ, and BJ higher value = better performance
- Opposite for running events (40YD, SD, and 3CD)!
- Solution: use avg speed (yd/s) for running events
- Larger values correspond with better performances in all events

[1] `http://nflcombineresults.com/nflcombinedata.php`

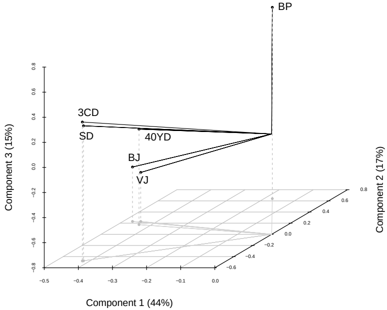
Analysis of Combine Data

- Estimate PDMs: correlation matrix and TPDM
- Account for uncertainty using nonparametric bootstrap
- Eigenvectors of both matrices provide for interesting interpretation
- PDM decompositions – 'triplots'

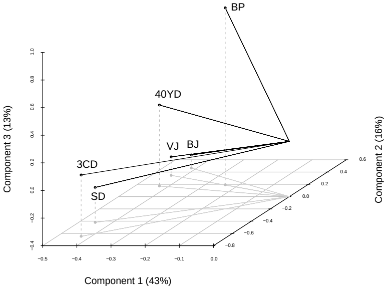
Sample PDMs


Correlation Matrix (All, Diagonal Omitted)
BP 0.13 (0.02) 0.14 (0.02) 0.19 (0.02) 0.1 (0.02) 0.1 (0.02)
0.13 (0.02) VJ 0.62 (0.01) 0.4 (0.02) 0.34 (0.02) 0.28 (0.02)
0.14 (0.02) 0.62 (0.01) BJ 0.48 (0.02) 0.32 (0.02) 0.34 (0.02)
0.19 (0.02) 0.4 (0.02) 0.48 (0.02) 40YD 0.31 (0.02) 0.37 (0.02)
0.1 (0.02) 0.34 (0.02) 0.32 (0.02) 0.31 (0.02) SD 0.57 (0.01)
0.1 (0.02) 0.28 (0.02) 0.34 (0.02) 0.37 (0.02) 0.57 (0.01) 3CD
1.0 0.8 0.6 0.4 0.2 0.0
Eigenvectors of Correlation Matrix
0.06 0 0.61 −0.65 0.07 −0.31 0.32
0.07 0.04 −0.17 0.39 −0.31 −0.58 0.62
0.1 −0.18 −0.42 −0.11 0.82 −0.3 0.1
0.15 0.73 −0.41 −0.37 −0.04 0.28 0.31
0.17 0.64 0.23 0.23 0.17 −0.47 −0.48
0.44 −0.18 −0.45 −0.47 −0.44 −0.42 −0.42
BP VJ BJ 40YD SD 3CD
0.5 0.0 −0.5
TPD Matrix (All, Diagonal Omitted)
BP 0.2 (0.04) 0.19 (0.04) 0.22 (0.04) 0.18 (0.03) 0.16 (0.02)
0.2 (0.04) VJ 0.57 (0.08) 0.38 (0.06) 0.36 (0.06) 0.31 (0.04)
0.19 (0.04) 0.57 (0.08) BJ 0.36 (0.05) 0.32 (0.05) 0.23 (0.03)
0.22 (0.04) 0.38 (0.06) 0.36 (0.05) 40YD 0.32 (0.05) 0.35 (0.06)
0.18 (0.03) 0.36 (0.06) 0.32 (0.05) 0.32 (0.05) SD 0.46 (0.07)
0.16 (0.02) 0.31 (0.04) 0.23 (0.03) 0.35 (0.06) 0.46 (0.07) 3CD
1.0 0.8 0.6 0.4 0.2 0.0
Eigenvectors of TPDM
0.06 0.01 −0.73 0.67 0 −0.01 0.1
0.1 0.13 0.26 0.2 −0.5 −0.56 0.55
0.12 −0.33 −0.09 −0.15 0.71 −0.53 0.28
0.13 0.88 −0.27 −0.31 0.19 −0.15 0.04
0.16 0.19 0.35 0.47 0.17 −0.42 −0.65
0.43 −0.24 −0.44 −0.42 −0.43 −0.45 −0.43
BP VJ BJ 40YD SD 3CD
0.5 0.0 −0.5

Tri Plots



BP
3CD
SD
40YD
BJ
VJ
Component 3 (15%)
Component 2 (17%)
Component 1 (44%)



BP
40YD
VJ
BJ
3CD
SD
Component 3 (13%)
Component 2 (16%)
Component 1 (43%)

Conclusions/Recommendations

- Results suggest four or five underlying physical attributes being assessed
- A revised set of four or five events may be sufficient
- We propose the following modifications:
  - An alternative upper body strength test (perhaps throwing a heavy medicine ball)
  - Keep BJ and eliminate VJ
  - Keep 40YD but retain measure of top end speed (perhaps last 10 yards) and intermediate times
  - Keep SD and eliminate 3CD
  - As an additional test, we propose an event that measures posterior chain power (perhaps heavy overhead throw for distance)

Case Study 2

Case Study 1

- Explore multivariate dependence in NFL Combine event performances
- Consider dependence in bulk of data and in multivariate tail

Case Study 2 (Work in progress):

- Identify NFL prospect data (college, Combine, etc.) that have high level of asymptotic dependence with NFL outcomes
- Our plan:
  - Exploratory data analysis
  - Extend extreme value approach in Russell et al. (2016) – "optimizing tail dependence"

Optimizing Tail Dependence

- Classical Linear Regression: identify linear combination of covariates with highest possible (Pearson's) correlation with response
- Our method to optimize tail dependence: identify linear combination of covariates with highest possible asymptotic dependence with response
  - Quantify tail dependence via EDM (Larsson and Resnick, 2012)
  - Requires marginal transformations (typical in MV extremes)
  - Optimization in transformed space
  - Details: Russell et al. (2016)

Comparison with Other Works

- Others have similar goals:
  - Suggestions for modifying entire player evaluation process (Kuzmits and Adams, 2008)
  - Predict TE success (Mulholland and Jensen, 2014)
  - Predict QB success (Wolfson et al., 2011)
- Our aim is a bit different – we're looking at extremes!

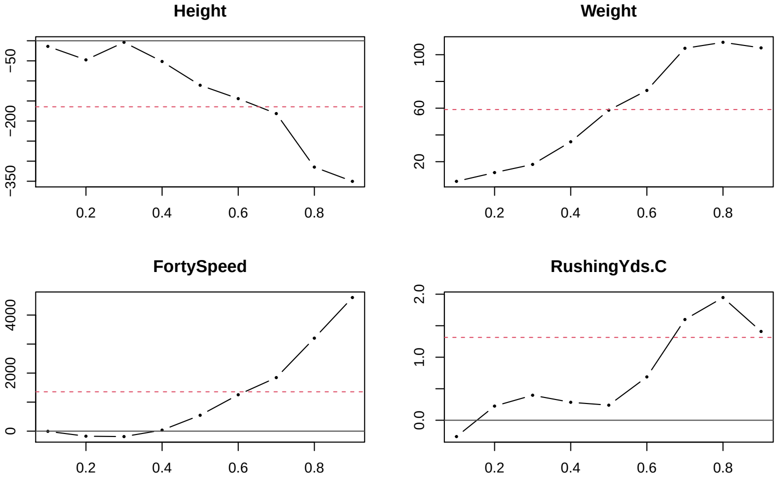
Data

- Positions: (for now)
  - RB
  - WR
- Response Variables:
  - Single Season
  - Career
- Predictor Variables:
  - College Data
  - NFL Combine Data

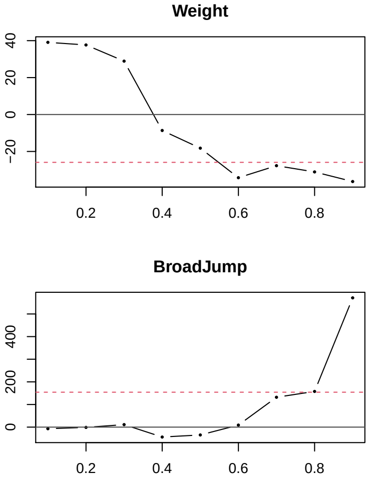
Exploratory Data Analysis: Quantile Regression

- Currently in exploratory analysis stage
- Classical linear regression: minimize SSE
- Quantile Regression: minimize sum of "check loss"
- Check loss function models τth conditional quantile of response ($0 < \tau < 1$)
- Quantile regression doesn't allow modeling as far into tail – ok for now

Preliminary Results: RB



Preliminary Results: WR



Discussion/Future Work

- Bulk of data vs. tail
- Quantile regression vs. our extreme value model
- Extending extreme value model used in Russell et al. (2016)
  - Improving optimization
  - Incorporating prior information

Overall Conclusion

- Conclusions based on extremes analysis may yield different insights (compared to traditional statistical analyses)
- Multivariate extremes analyses typically have different aims (compared to univariate)
- Please feel free to reach out: `brookr@clemson.edu`
- Questions?

References I

Adam, M. B. and Tawn, J. A. (2012). Bivariate extreme analysis of Olympic swimming data. *Journal of Statistical Theory and Practice*, 6(3):510–523.

Cooley, D. and Thibaud, E. (2019). Decompositions of dependence for high-dimensional extremes. *Biometrika*, 106(3):587–604.

Einmahl, J. H. and Magnus, J. R. (2008). Records in athletics through extreme-value theory. *Journal of the American Statistical Association*, 103(484):1382–1391.

Einmahl, J. H. J. and Smeets, S. G. W. R. (2011). Ultimate 100-m world records through extreme-value theory. *Statistica Neerlandica*, 65(1):32–42.

Gomes, D. T. and Henriques-Rodrigues, L. (2019). Swimming performance index based on extreme value theory. *International Journal of Sports Science & Coaching*, 14(1):51–62.

References II

Grycmann, P., Maszczyk, A., Socha, T., Gołaś, A., Wilk, M., Zajac, T., and Przednowek, K. (2015). Modelling analysis and prediction of women javelin throw results in the years 1946–2013. *Biology of Sport*, 32(4):345.

Henriques-Rodrigues, L., Gomes, M. I., and Pestana, D. (2011). Statistics of extremes in athletics. *Revstat*, 9(2):127–153.

Kuzmits, F. E. and Adams, A. J. (2008). The NFL combine: does it predict performance in the National Football League? *The Journal of strength & conditioning research*, 22(6):1721–1727.

Larsson, M. and Resnick, S. I. (2012). Extremal dependence measure and extremogram: The regularly varying case. *Extremes*, 15(2):231–256.

Mulholland, J. and Jensen, S. T. (2014). Predicting the draft and career success of tight ends in the National Football League. *Journal of Quantitative Analysis in Sports*, 10(4):381–396.

References III

Noubary, R. D. (2010). Tail modeling, track and field records, and Bolt's effect. *Journal of Quantitative Analysis in Sports*, 6(3).

Pieramati, C., Giontella, A., Sarti, F. M., and Silvestrelli, M. (2011). Assessment of limits for racing speed in the italian trotter population. *Italian Journal of Animal Science*, 10(4):e53.

Robbins, D. W. (2012). Relationships between National Football League combine performance measures. *The Journal of Strength & Conditioning Research*, 26(1):226–231.

Robbins, D. W. and Goodale, T. (2012). Evaluation of the physical test battery implemented at the National Football League Combine. *Strength & Conditioning Journal*, 34(5):1–10.

Russell, B. T., Cooley, D. S., Porter, W. C., Reich, B. J., and Heald, C. L. (2016). Data mining to investigate the meteorological drivers for extreme ground level ozone events. *Ann. Appl. Stat.*, 10(3):1673–1698.

References IV

Russell, B. T. and Hogan, P. (2018). Analyzing dependence matrices to investigate relationships between National Football League Combine event performances. *Journal of Quantitative Analysis in Sports*, 14(4):201–212.

Sibuya, M. (1960). Bivariate extreme statistics, I. *Annals of the Institute of Statistical Mathematics*, 11(2):195–210.

Spearing, H., Tawn, J., Irons, D., Paulden, T., and Bennett, G. (2021). Ranking, and other properties, of elite swimmers using extreme value theory. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(1):368–395.

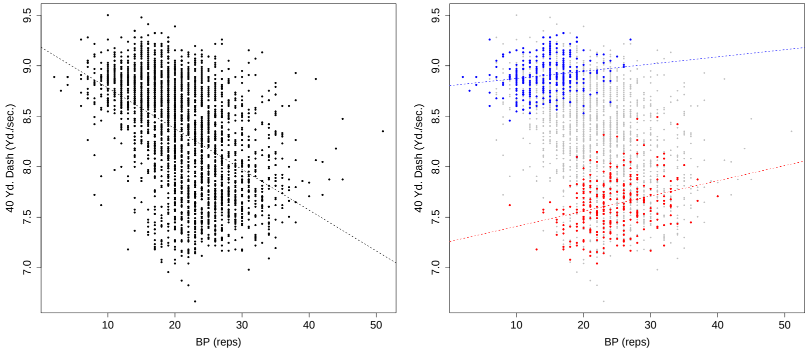
Wolfson, J., Addona, V., and Schmicker, R. H. (2011). The quarterback prediction problem: Forecasting the performance of college quarterbacks selected in the NFL draft. *Journal of Quantitative Analysis in Sports*, 7(3).

Simpson's Paradox: BP vs 40YD

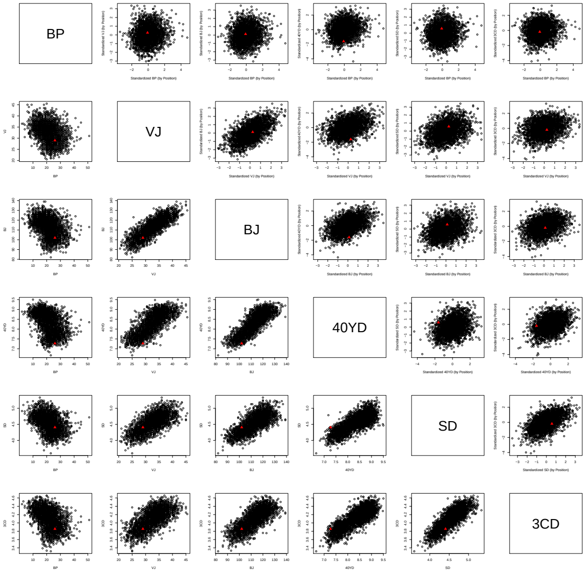
Pearson's correlation coefficients by position:

Pos.	$\hat{\rho}$	n	Lower	Upper
All	-0.48	2,647	-0.51	-0.45
C	0.22	117	0.04	0.39
CB	0.16	293	0.05	0.27
DE	0.20	275	0.08	0.31
DT	0.21	247	0.09	0.33
FS	0.17	137	0.00	0.33
ILB	0.07	134	-0.10	0.24
OG	0.22	207	0.09	0.35
OLB	0.24	241	0.12	0.36
OT	0.28	269	0.17	0.39
RB	0.19	230	0.06	0.31
SS	0.20	98	0.00	0.38
TE	0.23	194	0.09	0.36
WR	0.09	206	-0.05	0.22

Simpson's Paradox: BP vs 40YD



Transformed Data



Traditional Principal Components Analysis Example

- Use R mtcars data set
- Contains data on 32 car makes (from 1970s)
- We'll look at variables: mpg, disp, hp, drat, wt, and qsec
- Correlation matrix for these data:

	mpg	disp	hp	drat	wt	qsec
mpg	1.00	-0.85	-0.78	0.68	-0.87	0.42
disp	-0.85	1.00	0.79	-0.71	0.89	-0.43
hp	-0.78	0.79	1.00	-0.45	0.66	-0.71
drat	0.68	-0.71	-0.45	1.00	-0.71	0.09
wt	-0.87	0.89	0.66	-0.71	1.00	-0.17
qsec	0.42	-0.43	-0.71	0.09	-0.17	1.00

Traditional Principal Components Analysis Example

- Use R prcomp function to perform PCA analysis (eigendecomposition of correlation matrix)
- Eigenvectors give the six PCs

	PC1	PC2	PC3	PC4	PC5	PC6
mpg	-0.46	0.06	-0.19	0.78	-0.11	-0.35
disp	0.47	-0.06	0.10	0.60	0.29	0.57
hp	0.43	0.36	0.15	0.12	-0.81	-0.05
drat	-0.37	0.44	0.80	0.02	0.14	0.11
wt	0.44	-0.30	0.42	0.10	0.23	-0.69
qsec	-0.25	-0.76	0.34	0.04	-0.42	0.24

Traditional Principal Components Analysis Example

- Use R prcomp function to perform PCA analysis (eigendecomposition of correlation matrix)
- Eigenvalues give the variability for which each PC accounts

	PC1	PC2	PC3	PC4	PC5	PC6
St Dev	2.0463	1.0715	0.5774	0.3929	0.3533	0.2280
Prop of Var	0.6979	0.1913	0.0556	0.0257	0.0208	0.0087
Cum Prop	0.6979	0.8892	0.9448	0.9705	0.9913	1.0000

Traditional Principal Components Analysis Example

Biplot: