Penalized Utility Posterior Summaries

Carlos M. Carvalho (UT Austin)
P. Richard Hahn (Chicago Booth)
Rob McCulloch (Chicago Booth)
David Puelz (UT Austin)
Beatrix Jones (Massey, NZ)

Ohio State, Nov 2015

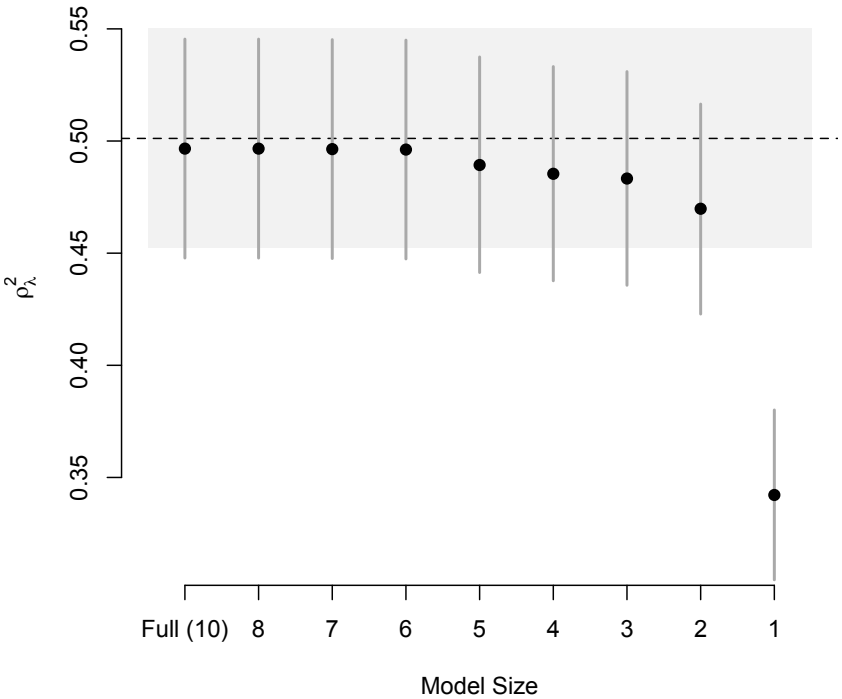
Bayes Two Step

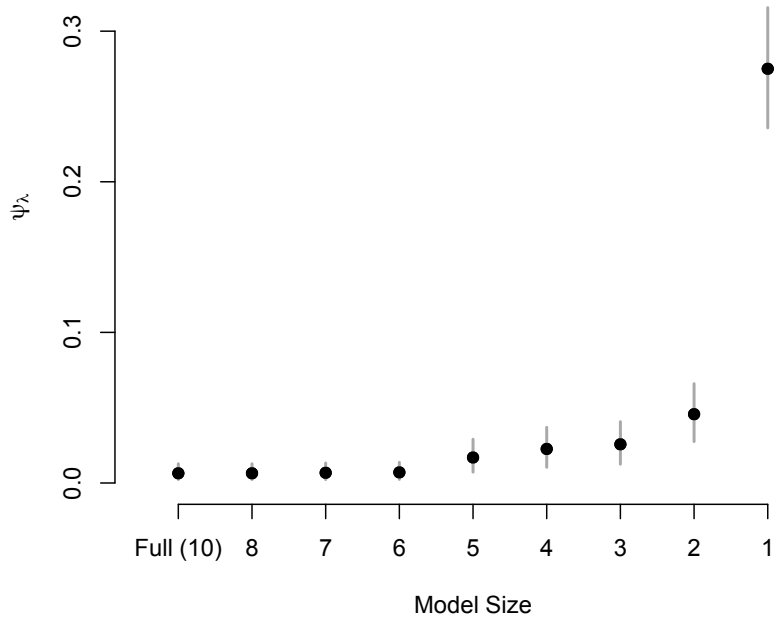
1. [Learning Step]: Choose a model and a prior... get your posterior

2. [Summary Step]: Choose a Utility. Penalize it to encourage parsimony. Optimize. Get your summary!

This sounds embarrassingly simple... and it is!

Decoupling Shrinkage and Selection
Posterior Summary for Regression Models [1]





- a posterior variable summary which distills a full posterior distribution into a *sequence of sparse predictors*
- Step 1 gives us the gray area... Step 2 provides a sequence of summaries

[1] *Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective – JASA 2015*

Linear Model Selection... What do we usually do?

What options are available for obtaining truly sparse summaries?

- Bayesians: The Median Probability Model
  (Barbieri and Berger, 2004) or the Highest Probability Model
- Everyone else: thresholding or LASSO-like methods
  (penalized likelihood selection methods)

Separating Priors from Utilities

- Our view: subset selection/summarization is a decision problem. We need a suitable loss function, not a more clever prior.

- In general, we argue that selection is associated with utility choices that balance predictive ability and parsimony. In linear models, a widely applicable loss function is defined as:

$$\mathcal{L}\left(\tilde{Y}, \gamma\right) = n^{-1}||\tilde{Y} - \mathbf{X}\gamma||_2 + \lambda||\gamma||_0$$

This is what we going to optimize in step 2 based on what we leaned about $\tilde{Y}$ in step 1.

Take Expectation under the posterior (step 1)...

$$
\begin{aligned}
\mathsf{E}_{\tilde{Y}}\mathcal{L}(\tilde{Y},\gamma) &= \mathsf{E}_{\theta}\mathsf{E}_{\tilde{Y}|\theta}\lambda||\gamma||_0 + ||\tilde{Y} - \mathbf{X}\gamma||_2, \\
&= \lambda||\gamma||_0 + \mathsf{E}_{\theta}\mathsf{E}_{\tilde{Y}|\theta}||\mathbf{X}\beta + \tilde{\epsilon} - \mathbf{X}\gamma||_2, \\
&= \lambda||\gamma||_0 + \mathsf{E}_{\theta}\mathsf{E}_{\tilde{\epsilon}|\sigma^2} \quad \tilde{\epsilon}^t\tilde{\epsilon} + 2\tilde{\epsilon}^t(\mathbf{X}\beta - \mathbf{X}\gamma) + ||\mathbf{X}\beta - \mathbf{X}\gamma||_2, \\
&= \lambda||\gamma||_0 + n\bar{\sigma}^2 + ||\mathbf{X}\bar{\beta} - \mathbf{X}\gamma||_2 + \text{constant} \\
&\propto \lambda||\gamma||_0 + ||\mathbf{X}\bar{\beta} - \mathbf{X}\gamma||_2
\end{aligned}
$$

Enter the LASSO

The DSS expected loss can be written as:

$$\mathrm{E}_{\tilde{Y}}\mathcal{L}(\tilde{Y},\gamma) = \lambda||\gamma||_0 + ||\mathbf{X}\bar{\beta} - \mathbf{X}\gamma||_2.$$

The counting penalty $||\gamma||_0$ is hard to handle computationally...as an initial approximation, we consider the 1-norm instead.

$$\mathrm{E}_{\tilde{Y}}\mathcal{L}(\tilde{Y},\gamma) = \lambda||\gamma||_1 + ||\mathbf{X}\bar{\beta} - \mathbf{X}\gamma||_2.$$

What does this look like?!!
Yes, the this is precisely the LASSO objective function using $\hat{Y}_{Bayes} = \mathbf{X}\bar{\beta}$ *in place of* Y*.*

The Two Step Recipe for Linear Models (DSS)

Step 1. Using your favorite prior for (β, σ^2) obtain the posterior and determine $\hat{Y}_{Bayes} = \mathbf{X}\bar{\beta}$.

Step 2. Feed $\hat{Y}_{Bayes}$ to `LARS` and get our optimal action, i.e. the subset of point estimates β_λ

$$\beta_\lambda \equiv \text{arg min}_\gamma \ \ \lambda||\gamma||_1 + n^{-1}||\mathbf{X}\bar{\beta} - \mathbf{X}\gamma||_2^2.$$

The "Savage/Lindley Bayesian" is done!
Everyone else has to figure out what λ should be...

Posterior Summary Plots

How much predictive deterioration is a result of sparsification? Practically, how different is β_λ from the optimal $\bar{\beta}$?

1. Variation Explained

$$\rho^2_\lambda = \frac{n^{-1}||\mathbf{X}\beta||^2}{n^{-1}||\mathbf{X}\beta||^2 + \sigma^2 + n^{-1}||\mathbf{X}\beta - \mathbf{X}\beta_\lambda||^2}$$

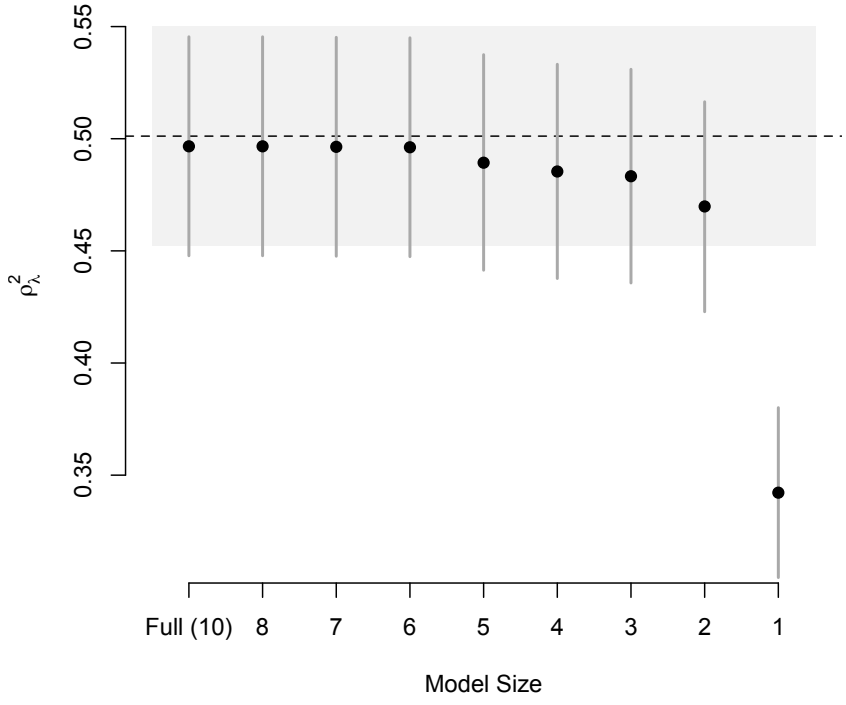
2. Excess Error

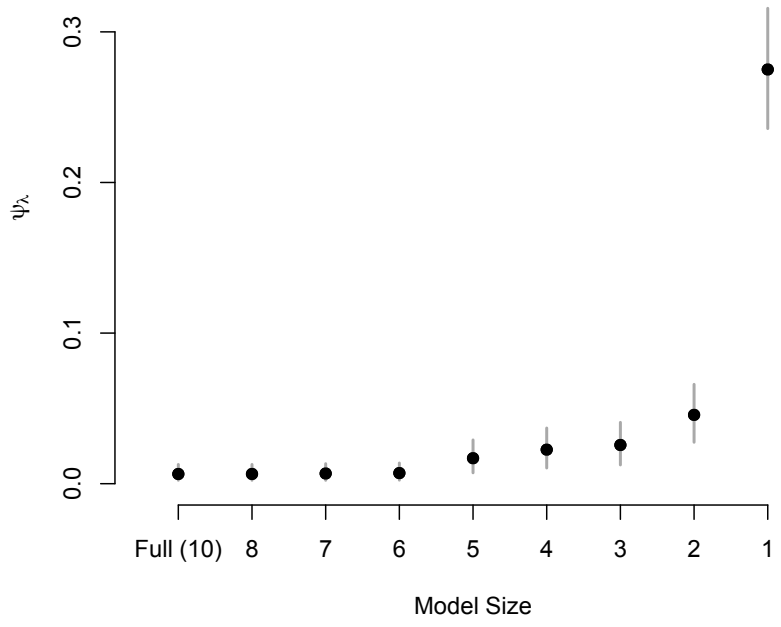
$$\psi_\lambda = \sqrt{n^{-1}||\mathbf{X}\beta_\lambda - \mathbf{X}\beta||^2 + \sigma^2} - \sigma.$$

These quantities are random variables and are evaluated using the available posterior draws.

Posterior Summary Plots

"Find the smallest model such that with probability (`blank`*) I give up less than (*`blank`*) in predictive ability."*





- The minimization of the DSS loss function via `LARS` provides a solution path for the action γ for all values of λ
- In essence, we can summarize the entire posterior, a very complex object, into a sequence of potential summaries... we just need to figure out where to stop!

DSS vs. MPM

	DSS(5)	DSS(4)	DSS(3)	DSS(2)	DSS(1)	MPM (`Robust`)
Age	–	–	–	–	–	0.08
Sex	•	–	–	–	–	**0.98**
BMI	•	•	•	•	•	**0.99**
MAP	•	•	•	–	–	**0.99**
TC	–	–	–	–	–	**0.66**
LDL	–	–	–	–	–	0.46
HDL	•	•	–	–	–	**0.51**
TCH	–	–	–	–	–	0.26
LTG	•	•	•	•	–	**0.99**
GLU	–	–	–	–	–	0.13

Similar, but not the same.

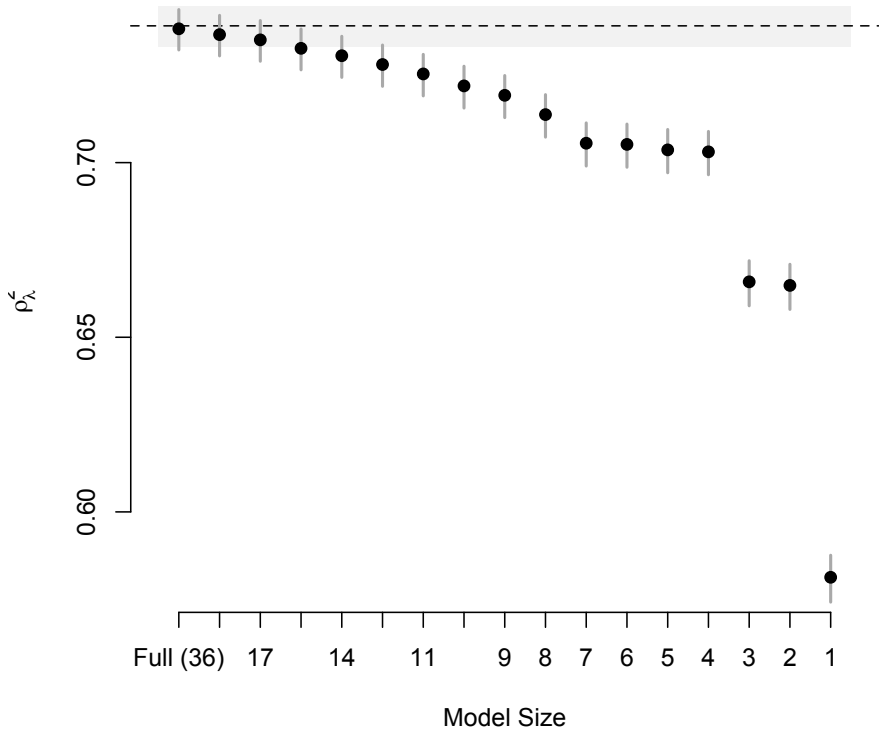
Example: Car Prices... $n = 20k$, $p \approx 100$

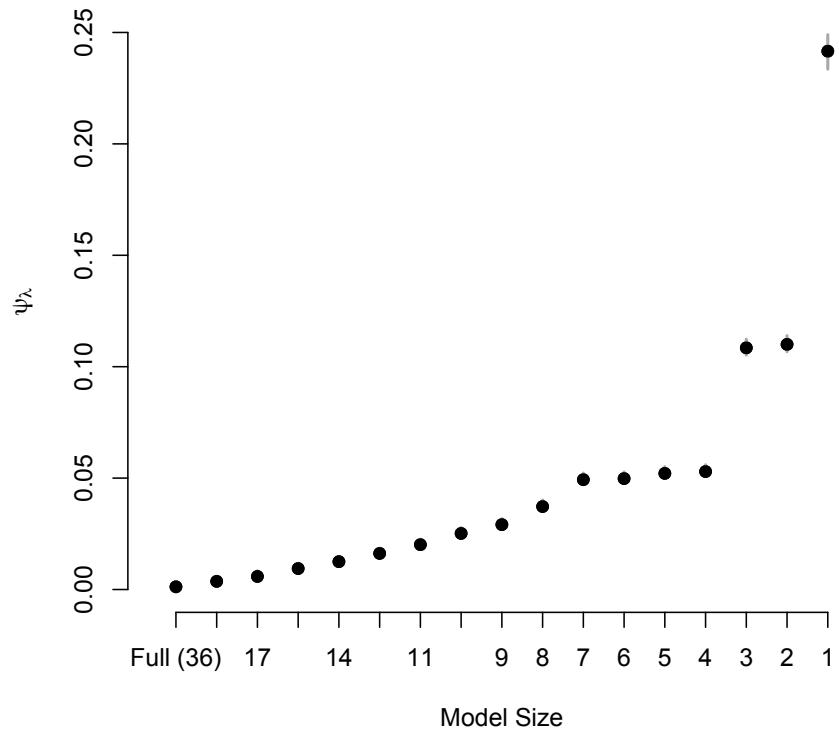
This is my motivation for sparsity... what a mess!

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -8.211e+06  8.519e+04 -96.383  < 2e-16 ***
X.trim350                2.657e+04  1.160e+04   2.290 0.022024 *
X.trim400               -1.275e+04  1.649e+04  -0.773 0.439459
X.trim420                4.954e+04  1.178e+04   4.205 2.62e-05 ***
X.trim430                2.662e+04  1.177e+04   2.263 0.023662 *
X.trim500                2.935e+04  1.177e+04   2.494 0.012623 *
X.trim550               -4.942e+03  1.078e+04  -0.458 0.646705
X.trim55 AMG             2.823e+04  1.178e+04   2.397 0.016542 *
X.trim600                4.477e+03  1.079e+04   0.415 0.678100
X.trim63 AMG             4.445e+04  1.080e+04   4.117 3.84e-05 ***
X.trim65 AMG             6.142e+03  1.083e+04   0.567 0.570524
X.trimunsp               2.666e+04  1.081e+04   2.466 0.013657 *
X.conditionNew           3.513e+04  2.284e+02 153.819  < 2e-16 ***
X.conditionUsed         -4.337e+03  1.993e+02 -21.758  < 2e-16 ***
X.isOneOwnert           -5.043e+02  1.725e+02  -2.924 0.003459 **
X.mileage               -1.324e-01  2.522e-03 -52.488  < 2e-16 ***
X.year                   4.103e+03  4.224e+01  97.134  < 2e-16 ***
X.colorBlack            -4.381e+02  6.660e+02  -0.658 0.510685
X.colorBlue             -6.830e+02  7.000e+02  -0.976 0.329230
X.colorBronze            3.997e+03  3.460e+03   1.155 0.247937

Residual standard error: 10740 on 39391 degrees of freedom
Multiple R-squared:  0.9429,Adjusted R-squared:  0.9428
F-statistic:  9706 on 67 and 39391 DF,  p-value: < 2.2e-16
```

Example: Car Prices... $n = 20k$, $p \approx 40$





X sequence: year, trim430, trim500, mileage, displacement, trimX1, trimX2, trimX3...

Common question #1

Q: Why not just run lasso and be done with it?

A:

1. For what λ?

2. By "preconditioning", we tend to get sparser solutions.

3. We don't have to use the same design matrix as the observed data.

4. DSS makes it kosher for a Bayesian to use the LASSO (or forward selection!!) and provides a natural stoping rule.

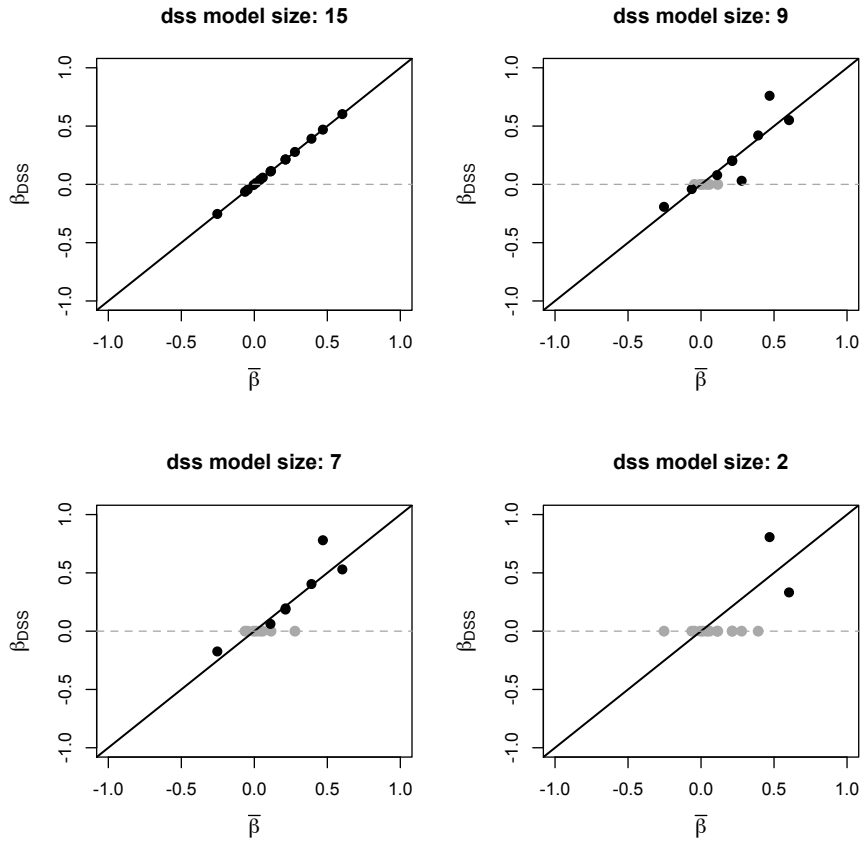
Common question #2

Q: Won't you get unwanted "double shrinkage"?

A: By using a version of the adaptive lasso (using posterior draws to determine the adaptive penalty) we appear not to.

$$\beta_\lambda \equiv \text{arg min}_\gamma \sum_j \frac{\lambda}{|\bar{\beta}_j|}|\gamma_j| + n^{-1}||\mathbf{X}\bar{\beta} - \mathbf{X}\gamma||_2^2. \qquad (1)$$

Minimal double shrinkage



We actually observe "re-inflation" (ℓ_0 approximation in action).

Common question #3

Q: Does this work in the $p > n$ scenario?

A: DSS can be applied to summarize a prior, so it trivially works if $p > n$!! All it is required are predictions at p locations in X space...

In fact, this is one of my favorites insights about DSS: it can help is summarizing complex prior assumptions in meaningful, interpretable ways!

Common question #4

Q: Can you do anything other than linear models?

A: Glad you asked!

DSS for Nonparametric Function Estimation

In the linear case we had...

$$\mathsf{E}_\theta \mathcal{L}(\theta, \gamma) = \lambda ||\gamma||_0 + ||\mathbf{X}\bar{\beta} - \mathbf{X}\gamma||_2^2$$

in the non-linear case we could move to this...

$$\mathsf{E}_\theta \mathcal{L}(\theta, \gamma) = \lambda ||\gamma||_0 + ||\widehat{f(\mathbf{X})} - \mathbf{X}\gamma||_2^2$$

or, (abusing notation!) change the action set and get to...

$$\mathsf{E}_\theta \mathcal{L}(\theta, \gamma) = \lambda(\#\ \text{var}) + ||\widehat{f(\mathbf{X})} - \gamma(\mathbf{X})||_2^2$$

Now that we turned the problem into an optimization, we have lots of options to make this work!

The General Recipe

(1) *fit (inference)*:
Choose a model for $f(X)$... do this once with all the X's.

In a Bayesian sense, this is the "true" posterior.
You do not corrupt your prior to search for simple models.

(2) *fit the fit (decision)*:
Use utility (you want a simple model) to find simple models that approximate (1) well.

Key: need fast, flexible ways to fit the fit.

Key: inference is all in (1),
you do not worry about over-fitting!!
this makes things very different !!!.

All uncertainty comes from the posterior in (1).

The General Recipe

$$Y = f(x) + \epsilon, \quad \epsilon \sim N(0, \sigma^2).$$

(1) *fit*:

Get posterior draws of f and σ.

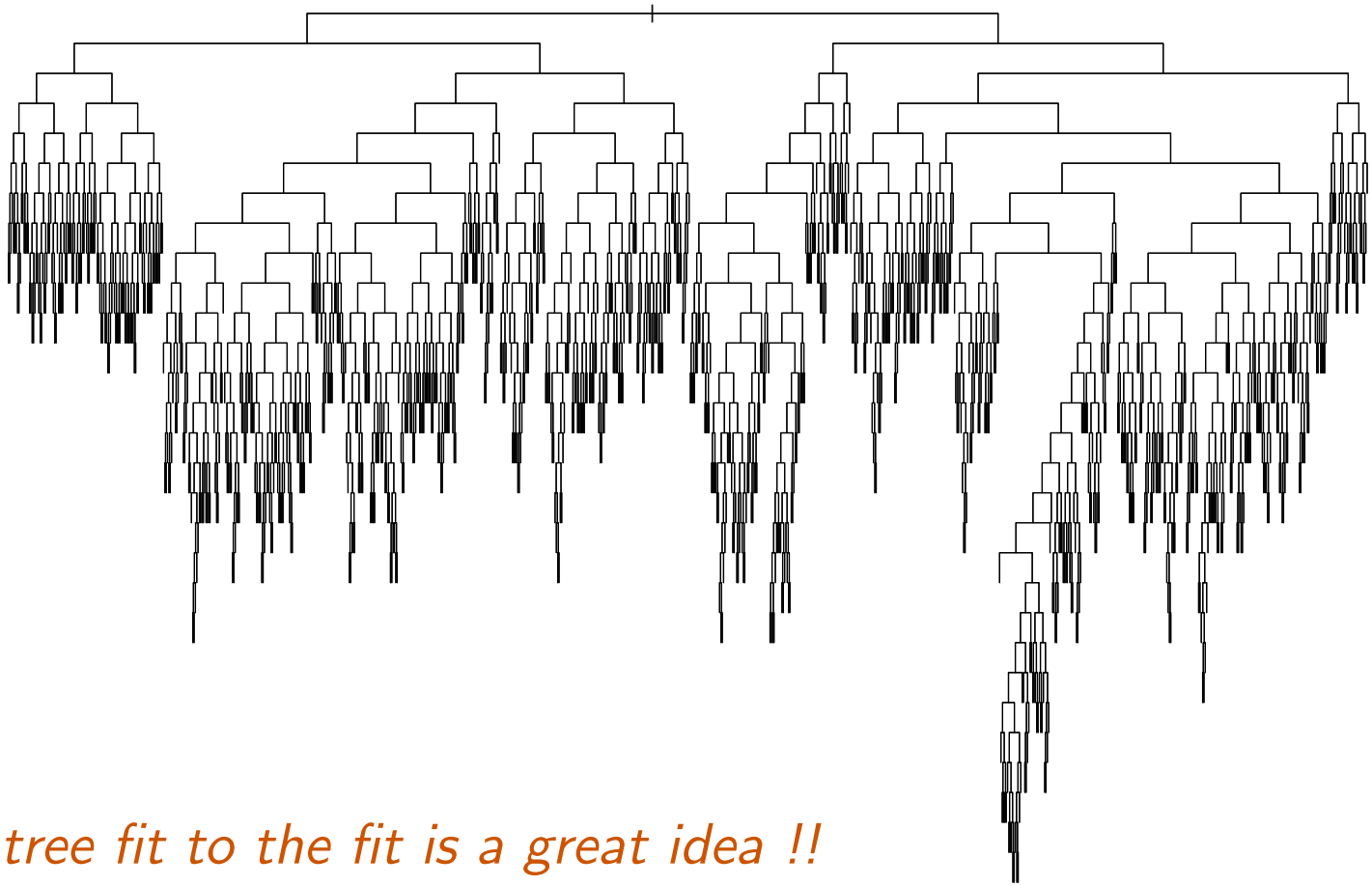
(2) *fit the fit*:

Using subsets of the x variables,
fit very large single tree models to draws of f from (1).
we use forward and backward selection as a heuristic.

Note:

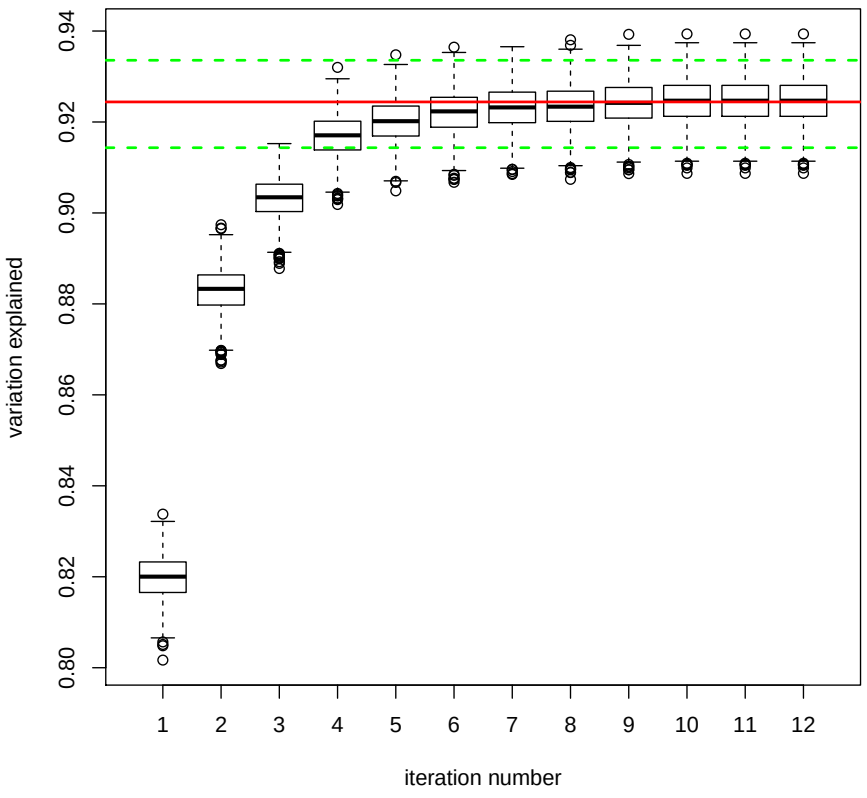
Single trees can be fit very fast and are very flexible.
No inference in (2)! No cross-validation.

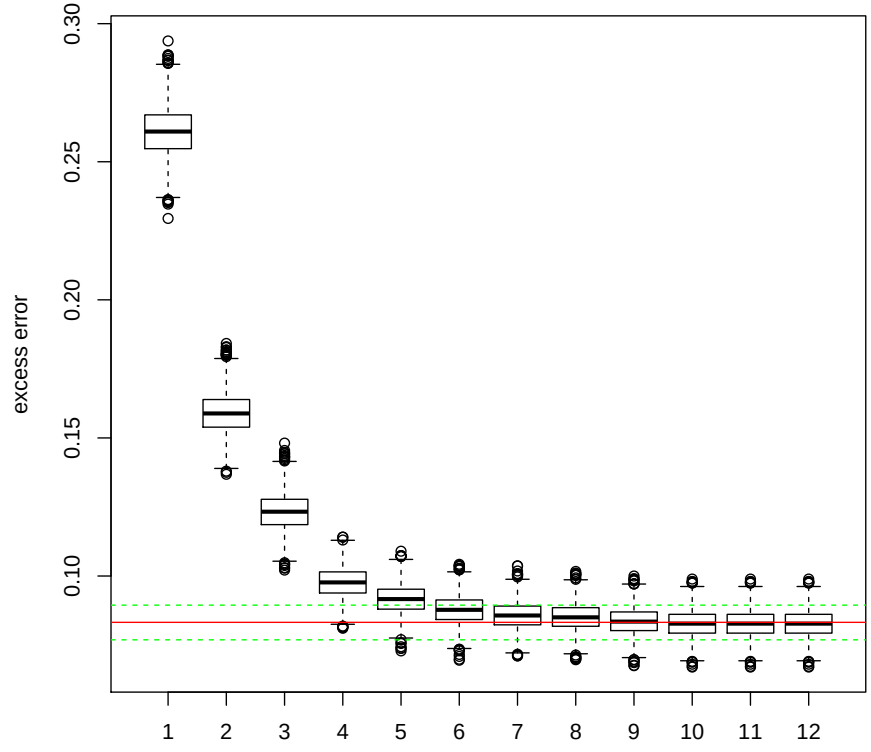
A big tree fit to the data is a terrible idea...



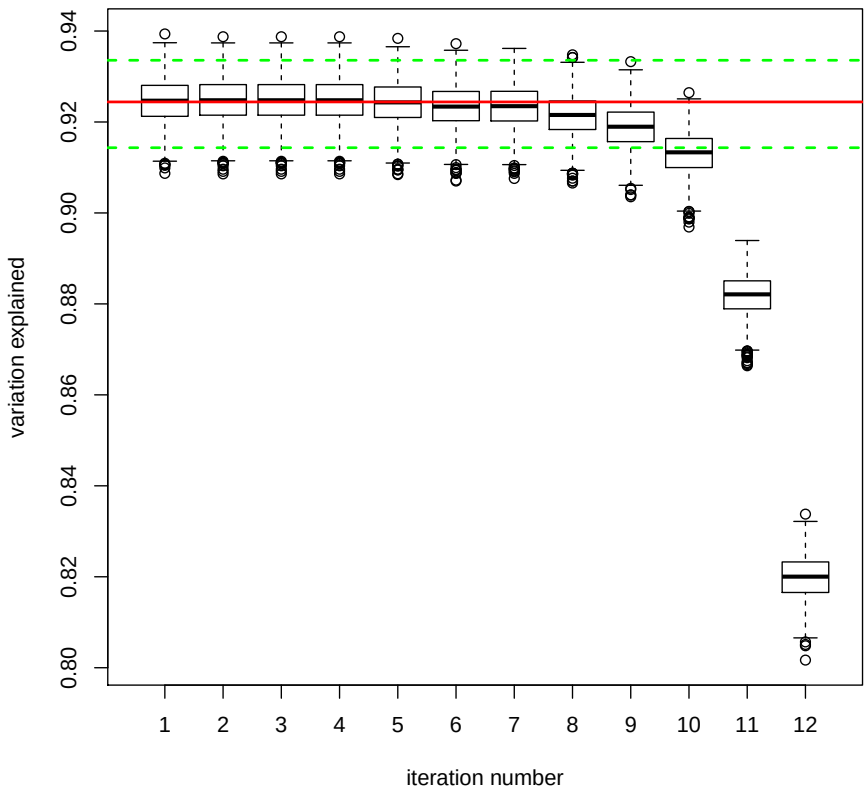
A big tree fit to the fit is a great idea !!

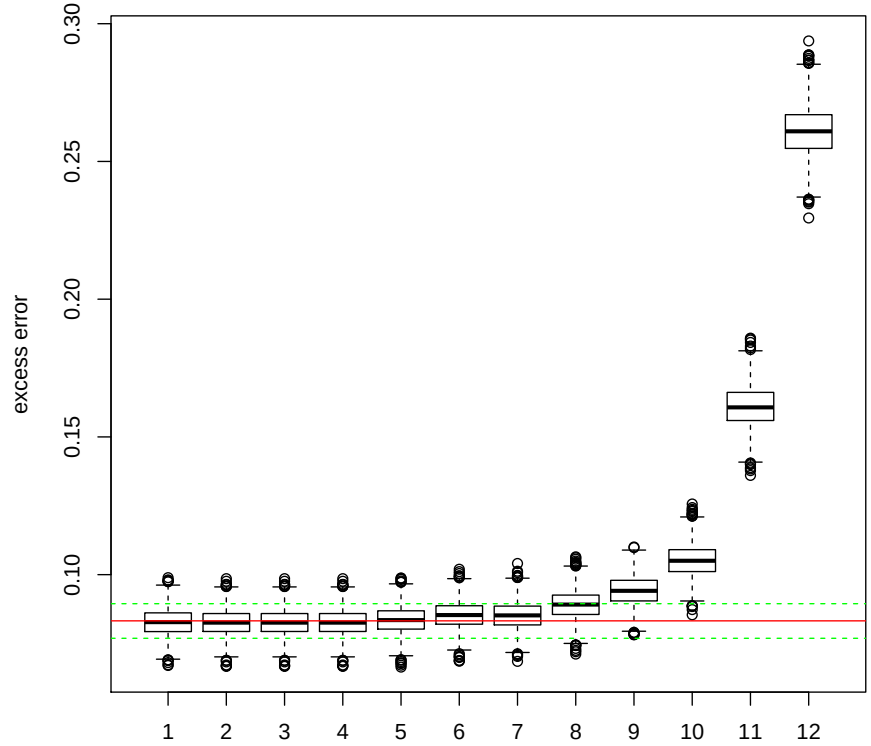
Examples: Boston – Forward Selection



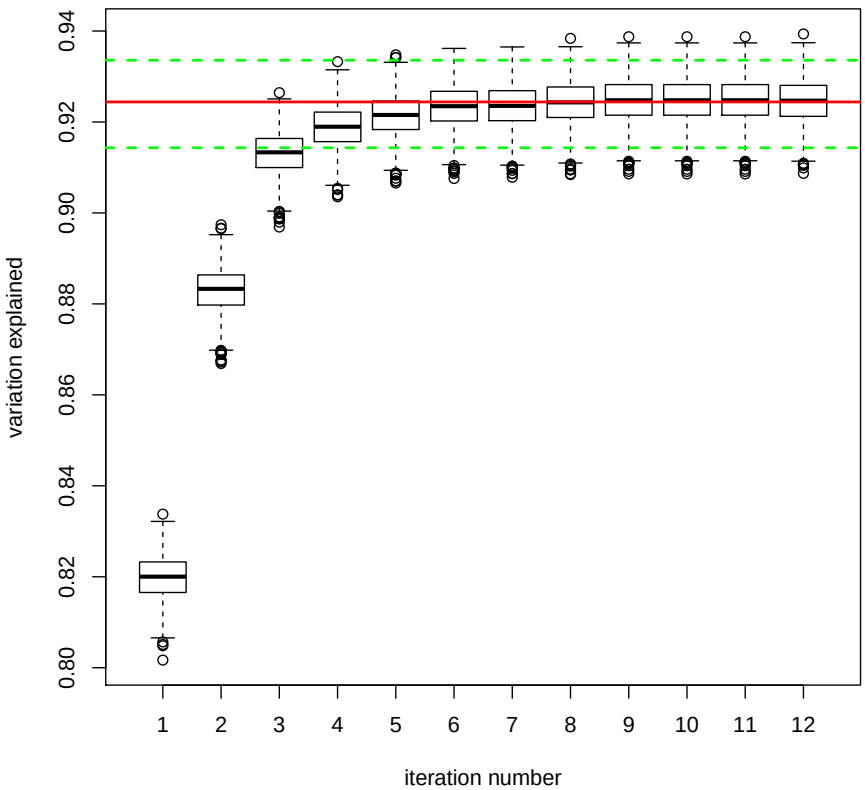


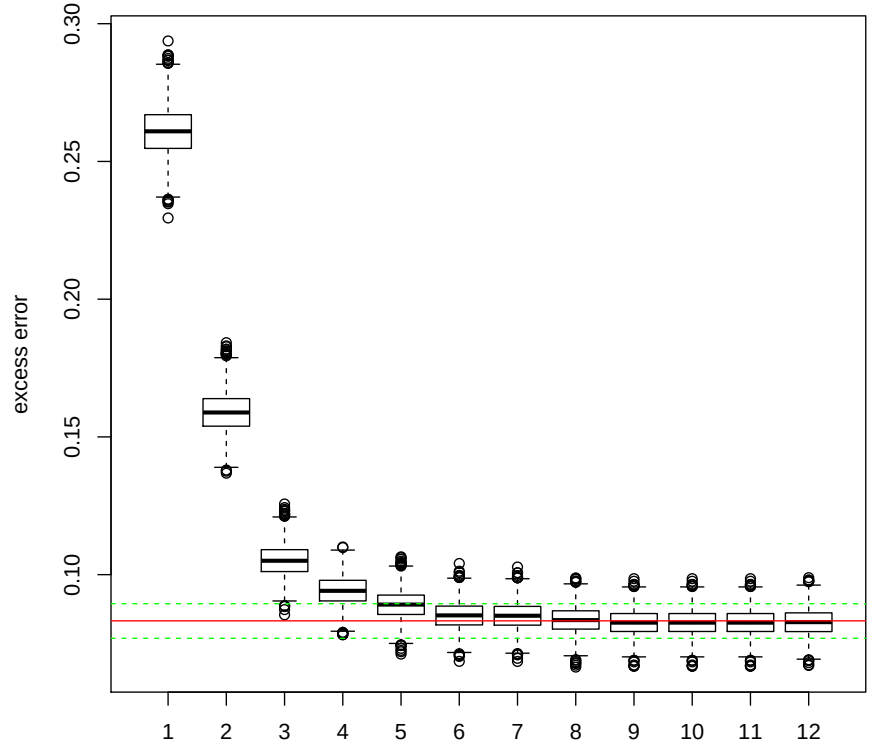
Examples: Boston – Backwards Selection





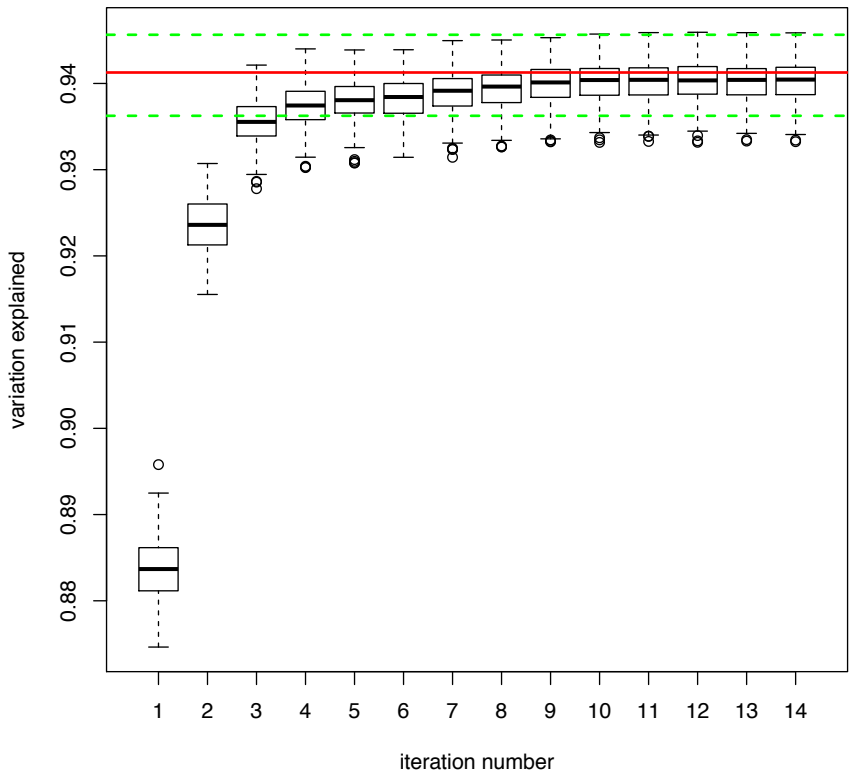
Examples: Boston – All Subsets

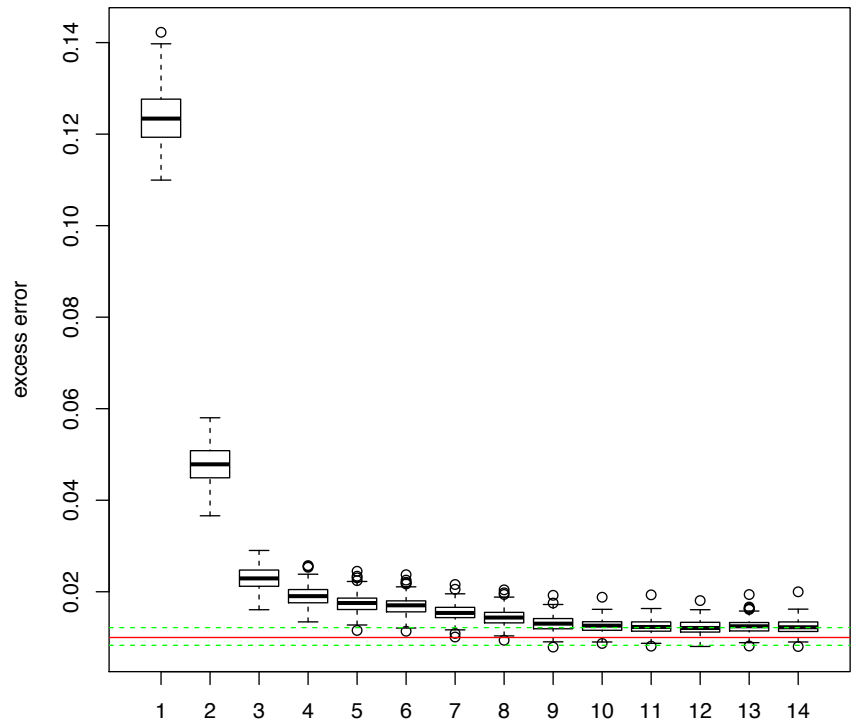




Examples: Car Prices... $n = 20k$, $p \approx 100$

For each subset $R^2 = \text{cor}(\hat{f}_{\text{tree}}, f_j)^2$ where j indexes draws from the posterior.





Even though we are non-parametric, with our big n we are very sure we have found a subset that does the job!!!

Examples: Car Prices... $n = 20k$, $p \approx 100$

Forward stepwise:

```
*****variables at iteration  1
[1] "year"
*****variables at iteration  2
[1] "year"  "trim8"
*****variables at iteration  3
[1] "year"    "trim8"   "mileage"
*****variables at iteration  4
[1] "year"    "trim8"   "mileage" "trim12"
*****variables at iteration  5
[1] "year"          "trim8"         "mileage"       "trim12"
[5] "displacement9"
*****variables at iteration  6
[1] "year"          "trim8"         "mileage"       "trim12"
[5] "displacement9" "fuel2"
*****variables at iteration  7
[1] "year"          "trim8"         "mileage"       "trim12"
[5] "displacement9" "fuel2"         "featureCount"
*****variables at iteration  8
[1] "year"          "trim8"         "mileage"       "trim12"
[5] "displacement9" "fuel2"         "featureCount"  "condition3"
*****variables at iteration  9
[1] "year"          "trim8"         "mileage"       "trim12"
[5] "displacement9" "fuel2"         "featureCount"  "condition3"
[9] "condition2"
```

Now I know that with just the 7 variables:

year, mileage, featureCount, trim8, trim12, displacement9, fuel2
(where fuel2 is level 2 dummy for fuel categorical variable),

I can get as good a prediction as I get from using 90!!!

Without *making any assumptions about the functional form.*

First three variables might be obvious, next 4 are not...

And it was pretty easy to do and to understand.

Satisficing

Statistical uncertainty helps our cause...it makes finding a good-enough model easier. For example, allows us to use forward or backward selection without the standard inferential worries...

Provided we took great care in our initial modeling, and take seriously our posterior uncertainty, we can use that uncertainty to ease the burden of our optimization step.

The utility function then effectively serves to "break ties". People have been using priors (improperly) in this role for years.

More Extensions...

The DSS framework can be extended to other modeling set-ups. Using log-likelihoods as measures of predictive performance we can work, among others, with

1. GLMs:

$$\mathcal{L}(\tilde{Y}, \gamma) = \lambda||\gamma||_0 - n^{-1} \log \left[f(\tilde{Y}, \mathbf{X}, \gamma) \right]$$

2. Covariance Selection:

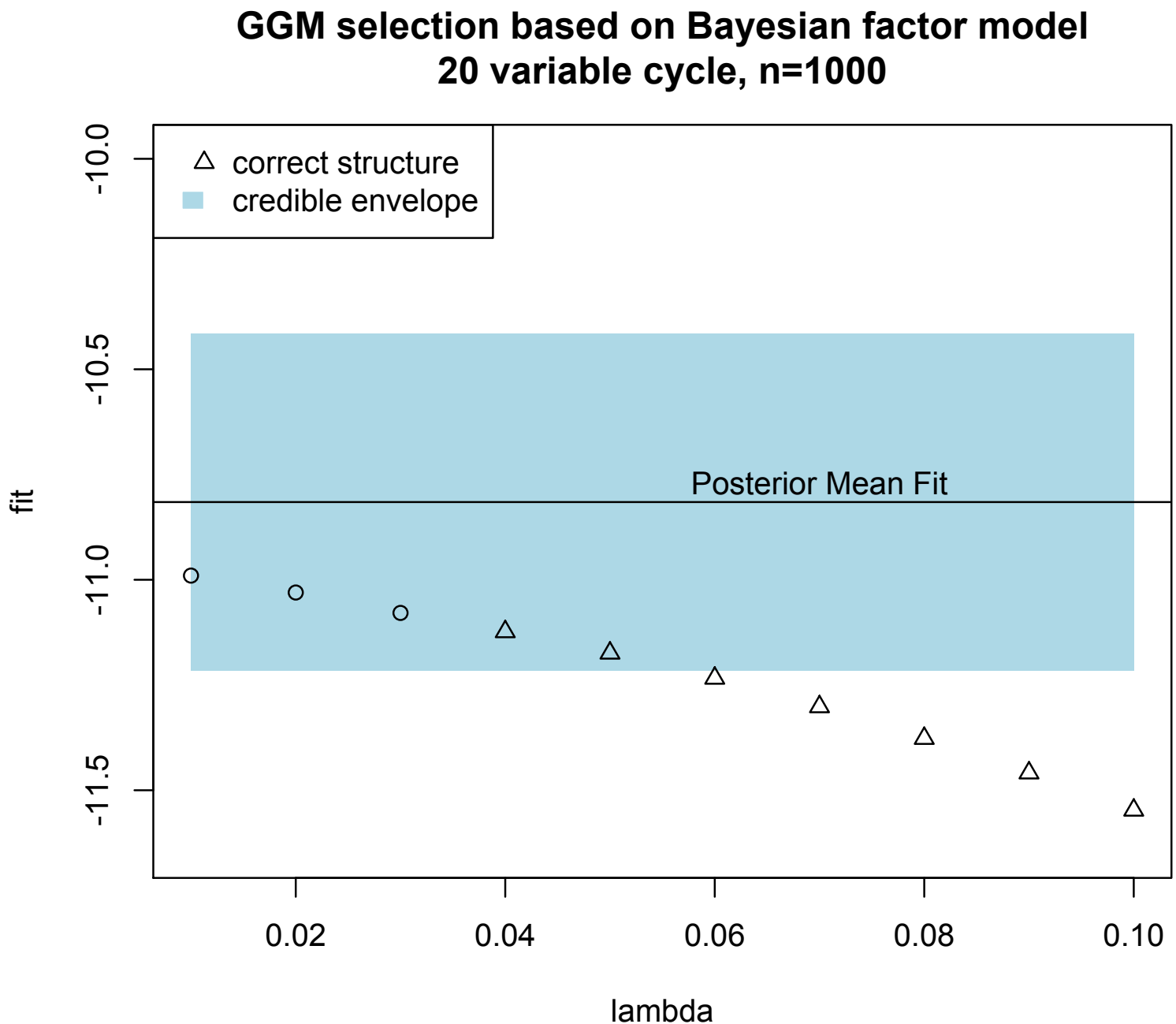
$$\mathcal{L}(\tilde{\mathbf{X}}, \mathbf{\Gamma}) = \lambda||\mathbf{\Gamma}||_0 - \log\det(\mathbf{\Gamma}) - \text{tr}(n^{-1}\tilde{\mathbf{X}}\tilde{\mathbf{X}}'\mathbf{\Gamma})$$

In both cases, the expectation step lead to close form expressions that can be optimized via standard procedures that approximate $||\cdot||_0$ penalty.

Covariance Selection (GGM) example...

- Simulated structure: 20 variable cycle;
- Prior: "Bayesian factor model" prior for Σ;
- No need for decomposability restriction;
- Graphical Lasso for optimization step;
- within seconds we get the right answer!

Covariance Selection (GGM) example...



Selection an optimal ETF Portfolio

This is an application where we seek to find the simplest (smallest) optimal portfolio for long-run investors based solely on passive funds (ETFs).

Given past returns on a set of target asset, $\{R_j\}_{j=1}^{q}$, and a collection of exchange traded funds (ETFs), $\{X_i\}_{i=1}^{p}$, write:

$$R_j = \beta_{j1}X_1 + \cdots + \beta_{jp}X_p + \epsilon_j, \quad \epsilon_j \sim \mathrm{N}(0, \sigma^2).$$

In real life $p \approx 200$. Do we need this many to recapitulate the co-movement of the target assets?

Selection an optimal ETF Portfolio

Step 1. Run an empirical factor model (APT model) using a multivariate application to SSVS:

$$\mathbf{R_t} = \beta \mathbf{X_t} + \epsilon_t$$

and a model for

$$\mathbf{X_t} \sim N(\mu_x, \Sigma_X)$$

Define

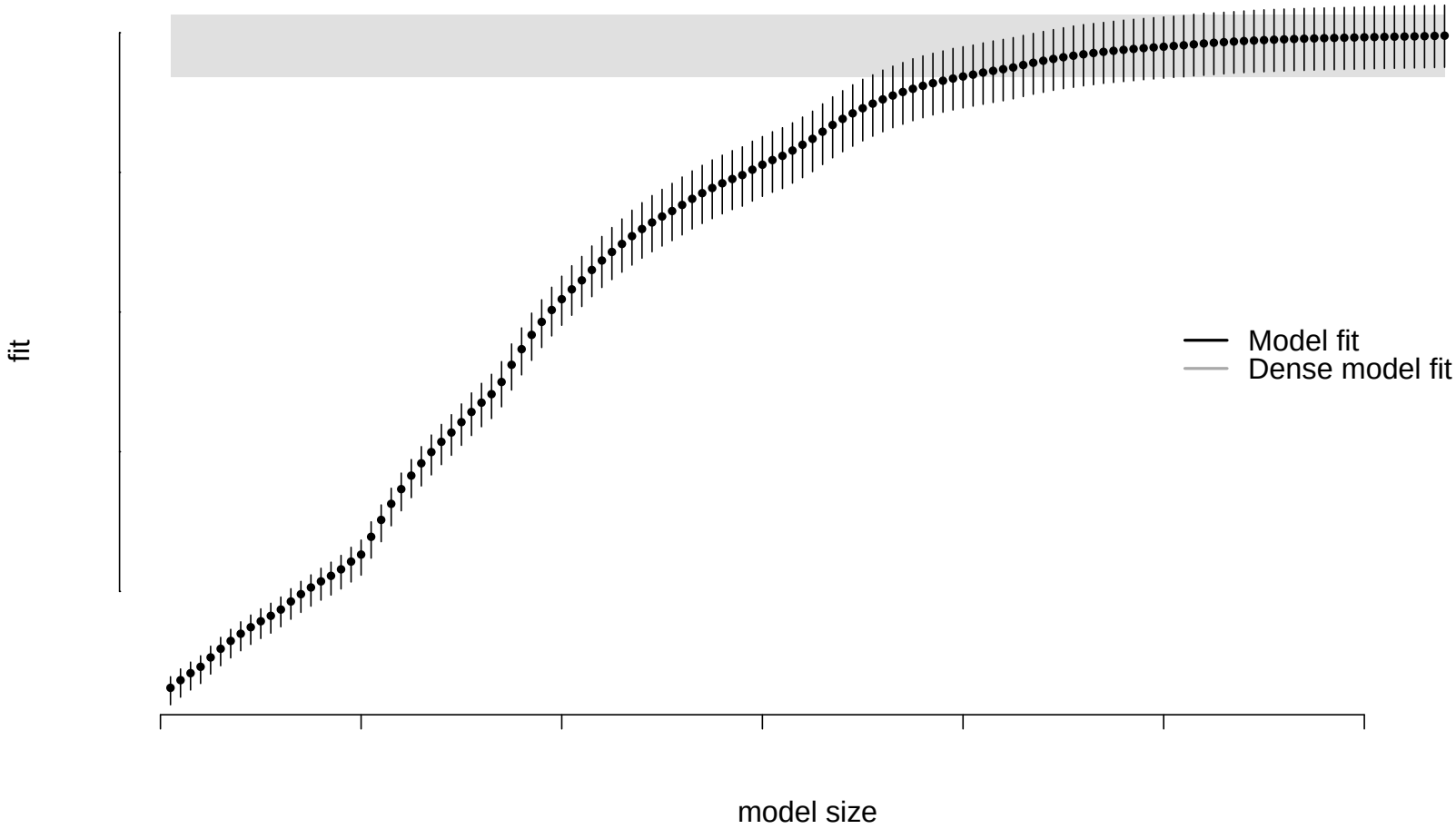
$$\mathbf{\Gamma}^{-1} = \mathbf{\Sigma} = \begin{pmatrix} \beta\Sigma_X\beta' + \Psi & \beta\Sigma_X \\ \Sigma_X\beta' & \Sigma_X \end{pmatrix}$$

Step 2:

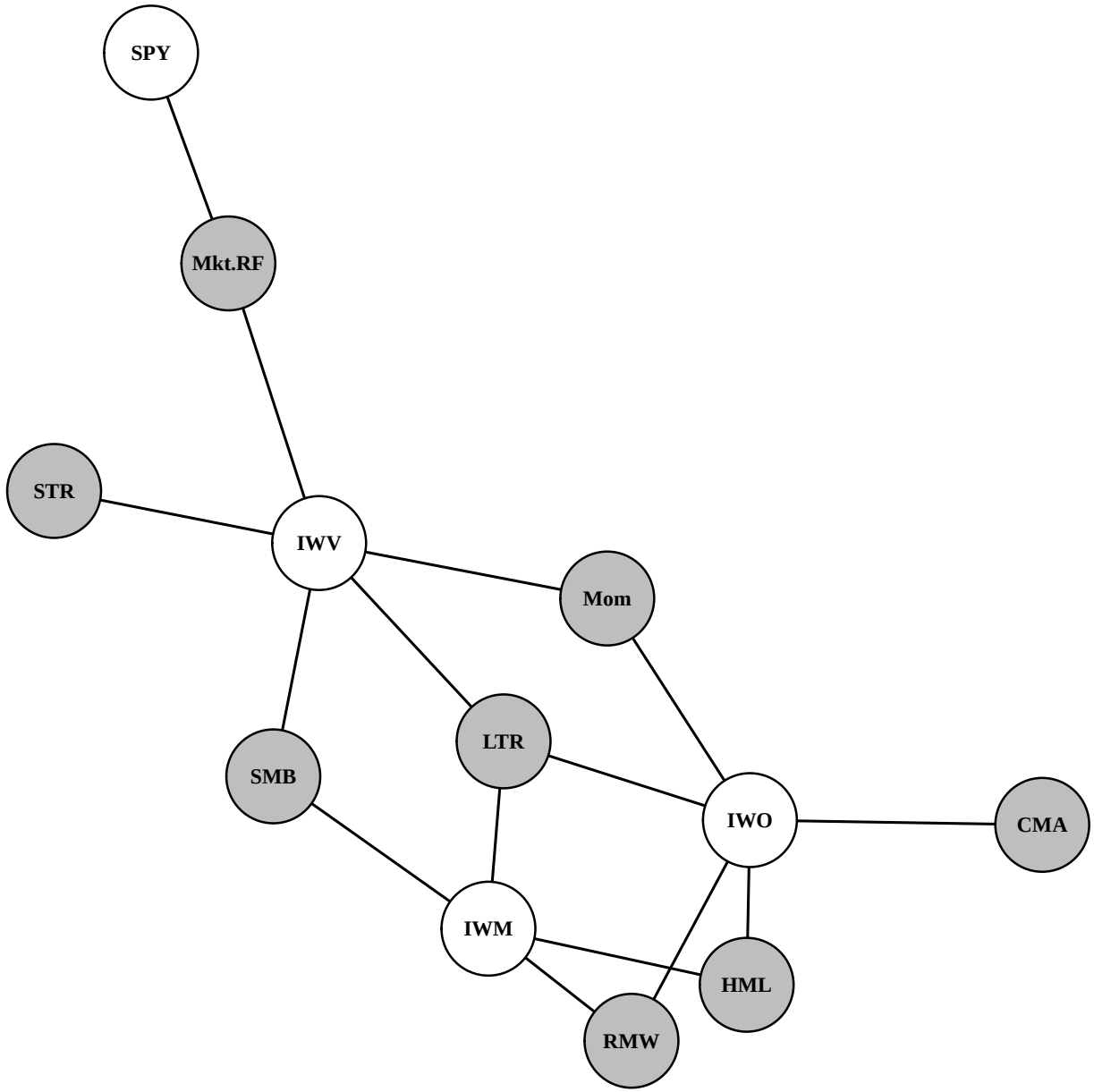
$$\mathcal{L}(\tilde{\mathbf{Q}}, \mathbf{\Gamma}) = \lambda||\mathbf{\Gamma}||_0 - \log\det(\mathbf{\Gamma}) - \text{tr}(n^{-1}\tilde{\mathbf{Q}}\tilde{\mathbf{Q}}'\mathbf{\Gamma})$$

Where we are penalizing just the diagonal block $\mathbf{\Gamma}$... (not quite) Think about this an a multivariate regression extension to the linear model where the future predictors are random...

Selection an optimal ETF Portfolio



Selection an optimal ETF Portfolio



Closing Comments

- DSS/Two Steps idea is an operation that provides practically meaningful solutions to the posterior summarization process.
- Utility functions can enforce inferential preferences that are not prior beliefs. (statistical versus practical significance)
- In essence, we transform a very complex object, the posterior (or the prior), into a "simple" sequence of potential solutions

Closing Comments

- In the linear set up, DSS makes useful to Bayesians the very popular optimization toolset: LASSO, stepwise selection, etc... it can be extended to a variety of situations where we currently struggle with summarizing posteriors (GLMs, graphs, trees, etc...)!

- Posterior draws can be used to benchmark candidate summaries.

- This is just good-old decision theory using modern tools...

Redemption

Re George and McCulloch, "Stochastic Search Variable Selection".

Jay Kadane (circa 1992):

"You are confusing your prior with your utility function."

Not any more!!!!

References...

- *"Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective." JASA 2015* (with Hahn)
- *"Optimal ETF selection for passive investing."* (with Hahn and Puelz)
- *"Variable selection in nonlinear regression with penalized predictive utility."* (with Hahn and McCulloch)
- *"Decoupled shrinkage and selection for Gaussian graphical models."* (with Hahn and Jones)