Implications of studies of SNP genotype covariance for prediction from genotypes

- Most of the genetic variance of complex traits is accounted for by additive (linear) effects of common SNPs
- It is therefore possible in principle to predict these traits from linear models based on common SNPs
- Most SNP-trait effects are so polygenic (tiny effects of many SNPs) that to learn such models from genotype-trait associations may require unrealistically large sample sizes

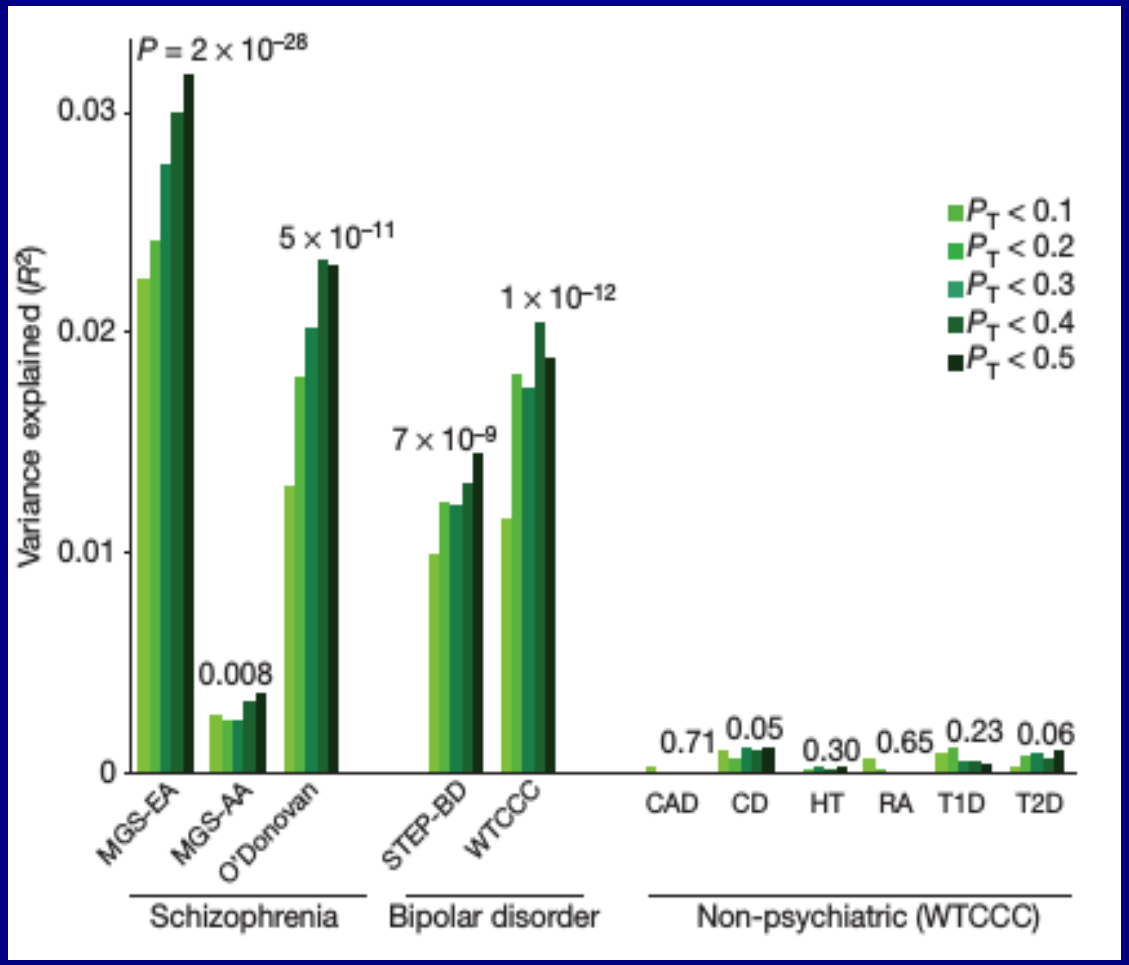
Prediction from high-dimensional data

- Dimensionality reduction
  - Principal components analysis (PCA) for correlated variables
  - Summary scores learned from data
- Non-parametric methods
  - learn a function (kernel) that evaluates the similarity between pairs of observations
- Sparse priors (e.g. LASSO regression)
  - encode prior belief that effects are mostly small or near zero
  - sparsity parameter can be learned from data

Using allele scores to predict outcome

- Allele scores can be computed from summary results of a GWAS
- (1) filter SNPs to select those that have p-value below some threshold
  - can be less stringent than the conventional threshold for declaring genome-wide significance
- (2) Calculate individuals'scores as sum of filtered SNP genotypes weighted by the regression coefficients
  - Use this score as a predictor

International Schizophrenia Consortium (2009): prediction of schizophrenia in independent samples from allele scores



LASSO regression

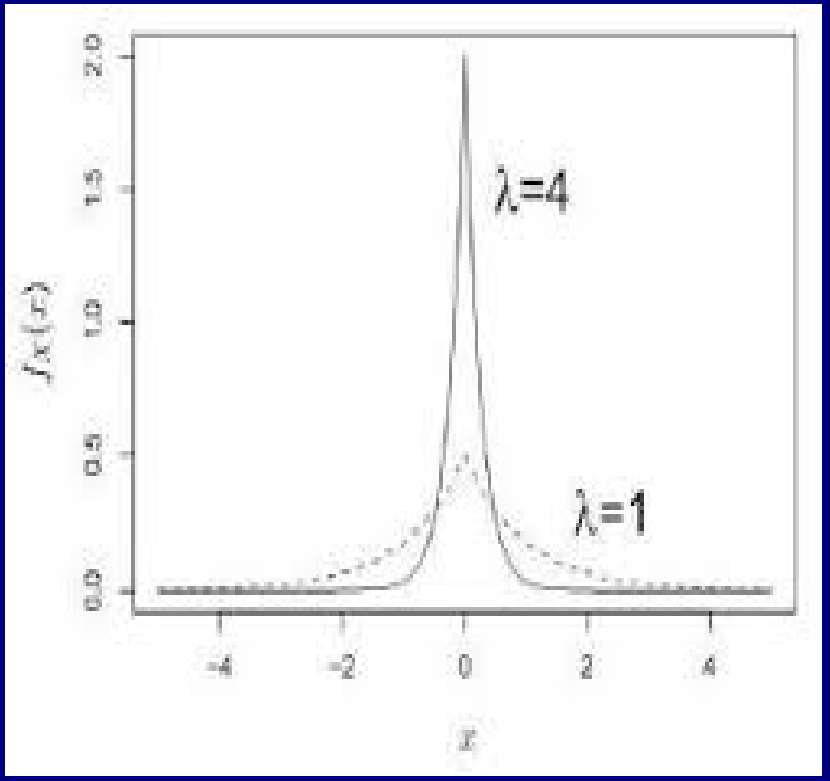
- Least Absolute (value) Shrinkage and Selection Operator
- Standard regression programs maximize log-likelihood (probability of data given model) as a function of regression coefficients $\boldsymbol{\beta}$
- LASSO regression maximizes log-likelihood - $\lambda \Sigma | \beta_i |$, where λ is a parameter controlling sparsity
  - best value of λ is learned by cross-validation against withdrawn observations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 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.

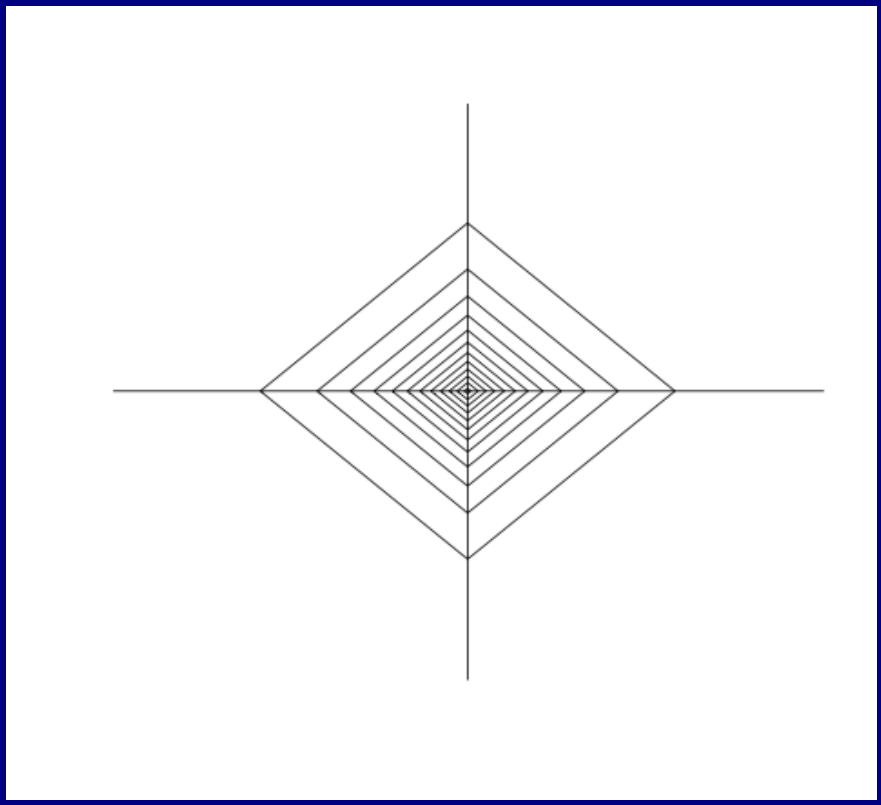
LASSO regression and the double exponential prior

- 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

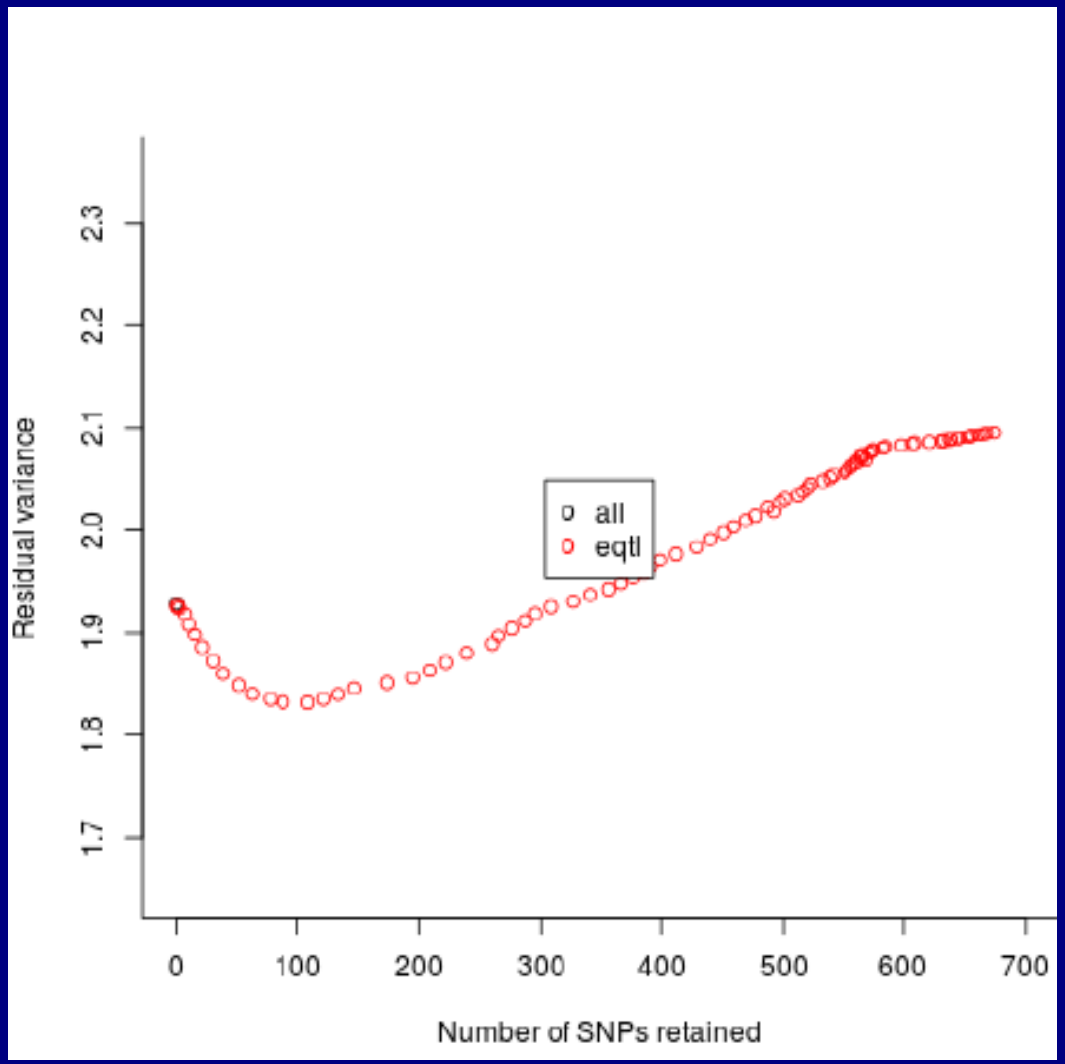
- Contour plot of 2D probability density looks like pyramid
  - Contour plot of gaussian density would be concentric circles
- Density varies inversely with sum of absolute values of effect parameters



Why do we need to use cross-validation to learn and evaluate a predictive model?

- To evaluate predictive performance
  - on test data that have never been used to learn the model
  - no need for a separate validation study (but may be hard to convince reviewers / regulatory agencies of this)
- To tune the learning algorithm
  - Optimal number of variables to retain
  - More generally, learn parameters that control how much the model adapts to the data
  - Models that adapt too much will *overfit*

Using cross-validation to learn the number of SNPs retained (controlled by sparsity parameter) by LASSO regression



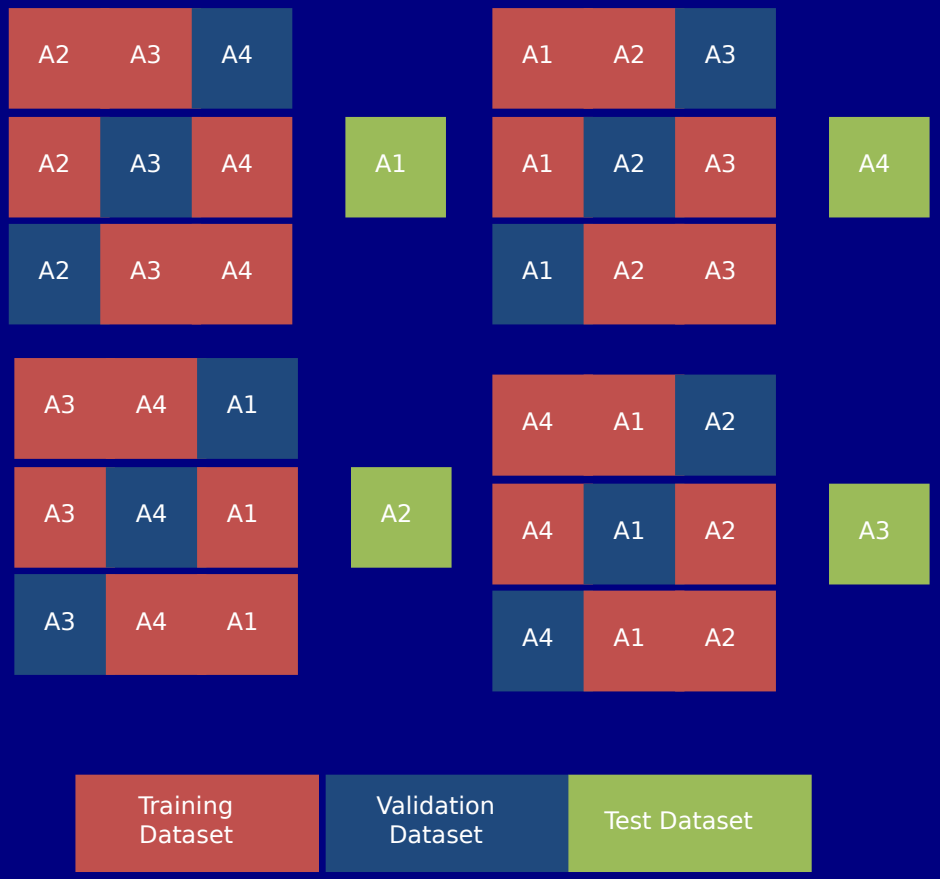
N-fold cross-validation

- 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

Nested cross-validation



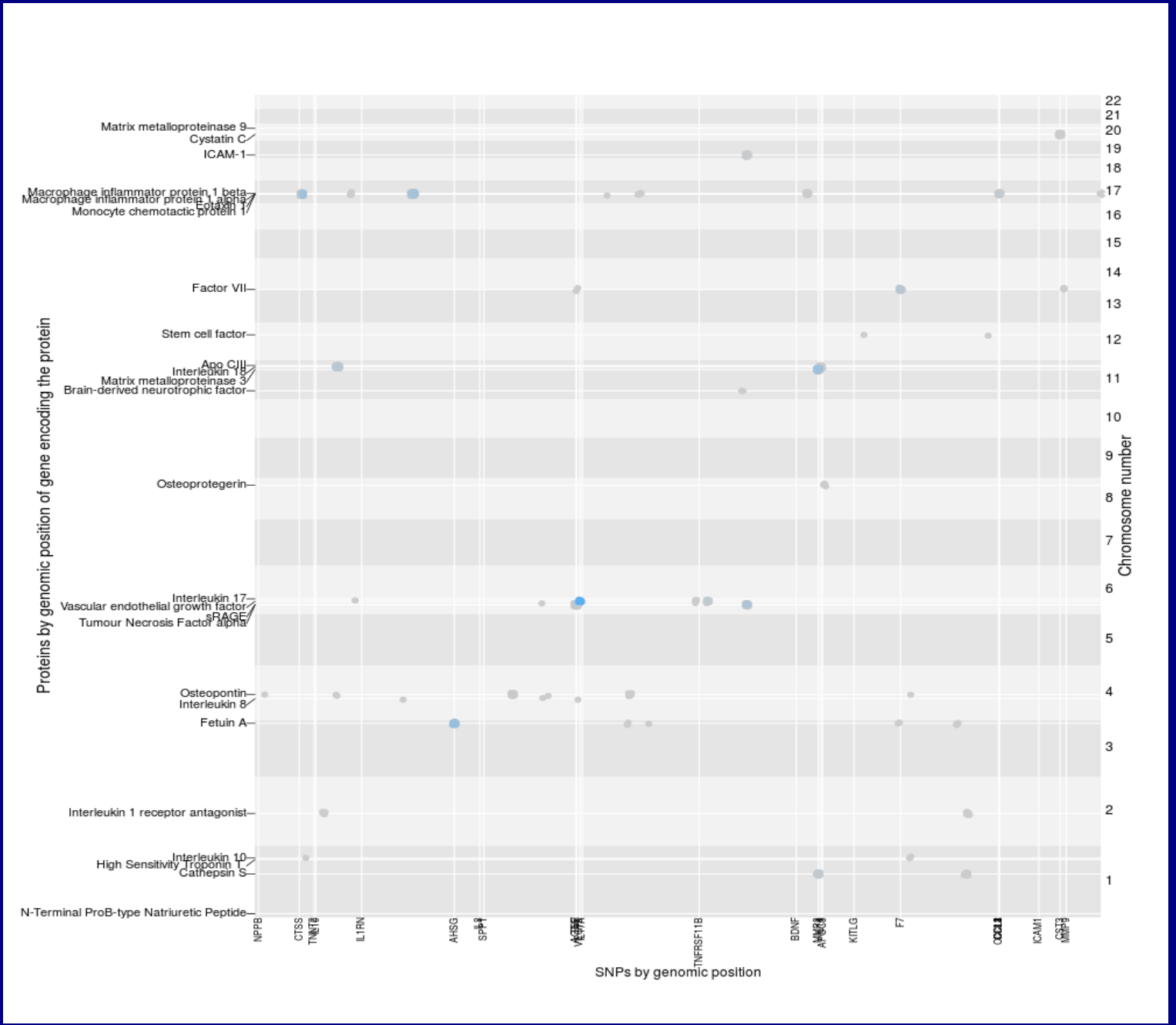
If we are using cross-validation to tune the model and also to evaluate predictive performance, we need *nested* cross-validation.

Inner folds are used to tune the model (e.g. learn the optimal setting of the LASSO penalty parameter

Outer folds are used to evaluate performance of the tuned model

With 10-fold cross-validation, nested cross-validation requires 100 model fitting runs

GWAS of 25 protein biomarkers



Genotypic prediction of 25 protein biomarkers

Protein	Proportion of variance explained by allele score, filtered at 50000 SNPs	Proportion of variance explained by LASSO	Number of SNPs retained in LASSO (mean over training folds)
FETUIN	0.071	0.327	140
VEGF	0.003	0.250	48.3
MCP-1	0.002	0.229	23.1
OPN	0.053	0.217	443
EOTAXIN-1	0.008	0.198	42.1
MIP1BETA	0.020	0.143	90.2
IL-1 RA	0.021	0.113	23.2
MMP9	0.017	0.093	919.8
FACTORVII	0.003	0.089	25.8
IL1ALPHA	0.005	0.084	17.1
TNFALPHA	0.002	0.071	13.4
MMP3	0.002	0.069	15
IL-18	0.026	0.065	78.1
IL-10	0.005	0.060	1049.1

The future of genotypic prediction

- Allele scores can be computed from summary level meta-analyses which are available for very large datasets
- LASSO predictors outperform allele scores but constructing them requires access to individual-level data
- Genotypic effects on biomarkers are more oligogenic than effects on disease
  - Can learn genotypic predictors of biomarkers from cross-sectional studies, then use them as “features” to construct disease predictors

Using genetic variation to infer causal biomarker-disease associations Bayesian instrumental variable analysis

- "-omic" epidemiology yields many phenotypic *biomarkers* that predict outome
  - metabolic measurements, gene expression levels, serum proteins
- We want to infer which biomarker-disease associations are causal
  - possible therapeutic targets
  - as surrogate end-points in early-stage clinical trials

Classical epidemiological approach to inferring causation from an exposure-disease association

- Measure all likely confounders: factors that are independently associated with outcome
- Test whether exposure-disease association persists after adjusting for these confounders
- Control of confounding is likely to fail with biomarkers because the likely confounders are unknown or difficult to measure
  - for instance raised cytokine levels predict age-related cognitive impairment – but are affected by underlying disease processes

Control of confounding: epidemiology faces its limits

- Standard methods for control of confounding in epidemiological studies are likely to fail if the exposure under study is:-
- A biomarker: e.g. an inflammatory marker
  - Association with outcome may be confounded by unknown metabolic/physiologic factors
- A health-seeking behaviour: e.g. use of vitamin E supplements, post-menopausal oestrogen
  - Association with outcome may be confounded by other health-seeking behaviours

Why does control of confounding fail for "endogenous" variables?

- Biomarkers:
  - confounders are unknown
  - Temporal sequence from exposure to outcome is difficult to establish: reverse causation is possible
- Behavioural factors
  - confounding is likely to be strong for a disease/outcome where risk can be modified by "lifestyle" factors
  - Measurement of exposure is often biased

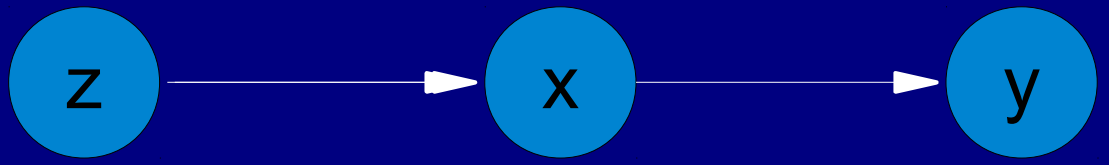
Instrumental variable analysis

- Identify an “instrument” that perturbs the exposure of interest (usually a biomarker or behavioural factor)
- Assumptions:-
  - Effect of instrument on outcome is unconfounded
  - Any effect of instrument on outcome is mediated through the intermediate variable.
  - Effects of setting different levels of exposure are independent of the instrument

Instrumental variable analysis in economics

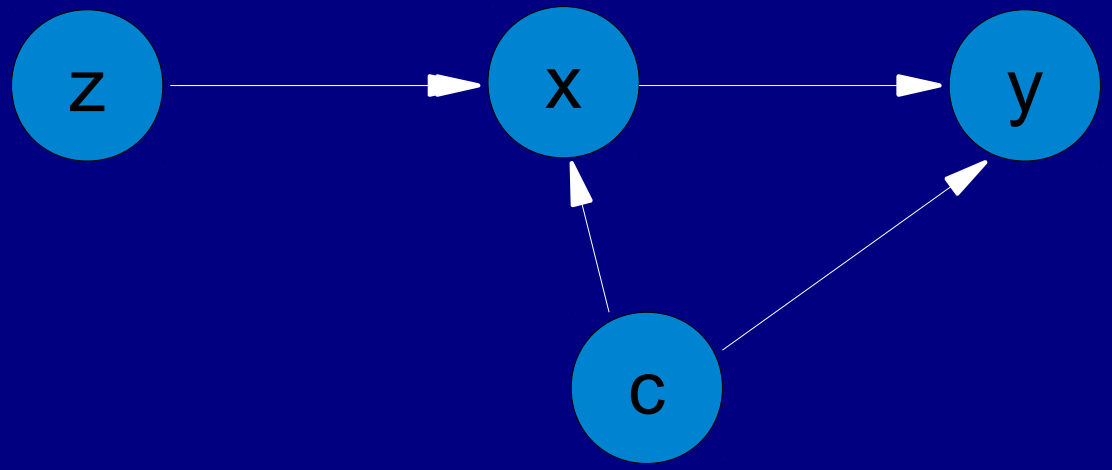
- Economists want to infer the effects of “endogenous” (intermediate) variables that are likely to be confounded
- Example
  - age at leaving school is an “endogenous variable” that predicts lifetime earnings
  - variation in statutory school-leaving age can be used as an instrument
  - can estimate the causal effect of extra year's school on outcome

Conditional independence in graphs



- Rules of conditional probability give $P(x, y, z) = p(x \mid z)\, p(y \mid x)\, p(x)$
- z and y are dependent, but conditionally independent given x

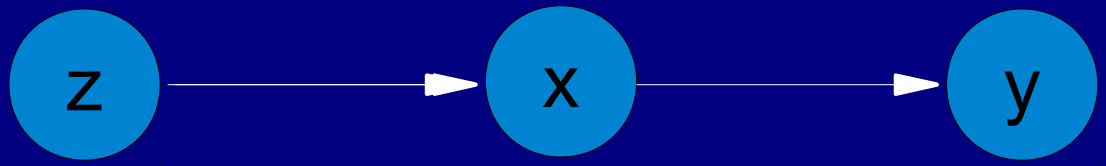
Graphical definition of a confounder



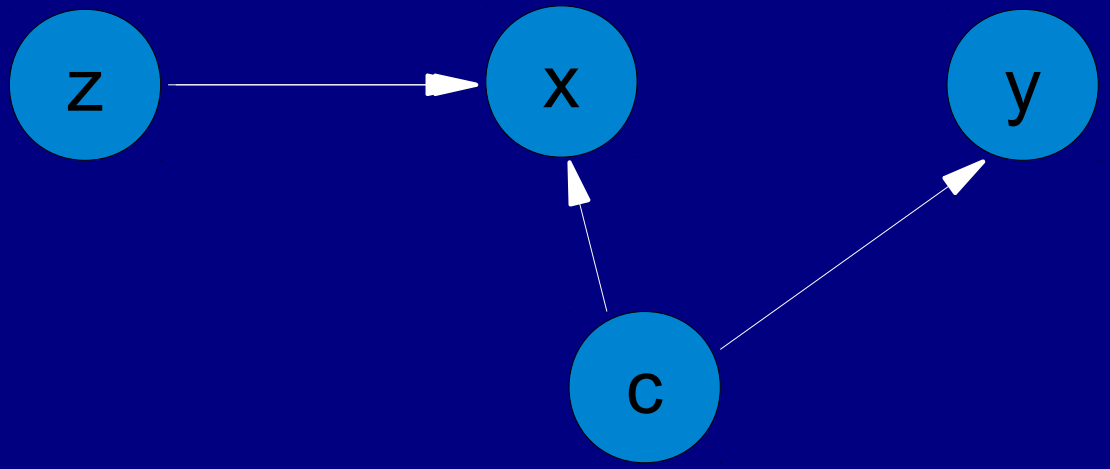
- Confounder of association between x and y is any variable on a pathway from which information flows to x and y
- information flow is defined by introducing a *do*-operator (equivalent to a latent instrumental variable z)
  - Information cannot flow backwards in time

Inferring causation from conditional independence in graphs

- Causal relationship: y depends on z



- Confounding: no information flow between z and y



Classical approaches to causal inference

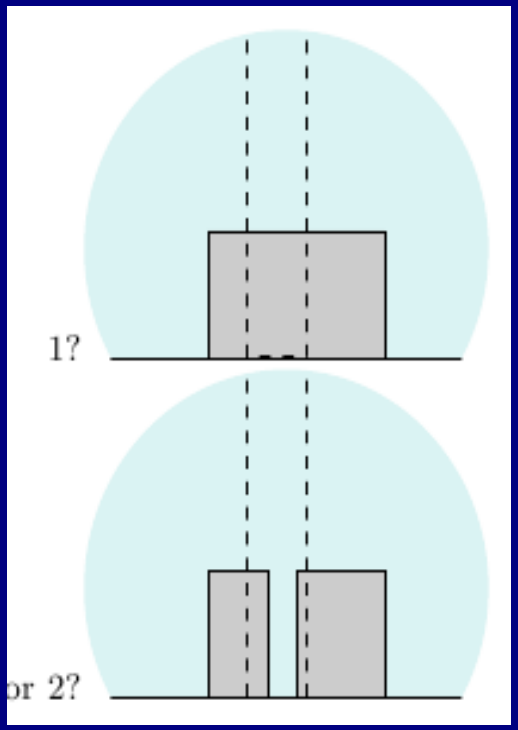
- Experimentalists: causal relationships can be inferred only from randomized intervention
- *Structural causal model* (Pearl): causation can be inferred if one of three conditions holds
  - an instrumental variable has been measured (randomization is a special case)
  - all confounders have been measured (back-door criterion)
  - an unconfounded variable on the causal pathway has been measured (front-door criterion

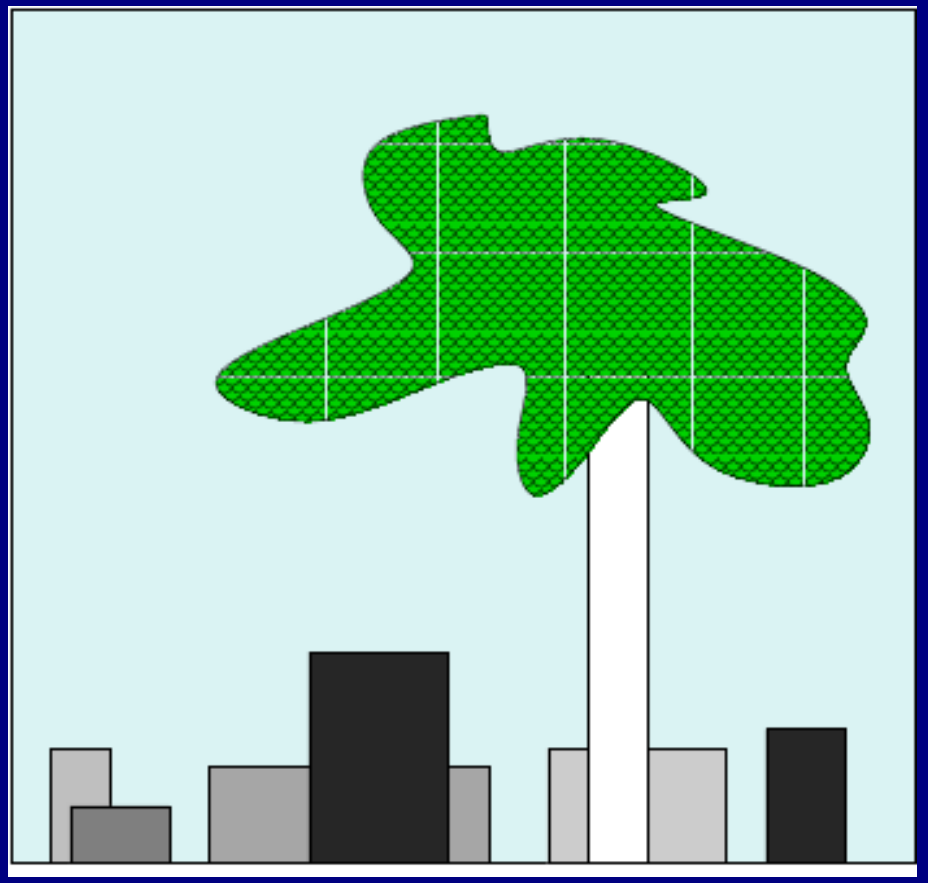
Bayesian approach to causal inference

- Simple explanations, involving only a few parameters that are learned from the data, are *more probable* than explanations that invoke many parameters learned from the data
  - The probability of data given the model is the *likelihood* of the model
  - The weight of evidence favouring one model over another is the difference between the logs of their likelihoods (or equivalently the log-likelihood ratio
  -

How Bayesian hypothesis testing favours the simplest explanation that fits the data: Mackay 2003

- How many boxes are behind the tree?





Smoking and lung cancer debate in 1950s

- Classical statisticians' argument:
  - any inference of causation from observational data is unreliable
  - how do you know that all relevant confounders have been measured?
- Epidemiologists' argument
- even without experimental confirmation, evidence from observational studies can strongly favour causation

Bradford Hill criteria: how to infer causation where classical criteria are not met

- Strength of association
- Temporal sequence
- Consistency
- Biological plausibility
- Coherence
- Specificity in the causes
- Dose-response relationship
- Experimental evidence
- Analogy

“Mendelian randomization”: instrumental variable analysis with genetic instruments

- Find genes in which variation perturbs levels of the biomarker. Compare effects on outcome of
  - genetic perturbation of the biomarker
  - non-genetic variation of the biomarker
- Example:
  - raised plasma fibrinogen predicts cardiovascular disease
  - genotype in the beta-fibrinogen gene predicts fibrinogen levels
  - genotypic effects on fibrinogen levels do not predict cardiovascular disease

Assumptions underlying instrumental variable analysis with genetic instruments

- Effect of genotype on outcome is unconfounded
  - guaranteed by Mendel's laws, if population stratification is controlled
- Effect of genotype on outcome is mediated only through the intermediate phenotype (no *pleiotropy*)
- To be able to generalize: effects on outcome of different settings of the biomarker are independent of the instrument
  - no *developmental compensation* / *channelling*

Instrumental variable estimate of causal effect size

- Two-stage method
- regress intermediate variable x on genotype g to evaluate predicted values of x given genotype: $\langle x|g \rangle$
  - can use cross-sectional studies with measurements of x and g only.
- regress outcome *y* on the predicted values $\langle x|g \rangle$: valid for logistic regression
  - can use case-control studies with measurements of y and g only

Is the association of alcohol intake with colorectal cancer causal? (Wang 2010)

- ALD2 polymorphism: 487Lys allele impairs metabolism of acetaldehyde, causing flushing and other symptoms in response to alcohol.
- allele frequency ~ 0.25 in East Asia
- Meta-analysis of 1960 cases and 3163 controls in Japan and China
- Odds ratio for colorectal cancer was 1.3 in Glu/Glu homozygotes versus Lys/Lys homozygotes

Do raised homocysteine levels cause coronary heart disease?

- Raised homocysteine levels are associated with CHD
- TT genotype in the MTHFR gene is associated with reduced folate-dependent enzyme activity and with 20% higher homocysteine levels
- Clarke 2012: in meta-analysis of 48175 CHD cases and 67961 controls, odds ratio for CHD associated with TT genotype was 1.02

Does being fat cause psychological distress? (Lawlor 2011)

- BMI and waist/hip ratio are associated with questionnaire measures of psychological distress in the population
  - 53221 adults in Copenhagen
  - odds ratio 1.1 for “not accomplishing very much” for increase of 1 SD in BMI or WHR
- 2 SNPs in *FTO* (fat mass & obesity-related) and *MC4R* (melanocortin 4 receptor) as instruments for adiposity effect
  - Causal odds ratios 0.6 (0.46-0.89) for effect of BMI and 0.5 (0.25-0.94) for WHR

Are effects of FTO and MC4R variants on psychological distress mediated only through adiposity?

- Compare causal effects estimated using 2 instruments
- “Over-identification” test: test for residual associations of SNPs with psychological distress in a model that includes predicted level of adiposity given genotype.
- Multiple instruments allow you to test the assumption of no pleiotropy in a standard instrumental variable analysis

Summary: standard approach to exploiting Mendelian randomization

- For the intermediate trait of interest, try to find one or two genes that have moderately large effect, where pleiotropic effects can reasonably be excluded
- Calculate the predicted values of the trait given genotype in a cross-sectional study
- In a large case-control study, test for dependence of outcome on the predicted value given genotype
- If there are > 2 instruments, test for pleiotropy

Summary: future methods for exploiting Mendelian randomization

- Most genetic effects are polygenic: construct predictions of trait or biomarker from many SNPs (not necessarily intragenic)
  - Can use cross-sectional studies of genotype-trait associations
- Model causal effects on outcome using genotypic predictors of traits/biomarkers
  - Can use large case-control studies of outcome
- Where genotypic predictor is associated with outcome, evaluate causality or pleiotropy as alternative explanations

Association of LASSO genotypic predictors with CVD in an independent sample

protein	beta	pvalue
Factor VII	61.9	0.0001
SCF	-441.3	0.03
BDNF	509.6	0.03
EOTAXIN1	35.1	0.03
MCP1	10.5	0.07
MMP9	-267.3	0.17
IL8	72.3	0.18

- Supports causal explanation of observed association of Factor VII with CVD