Course 4 · Week 1 — Validation, regularisation, multivariate

Cheatsheet — biostats_courses

R. Heller

Cross-validation

Scheme	Use
k-fold (k = 5 or 10)	baseline, fast
Repeated k-fold	reduces Monte Carlo noise
Leave-one-out	small n, high variance
Stratified k-fold	imbalanced class labels
Grouped / blocked	correlated units (patients, sites)
Nested	honest evaluation of tuned models

Nested CV is the only defensible way to estimate generalisation error when you are also tuning hyperparameters.

Regularisation

```
library(glmnet)
x <- model.matrix(y ~ . - 1, df); y <- df$y
fit <- cv.glmnet(x, y, alpha = 1)     # alpha = 1 = lasso; 0 = ridge
coef(fit, s = "lambda.1se")           # sparse solution
plot(fit)                              # CV curve
```

- `lambda.min`: empirical minimum of CV error.
- `lambda.1se`: more regularised, simpler, within 1 SE of min.
- Elastic net: `alpha` between 0 and 1 blends ridge and lasso.

Multivariate workhorses

Method	Purpose
PCA	variance → low-d summary; `prcomp(x, scale. = TRUE)`
FA	latent-variable modelling; `psych::fa`
CCA	correlation structure between two sets
LDA	supervised low-d projection; `MASS::lda`

```
pc <- prcomp(x, scale. = TRUE)
summary(pc)   # proportion of variance per PC
```

Scale before PCA when variables are on different units.

Clustering

```
km <- kmeans(scale(x), centers = 4, nstart = 25)
hc <- hclust(dist(scale(x)), method = "ward.D2")
library(mclust); bic <- mclustBIC(x)      # model-based, BIC-selected
```

- Choose k by silhouette, gap statistic, or BIC.
- Scale first. K-means assumes round clusters of similar size.

Non-linear embeddings

```r
library(uwot)
umap_xy <- umap(scale(x), n_neighbors = 15, min_dist = 0.1)

library(Rtsne)
ts <- Rtsne(scale(x), perplexity = 30)$Y
```

Embeddings are *visualisations*, not distances you should input into downstream stats. Global geometry is often distorted.

Decision rule for Week 1

- Lots of predictors, small n → lasso; report `lambda.1se`.
- Exploring structure → PCA + UMAP on scaled data.
- Clustering for biology → model-based + sensitivity to k.
- Tuning anything → nested CV.

Common pitfalls

- Selecting features on full data, then estimating error by k-fold.
- PCA before scaling when features are on different units.
- Interpreting UMAP distances as real distances.
- Using AUC on rank-ordered hold-out error as if it were CV error.

Further reading

- Hastie, Tibshirani, Friedman, *The Elements of Statistical Learning*.
- James et al., *ISLR2*.