Variable Selection

Jacob M. Montgomery

2017

Variable/feature selection

The problem of variable selection

- With p variables, there are 2^p possible models

The problem of variable selection

- With p variables, there are 2^p possible models
- How can we choose the "right" one?

The problem of variable selection

- With p variables, there are 2^p possible models
- How can we choose the "right" one?
- Various approaches include:

The problem of variable selection

- With p variables, there are 2^p possible models
- How can we choose the "right" one?
- Various approaches include:
  1. Use theory
  2. Use model fit
  3. Regularization (e.g., ridge)
  4. Don't
- Do *not* use stepwize regression (or any proxy)

Variable selection is even harder than you think

- Variables that are highly correlated are hard to think about
- Computationally "hard" in that the problem is generally NP-complete
- What if you have more variables than observations?
- What if missingness varies by variable?

Model fit: Things to try and avoid

- Backward Elimination
  - Start with all variables
  - Remove predictors with high p-values
  - Repeat until everything is significant

Model fit: Things to try and avoid

- Backward Elimination
  - Start with all variables
  - Remove predictors with high p-values
  - Repeat until everything is significant

- Forward selection
  - Start with no variables
  - Add all variables one at a time and choose the one that is most significant
  - Continue until no variables can be added

Model fit: Things to try and avoid

- Backward Elimination
  - Start with all variables
  - Remove predictors with high p-values
  - Repeat until everything is significant
- Forward selection
  - Start with no variables
  - Add all variables one at a time and choose the one that is most significant
  - Continue until no variables can be added
- Stepwise
  - Some combination of these two
  - Often done in case a variable was "lost" at early stages but would be significant later.

Why this is bad

- Can easily miss the optimal model since you are not considering anything like the entire model space.
- **p-values don't mean what you think they mean**
- Tends to lead to models that over-estimate the significance of the variables that remain
- Just. Don't. Do. It.

A little better: Criteria-based

- Choose some model fit criteria (e.g., AIC, BIC, adjusted R-squared, etc.) and choose the model that fits best.
- Risks over-fitting, but you can hope the complexity penalty will help.

A little better: Criteria-based

- Choose some model fit criteria (e.g., AIC, BIC, adjusted R-squared, etc.) and choose the model that fits best.
- Risks over-fitting, but you can hope the complexity penalty will help.
- Different "search" algorithms have been proposed for only considering "likely" models when all 2^p are infeasible.

Problems and recommendations

Problems:

- Overfitting is very likely, and in any case finding a model that fits the data we have in hand is not the goal.
- You can get widely different answers based on choices of criteria

Problems and recommendations

Problems:

- Overfitting is very likely, and in any case finding a model that fits the data we have in hand is not the goal.
- You can get widely different answers based on choices of criteria

Recommendations:

- Theory (and literature) first.
- Can use automated criteria methods, but only as a first step.
- If your subsantive results change, you are in a danger zone.
- Radically different implications from seemingly inocuous changes in control variables mean you have serious post-treatment/colinearity problems.

Regularization 1: Lasso and Ridge

Regularization 2: Elastic Net

- Lasso can handle at most n variables.
- It tended to choose only one of any set of related variables.
- It is dominated by Ridge in $n > p$ situations where there are many correlated predictors.

Regularization 2: Elastic Net

- Lasso can handle at most n variables.
- It tended to choose only one of any set of related variables.
- It is dominated by Ridge in $n > p$ situations where there are many correlated predictors.
- Solution is to combine ridge and Lasso

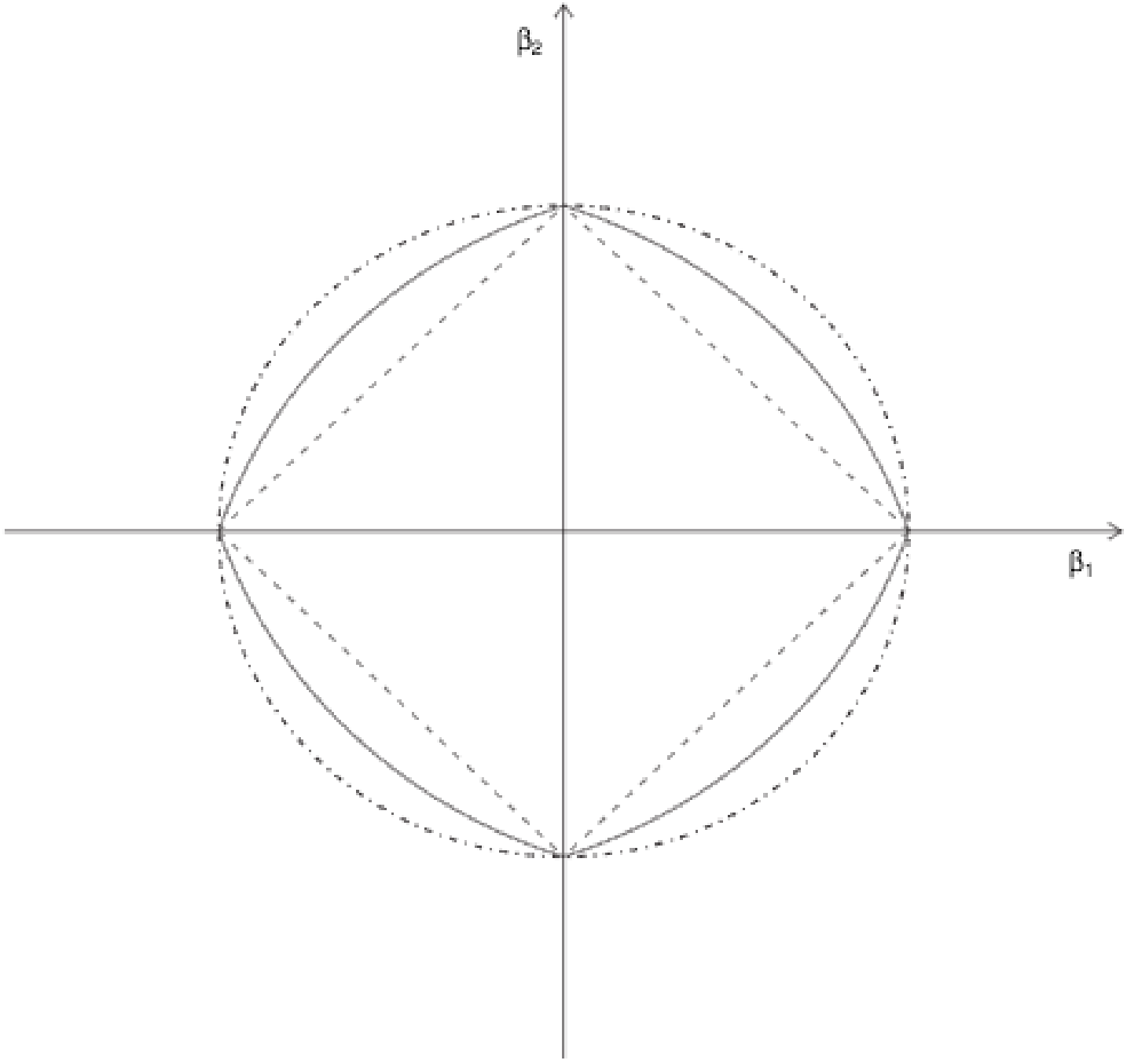
$$\text{argmin} \left\{ \sum (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \sum_{j=1}^{p} \lambda_2 |\beta_j|^2 + \sum_{j=1}^{p} \lambda_1 |\beta_j| \right\}$$

Regularization 2: Elastic Net

- Lasso can handle at most n variables.
- It tended to choose only one of any set of related variables.
- It is dominated by Ridge in $n > p$ situations where there are many correlated predictors.
- Solution is to combine ridge and Lasso

$$\text{argmin} \left\{ \sum_{i=1}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \sum_{j=1}^{p} \lambda_2 |\beta_j|^2 + \sum_{j=1}^{p} \lambda_1 |\beta_j| \right\}$$

- Equivalent to: $\text{argmin}|y - X\beta|^2$ where $(1-\alpha)|\beta| + \alpha|\beta|^2 \leq t$ and $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$


β_2
β_1

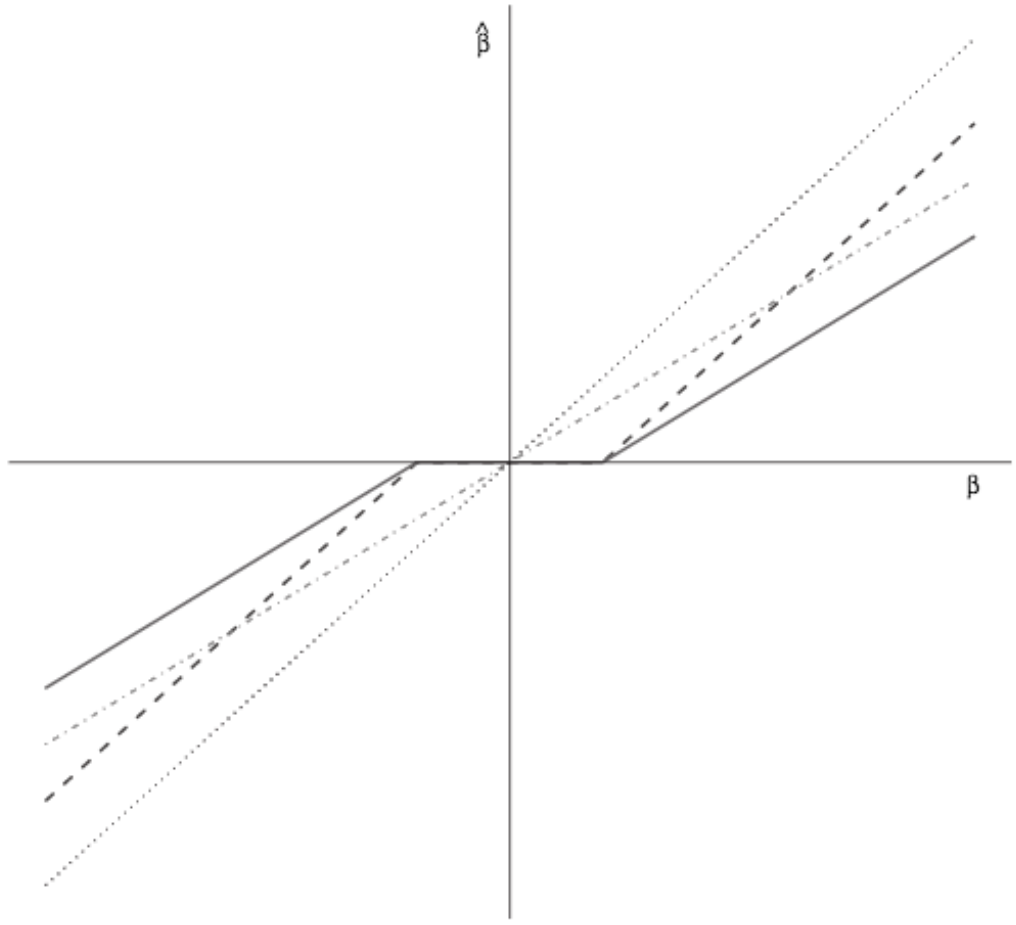


Fig. 2. Exact solutions for the lasso (-------), ridge regression (· - · - · -) and the naïve elastic net (———) in an orthogonal design (·····, OLS): the shrinkage parameters are $\lambda_1 = 2$ and $\lambda_2 = 1$

Additional notes

- The non-naive estimates work to get rid of the double shrinkage by gettig rid of the ridge shrinkage:

$$\hat{\beta}^{EN} = (1 + \lambda_2)\hat{\beta}^{Naive}$$

Additional notes

- The non-naive estimates work to get rid of the double shrinkage by gettig rid of the ridge shrinkage:

$$\hat{\beta}^{EN} = (1 + \lambda_2)\hat{\beta}^{Naive}$$

- Algorithm is a version of LARS (Least Angle Regression) and works as:
  1. Start with all coefficients at zero
  2. Find the predictor most correlated with the outcome, and increase coefficient estimate until some other variable more correlated with the residuals.
  3. Proceeds in a direction "equiangular" between the two until a third added.
  4. Repeat.
- Thus, in a weird way is related to the the other procedures

Additional notes

- The non-naive estimates work to get rid of the double shrinkage by gettig rid of the ridge shrinkage:

$$\hat{\beta}^{EN} = (1 + \lambda_2)\hat{\beta}^{Naive}$$

- Algorithm is a version of LARS (Least Angle Regression) and works as:
  1. Start with all coefficients at zero
  2. Find the predictor most correlated with the outcome, and increase coefficient estimate until some other variable more correlated with the residuals.
  3. Proceeds in a direction "equiangular" between the two until a third added.
  4. Repeat.
- Thus, in a weird way is related to the the other procedures
- Can use the same package as just discusssed.

Don't select variables: Bayesian Model Averaging

Fearon and Laitin (2003) wish to model civil conflict

Constant	-6.731	-7.019
	(0.736)	(0.751)
Prior war	-0.954	-0.916
	(0.314)	(0.312)
Per capita income	-0.344	-0.318
	(0.072)	(0.071)
log (population)	0.263	0.272
	(0.073)	(0.074)
log (% mountainous)	0.219	0.199
	(0.085)	(0.085)
Noncontiguous state	0.443	0.426
	(0.274)	(0.272)
Oil exporter	0.858	0.751
	(0.279)	(0.278)
New state	1.709	1.658
	(0.339)	(0.342)
Instability	0.618	0.513
	(0.235)	(0.242)
Democracy (Polity IV)	0.021	
	(0.017)	
Ethnic fractionalization	0.166	0.164
	(0.373)	(0.368)
Religious fractionalization	0.285	0.326
	(0.509)	(0.506)
Anocracy		0.521
		(0.237)
Democracy		0.127
		(0.304)
Wars in neighboring countries		

"When we add dummy variables for countries that have an ethnic or religious majority and a minority of at least 8% of the country's poulation, both are incorrectly signed and neither comes close to statistical significance. This finding does not depend on which other variables are included in the model."

> *"When we add dummy variables for countries that have an ethnic or religious majority and a minority of at least 8% of the country's poulation, both are incorrectly signed and neither comes close to statistical significance. This finding does not depend on which other variables are included in the model."*

We often want to show that our results are "robust" to modeling choices:

- Online appendix includes 18 additional tables
- At least 74 possible explnatory variables discussed
- $2^{74} = 2 \times 10^{22} = 20$ sextillion

The basics

The model:

- Let $\mathbf{X}$ denote the $n \times p$ matrix of all the independent variables theorized to be predictors of outcome Y.
- $Y = \mathbf{X}\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2\mathbf{I})$.
- $\mathcal{M} = [\mathcal{M}_1, \mathcal{M}_2, \ldots \mathcal{M}_q]$ where $q = 2^p$.

The basics

The model:

- Let $\mathbf{X}$ denote the $n \times p$ matrix of all the independent variables theorized to be predictors of outcome Y.
- $Y = \mathbf{X}\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2\mathbf{I})$.
- $\mathcal{M} = [\mathcal{M}_1, \mathcal{M}_2, \ldots \mathcal{M}_q]$ where $q = 2^p$.

Priors:

- $\pi(\beta_k|M_k)$
- $\pi(\sigma|M_k)$.
- We also put a prior probability $\pi(M_k)$ on the model space.

BMA as an Hierarchical Mixture model (1)

$$\mathcal{M}_k \sim \pi(\mathcal{M}_k). \tag{1}$$

BMA as an Hierarchical Mixture model (1)

$$\mathcal{M}_k \sim \pi(\mathcal{M}_k). \tag{1}$$

$$\sigma|\mathcal{M}_k \sim \pi(\sigma|\mathcal{M}_k) \tag{2}$$

BMA as an Hierarchical Mixture model (1)

$$\mathcal{M}_k \sim \pi(\mathcal{M}_k). \tag{1}$$

$$\sigma|\mathcal{M}_k \sim \pi(\sigma|\mathcal{M}_k) \tag{2}$$

$$\beta_\omega|\mathcal{M}_k \sim \pi(\beta_\omega|\mathcal{M}_k, \sigma^2). \tag{3}$$

where $\Omega = \omega_1, \ldots, \omega_p$ represents a vector of zeroes and ones indicating the inclusion (or exclusion) of variables in model M_k.

BMA as an Hierarchical Mixture model (1)

$$\mathcal{M}_k \sim \pi(\mathcal{M}_k). \tag{1}$$

$$\sigma|\mathcal{M}_k \sim \pi(\sigma|\mathcal{M}_k) \tag{2}$$

$$\beta_\omega|\mathcal{M}_k \sim \pi(\beta_\omega|\mathcal{M}_k, \sigma^2). \tag{3}$$

where $\Omega = \omega_1, \ldots, \omega_p$ represents a vector of zeroes and ones indicating the inclusion (or exclusion) of variables in model M_k.

$$Y|\beta, \sigma, \mathcal{M}_k \sim N(X_\omega \beta_\omega, \sigma^2 \mathbf{I}) \tag{4}$$

BMA as an Hierarchical Mixture model (2)

$$E(\beta_k|Y) = \sum_{k=0}^{q} p(\mathcal{M}_k|Y)E(\beta_k|\mathcal{M}_k, Y) \tag{5}$$

BMA as an Hierarchical Mixture model (2)

$$E(\beta_k|Y) = \sum_{k=0}^{q} p(\mathcal{M}_k|Y)E(\beta_k|\mathcal{M}_k, Y) \tag{5}$$

$$p(\beta_k|Y) = \sum_{k=0}^{q} p(\mathcal{M}_k|Y)p(\beta_k|\mathcal{M}_k, Y) \tag{6}$$

BMA as an Hierarchical Mixture model (2)

$$E(\beta_k|Y) = \sum_{k=0}^{q} p(\mathcal{M}_k|Y)E(\beta_k|\mathcal{M}_k, Y) \tag{5}$$

$$p(\beta_k|Y) = \sum_{k=0}^{q} p(\mathcal{M}_k|Y)p(\beta_k|\mathcal{M}_k, Y) \tag{6}$$

$$p(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_{k=0}^{q} p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)} \tag{7}$$

BMA as an Hierarchical Mixture model (2)

$$E(\beta_k|Y) = \sum_{k=0}^{q} p(\mathcal{M}_k|Y)E(\beta_k|\mathcal{M}_k, Y) \tag{5}$$

$$p(\beta_k|Y) = \sum_{k=0}^{q} p(\mathcal{M}_k|Y)p(\beta_k|\mathcal{M}_k, Y) \tag{6}$$

$$p(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_{k=0}^{q} p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)} \tag{7}$$

$$p(Y|\mathcal{M}_k) = \int\int p(Y|\beta_\omega, \sigma^2, \mathcal{M}_k)\pi(\beta_\omega|\sigma^2, \mathcal{M}_k)\pi(\sigma^2|\mathcal{M}_k)\, d\beta_\omega\, d\sigma^2 \tag{8}$$

Bayesian model averaging

- Y is a $n \times 1$ vector of outcomes
- X is an $n \times p$ matrix
- $Y = X\beta + \epsilon$
- $\epsilon \sim N(0, \sigma^2 I)$
- $q = 2^p$
- $\mathcal{M} = [\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_q]$

Priors and likelihood:

- $\mathcal{M}_k \sim \pi(\mathcal{M}_k)$
- $\sigma^2|\mathcal{M}_k \sim \pi(\sigma^2|\mathcal{M}_k)$
- $\beta_\omega|\sigma^2, \mathcal{M}_k \sim \pi(\beta_\omega|\sigma^2, \mathcal{M}_k)$
- $\Omega = [\omega_1, \ldots, \omega_p]$ is a binary vector indicating inclusion.

Priors and likelihood:

- $\mathcal{M}_k \sim \pi(\mathcal{M}_k)$
- $\sigma^2|\mathcal{M}_k \sim \pi(\sigma^2|\mathcal{M}_k)$
- $\beta_\omega|\sigma^2, \mathcal{M}_k \sim \pi(\beta_\omega|\sigma^2, \mathcal{M}_k)$
- $\Omega = [\omega_1, \ldots, \omega_p]$ is a binary vector indicating inclusion.

$$p(Y|\beta_\omega, \sigma^2, \mathcal{M}_k) \sim N(X_\omega \beta_\omega, \sigma^2 I)$$

Priors and likelihood:

- $\mathcal{M}_k \sim \pi(\mathcal{M}_k)$
- $\sigma^2|\mathcal{M}_k \sim \pi(\sigma^2|\mathcal{M}_k)$
- $\beta_\omega|\sigma^2, \mathcal{M}_k \sim \pi(\beta_\omega|\sigma^2, \mathcal{M}_k)$
- $\Omega = [\omega_1, \ldots, \omega_p]$ is a binary vector indicating inclusion.

$$p(Y|\beta_\omega, \sigma^2, \mathcal{M}_k) \sim N(X_\omega \beta_\omega, \sigma^2 I)$$

$$p(Y|\mathcal{M}_k) \sim \int\int p(Y|\beta_\omega, \sigma^2, \mathcal{M}_k)\pi(\beta_\omega|\sigma^2, \mathcal{M}_k)\pi(\sigma^2|\mathcal{M}_k)d\beta_\omega d\sigma^2$$

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}$$

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}$$

With this, we can easily create other quantities of interest as weighted sums. For example,

$$E(\beta_k|Y) = \sum_{k=0}^{q} P(\mathcal{M}_k|Y)E(\beta|M_k, Y)$$

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}$$

$$P(\mathcal{M}_k|Y) = \frac{p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_k p(Y|\mathcal{M}_k)\pi(\mathcal{M}_k)}$$

With this, we can easily create other quantities of interest as weighted sums. For example,

$$E(\beta_k|Y) = \sum_{k=0}^{q} P(\mathcal{M}_k|Y)E(\beta|M_k, Y)$$

Posterior computation

The great advantage of these priors is that we can quickly compute Bayes factors *exactly*.

- This makes it possible to "sample" all possible models.
- Two options are Bayesian adaptive sampling (BAS) and leaps and bounds Algorithm.
- The paper includes a detailed discussion of available software.

An aside on Bayes factors

Let's say there are just two models, and want to know the posterior probability of some model k:

$$P(M_k|\mathbf{D}) = \frac{P(\mathbf{D}|M_k)\pi(M_k)}{P(\mathbf{D}|M_1)\pi(M_1) + P(\mathbf{D}|M_2)\pi(M_2)}$$

$$P(M_k|\mathbf{D}) = \frac{P(\mathbf{D}|M_k)\pi(M_k)}{P(\mathbf{D}|M_1)\pi(M_1) + P(\mathbf{D}|M_2)\pi(M_2)}$$

$$P(M_k|\mathbf{D}) = \frac{P(\mathbf{D}|M_k)\pi(M_k)}{P(\mathbf{D}|M_1)\pi(M_1) + P(\mathbf{D}|M_2)\pi(M_2)}$$

This implies that:

$$\frac{P(M_1|\mathbf{D})}{P(M_2|\mathbf{D})} = \frac{P(\mathbf{D}|M_1)\pi(M_1)}{P(\mathbf{D}|M_2)\pi(M_2)}$$

The limiting factor in BMA approaches has always been a combination of:

- an *extremely* high dimensional space
- intractable integrals.

Early work used "approximations" of the Bayes factors as a solution:

- $BF(M_k|M_0) = BIC_k = -2\log(L_k - L_0) + p_k \log n$
- $BF(M_K|M_0) = AIC_k = -2\log(L_k - L_0) + 2p$

Clyde (2003) and Clyde and George(2004) summarize a more comprehensive approach.

$$\pi(\mathcal{M}_k) = \gamma^{p_\omega}(1-\gamma)^{p-p_\omega} \tag{9}$$

Clyde (2003) and Clyde and George(2004) summarize a more comprehensive approach.

$$\pi(\mathcal{M}_k) = \gamma^{p_\omega}(1-\gamma)^{p-p_\omega} \tag{9}$$

Priors for posterior calculations

Prior	*Formulation*
g-prior	$\pi(\beta_\omega \mid \mathcal{M}_k, \sigma^2) \sim N_{p_\omega}(0, g\sigma^2(X'_\omega X_\omega)^{-1})$
	$\pi(\beta_0, \sigma^2 \mid \mathcal{M}_k) \propto 1/\sigma^2$

Hyper-priors for g

Prior	*Formulation*
Hyper-g	$\pi(g) = \frac{a-2}{2}(1+g)^{\frac{a}{2}} \text{ if} g > 0$
ZS	$\pi(g) = \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-\frac{n}{2g}}$

Example: g-prior

$$p(\mathcal{M}_k|Y) = \frac{B[\mathcal{M}_k : \mathcal{M}_0]}{\sum_{k=0}^{q} B[\mathcal{M}_k : \mathcal{M}_0]} \tag{10}$$

Where:

$$B[\mathcal{M}_k : \mathcal{M}_0] = (1+g)^{(n-p_k-1)/2} \times (1+g(1-R_k^2))^{-(n-1)/2}$$

For coefficient j:

$$E(\beta_j|Y) = \sum_{k=0}^{q} p(\mathcal{M}_k|Y)E(\beta_j|\mathcal{M}_k, Y) \tag{11}$$

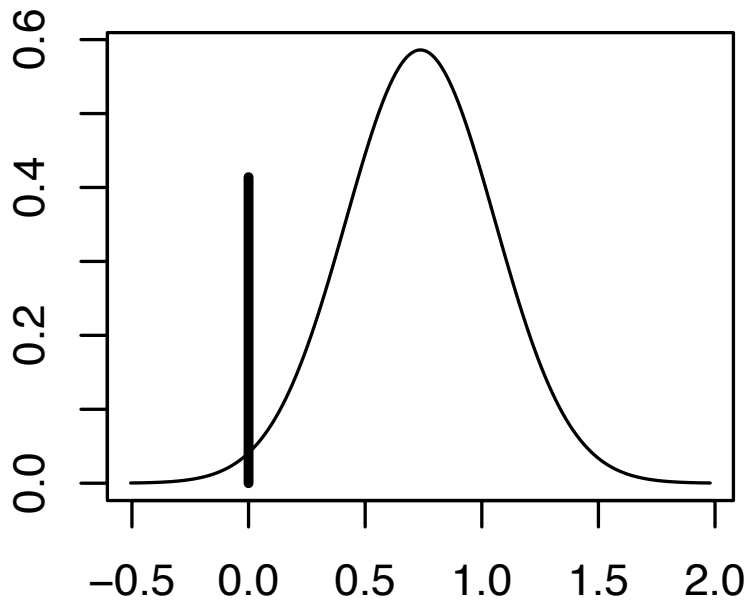
Where:

$$E(\beta_j|\mathcal{M}_k) = \frac{g}{g+1}\hat{\beta}_{j,OLS}$$

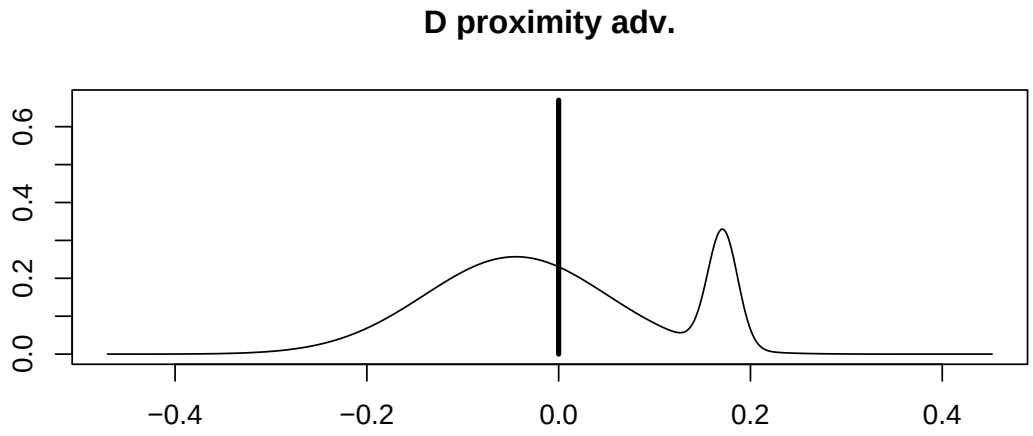
Posterior coefficient plots

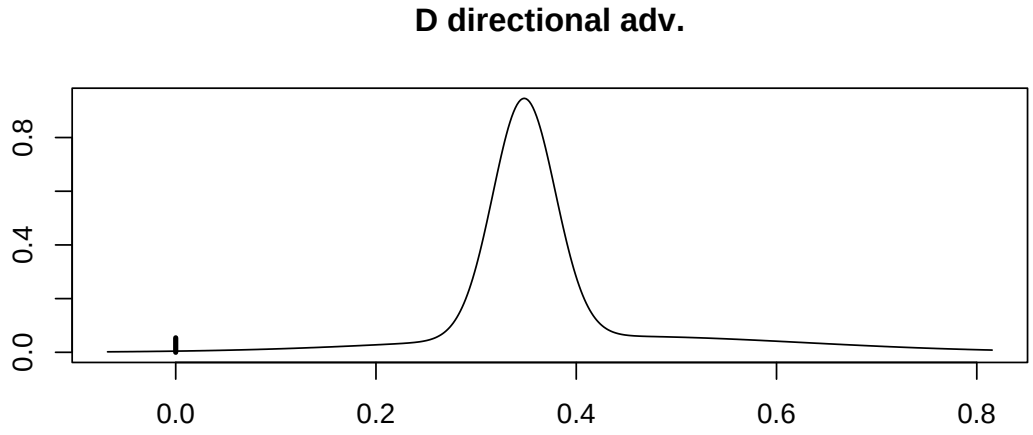
Allows us to focus on two questions separately:

- Does the variable contribute to the model's explanatory power?
- Is the variable correlated with unexplained variance when it is included?



Beware the t-test





Adams, Bishin, and Dow (ABD)

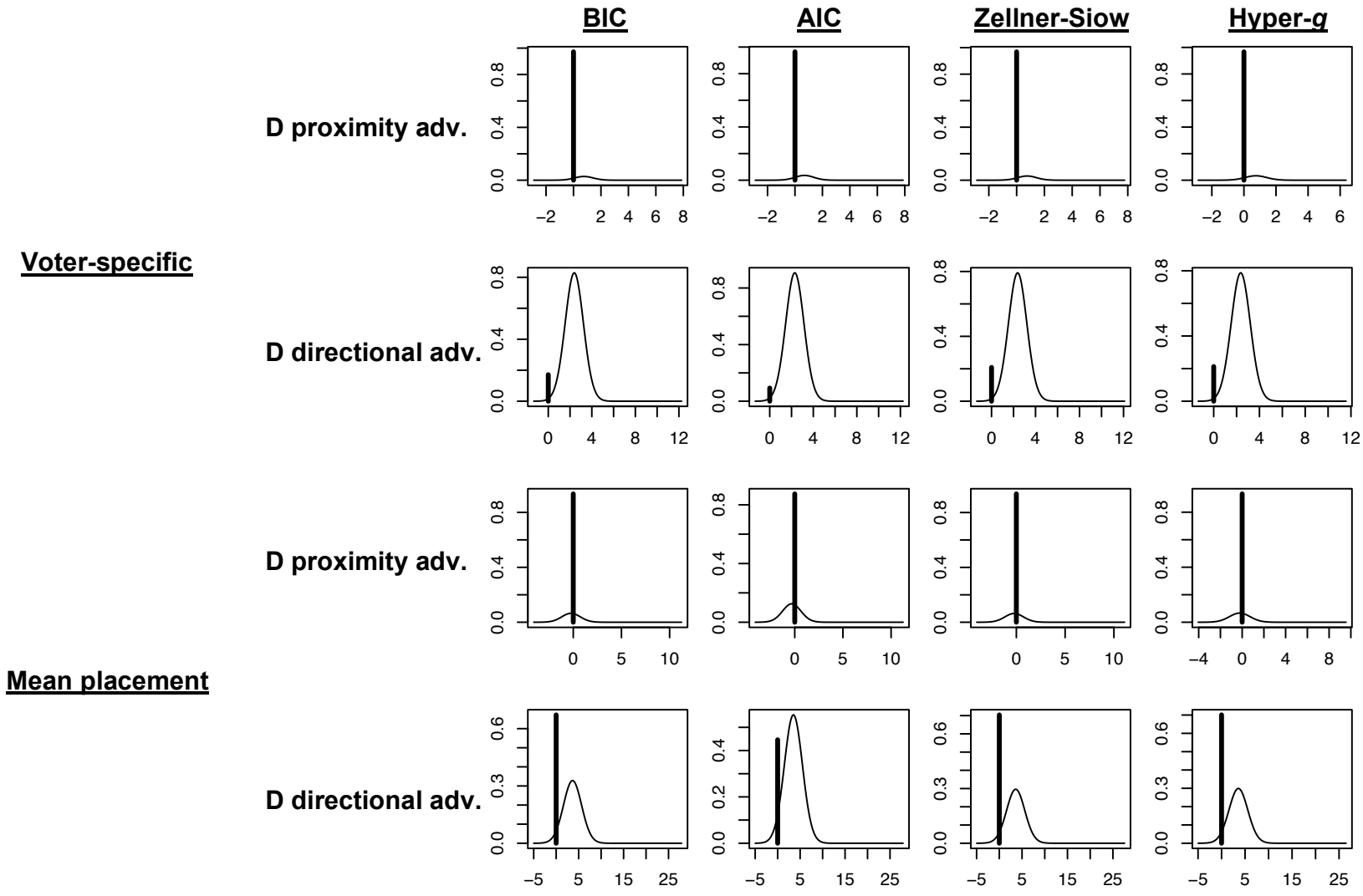
Adams, Bishin, and Dow (2004) compare directional to proximity voting in the context of U.S. Senate elections from 1988-1992. The construct an OLS analysis at the campaign level.

- DV: % of 2-party vote won by Democrat
- Main explanatory variables: Aggregate measures of the advantage of the Democratic candidate under the proximity and directional modeling assumptions.
- Controls: Partisan registration advantage, spending, incumbency, and candidate quality.
- They present total of four models presented.

Adams, Bishin, and Dow (ABD)

	ABD (2004) replication			
	Reduced model		**Full model**	
	Mean cand.	Voter-specific	Mean cand.	Voter-specific
	Mean (SD)	*Mean (SD)*	*Mean (SD)*	*Mean (SD)*
D proximity adv.	-3.137	-2.881	-3.053	-2.007
	(1.593)	(1.270)	(1.315)	(1.056)
D directional adv.	11.146	5.729	7.953	4.177
	(3.376)	(1.620)	(2.854)	(1.356)
D incumbency adv.	6.376	6.609	1.060	1.139
	(1.092)	(1.054)	(1.201)	(1.189)
D quality adv.	5.972	5.035	3.117	2.378
	(1.400)	(1.384)	(1.240)	(1.216)
D spending adv.			0.270	0.265
			(0.041)	(0.040)
D partisan adv.			0.055	0.060
			(0.054)	(0.054)
Constant	54.759	52.786	53.309	52.028
	(1.325)	(0.892)	(1.155)	(0.758)
N	95	95	94	94

BMA (hyper-g)				
Mean cand.			Voter-specific	
	Cond. mean (SD)	$P(\beta \neq 0)$	*Cond. mean (SD)*	$P(\beta \neq 0)$
D proximity adv.	-0.222	0.067	0.745	0.033
	(0.954)		(0.677)	
D directional adv.	3.576	0.299	2.363	0.787
	(2.015)		(0.840)	
D incumbency adv.	1.607	0.194	1.295	0.159
	(1.242)		(1.237)	
D quality adv.	2.964	0.599	2.740	0.541
	(1.246)		(1.223)	
D spending adv.	0.3238	1.00	0.314	1.00
	(0.041)		(0.037)	
D partisan adv.	0.0749	0.201	0.0661	0.181
	(0.057)		(0.055)	
Constant	51.381	1.00	51.560	1.00
	(1.032)		(0.764)	
N	94		94	


BIC
AIC
Zellner-Siow
Hyper-g
Voter-specific
D proximity adv.
D directional adv.
D proximity adv.
Mean placement
D directional adv.