Variable selection

David Rossell[1] and Omiros Papaspiliopoulos[2]

[1]: University of Warwick (UK)
[2]: ICREA and UPF

Outline

Why bother?

Goal: predict outcome t_i using $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ $(i = 1, \ldots, n)$

$$t_i = f(\mathbf{x}_i) + \epsilon_i, \text{ where } E(\epsilon_i) = 0, V(\epsilon_i) = q \text{ indep.}$$

Key: often only a subset of $\mathbf{x}_i$ really has an effect on t_i

1. Easier interpretation (clearly)
2. If p large, irrelevant $\mathbf{x}_i$'s $\rightarrow$ better predictions (as we shall see)
3. Higher robustness to outliers in $\mathbf{x}_i$
4. Cost of recording predictors
5. Practical considerations: ease of use, more convincing to practicioners, faster computations...

Specially important when p is large! (possibly $p \gg n$)

Bias/variance trade-off

Mean error to predict t_{n+1} at $\mathbf{x}_{n+1}$. $E((t_{n+1} - \hat{f}(\mathbf{x}_{n+1}))^2) =$

$$E((t_{n+1} - f(\mathbf{x}_{n+1}) + f(\mathbf{x}_{n+1}) - \hat{f}(\mathbf{x}_{n+1}))^2) =$$
$$E\left((t_{n+1} - f(\mathbf{x}_{n+1}))^2\right) + E\left((f(\mathbf{x}_{n+1}) - \hat{f}(\mathbf{x}_{n+1}))^2\right) + 0 =$$
$$q + \text{Bias}^2(\hat{f}(\mathbf{x}_{n+1})) + \text{Var}(\hat{f}(\mathbf{x}_{n+1}))$$

1. q is outside our control
2. Bias and variance we can hope to improve
   - Adding any variable increases variance
   - Dropping necessary variables increases bias

For those curious about details, the cross-product is 0 because

$$E\left((t_{n+1} - f(x_{n+1}))(f(x_{n+1}) - \hat{f}(x_{n+1}))\right) =$$

$$\cancel{f^2(x_{n+1})} - E(t_{n+1}\hat{f}(x_{n+1})) - \cancel{f^2(x_{n+1})} + f(x_{n+1})E(\hat{f}(x_{n+1})) =$$

$$-\cancel{f(x_{n+1})E(\hat{f}(x_{n+1}))} - E(\epsilon_{n+1}\hat{f}(x_{n+1})) + \cancel{f(x_{n+1})E(\hat{f}(x_{n+1}))} =$$

$$-E(\epsilon_{n+1})E(\hat{f}(x_{n+1})) = 0$$

We only used that

- ϵ_{n+1} independent of $\hat{f}(x_{n+1})$
- $E(\epsilon_{n+1}) = 0$

Example: $p = 2$ linear regression

$$\text{Let } \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} \\ \dots & \dots \\ x_{n1} & x_{n2} \end{pmatrix}; \mathbf{X}^T\mathbf{X} = \begin{pmatrix} s_{11} & s_{12} \\ s_{12} & s_{22} \end{pmatrix}$$

- Model 1: $t_i = w_1 x_{i1} + \epsilon_i$. Then $\text{Var}(\hat{w}_1) = q/s_{11}$
- Model 2: $t_i = w_1 x_{i1} + w_2 x_{i2} + \epsilon_i$

$$\text{Cov}\begin{pmatrix} \hat{w}_1 \\ \hat{w}_2 \end{pmatrix} = q(\mathbf{X}^T\mathbf{X})^{-1}$$

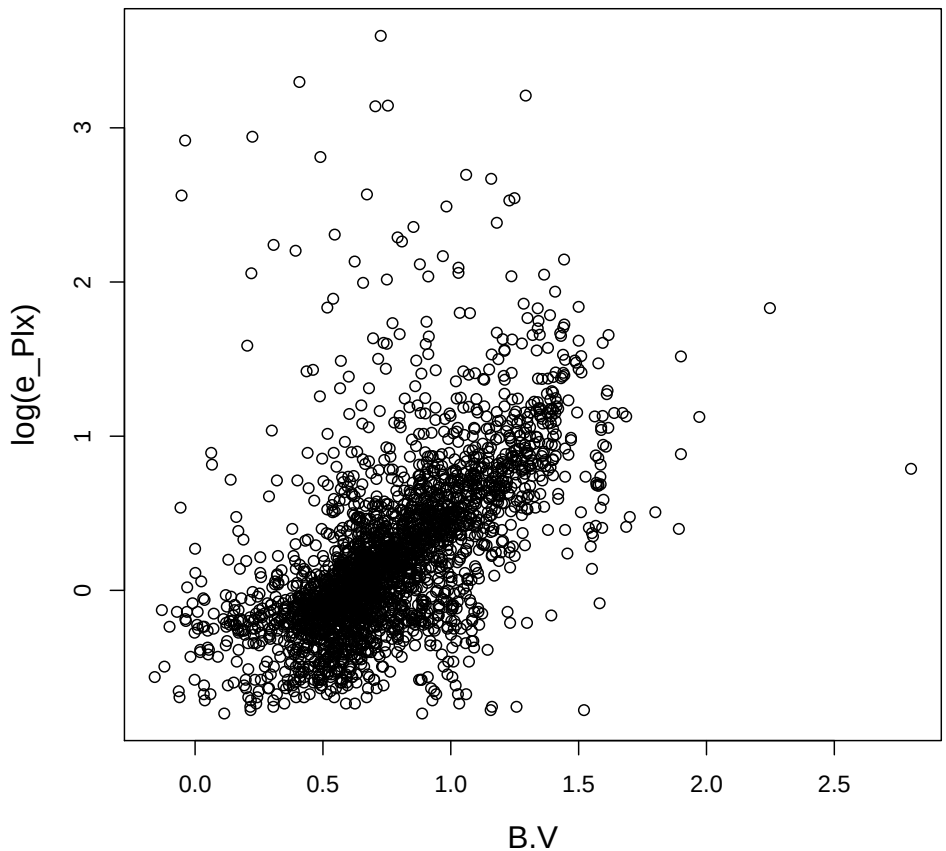
$$\text{Var}(\hat{w}_1) = q\frac{s_{22}}{(s_{11}s_{22} - s_{12}^2)} = q\frac{1}{s_{11} - s_{12}^2/s_{22}} > \frac{q}{s_{11}}$$

If truly $w_2 = 0$ then bias=0 for both models, but $E((t_{n+1} - \hat{t}_{n+1})^2)$ will tend to be smaller under Model 2

Example: Hipparcos star dataset

(astrostatistics.psu.edu/datasets/HIP_star.html)

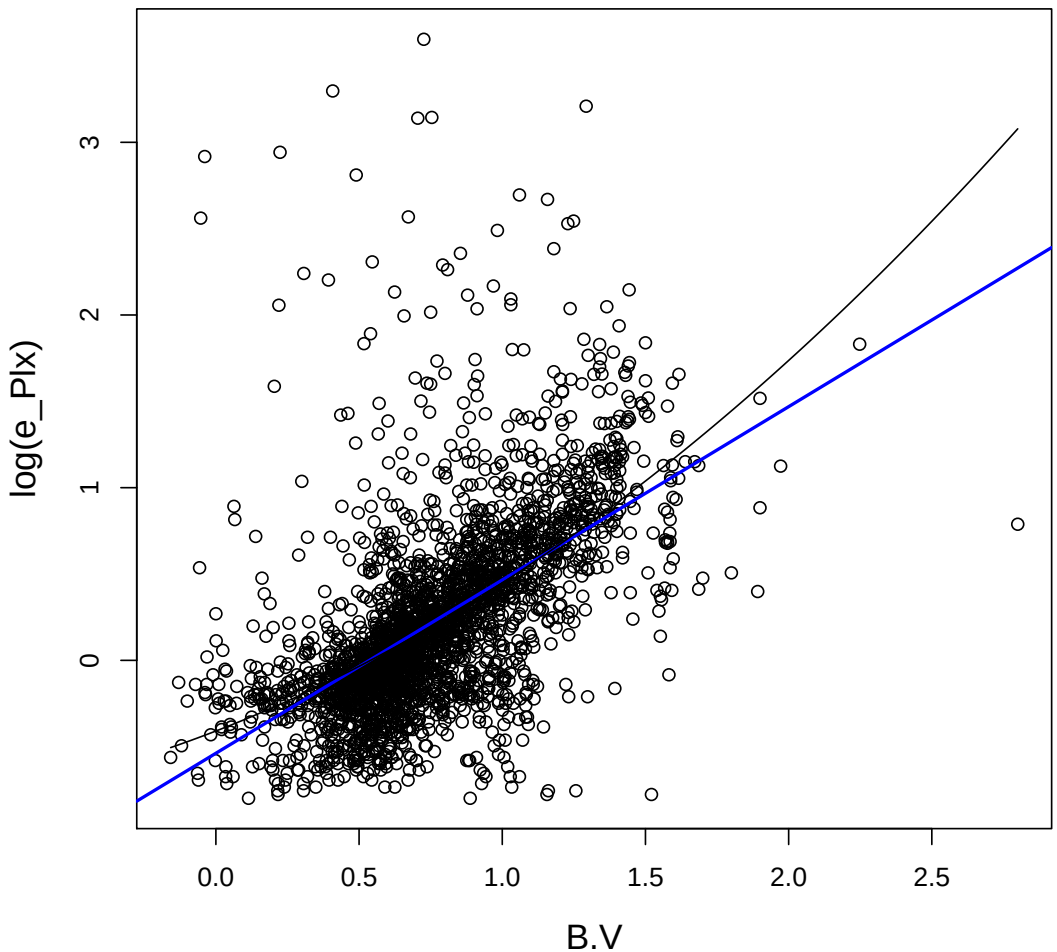
- B-V: color of star (brightness); e_PLx: error in measuring distance



Linear & quadratic fits

Model 1: $t_i = w_0 + w_1 x_{i1} + \epsilon_i$

Model 2: $t_i = w_0 + w_1 x_{i1} + w_2 x_{i1}^2 + \epsilon_i$



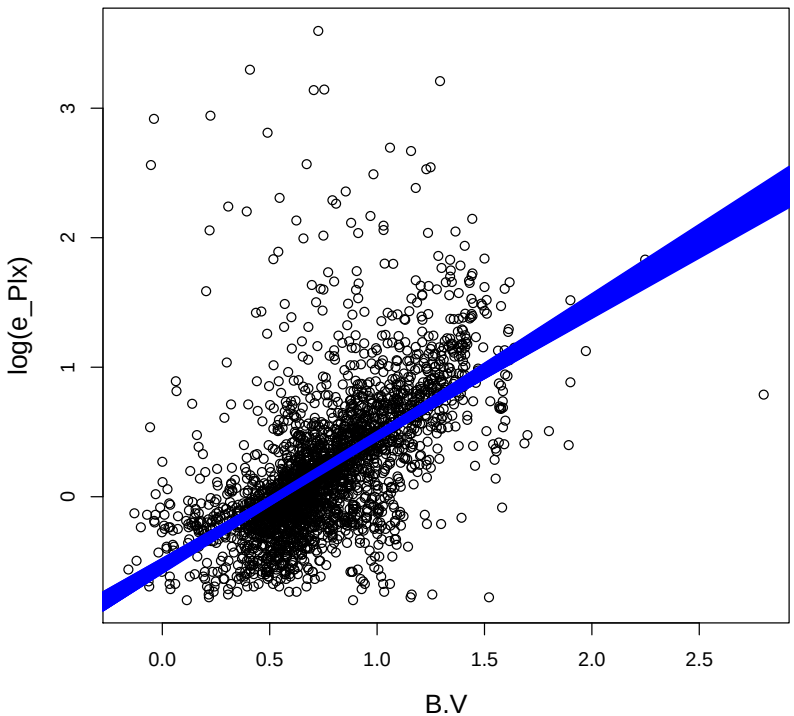
P-values

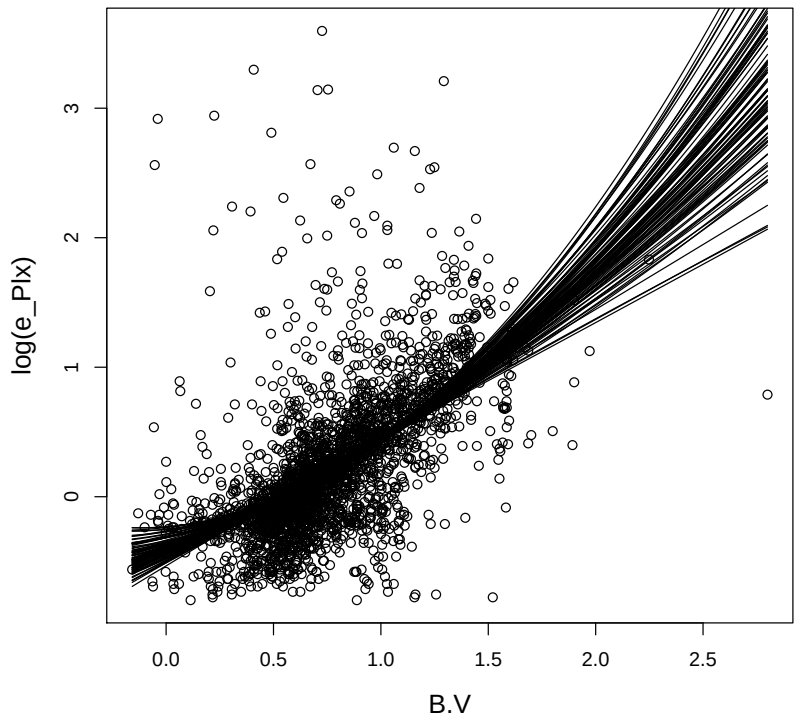
$\hat{w}_1$: 2.8E-12

$\hat{w}_2$: 3.42E-5

Let's assess the stability of the predictions (bootstrap)

1. Sample n observations with replacement
2. Fit linear & quadratic models





Suppose you want to predict at $x_{n+1} = 1.0$. And at $x_{n+1} = 2.0$?

Colon cancer dataset

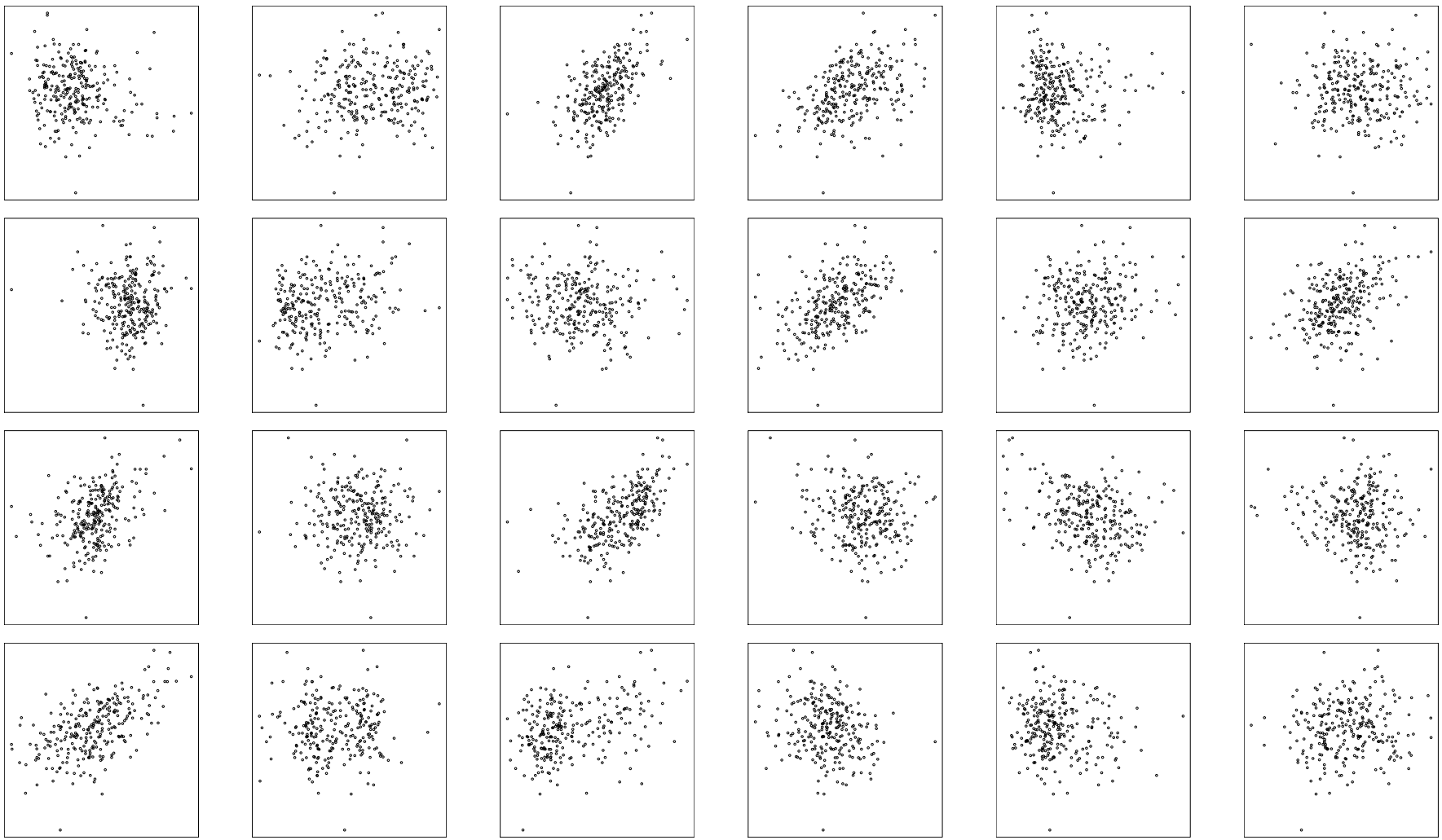
Calon et al (Cancer Cell, 2012)

- $n = 262$ colon cancer patients
- t_i: expression of gene TGFB in patient i
- $\mathbf{x}_i$: expression of $p \approx 20,000$ genes

Goals: TGFB is very important for cancer development, and there are experimental drugs to inhibit it

1. Understand how other genes are related to TGFB
2. Predict TGFB from a few genes (*e.g.* to identify patients potentially benefiting from TGFB inhibitors)

Scatterplot TGFB vs. X_1-X_{24}



Linear model appears reasonable

- No obvious outliers
- No obvious non-linear trends

In practice a better exploratory analysis would be needed: principal components, multivariate outliers/influential points...

How to select variables? You've already seen penalized likelihood (LASSO etc.). We'll focus on the Bayesian counterpart

Outline

Bayesian model selection

Let $M_1, \ldots, M_K$ be a collection of possible models for $\mathbf{t}$

- Likelihood $p(\mathbf{t} \mid \mathbf{w}_k, q_k, M_k)$
- Prior on parameters $p(\mathbf{w}_k, q_k \mid M_k)$
- Prior on models $p(M_k)$

We may obtain posterior model probabilities

$$p(M_k \mid \mathbf{t}) = \frac{p(\mathbf{t} \mid M_k)p(M_k)}{p(\mathbf{t})} \propto p(\mathbf{t} \mid M_k)p(M_k)$$

where as usual

$$p(\mathbf{t} \mid M_k) = \int \int p(\mathbf{t} \mid \mathbf{w}_k, q_k, M_k)p(\mathbf{w}_k, q_k \mid M_k)d\mathbf{w}_k dq_k$$

Bayesian model averaging

Prediction: we may predict t_{n+1} with

$$E(t_{n+1} \mid \mathbf{t}) = \sum_{k=1}^{K} E(t_{n+1} \mid \mathbf{t}, M_k) p(M_k \mid \mathbf{t})$$

- Considers uncertainty in what is the "right" model
- Using mean inherently means we care about quadratic losses

Estimation: if $\mathbf{w} = \mathbf{w}_k$ has a common meaning across models

$$E(\mathbf{w} \mid \mathbf{t}) = \sum_{k=1}^{K} E(\mathbf{w} \mid \mathbf{t}, M_k) p(M_k \mid \mathbf{y})$$

Bayesian variable selection

Consider models $M_1, \ldots, M_K$ with corresponding predictors $\mathbf{X}_1, \ldots, \mathbf{X}_K$ being subsets of $\mathbf{X}$

Example: $t_i \sim N(w_1 x_{i1}, 1)$

$$M_1 : w_1 = 0$$
$$M_2 : w_1 \neq 0$$

Example: $t_i \sim N(w_1 x_{i1} + w_2 x_{i2}, 1)$

$$M_1 : w_1 = 0, w_2 = 0$$
$$M_2 : w_1 = 0, w_2 \neq 0$$
$$M_3 : w_1 \neq 0, w_2 = 0$$
$$M_4 : w_1 \neq 0, w_2 \neq 0$$

Bayesian variable selection

Linear model likelihood: $p(\mathbf{t} \mid \mathbf{w}, M_k) = N(\mathbf{t}; \mathbf{X}\mathbf{w}, q\mathbf{I})$

- $\mathbf{t}$: vector with response $(n \times 1)$
- $\mathbf{X}$: matrix with predictors $(n \times p)$
- M_k includes subset of p variables ($K = 2^p$ models)
- $\mathbf{w}$: coefficients $(p \times 1)$ with 0 entries indicated by M_k
- q: residual variance

Let $\mathbf{w}_k$ be non-zero coefficients under M_k

- Prior on parameters: $p(\mathbf{w}_k, q \mid M_k)$
- Prior on models: $p(M_k)$

Prior on parameters $p(\mathbf{w}_k, q \mid M_k)$

$p(\mathbf{w}_k, q_k \mid M_k)$ can take any form and may incorporate prior beliefs or information. But we need to compute

$$p(\mathbf{t} \mid M_k) = \int \int p(\mathbf{t} \mid \mathbf{w}_k, q_k, M_k) p(\mathbf{w}_k, q_k \mid M_k) d\mathbf{w}_k dq_k$$

When 2^p is large closed-form expressions are convenient. One option is to use conjugate priors

- $p(\mathbf{w}_k \mid q, M_k) = N(\mathbf{w}_k; \mathbf{0}, q\mathbf{D})$
- $p(q) = \text{IG}\left(\frac{a_q}{2}, \frac{b_q}{2}\right)$ (equivalently $\frac{1}{q} \sim \text{G}(\frac{a_q}{2}, \frac{b_q}{2})$)

We need to set $\mathbf{D}, a_q, b_q$. To gain intuition, set $\mathbf{D} = g(\mathbf{X}^T\mathbf{X})^{-1}$ and let's look at the resulting expressions

Posterior model probabilities

$$p(\mathbf{t} \mid M_k) = \frac{p(\mathbf{t} \mid M_k)p(M_k)}{\sum_{j=1}^{K} p(\mathbf{t} \mid M_j)p(M_j)} = \frac{\mathrm{BF}_{k1}(\mathbf{t})p(M_k)}{\sum_{j=1}^{K} \mathrm{BF}_{j1}(\mathbf{t})p(M_j)}$$

where $\mathrm{BF}_{k1}(\mathbf{t})$ is Bayes factor between M_k and M_1 (no variables model)

$$\frac{p(\mathbf{t} \mid M_k)}{p(\mathbf{t} \mid M_1)} = (1+ng)^{-\frac{d_k}{2}} \left(1 + ng\left(1 + \frac{\tilde{\mathbf{w}}_k^T \mathbf{X}_k^T \mathbf{X}_k \tilde{\mathbf{w}}_k}{b_q + \widetilde{SSR}_1}\right)^{-1}\right)^{-\frac{a_q+n}{2}}$$

- $\tilde{w}_k = E(w_k \mid \mathbf{t}, M_k)$ is the posterior mean
- $\widetilde{SSR}_1$ the sum of squared residuals
- a_q is "prior sample size" (little influence for moderate n)
- b_q is prior guess for SSR (little influence for moderate n)
- g can be more influential

Default priors

Prior on q

- Often $a_q = b_q = 0.001$ (or other small value)
- Some authors set $a_q = b_q = 0$. An improper prior but works OK

Prior on $\mathbf{w}_k$

- $(\mathbf{X}^T\mathbf{X})/q$ is the information given by n observations, thus

$$\mathbf{w}_k \sim N(\mathbf{0}, gq(\mathbf{X}^T\mathbf{X})^{-1})$$

  with $g = n$ contains as much info as 1 observation (Unit Information Prior)

- $\mathbf{w}_k \sim N(\mathbf{0}, gq\mathbf{I})$ with $g = 1$ interpreted as info from 1 observation from earlier experiment with uncorrelated, unit variance predictors

Lots of literature on how to set g

- Treat as unknown parameter: hyper-g prior $p(g)$
- Empirical Bayes: estimate g from the data. Feasible, but may run into consistency problems. Care is needed...
- Frequentist calibration
- ...

There is a common perception that results are hugely sensitive to g, usually not true for g in a "reasonable range"

For $g = 0$ and $g = \infty$ we do run into trouble (Jeffreys-Lindley-Bartlett paradox), but these are silly values. For simplicity we will focus on Unit Information Prior ($g = n$)

Example: Hipparcos star dataset ($n = 2678$)

Star brightness (t_i) vs error in measuring distance (x_{i1})

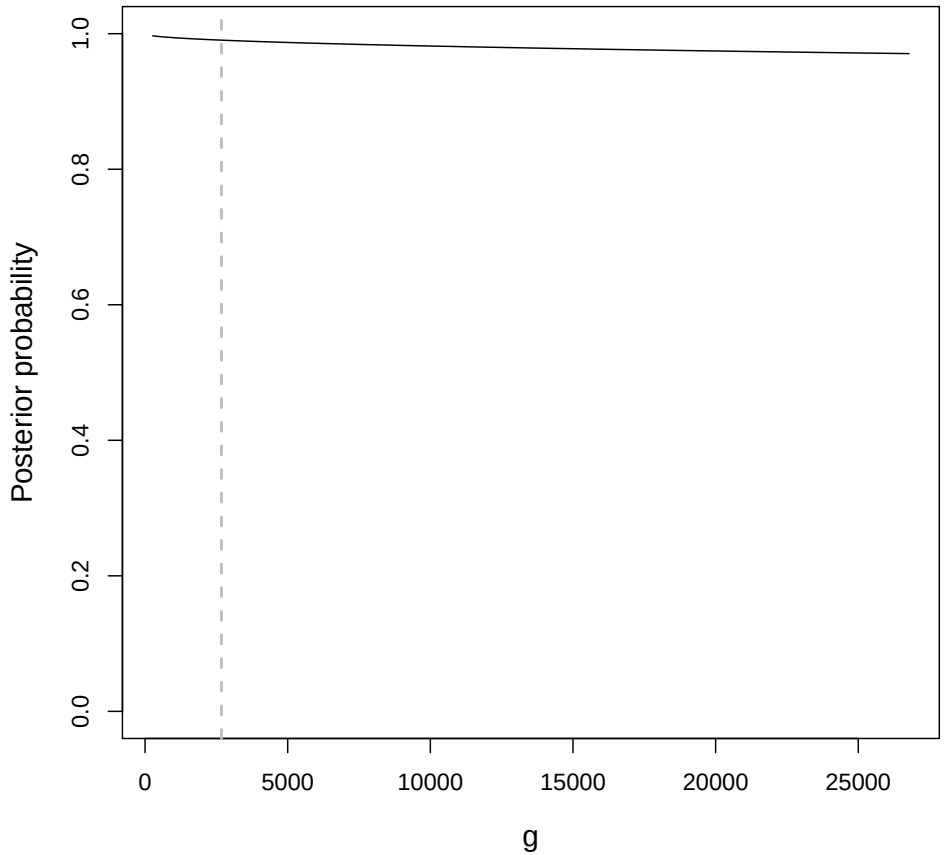
$$t_i = w_1 x_{i1} + w_2 x_{i1}^2$$

Four possible models

$$M_1 : w_1 = 0, w_2 = 0$$
$$M_2 : w_1 = 0, w_2 \neq 0$$
$$M_3 : w_1 \neq 0, w_2 = 0$$
$$M_4 : w_1 \neq 0, w_2 \neq 0$$

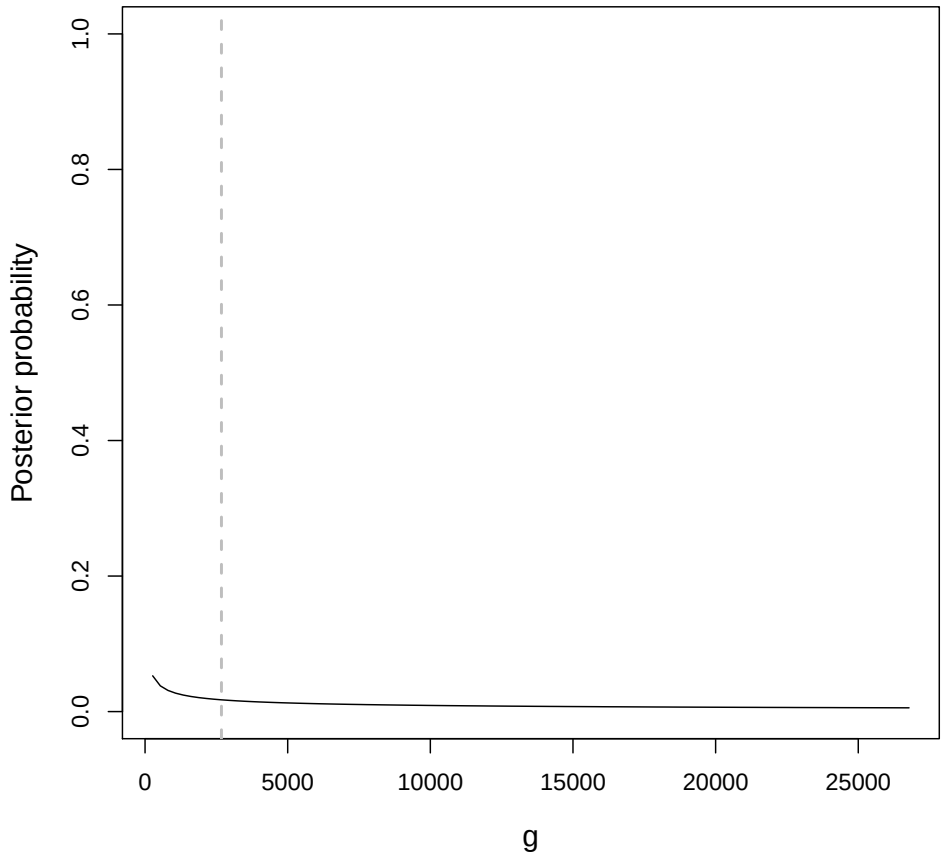
Set $P(M_1) = P(M_2) = P(M_3) = P(M_4) = \frac{1}{4}$

Consider $g \in (n/10, \ldots, 10n)$. Report $P(w_1 \neq 0, w_2 \neq 0 \mid \mathbf{t})$



Results not sensitive to g, but $n = 2,678$.

Randomly select $n = 200$ observations. Again $g \in (n/10, \ldots, 10n)$



- Smaller n: we do not detect that x_{i1}^2 is needed
- Conclusions still robust to g

Prior on model space

If no prior info available, let $p(M_k)$ depend only on $d_k = \sum_{j=1}^{p} I(w_j \neq 0)$ (number of predictors in M_k)

Common choices

1. Equal prior probability for all models (Uniform)

$$p(M_k) = \frac{1}{K}$$

2. Set $p(w_j \neq 0) = \pi$ indep. across $j = 1, \ldots, p$ (*e.g.* $\pi = 0.5$)

$$d_k \sim \text{Binom}(p, \pi)$$

3. Beta-Binomial(1,1): equal prior probabilities $d_k \sim \text{Unif}\{0, \ldots, p\}$
   Note: equivalent to $d_k \sim \text{Binom}(p, \pi), \pi \sim \text{Unif}(0, 1) = \text{Beta}(1, 1)$

Prior on model space

Example

Model	d_k	Unif	Bin$(2, 0.5)$	BetaBin
$w_1 = 0, w_2 = 0$	0	$1/4$	$1/4$	$1/3$
$w_1 = 0, w_2 \neq 0$	1	$1/4$	$1/4$	$1/6$
$w_1 \neq 0, w_2 = 0$	1	$1/4$	$1/4$	$1/6$
$w_1 \neq 0, w_2 \neq 0$	2	$1/4$	$1/4$	$1/3$

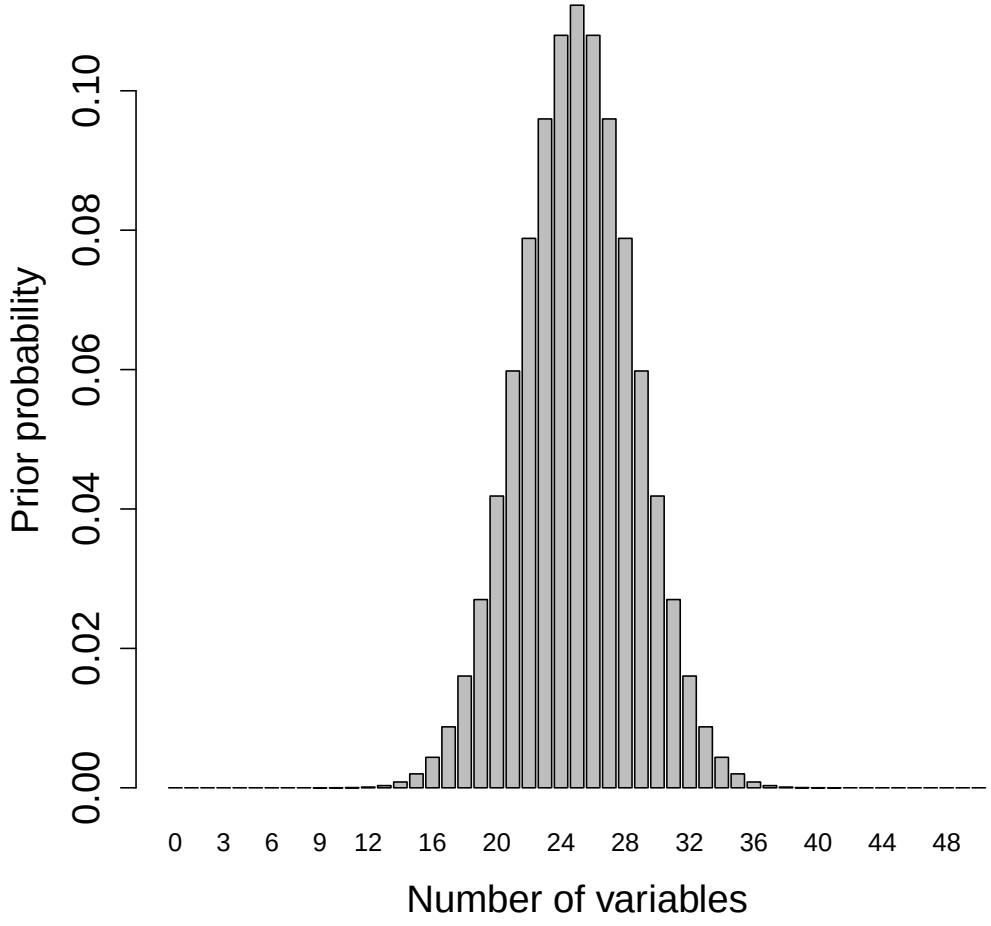
Note that we used two strategies

1. Set $p(M_k)$, then work out implied $p(d_k)$
2. Set $p(d_k)$, then split prob across models of equal size

But sometimes both are equivalent

1. Uniform $p(M_k)$ implies $p(d_k) \propto \binom{p}{d_k}$
2. Binom$(p, 0.5)$ implies $p(d_k) = \binom{p}{d_k} 0.5^{d_k} 0.5^{p-d_k} = \binom{p}{d_k} 0.5^p \propto \binom{p}{d_k}$

Example: suppose $p = 50$. Uniform $p(M_k)$ implies $P(d_k)$



Uniform & Binom($p, \pi = 0.5$) favour mid-size models

- Alternative 1: set $\pi < 0.5$ (e.g. from subject-matter considerations)
- Alternative 2: $d_k \sim$Beta-Binomial(1,1)

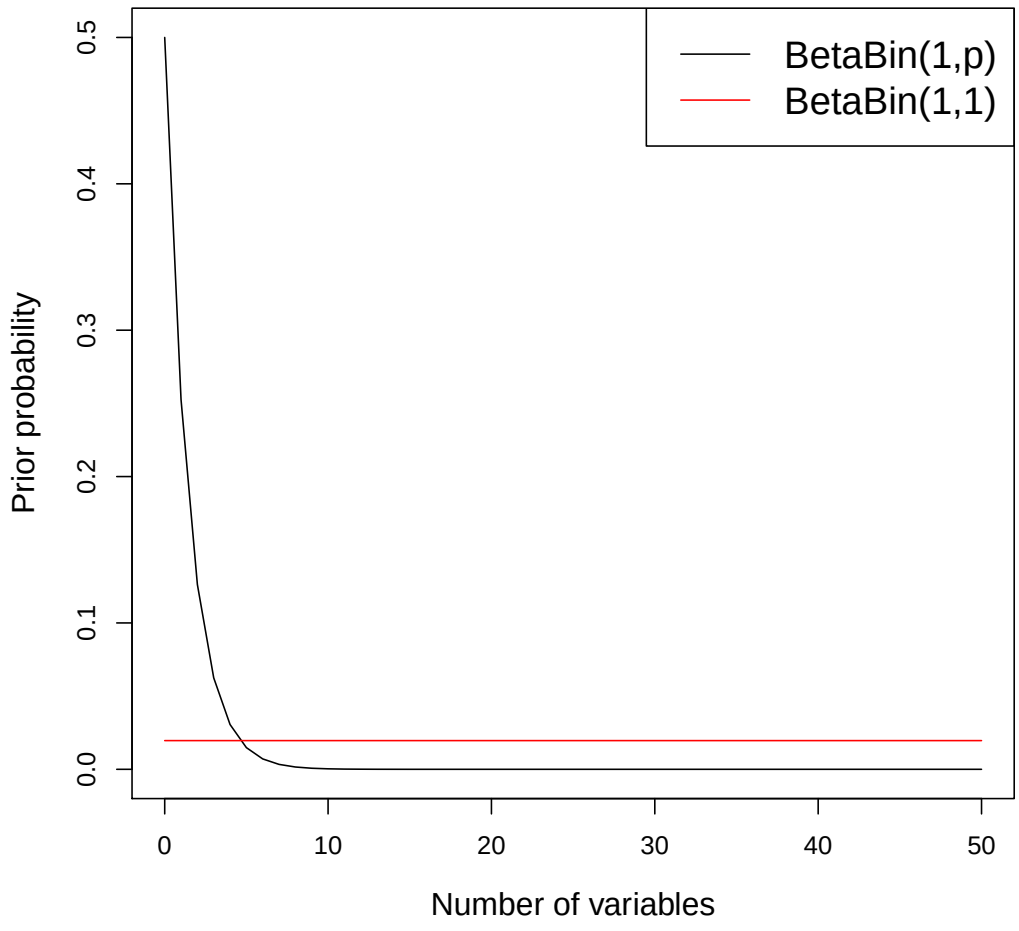
$$p(d_k) = \frac{1}{p+1} \text{ for } d_k = 0, \ldots, p$$

$$p(M_k) = \frac{1}{p+1} \binom{p}{d_k}^{-1}$$

- Alternative 3: $d_k \sim$ Poisson
- ...

Important: if p fixed and $n \to \infty$, the influence of $p(M_k)$ vanishes and $p(M_k \mid \mathbf{y}) \to 1$ for the data-generating model

When $p \gg n$ (e.g. $p = O(e^n)$) there's theory suggesting that $p(d_k)$ decrease exponentially with d_k (Castillo, van der Vaart et al 2012)



Colon cancer example ($n = 262$)

Consider first $p = 20$ variables

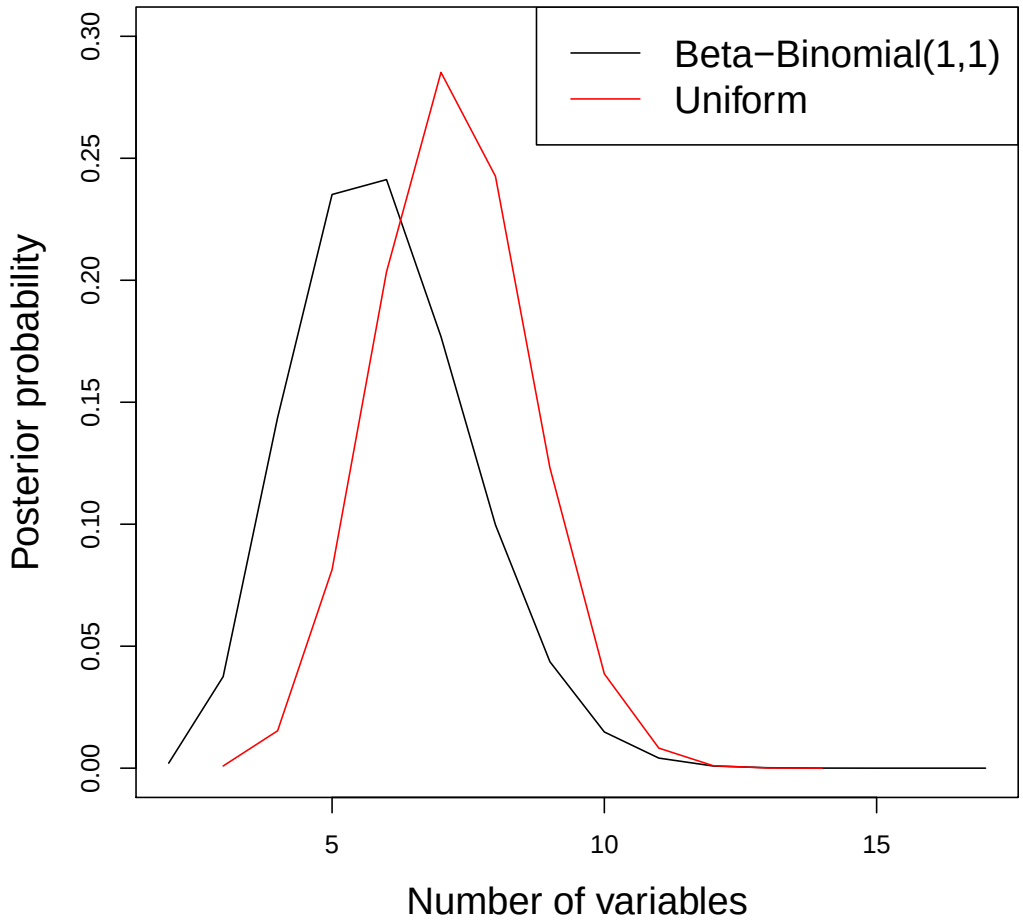
- $p(\mathbf{w}_k \mid M_k) \sim N(\mathbf{0}, nq(\mathbf{X}_k^T \mathbf{X}_k)^{-1})$ (Unit Information Prior)
- $p(M_k)$ either Uniform or Beta-Binomial(1,1)

We can enumerate all $2^p = 1,048,576$ models and compute

- $p(M_k \mid \mathbf{t})$ (Posterior model probabilities)
- $p(d_k \mid \mathbf{t})$ (Posterior distribution of model size)

Colon cancer example ($n = 262$)

Let's look at $p(d_k \mid \mathbf{t})$ (*i.e.* did we learn number of necessary variables?)



Top 5 models under Uniform $p(M_k)$

Variables	$p(M_k \mid \mathbf{t})$
3,11,15,16,17,19	0.0084
3,15,16,17,19	0.0057
1,3,11,13,15,19	0.0045
1,3,11,15,16,17,19	0.0045
3,11,13,15,19	0.0043

Top 5 models under Beta-Binomial(1,1)

Variables	$p(M_k \mid \mathbf{t})$
3,11,15,19	0.0284
3,15,16,17,19	0.0207
3,15,19	0.0205
15,16,17,19	0.0177
3,15,16,19	0.0171

- We reflect uncertainty in model choice
- Some clear candidates, others not so clear. What to do?

Final thoughts on $p(M_k)$

Throughout we assumed that variables are exchangeable, but sometimes this is not fully reasonable

- Temporal or spatial structure: environmental data, images etc.
- Networks: subsets of genes belong to networks/pathways, collaborate to perform biological function etc.
- Hierarchical structure: text data with different types of words (names/verbs/adjectives) etc.

Framework can easily accommodate this

- Let $\gamma_i = I(w_i \neq 0)$ be variable inclusion indicators
- Set dependent prior probability model for γ_i

Example: Let z_i be the group for variable i. Set $P(\gamma_i = 1) = \pi_{z_i}$ (possibly also $p(\pi_{z_i})$)

Outline

For now assume we can enumerate all 2^p models. Sometimes there is a clear winner, *e.g.* in hipparcos star data

Variables	$p(M_k \mid \mathbf{t})$
1,2	0.9904
1	0.0095
2	<1E-5
-	<1E-5

Sometimes not, *e.g.* in colon cancer data

Variables	$p(M_k \mid \mathbf{t})$
3,11,15,19	0.0284
3,15,16,17,19	0.0207
3,15,19	0.0205
15,16,17,19	0.0177
...	...

Several ways to proceed, depending on what our goals are

Explanatory models

Goal: report variables that "truly" have an effect on $\mathbf{t}$

Option 1: HPM (highest probability model(s))

Pros:

- Simple
- Takes into account correlations between predictors

Cons:

- If post prob small, unsure that this is the "right" model
- Need to enumerate all models, else not sure which is the HPM

Option 2: report variables with $P(\gamma_i = 1 \mid \mathbf{y}) > s$ for some threshold s

Pros:

- Simple
- If we cannot enumerate all models, $P(\gamma_i = 1 \mid \mathbf{t})$ often easier to estimate than $p(M_k \mid \mathbf{t})$

Cons:

- Does not take correlations between predictors into account
- Chosen model could have low $p(M_k \mid \mathbf{t})$

Issue: s should be a "large" value, but how large? We can use the *Bayesian False Discovery Rate* (FDR), but first let's see an example

Simulated example

Set $n = 50$, $p = 3$, $\mathbf{w} = (1, 0, 0)$, $q = 1$, $\text{Cov}(\mathbf{x}_i) = \begin{pmatrix} 1 & 0.99 & 0.99 \\ 0.99 & 1 & 0.99 \\ 0.99 & 0.99 & 1 \end{pmatrix}$

Results

Variables	$p(M_k \mid \mathbf{t})$
1	0.339
3	0.316
2	0.193
1,3	0.050
1,2	0.048
2,3	0.044
1,2,3	0.007
-	$< 10^{-7}$

Variable inclusion probabilities

$P(\gamma_1 = 1 \mid \mathbf{t}) = 0.446$
$P(\gamma_2 = 1 \mid \mathbf{t}) = 0.294$
$P(\gamma_3 = 1 \mid \mathbf{t}) = 0.419$

Bayesian FDR

Denote our decision by $g_i = \begin{cases} 1, \text{ if } p(\gamma_i = 1 \mid \mathbf{t}) \geq s \\ 0, \text{ if } p(\gamma_i = 1 \mid \mathbf{t}) < s \end{cases}$

We can summarize our decisions as

	Truth	
	$\gamma_i = 0$	$\gamma_i = 1$
$g_i = 0$	TN	FN
$g_i = 1$	FP	TP

False discovery proportion (FDP)

$$\text{FDP} = \frac{FP}{FP + TP} = \frac{\sum_{i=1}^{p} g_i(1 - \gamma_i)}{\sum_{i=1}^{p} g_i}$$

Bayesian FDR

For any given s, only γ_i is random

$$E(\text{FDP} \mid \mathbf{t}) = \frac{\sum_{i=1}^{p} g_i E(1-\gamma_i \mid \mathbf{t})}{\sum_{i=1}^{p} g_i} = \frac{\sum_{i=1}^{p} g_i P(\gamma_i = 0 \mid \mathbf{t})}{\sum_{i=1}^{p} g_i}$$

That is, Bayesian FDR= mean $P(\gamma_i = 0 \mid \mathbf{t})$ over included variables

Recipe to control Bayesian FDR$< \alpha$

1. Order $P(\gamma_i = 1 \mid \mathbf{t})$ in decreasing order
2. Keep including variables until $E(\text{FDP} \mid \mathbf{t}) > \alpha$

Example: colon cancer data

Recall that

Variables	$p(M_k \mid \mathbf{t})$
3,11,15,19	0.0284
3,15,16,17,19	0.0207
3,15,19	0.0205
15,16,17,19	0.0177
...	...

Order variables according to marginal inclusion probability

Variable	$P(\gamma_i = 1 \mid \mathbf{t})$	$E(\text{FDP} \mid \mathbf{y})$
19	0.980	0.020
15	0.938	0.041
3	0.709	0.124
11	0.531	0.210
...	...	...

Predictive models

Goal: predict t_{n+1} or estimate $\mathbf{w}$ as accurately as possible

Given our posterior $P(\mathbf{w} \mid \mathbf{t})$, L_2 error minimized by BMA

$$E(w_i \mid \mathbf{t}) = \sum_{k=1}^{K} E(w_i \mid M_k, \mathbf{t}) p(M_k \mid \mathbf{t})$$

$$E(t_{n+1} \mid \mathbf{t}) = \sum_{i=1}^{p} E(w_i \mid \mathbf{t}) x_{n+1,i}$$

- In principle, this is our best estimate
- Includes all variables, but some have negligible effect

BMA shrinkage example

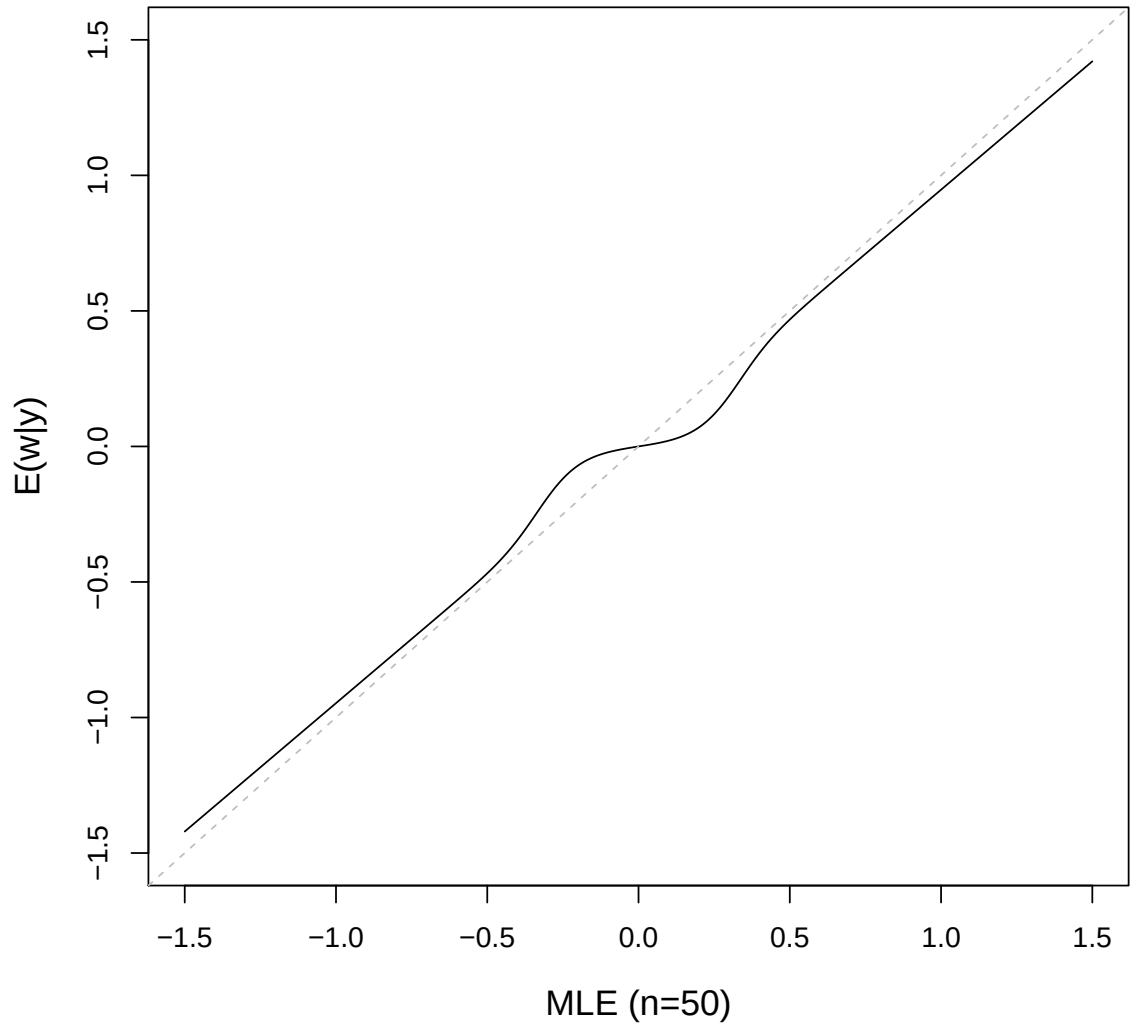
Consider the vanilla setting

- $M_1 : t_i \sim N(0, 1)$
- $M_2 : t_i \sim N(w, 1)$, $w \sim N(0, 1)$
- $p(M_1) = p(M_2) = 0.5$

Trivially, $E(w \mid \mathbf{t}) = E(w \mid M_2, \mathbf{t}) p(M_2 \mid \mathbf{t})$

- $E(w \mid M_2, \mathbf{t})$ linear shrinkage estimator (Ridge regression)
- $p(M_1 \mid \mathbf{t})$ non-linear shrinkage

Here $E(w \mid \mathbf{t})$ is a function of the least squares $\hat{w} = \sum_{i=1}^{n} t_i / n$



Predictive models

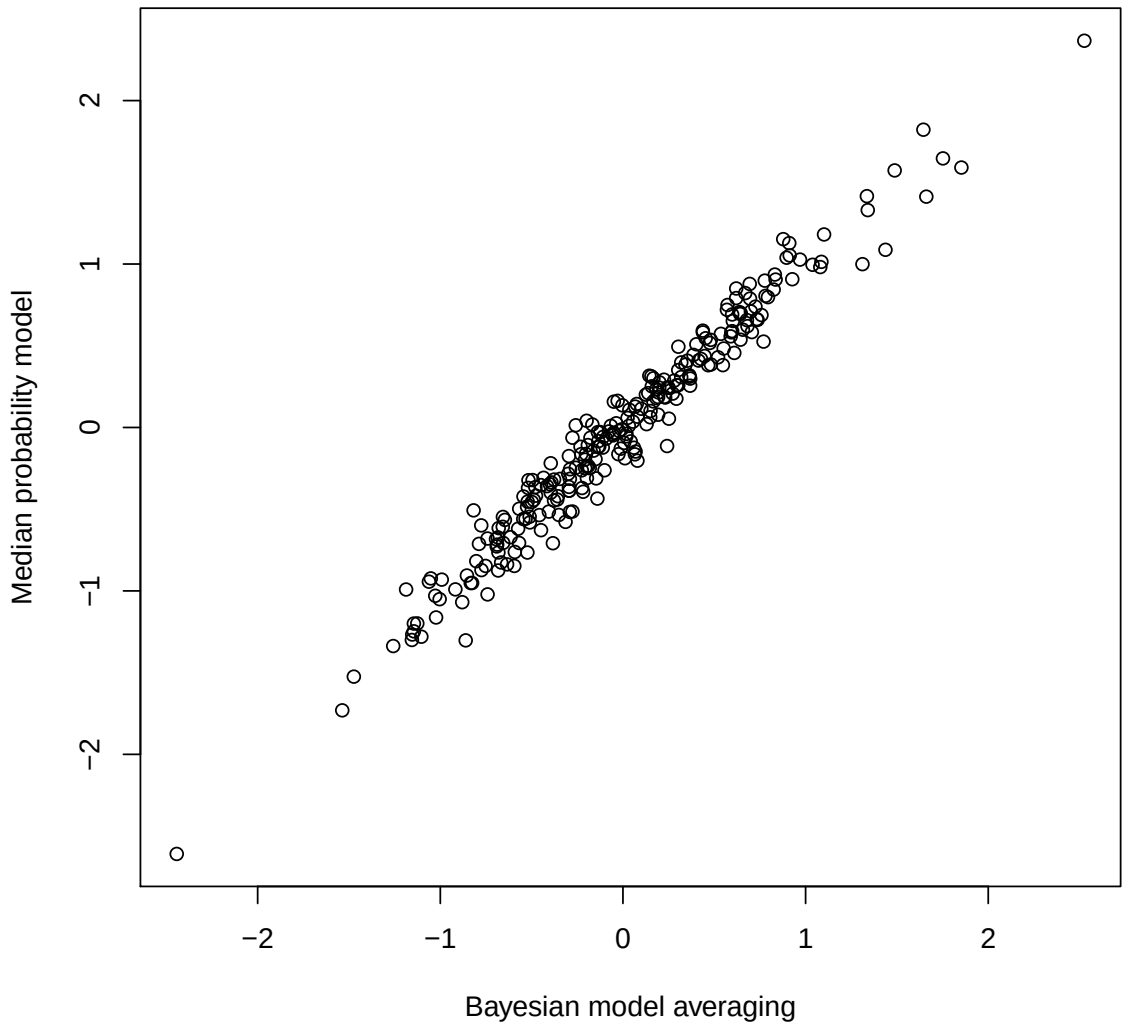
BMA may be impractical (requires all variables). Alternative: approximate BMA using a subset of the $\mathbf{x}$'s

- Option 1. Median probability model. Select variables with $P(\gamma_i = 1 \mid \mathbf{t}) > 0.5$, in some situations this approximates BMA estimate (Barbieri & Berger, 2004)
  - Simple, fast
  - May run into trouble when $\mathbf{x}$'s highly correlated
- Option 2. Enumerate all models and choose that yielding closest predictions to BMA (on average)
  - Requires enumerating all models
  - Need to define "closest" (L_2, Kullback-Leibler etc.)
  - Average with respect to what? (model-based, cross-validation...)
  - No problems with correlated $\mathbf{x}$'s

Interesting open research problem!

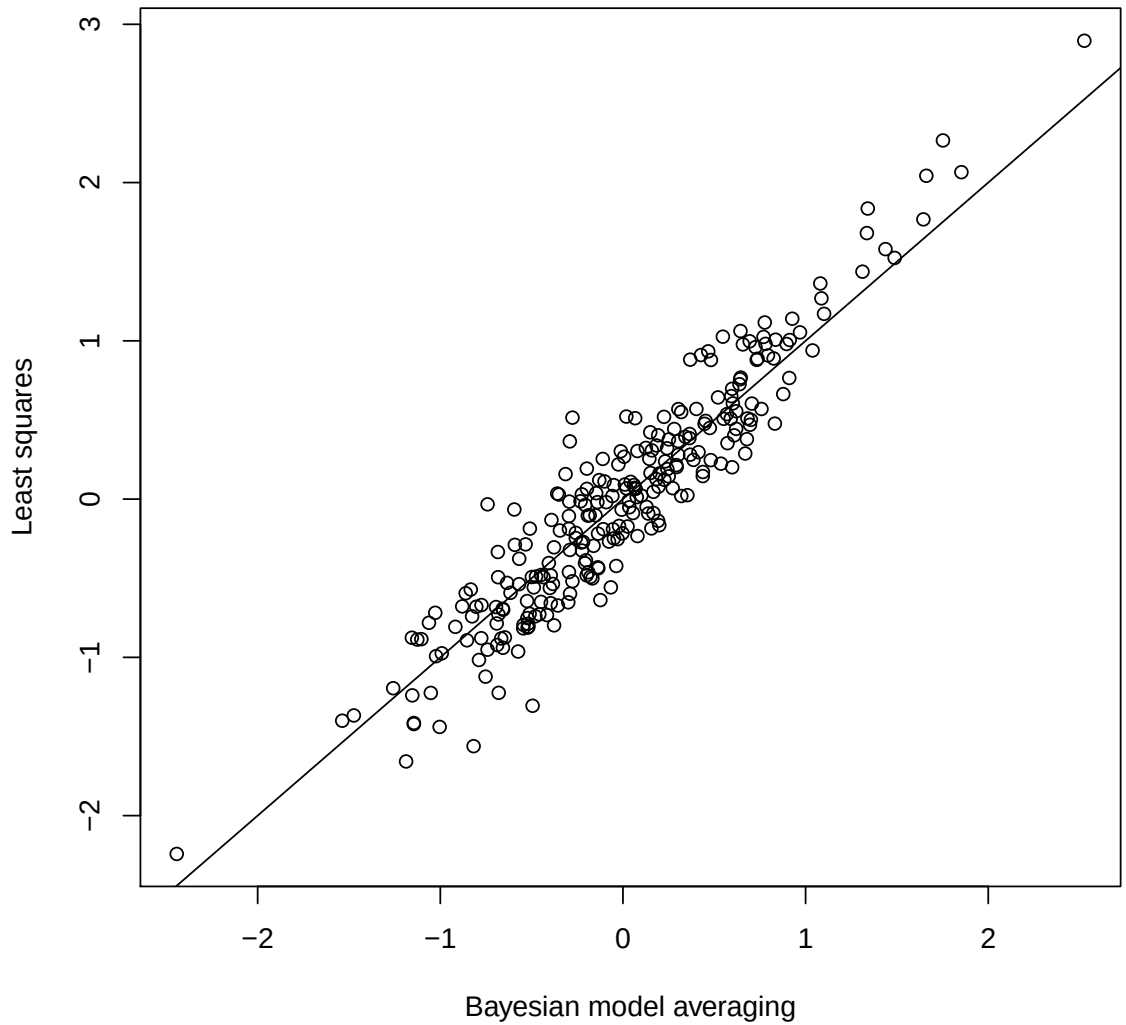
Example: colon cancer data ($n = 262$)

Compare predictions from BMA vs. median probability model



Example (continued)

BMA vs. least squares from full model



Example (continued)

If we only cared about prediction, was it worth the effort? Leave-one-out cross-validated R^2 between t_i and $\hat{t}_i$

	R^2
BMA	0.39
Full model	0.46

Now consider $p = 172$ variables (remember we have $\approx 20,000$)

	R^2
BMA	0.56
Full model	0.37

BMA introduces bias to reduce variance

- It can hurt us if p small
- Essential if p large