Molecular pathology: predictive modelling

Paul McKeigue

Learning objectives

- Measuring performance of a predictor
  - Sensitivity, specificity, likelihood ratio,
  - area under ROC curve, test log-likelihood (for a probabilistic predictor)
- Fitting/learning predictive models
  - Epidemiological study designs and for learning and evaluating predictors
  - Classical statistical models for prediction:linear regression, logistic regression
  - Advantages and disadvantages of different approaches for biomarker selection and construction of predictors
  - Methods for constructing models from high-dimensional data
- Validation and cross-validation
  - Why use cross-validation?
  - How to use cross-validation to it to evaluate and compare predictive performance of models

Measuring performance of a predictor

- Sensitivity and specificity
  - Likelihood ratio = sensitivity / (1 – specificity)
  - Can be estimated from case-control study
- Prior odds x likelihood ratio = posterior odds
  - Why tests that are useful in clinical medicine perform badly when used for screening low-risk individuals
- ROC curve for a continuous score
  - plot sensitivity against (1 – specificity)

Likelihood of model given data

- Deviance = -2 x $\log_e$ likelihood

Epidemiological study designs

- Cohort (prospective):
  - can study multiple outcomes
- Case-control:
  - quicker, greater statistical power for equivalent outlay of resources
  - if risk factors can be measured retrospectively
- Nested case-control:
  - if you have a cohort with tissue samples stored at baseline, and biomarker measurements are expensive

Fitting/learning predictive models

- Classical statistical models for binary outcomes
  - Logistic regression
  - Linear discriminant function analysis:
    - assumes multivariate gaussian distributions of predictors within cases and controls
- Problems of modelling high-dimensional data
  - number of variables >> number of observations
  - Overfitting
    - With enough parameters that adapt to the data, model will fit the data but fail to predict new data

Methods for high-dimensional data

- Dimension reduction
  - where many variables are correlated
  - principal components analysis is simplest method
- Regression models with constraint on number of retained predictors and shrinkage of effect sizes
  - Forward stepwise regression with stopping rule to limit number of variables added to model
  - Penalized regression: LASSO, ridge regression
- Non-parametric (kernel-based) methods
  - Learn a function (kernel) that evaluates pairwise similarity between observations
  - Black-box predictor: not interpretable

LASSO regression

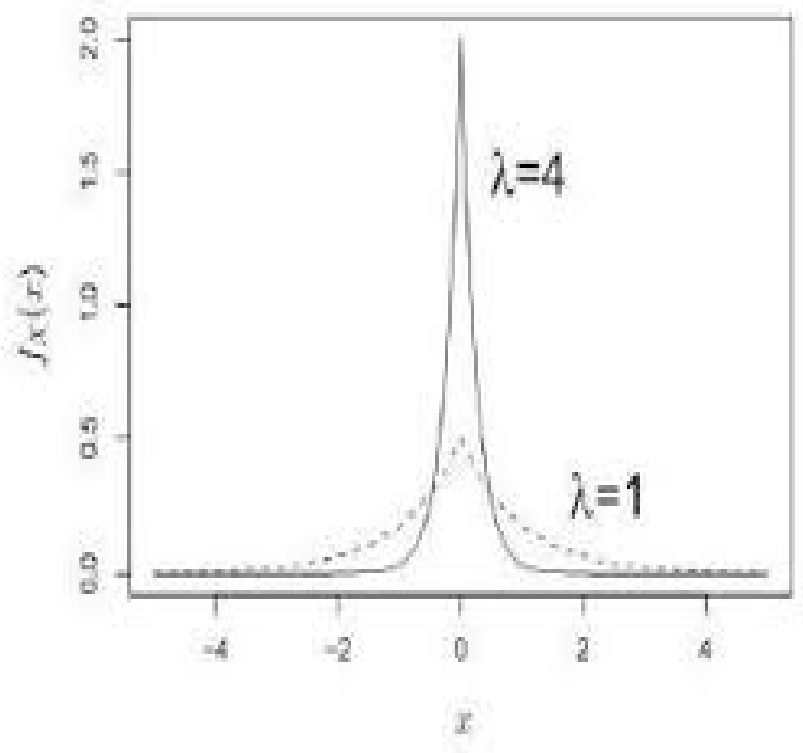
- Least Absolute (value) Shrinkage and Selection Operator
- Standard regression programs maximize the log-likelihood (probability of data given model) as a function of the regression coefficients $\boldsymbol{\beta}$
- LASSO regression maximizes the log-likelihood minus $\boldsymbol{\Sigma} \mid \boldsymbol{\beta}_i \mid$ (the sum of the absolute values of the regression coefficients) multiplied by a *sparsity parameter* λ
- With large values of λ, most regression coefficients will go to zero when the model is fitted to the data, and those that are retained will be shrunk towards zero
  - best value of λ is learned by cross-validation against withdrawn observations
  - value of λ determines how many variables are retained in the model (non-zero coefficients)
- Bayesian interpretation
  - LASSO regression is equivalent to specifying a prior belief that large effects are less probable than small effects, and that many effects are close to zero
    - double exponential priors on the regression coefficients)

Bayesian interpretation of LASSO regression

- LASSO regression is equivalent to specifying a prior belief that large effects are less probable than small effects, and many effects are close to zero
  - Specifically, the LASSO penalty is equivalent to double exponential priors on the regression coefficients)
  - λ is a scale parameter that controls the strength of the prior: large values force regression coefficients towards zero.

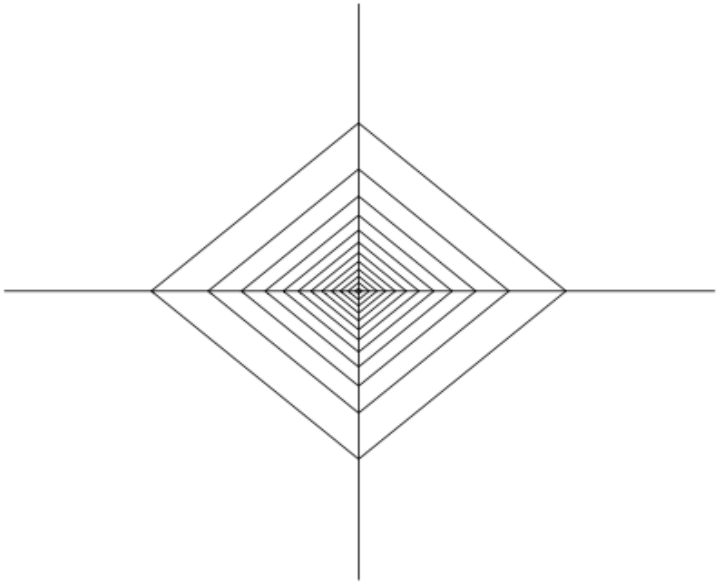
LASSO regression and the double exponential prior

- Parameter λ specifies the strength of the prior (penalty for large effect sizes)
  - learned from data by cross-validation



How double exponential prior encodes sparsity

- Contour plot of 2D probability density looks like pyramid
  - Contour plot of gaussian density would be concentric circles
- Density varies inversely with sum of absolute values of effect parameters



Why use cross-validation?

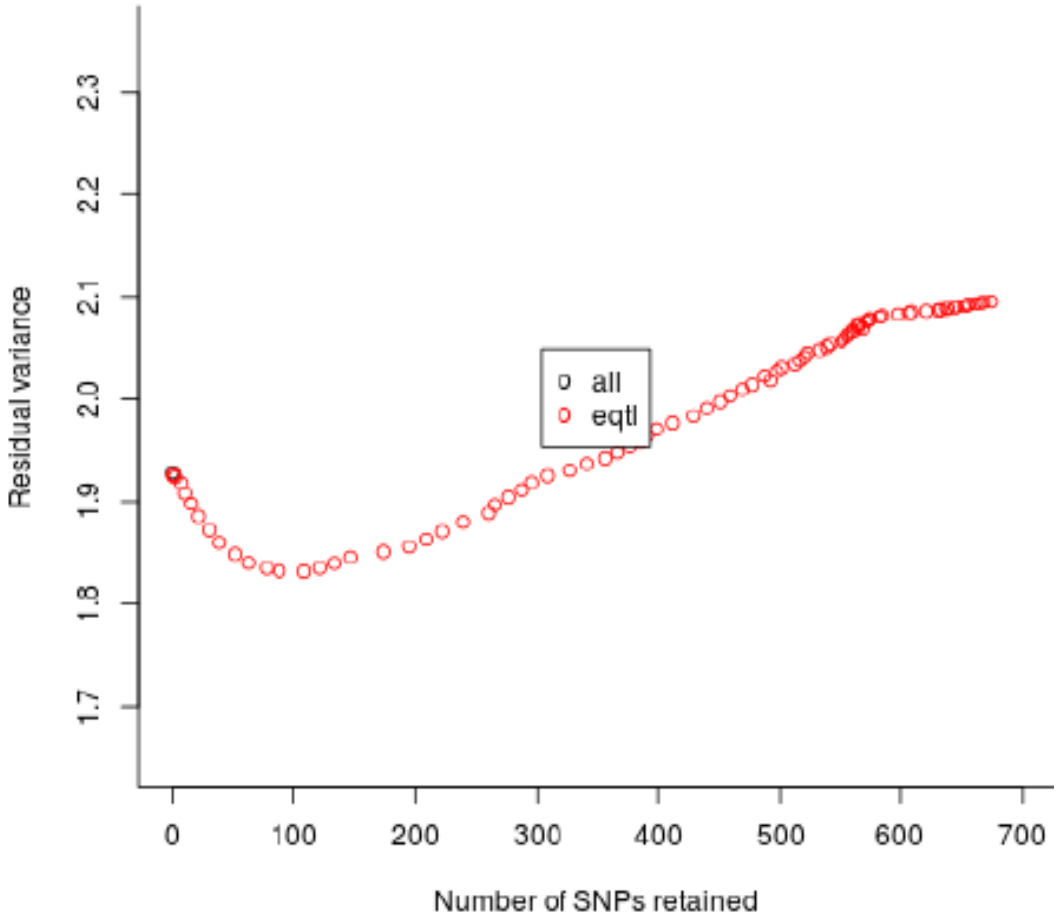
Predictive performance must be evaluated on data not used to learn the model

- Cross-validation allows you to use all the data for both training and testing
  - More efficient than a single test-training split
- Can tune learning algorithms for optimal prediction
  - Number of variables to retain
  - Evaluating performance of predictive model

Using cross-validation is used to evaluate predictive performance

- Split dataset into N equal test folds
  - For each test fold, the remaining (N-1)/N fraction is the training fold
- Fit model to each training fold, and calculate predictor (e.g. probability of disease) for the observations in the corresponding test fold
- Evaluate predictive performance by comparing observed with predictive status over all test folds
  - Area under ROC (uncalibrated prediction)
  - test log-likelihood (calibrated)

Using cross-validation to learn number of SNPs retained by LASSO regression



N-fold cross-validation

- Partition dataset into N disjoint *test folds*
    - For each test fold, all other observations are the corresponding *training set*
- For each test/training fold
    - a model is fitted to the training fold and predictions are evaluated on the test fold
    - Predictive performance is evaluated by summing over all test folds
        - For each observation, can compare observed value with value predicted from model fitted to the corresponding training fold
        - Can compute area under ROC curve

Cross-validation compared with a conventional test/training split

With 4-fold cross-validation, each observation appears in one test fold and in 3 training folds

Tumor Marker Utility Grading System (Am Soc Clin Oncology 1996)

- Levels of evidence:-
  - Level I: prospective study specifically designed to test marker, or meta-analysis of level II or III studies
  - Level II: prospective trial in which marker study is secondary objective of trial protocol)
  - Level III: large retrospective studies
  - Levels IV, V: small retrospective studies or pilot data

Critical reading of a paper on a new biomarker-based prediction

- What was the outcome variable, how many biomarkers (P), how many individuals (N)
- What methods were used to control overfitting (unless P << N)?
- Was performance of biomarker-based prediction compared with performance of prediction from clinical data only?
- Was predictive performance evaluated on data that were not used to learn the model or to preselect relevant variables?
- Is the predictor generalizable or interpretable?

Relation of molecular pathology to stratified medicine and molecular epidemiology

- Molecular pathology:
  - using molecular biomarkers to diagnose disease
- Stratified medicine:
  - using (molecular) biomarkers to subtype disease and predict therapeutic response
- Molecular epidemiology:
  - using molecular biomarkers to study states of health and disease in populations

Biomarkers

- Biomarker is any standardized measurement that predicts a biological state: disease, disease subtype or therapeutic response
- Usually used for in vitro molecular measurements of molecules
  - imaging measurements are also biomarkers
- Most classic molecular biomarkers are univariate
- “-omic”platforms yield *high-dimensional* data
  - Number of variables >> number of individuals

Types of biomarker

- genotypic: SNPs, sequence data
- phenotypic:
  - Somatic DNA: in tumour or cell-free
  - gene expression: microarrays, custom RT-PCR kits, DNA methylation, micro-RNA
  - proteins and glycans
  - lipoproteins
  - small molecules: lipidomics, metabolomics
  - imaging
  - time-series clinical measurements

Discovering biomarkers – two complementary approaches

- (1) Identify candidate biomarkers from basic biology, and test them for association with outcome
  - Success in cancer: expression of the drug target is a candidate biomarker of drug response
- (2) Use -omic platforms to measure many biomarkers simultaneously and select those that predict outcome jointly or singly
  - Requires large sample sizes and methods for high-dimensional data
- Both approaches require tissue samples stored at baseline, and long-term follow-up
  - Electronic health records make follow-up cheaper

Statistical methods for molecular epidemiology

- Prediction from high-dimensional data
  - classical statistical methods fail when number of variables >> number of observations
- Learning how to cluster patients into disease subtypes, so as to optimize prediction
  - “mixtures of experts” model learns “soft” classification, in which each patient is may be a mixture of different disease subtypes

Prediction from high-dimensional data

- Dimensionality reduction
  - Clusters of correlated biomarkers can be replaced by a few weighted scores that contain the same information
  - Only useful where many biomarkers measure the same thing
- Sparsity-enforcing methods
  - Learn the optimal number of biomarkers to retain in a predictive model
- Non-parametric methods
  - Kernel-based methods learn a function that evaluates the similarity between pairs of observations
  - Deep learning – a new kind of neural network
  - These are “black-box” methods: don't select the most useful biomarkers

Cross-validation (repeated training/test splits) is used to learn the best tradeoff between complexity and fit

What is stratified medicine?

- Use of biomarkers + clinical data to stratify patients with a given diagnosis so as to select optimal treatment for individual patient
  - stratification may be on subtype of disease (e.g. cancer, rheumatoid arthritis)
  - or on factors that are not disease-related but influence drug response (e.g. individual variation in drug sensitivity)
  - Subtyping disease is more useful than “black box” prediction because disease subtype is likely predict response to drugs not yet discovered

Status of stratified medicine

- Long-established: antibiotic/antiviral sensitivity testing
- Adverse drug reactions: HLA-B*5701 allele predicts hypersensitivity to abacavir (Mallal 2002):
  - > 90% sensitivity for reaction confirmed by skin patch testing
- Cancer – several established applications based on gene expression in tumours
- Rheumatoid arthritis – still at research stage

Examples: stratification based on expression of single gene by tumour

- Many drugs for cancer target tyrosine kinase signalling pathways: epidermal growth factor receptor (EGFR), HER2/neu, B-raf, K-ras
- cetuximab: monoclonal antibody against EGFR
  - effective in colorectal cancer only if KRAS gene is not mutated (Karapetis 2008)
- trastzumab (herceptin): monoclonal antibody against HER2/neu receptor
  - effective against breast tumours that express HER2 (but retrospective analysis shows benefit in women reclassified as HER2-negative)
- vemurafenib: inhibits B-raf enzyme
  - effectiveagainst melanoma only if BRAF gene has V600 mutation
- imatinib: inhibits BCR-ABL tyrosine kinase
  - effective in myeloproliferative leukaemia / lymphoma only with PDGFR gene rearrangements or Philadelphia chromosome

Tumor Marker Utility Grading System (Am Soc Clin Oncology 1996)

- Levels of evidence:-
  - Level I: prospective study specifically designed to test marker, or meta-analysis of level II or III studies
  - Level II: prospective trial in which marker study is secondary objective of trial protocol)
  - Level III: large retrospective studies
  - Levels IV, V: small retrospective studies or pilot data

Stratification based on multivariate gene expression: breast cancer

- 2 assays licensed by FDA
  - Oncotype Dx: 21 genes measured by real-time polymerase chain reaction (RT-PCR)
  - Mammaprint: 70 genes measured by microarray
- Other commercialized assays:
  - Mammostrat: 5 genes
  - Breast Cancer Index: 2 genes
  - BreastOncPx: 14 genes
  - PAM50 Breast Cancer Intrinsic Classifier: 50 genes

Oncotype Dx

- for oestrogen-receptor-positive breast cancer
  - assay kit measures 21 genes
  - Score validated using archived tumour samples from a randomized trial
  - predicts recurrence, and may also predict response to adjuvant hormonal therapy and chemotherapy
    - identifies women who do not need chemo
    - Test costs $4000, manufacturer estimates it pays for itself but NICE disagrees

Mammaprint

- van t'Veer 2002: 78 patients with breast cancer
- 5000 “significantly regulated” probes selected from 25000 on microarray, ranked by correlation with time to recurrence
- clustering procedure
- optimal number of probes to retain determined by leave-one-out cross-validation
  - 70 genes retained
- initially validated against additional 19 patients
- Buyse 2006: validation in 307 patients
- score dichotomized~40/60 split low risk/high-risk
- hazard ratio ~ 2.1 for metastases after adjusting for clin/path risk classification
- Scoring algorithm is not published

Some questions about multivariate scores

- Does it matter that the genes used in different scores do not overlap much?
  - there may be an underlying structure (e.g. patients clustered into two disease subtypes) for which many different scoring systems are good enough to classify patients
- If underlying score is continuous users should not be given just a dichotomous classification
- Are these scores “inventions” and how far do IP rights extend?

Regulatory efforts for in vitro diagnostics

- FDA 2007: univariate “laboratory-developed tests” are subject to “enforcement discretion”
  - FDA enforces lab quality, but leaves interpretation to the clinician
- In vitro diagnostic “multivariate index” assays are not transparent to the clinician and should be regulated by FDA.
  - draft guidelines: prospective studies ideal, but retrospective studies using archived samples may sometimes be used
- FDA 2010: draft guidelines for multivariate index assays withdrawn
  - FDA to focus on broader framework for regulating “laboratory-developed tests”, including direct-to-consumer genetic testing

Economics of stratified medicine

- *Value-based pricing* now being introduced in the UK
  - National Institute for Clinical Excellence negotiates drug price based on the cost per quality-adjusted life-year (QALY) gained
- Number of patients treated with a given drug may fall but the value added per patient should rise
  - pricing per dose is economically inefficient
  - marginal cost of extra dose is low, but each dose is priced to recover development costs
- Alternatives to pricing per dose:-
  - site licensing at national or international level for unlimited use, paid for by governments
  - private sector competition for direct government funding of drug development costs, with no patent rights

The future of molecular pathology and stratified medicine

- Genome-wide genotypic profiles and high-dimensional phenotypic biomarker profiles will be collected routinely in clinical practice
  - Not just one-off snapshots, but time-series data will be available
- Linkage of these biomarker data to electronic health records, with individual consent, will allow models to be developed for risk prediction and disease subtyping
- Single-biomarker diagnostics will be supplanted by multivariate diagnostics
  - Risk prediction and disease subtyping will rely on machines: too much data for humans to interpret
- Regulating these multivariate in vitro diagnostics will be a challenge.