

Comparing Bayesian and non-Bayesian Strategies for Variable Selection in Regression Models with Missing Covariates

Thomas R. Belin

UCLA Department of Biostatistics

with

Xiaowei Yang (Ph.D. 2002, now UC Davis)

Gang Liu (Ph.D. 2007, now Google, Inc.)

W. John Boscardin (now UCSF)

Variable selection strategies

- **Stepwise strategies using significance tests**
(forward selection, backward elimination, stepwise procedures)
- **Criterion-based strategies with penalty to favor parsimony**
(e.g., AIC, adjusted R^2)
- **Bayesian variable selection strategies**
(smooth coefficients using informative prior distribution,
“spike and slab” mixture for coefficients b/w zero and non-zero values,
mixture for coefficients b/w tight and diffuse distribution around zero)

Notation

Multivariate
normal data:

$$\mathbf{D} = \begin{bmatrix} Y_1 & x_{11} & \dots & x_{1p} \\ Y_2 & x_{21} & \dots & x_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ Y_n & x_{n1} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{d}_1^T \\ \cdot \\ \cdot \\ \mathbf{d}_n^T \end{bmatrix} \quad \text{with } \mathbf{d}_i \sim N_q(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (q = p + 1)$$

Linear regression
model:

$$\mathbf{y} = \alpha + \sum_{j=1}^p \gamma_j \mathbf{X}_j \beta_j + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$$

Indicator variables
for including predictors:

$$\gamma_j = \begin{cases} 1 & \text{if } \mathbf{X}_j \text{ is selected} \\ 0 & \text{Otherwise} \end{cases}$$

Vector of indicators:

$$\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$$

Stochastic Search Variable Selection (SSVS) (George and McCulloch 1993 JASA)

Linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2 \mathbf{I})$$

Hierarchical prior distributions

$$\beta_j | \gamma_j \sim (1 - \gamma_j)N(0, \tau_j^2) + \gamma_j N(0, c_j^2 \tau_j^2)$$

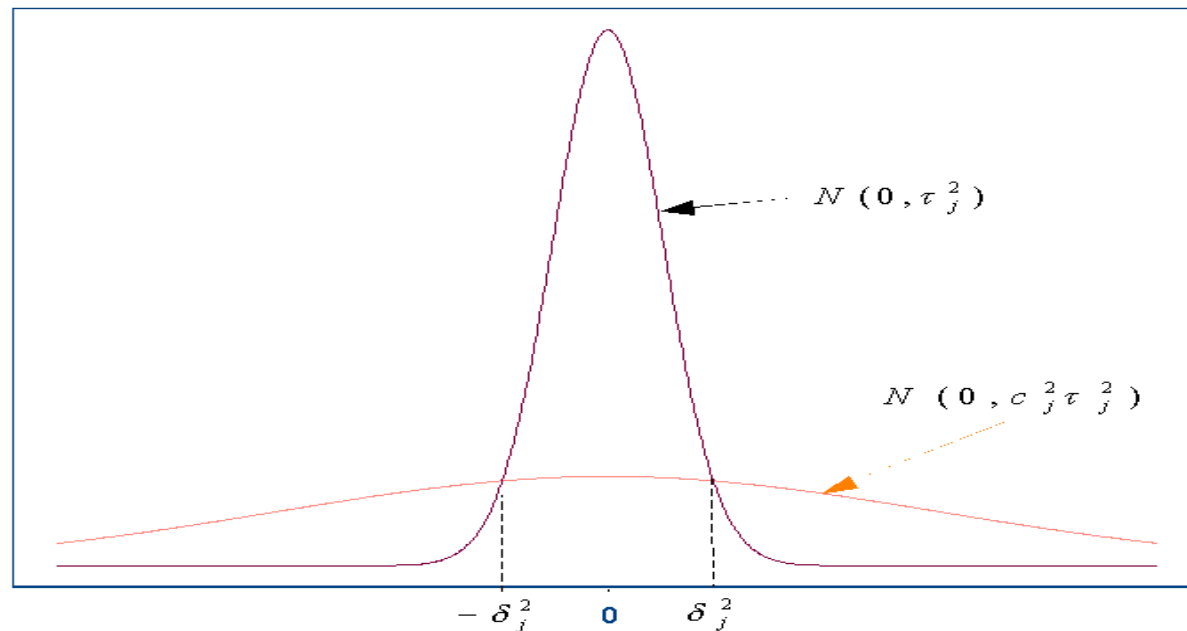
$$\sigma^2 | \boldsymbol{\gamma} \sim IG(v/2, v\lambda_{\boldsymbol{\gamma}}/2)$$

$$p(\boldsymbol{\gamma}) \sim \prod_{j=1}^p \omega_j^{\gamma_j} (1 - \omega_j)^{(1 - \gamma_j)}$$

$$(\omega_j = p(\gamma_j = 1))$$

SSVS: Conceptual framework

$$\beta_j | \gamma_j \sim (1 - \gamma_j)N(0, \tau_j^2) + \gamma_j N(0, c_j^2 \tau_j^2)$$



Prior specification for SSVS

- Choices for τ_j^2 and c_j^2

Following George and McCulloch (1993), can consider alternatives based on points of intersection, ratios at 0 of marginal densities: $(\beta_j | \sigma_{\beta_j}, \gamma_j = 0) \sim N(0, \sigma_{\beta_j}^2 + \tau_j^2)$ and

$$(\beta_j | \sigma_{\beta_j}, \gamma_j = 1) \sim N(0, \sigma_{\beta_j}^2 + c_j^2 \tau_j^2)$$

- Choices for $p(\gamma_j = 1) = \omega_j$
 - Equipoise: all $\omega_j = 0.5$
 - Other subjective considerations

Alternative Strategy I: Impute, Then Select (ITS)

- Step 1: Create multiple imputations under multivariate normal model (e.g., SAS PROC MI, norm)
- Step 2: Perform Bayesian variable selection (SSVS) on multiply imputed data sets
- Step 3: Combine results using multiple-imputation combining rules
 - Averaging over vector of inclusion indicators yields posterior probability of model $\gamma = (\gamma_1, \dots, \gamma_p)$

Alternative Strategy II: Simultaneously Impute and Select (SIAS)

- Step 1: Missing-data imputation under normal model:

$$\mathbf{d}_{i(mis)}^{(t+1)} \sim p(\mathbf{d}_{i(mis)} \mid \mathbf{d}_{i(obs)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\Sigma}^{(t)})$$

- Step 2: MCMC parameter generation (with normal inverse-Wishart prior)

$$(\boldsymbol{\mu}_x^{(t+1)}, \boldsymbol{\Sigma}_x^{(t)}) \sim P(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x \mid \mathbf{X}_{(mis)}^{(t+1)}, \mathbf{X}_{(obs)})$$

- Step 3: Variable selection (SSVS) on current data, parameters: $(\gamma^{(t+1)}, \alpha^{(t+1)}, \sigma^{2(t+1)}, \boldsymbol{\beta}^{(t+1)}) \sim P(\gamma, \alpha, \sigma^2, \boldsymbol{\beta} \mid \mathbf{D}_{(obs)}, \mathbf{D}_{(mis)}^{(t+1)})$
- Step 4: Reparameterization to facilitate imputation

$$\boldsymbol{\mu}_y = \alpha + \boldsymbol{\mu}_x \boldsymbol{\beta}$$

$$\sigma_y^2 = \sigma^2 + \boldsymbol{\Sigma}_{xy} \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\Sigma}_{yx}$$

$$\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_x \boldsymbol{\beta}$$

Application: Los Angeles County Foster Care Study

- Study mental-health concerns, service use among children in foster care (Zima et al. 2000 *J Behav Hlth Services Res*)
- Outcome: # office visits
- Predictors : age, sex, ethnicity (white/black/other), parent's education (yrs), monthly gov't benefit, # caseworker home visits, time in placement (yrs), # placements, type of placement (family vs. group home), child behavior checklist (CBCL) total score, child depression inventory (CDI) total score, Child and Adolescent Functional Assessment Scale (CAFAS) total score
- 47 complete cases among 77 individuals

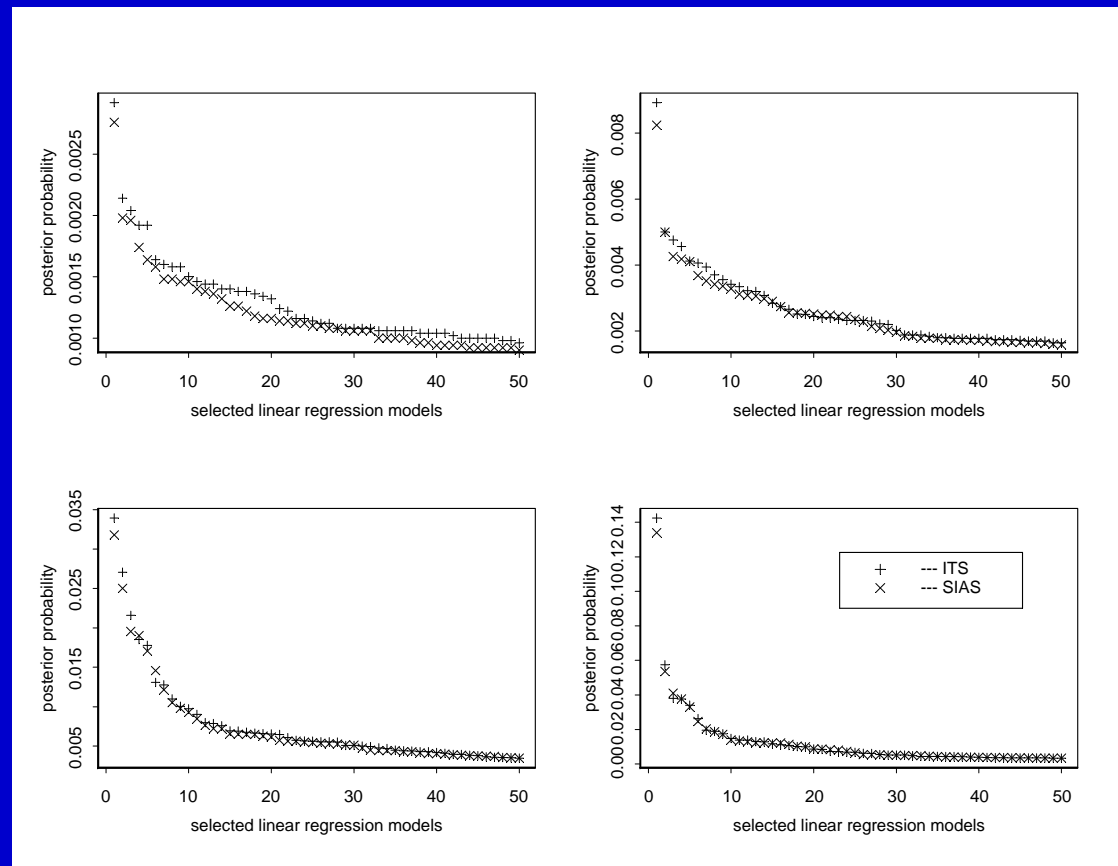
Posterior probabilities $p(\gamma|\mathbf{D})$ for top 5 models selected by ITS under alternative priors

	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (1, 5)$	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (1, 10)$
Variables Selection (top 5 models)	NUMVISIT WHITE (0.29%) NUMVISIT TYPEPLAC WHITE (0.21%) AGE NUMVISIT WHITE (0.20%) NUMVISIT SEX WHITE (0.19%) NUMVISIT BENEFIT WHITE (0.19%) NUMVISIT BLACK WHITE (0.16%)	NUMVISIT WHITE (0.89%) WHITE (0.50%) AGE NUMVISIT WHITE (0.48%) NUMVISIT (0.46%) NUMVISIT TYPEPLAC WHITE (0.41%) NUMVISIT BENEFIT WHITE (0.41%)
	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (10, 100)$	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (10, 500)$
Variables Selection (top 5 models)	NUMVISIT WHITE (3.4%) AGE NUMVISIT WHITE (2.7%) AGE NUMVISIT SEX WHITE (2.2%) NUMVISIT SEX WHITE (1.9%) NUMVISIT TYPEPLAC WHITE (1.8%) NUMVISIT BLACK WHITE (1.3%)	NUMVISIT WHITE (14%) AGE NUMVISIT WHITE (6.0%) NUMVISIT (3.9%) NUMVISIT TYPEPLAC WHITE (3.8%) NUMVISIT SEX WHITE (3.4%) NUMVISIT BLACK WHITE (2.6%)

Posterior variable selection probabilities $p(\gamma_j = 1 | \mathbf{D})$ under ITS for foster-care data

	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (1, 5)$	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (1, 10)$	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (10, 100)$	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (10, 500)$
AGE	.41 (.009)	.34 (.009)	.46 (.036)	.28 .033
T_TOT	.33 (.004)	.27 (.004)	.16 (.005)	.08 .003
CAFASTOT	.34 (.006)	.27 (.006)	.17 (.010)	.09 .006
CDL_TOT	.34 (.004)	.28 (.005)	.18 (.006)	.09 .004
NUMPLAC	.33 (.003)	.26 (.003)	.15 (.005)	.07 .003
TIMEPLAC	.34 (.009)	.27 (.009)	.18 (.020)	.09 .012
NUMVISIT	.65 (.027)	.63 (.034)	.96 (.026)	.94 .053
BENEFIT	.34 (.006)	.28 (.006)	.21 (.023)	.11 .015
P_EDU	.34 (.003)	.27 (.002)	.18 (.005)	.09 .002
SEX	.42 (.010)	.35 (.013)	.42 (.039)	.25 .036
TYPEPLAC	.39 (.009)	.33 (.009)	.33 (.041)	.20 .034
BLACK	.39 (.003)	.33 (.003)	.29 (.003)	.16 .003
WHITE	.64 (.046)	.63 (.059)	.81 (.079)	.74 .115

Ordered posterior model probabilities for top 50 models selected by ITS, SIAS



Simulation study

- True model: $Y = 1.0 * X_1 + 2.0 * X_2 + 1.0 * X_6 + 2.0 * X_7$
- Correlation structure of explanatory variables:

$$(\mathbf{X}_1, \dots, \mathbf{X}_{10})^T \sim N \left(\begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho & \dots & \dots & \rho \\ \rho & 1 & \dots & \dots & \rho \\ \cdot & \cdot & \dots & \dots & \cdot \\ \rho & \rho & \dots & \dots & \rho \\ \rho & \rho & \dots & \dots & 1 \end{bmatrix} \right) \quad \rho = 0.1 \text{ or } 0.5$$

- Missing-data mechanism:
 - MCAR: $P(\mathbf{X}_j = \text{missing}) = 5\% \text{ or } 10\% \quad (j = 1, \dots, 10)$
 - MAR: $P(\mathbf{X}_j = \text{missing}) = 0.0 \quad (j = 1, \dots, 5)$
 $\text{logit}(P(\mathbf{X}_k = \text{missing})) = a_0 + a_1 \mathbf{X}_1 + a_2 \mathbf{X}_2 + a_3 \mathbf{X}_3 + a_4 \mathbf{X}_4 + a_5 \mathbf{X}_5$
 $(k = 6, 7, \dots, 10)$
- Sample size: $n = 100$
- Residual variance: $\sigma^2 = 2.5$

Summarizing simulation results

- Compare ITS, SIAS via

- Posterior variable selection probability:

$$\hat{P}(\gamma_j = 1 | \mathbf{D}_{(obs)}) = \frac{1}{100} \sum_{l=1}^{100} \hat{P}(\gamma_j^{(l)} = 1 | \mathbf{D}_{(obs)}^{(l)})$$

- “Signal-to-noise ratio (SNR)”:

$$\text{SNR} = \frac{\min_{j \in \{1,2,6,7\}} \{\hat{P}(\gamma_j = 1 | \mathbf{D}_{obs})\}}{\max_{j \in \{3,4,5,8,9,10\}} \{\hat{P}(\gamma_j = 1 | \mathbf{D}_{obs})\}}$$

- Compare ITS, SIAS, Stepwise via criterion reflecting number of wrongly included variables:

$$C_{VS}^{(l)} = \sum_{j=1}^{10} |\gamma_j^{(l)} - t_j^{(l)}|$$

$$C_{VS} = \frac{1}{100} \sum_{l=1}^{100} C_{VS}^{(l)}$$

$$SD_C = \sqrt{\frac{1}{99} \sum_{l=1}^{100} (C_{VS}^{(l)} - C_{VS})^2}$$

ITS vs. SIAS for $\rho = 0.1$:
posterior variable selection probs, SNR

		X_1	X_2	X_6	X_7	SNR
ITS	MCAR 5%	0.90	0.99	0.90	0.99	2.90
	MCAR 10%	0.85	0.99	0.86	0.99	3.04
	MAR 20%	0.90	0.99	0.86	0.99	2.53
	MAR 40%	0.81	0.99	0.76	0.99	2.30
SIAS	MCAR 5%	0.92	0.99	0.91	0.99	2.94
	MCAR 10%	0.88	0.99	0.88	0.99	3.03
	MAR 20%	0.90	0.99	0.90	0.99	2.81
	MAR 40%	0.82	0.99	0.85	0.99	2.36

**ITS vs. SIAS for $\rho = 0.5$:
posterior variable selection probs, SNR**

		X_1	X_2	X_6	X_7	SNR
ITS	MCAR 5%	0.79	0.99	0.80	0.99	2.32
	MCAR 10%	0.68	0.99	0.71	0.99	2.27
	MAR 20%	0.88	0.99	0.73	0.99	2.09
	MAR 40%	0.88	0.99	0.58	0.95	1.45
SIAS	MCAR 5%	0.80	0.99	0.80	0.99	2.35
	MCAR 10%	0.71	0.99	0.72	0.99	2.29
	MAR 20%	0.88	0.99	0.77	0.99	2.14
	MAR 40%	0.89	0.99	0.69	0.98	1.53

Average number of incorrectly selected variables (C_{VS})

		SIAS	ITS	Stepwise
$\rho=0.1$	MCAR 5%	0.38	0.41	0.85
	MCAR 10%	0.62	0.68	1.45
	MAR 20%	0.42	0.51	0.81
	MAR 40%	0.90	1.01	1.66
$\rho=0.5$	MCAR 5%	0.73	0.82	1.40
	MCAR 10%	0.99	1.05	2.06
	MAR 20%	0.68	0.77	1.21
	MAR 40%	1.14	1.30	1.98

Extension to logistic model (Liu 2007): Average number of incorrectly selected variables (C_{VS})

		SIAS	ITS	Stepwise
$\rho=0.1$	MCAR 20%	0.20	0.18	0.39
	MCAR 50%	0.22	0.21	0.73
	MAR	0.64	0.40	1.03
	NMAR	0.52	0.39	0.94
$\rho=0.5$	MCAR 20%	1.44	1.46	2.03
	MCAR 50%	2.04	2.03	3.19
	MAR	2.70	2.65	3.17
	NMAR	2.62	2.47	3.15

Comments

- SIAS tends to slightly outperform ITS
- Both SIAS, ITS substantially outperform Stepwise
- Similar patterns, e.g., when covariance matrix unstructured, correlations uniform within $[0.1, 0.5]$
- Greater correlation among covariates leads to more precise imputation of missing values, but collinearity among covariates blurs distinctions between predictors in variable selection
- ITS less cumbersome to implement
- Possibility of adaptive method to finesse awkwardness of prior specification?

Posterior variable selection probabilities $p(\gamma_j = 1 | \mathbf{D})$ under ITS for foster-care data

	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (1, 5)$	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (1, 10)$	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (10, 100)$	$(\frac{\hat{\sigma}_{\beta_j}}{\tau_j}, c_j) = (10, 500)$
AGE	.41 (.009)	.34 (.009)	.46 (.036)	.28 .033
T_TOT	.33 (.004)	.27 (.004)	.16 (.005)	.08 .003
CAFASTOT	.34 (.006)	.27 (.006)	.17 (.010)	.09 .006
CDL_TOT	.34 (.004)	.28 (.005)	.18 (.006)	.09 .004
NUMPLAC	.33 (.003)	.26 (.003)	.15 (.005)	.07 .003
TIMEPLAC	.34 (.009)	.27 (.009)	.18 (.020)	.09 .012
NUMVISIT	.65 (.027)	.63 (.034)	.96 (.026)	.94 .053
BENEFIT	.34 (.006)	.28 (.006)	.21 (.023)	.11 .015
P_EDU	.34 (.003)	.27 (.002)	.18 (.005)	.09 .002
SEX	.42 (.010)	.35 (.013)	.42 (.039)	.25 .036
TYPEPLAC	.39 (.009)	.33 (.009)	.33 (.041)	.20 .034
BLACK	.39 (.003)	.33 (.003)	.29 (.003)	.16 .003
WHITE	.64 (.046)	.63 (.059)	.81 (.079)	.74 .115

References

Yang X, Belin TR, Boscardin WJ. Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, 2005; 61: 498-506.

Liu G. Alternative methods for variable selection in generalized linear models with binary outcomes and incomplete covariates. Ph.D. dissertation, UCLA Department of Biostatistics, 2007.