Stochastic Matching Pursuit for Bayesian Variable Selection and Analysis of Supersaturated Design

Ray-Bing Chen

Institute of Statistics,
National University of Kaohsiung
http://www.stat.nuk.edu.tw/Ray-Bing/index.html

Joint with Ying Nian Wu (UCLA);
Chi-Hsiang Chu, Te-You Lai and
Jian-Zhong Weng (NUK)

1. Variable Selection Problems

1. Variable Selection Problems
2. Stochastic Variable Selection Methods

1. Variable Selection Problems
2. Stochastic Variable Selection Methods
3. Large n Small p Problems

1. Variable Selection Problems
2. Stochastic Variable Selection Methods
3. Large n Small p Problems
4. Small n Large p Problems

1. Variable Selection Problems
2. Stochastic Variable Selection Methods
3. Large n Small p Problems
4. Small n Large p Problems
5. Comparison with Conjugate Prior Assumption

1. Variable Selection Problems
2. Stochastic Variable Selection Methods
3. Large n Small p Problems
4. Small n Large p Problems
5. Comparison with Conjugate Prior Assumption
6. Analysis of and Supersaturated Design

Variable Selection

- Model: $Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$
  - $\mathbf{Y}$: the n-dimensional response vector
  - $\mathbf{X}_i$: the n-dimensional regressor vector
  - ε: white noise
- Find the "promising" model:

$$Y = \beta_1^* X_1^* + \cdots + \beta_q^* X_q^* + \varepsilon.$$

- $n > p$ (Large n Small p)
- $p > n$ (Small n Large p)

Variable Selection Methods

- How to find the "promising" variables, $X_1^*, \ldots, X_q^*$?
  - Stepwise procedures (Forward, Backward, Stepwise)

Variable Selection Methods

- How to find the "promising" variables, $X_1^*, \ldots, X_q^*$?
  - Stepwise procedures (Forward, Backward, Stepwise)
  - Cross-validation method (CV)

Variable Selection Methods

- How to find the "promising" variables, $X_1^*, \ldots, X_q^*$?
  - Stepwise procedures (Forward, Backward, Stepwise)
  - Cross-validation method (CV)
  - Information criteria (AIC, BIC, $\ldots$)

Variable Selection Methods

- How to find the "promising" variables, $X_1^*, \ldots, X_q^*$?
  - Stepwise procedures (Forward, Backward, Stepwise)
  - Cross-validation method (CV)
  - Information criteria (AIC, BIC, ...)
  - Lasso, Lars, Bayesian Lasso

Variable Selection Methods

- How to find the "promising" variables, $X_1^*, \ldots, X_q^*$?
  - Stepwise procedures (Forward, Backward, Stepwise)
  - Cross-validation method (CV)
  - Information criteria (AIC, BIC, ...)
  - Lasso, Lars, Bayesian Lasso
  - Two-stage method (Screening and Selection)

Variable Selection Methods

- How to find the "promising" variables, $X_1^*, \ldots, X_q^*$?
  - Stepwise procedures (Forward, Backward, Stepwise)
  - Cross-validation method (CV)
  - Information criteria (AIC, BIC, ...)
  - Lasso, Lars, Bayesian Lasso
  - Two-stage method (Screening and Selection)
  - Stochastic Search Variable Selection

Bayesian Variable Selection

- Stochastic Search Variable Selection (SSVS)
  - George and McCulloch (1993, 1997)
  - Chipman (1996) and Chipman et al. (1997)
  - Smith and Kohn (1996)
- Applications:
  - Supersaturated design: Beattie et al. (2002)
  - Signal processing: Wolfe et al. (2004), and Févotte and Godsill (2006)
  - Gene selection: Lee et al. (2003)
  - ...

Statistical Model

- Model:

$$Y = \mathbf{X}\beta + \epsilon$$

  - Y is an $n \times 1$ response vector.
  - $\mathbf{X} = [X_1, \ldots, X_p]$ is an $n \times p$ model matrix, and X_i is the i-th predictor variable or regressor.
  - $\beta = (\beta_1, \ldots, \beta_p)'$ is a $p \times 1$ vector of the unknown coefficients.
  - $\epsilon = (\epsilon_1, \ldots, \epsilon_n)'$ is an $n \times 1$ noise vector that follows $MN(\mathbf{0}, \sigma^2 I_n)$
- A $p \times 1$ vector of latent variables, $\gamma = (\gamma_1, \ldots, \gamma_p)'$:

$$\gamma_i = \begin{cases} 1, & X_i \text{ is selected;} \\ 0, & \text{otherwise.} \end{cases}$$

Stochastic Search Variable Selection

- George and McCulloch (1993)
- Prior assumptions:
  - $(\beta_i, \gamma_i), i = 1, 2, \ldots, p$ are assumed to be independent.
  - γ: $P(\gamma_i = 0) = p_i$, and $P(\gamma_i = 1) = 1 - p_i$.
  - β:

    $$[\beta_i | \gamma_i = 0] \sim N(0, \nu_{0i}), \text{ and } [\beta_i | \gamma_i = 1] \sim N(0, \nu_{1i}).$$

    Usually set $\nu_{1i} = c_i \nu_{0i}$ and $c_i \gg 1$, i.e. $\nu_{1i} \gg \nu_{0i}$.
  - σ: $\sigma^2 \sim IG(\nu/2, \nu\lambda/2)$.

Stochastic Search Variable Selection

- Use Gibbs sampling scheme to sample from $[\beta, \sigma, \gamma | Y]$

Stochastic Search Variable Selection

- Use Gibbs sampling scheme to sample from $[\beta, \sigma, \gamma | Y]$
- Iteratively sample from $[\beta | \gamma, Y, \sigma]$, $[\gamma | \beta, Y, \sigma]$, and $[\sigma | Y, \beta, \gamma]$.

Stochastic Search Variable Selection

- Use Gibbs sampling scheme to sample from $[\beta, \sigma, \gamma | Y]$
- Iteratively sample from $[\beta|\gamma, Y, \sigma]$, $[\gamma|\beta, Y, \sigma]$, and $[\sigma|Y, \beta, \gamma]$.
- The best subset of variables is selected according to the Monte Carlo samples of γ.

Stochastic Search Variable Selection

- Use Gibbs sampling scheme to sample from $[\beta, \sigma, \gamma|Y]$
- Iteratively sample from $[\beta|\gamma, Y, \sigma]$, $[\gamma|\beta, Y, \sigma]$, and $[\sigma|Y, \beta, \gamma]$.
- The best subset of variables is selected according to the Monte Carlo samples of γ.
- The most costly step:

$$[\beta|\gamma, Y, \sigma] \sim N(\sigma^{-2} A_\gamma \mathbf{X}' Y, A_\gamma),$$

  - $A_\gamma = (\sigma^{-2}\mathbf{X}'\mathbf{X} + D_\gamma^{-1} R^{-1} D_\gamma^{-1})^{-1}$.
  - R is the prior correlation matrix.
  - $D_\gamma^{-2} = \text{diag}[(a_1\nu_{01})^{-1}, \ldots, (a_p\nu_{0p})^{-1}]$ with $a_i = 1$ if $\gamma_i = 0$, and $a_i = c_i$ if $\gamma_i = 1$.

Stochastic Search Variable Selection

- Speed up the stochastic search variable selection process:
  - Cholesky decomposition for A_γ.
  - Sample γ componentwise, i.e. $[\gamma_i|\gamma_{-i}, Y]$.
  - When $\nu_{0i} = 0$, Geweke (1996) suggested to jointly draw (γ_i, β_i).
  - Choose conjugate prior for β. Sample $[\gamma|Y]$ directly. (Smith and Kohn, 1996, and George and McCulloch, 1997)

Componentwise Gibbs Sampler

- The prior of β: $\beta_i|\gamma_i \sim (1-\gamma_i)\delta_0 + \gamma_i N(0, \tau_i^2)$.

Componentwise Gibbs Sampler

- The prior of β: $\beta_i|\gamma_i \sim (1-\gamma_i)\delta_0 + \gamma_i N(0, \tau_i^2)$.
- Sample (γ_i, β_i) one at time conditioning on $(\gamma_{-i}, \beta_{-i})$.

Componentwise Gibbs Sampler

- The prior of β: $\beta_i|\gamma_i \sim (1-\gamma_i)\delta_0 + \gamma_i N(0, \tau_i^2)$.
- Sample (γ_i, β_i) one at time conditioning on $(\gamma_{-i}, \beta_{-i})$.
- Assume $\tau_i = \tau$.

Componentwise Gibbs Sampler

- The prior of β: $\beta_i|\gamma_i \sim (1-\gamma_i)\delta_0 + \gamma_i N(0, \tau_i^2)$.
- Sample (γ_i, β_i) one at time conditioning on $(\gamma_{-i}, \beta_{-i})$.
- Assume $\tau_i = \tau$.
- The key step:

$$z_i = \frac{P(Y|\gamma_i = 1, \{\beta_k, \forall\ k \neq i\})}{P(Y|\gamma_i = 0, \{\beta_k, \forall\ k \neq i\})} = \sqrt{\frac{\sigma_{i\star}^2}{\tau^2}} \exp\left\{\frac{r_i^2}{2\sigma_{i\star}^2}\right\},$$

where $\sigma_{i\star}^2 = \dfrac{\sigma^2\tau^2}{X_i' X_i \tau^2 + \sigma^2}$, $r_i = \dfrac{R_i' X_i \tau^2}{\sigma^2 + X_i' X_i \tau^2}$, and $R_i = Y - \Sigma_{k\neq i}\beta_k X_k$.

Componentwise Gibbs Sampler

- The prior of β: $\beta_i|\gamma_i \sim (1-\gamma_i)\delta_0 + \gamma_i N(0, \tau_i^2)$.
- Sample (γ_i, β_i) one at time conditioning on $(\gamma_{-i}, \beta_{-i})$.
- Assume $\tau_i = \tau$.
- The key step:

$$z_i = \frac{P(Y|\gamma_i = 1, \{\beta_k, \forall\ k \neq i\})}{P(Y|\gamma_i = 0, \{\beta_k, \forall\ k \neq i\})} = \sqrt{\frac{\sigma_{i\star}^2}{\tau^2}} \exp\left\{\frac{r_i^2}{2\sigma_{i\star}^2}\right\},$$

where $\sigma_{i\star}^2 = \dfrac{\sigma^2\tau^2}{X_i'X_i\tau^2 + \sigma^2}$, $r_i = \dfrac{R_i'X_i\tau^2}{\sigma^2 + X_i'X_i\tau^2}$, and $R_i = Y - \Sigma_{k\neq i}\beta_k X_k$.

- Set $p_i = \rho$. Then

$$P(\gamma_i = 1|\{\beta_k, \forall\ k \neq i\}, Y) = \frac{(1-\rho)z_i}{\rho + (1-\rho)z_i}.$$

The componentwise Gibbs sampler for variable selection

(I) Randomly select a variable X_i.

(II) Compute

$$z_i = \frac{p(Y|\gamma_i = 1, \{\beta_k, \forall\ k \neq i\})}{p(Y|\gamma_i = 0, \{\beta_k, \forall\ k \neq i\})} = \sqrt{\sigma^2_{i\star}/\tau^2}\,\exp\{\frac{r_i^2}{2\sigma^2_{i\star}}\}.$$

Then evaluate the posterior probability

$$P(\gamma_i = 1|\{\beta_k, \forall\ k \neq i\}, Y) = \frac{(1-\rho)z_i}{\rho + (1-\rho)z_i}.$$

(III) Sample γ_i from the above posterior probability. If $\gamma_i = 0$, then set $\beta_i = 0$, otherwise, sample $\beta_i \sim N(r_i, \sigma^2_{i\star})$. Go back to (I).

(IV) After a number of iterations of the above steps, compute the current residual vector, $Res = Y - \sum_i \beta_i X_i$. Then sample $\sigma^2 \sim IG(\frac{n+\nu}{2}, \frac{Res' Res + \nu\lambda}{2})$. Go back to (I).

Componentwise Gibbs Sampler

- Similar to the search algorithm of Geweke (1996) with the truncated normal prior distribution.

Componentwise Gibbs Sampler

- Similar to the search algorithm of Geweke (1996) with the truncated normal prior distribution.
- Two possible problems:

Componentwise Gibbs Sampler

- Similar to the search algorithm of Geweke (1996) with the truncated normal prior distribution.
- Two possible problems:
  - The variables are highly correlated.

Componentwise Gibbs Sampler

- Similar to the search algorithm of Geweke (1996) with the truncated normal prior distribution.
- Two possible problems:
  - The variables are highly correlated.
  - The residual variance is small.

Matching Pursuit

- Matching Pursuit (Mallat and Zhang, 1993): Suppose $\|X_i\|^2 = 1$. At each iteration,
  - Select X_j such that
$$j = \arg\max |\langle R, X_i \rangle|.$$
  - Updated $\beta_i \leftarrow \beta_i + \langle R, X_i \rangle$, and $R \leftarrow R - \langle R, X_i \rangle X_i$.

Matching Pursuit

- Matching Pursuit (Mallat and Zhang, 1993): Suppose $\|X_i\|^2 = 1$. At each iteration,
  - Select X_j such that
    $$j = \arg\max |\langle R, X_i\rangle|.$$
  - Updated $\beta_i \leftarrow \beta_i + \langle R, X_i\rangle$, and $R \leftarrow R - \langle R, X_i\rangle X_i$.
- Forward selection

Metropolized Matching Pursuit

- Metropolis scheme with a pair of reversible moves: addition and deletion moves based on
  $z_i = P(Y|\gamma_i = 1, \{\beta_k, \forall\ k \neq i\})/P(Y|\gamma_i = 0, \{\beta_k, \forall\ k \neq i\})$.

Metropolized Matching Pursuit

- Metropolis scheme with a pair of reversible moves: addition and deletion moves based on
  $z_i = P(Y|\gamma_i = 1, \{\beta_k, \forall\ k \neq i\})/P(Y|\gamma_i = 0, \{\beta_k, \forall\ k \neq i\})$.
- The larger z_i is, the more promising the variable X_i is.

Metropolized Matching Pursuit

- Metropolis scheme with a pair of reversible moves: addition and deletion moves based on $z_i = P(Y|\gamma_i = 1, \{\beta_k, \forall\ k \neq i\})/P(Y|\gamma_i = 0, \{\beta_k, \forall\ k \neq i\})$.
- The larger z_i is, the more promising the variable X_i is.
- Proposal for the next status:

Metropolized Matching Pursuit

- Metropolis scheme with a pair of reversible moves: addition and deletion moves based on
  $z_i = P(Y|\gamma_i = 1, \{\beta_k, \forall\ k \neq i\})/P(Y|\gamma_i = 0, \{\beta_k, \forall\ k \neq i\})$.
- The larger z_i is, the more promising the variable X_i is.
- Proposal for the next status:
  - Add or delete a variable.

Metropolized Matching Pursuit

- Metropolis scheme with a pair of reversible moves: addition and deletion moves based on $z_i = P(Y|\gamma_i = 1, \{\beta_k, \forall\ k \neq i\})/P(Y|\gamma_i = 0, \{\beta_k, \forall\ k \neq i\})$.
- The larger z_i is, the more promising the variable X_i is.
- Proposal for the next status:
  - Add or delete a variable.
  - Addition proposal: Sample a inactive variable with probability proportional to z_i.

Metropolized Matching Pursuit

- Metropolis scheme with a pair of reversible moves: addition and deletion moves based on
  $z_i = P(Y|\gamma_i = 1, \{\beta_k, \forall\ k \neq i\})/P(Y|\gamma_i = 0, \{\beta_k, \forall\ k \neq i\})$.
- The larger z_i is, the more promising the variable X_i is.
- Proposal for the next status:
  - Add or delete a variable.
  - Addition proposal: Sample a inactive variable with probability proportional to z_i.
  - Deletion proposal: Randomly select one active variable.

- Acceptance probability for addition move:

$$
\begin{aligned}
& p_{\text{accept}-\text{add}} \\
= \quad & \min\left[1, \frac{P(\gamma_i = 1|\{\beta_k, \forall\ k \neq i\}, Y)}{P(\gamma_i = 0|\{\beta_k, \forall\ k \neq i\}, Y)} \frac{p_{\text{delete}}}{p_{\text{add}}} \frac{1/(A+1)}{z_i / \sum_{j:\gamma_j=0} z_j}\right] \\
= \quad & \min\left[1, \frac{(1-\rho)}{\rho} \frac{p_{\text{delete}}}{p_{\text{add}}} \frac{\sum_{j:\gamma_j=0} z_j}{(A+1)}\right]. \qquad (1)
\end{aligned}
$$

- Acceptance probability of the deletion move:

$$
\begin{aligned}
& p_{\text{accept}-\text{delete}} \\
= \quad & \min\left[1, \frac{P(\gamma_i = 0|\{\beta_k, \forall\ k \neq i\}, Y)}{P(\gamma_i = 1|\{\beta_k, \forall\ k \neq i\}, Y)} \frac{p_{\text{add}}}{p_{\text{delete}}} \frac{z_i/(\sum_{j:\gamma_j=0} z_j + z_i)}{1/A}\right] \\
= \quad & \min\left[1, \frac{\rho}{(1-\rho)} \frac{p_{\text{add}}}{p_{\text{delete}}} \frac{A}{\sum_{j:\gamma_j=0} z_j + z_i}\right]. \qquad (2)
\end{aligned}
$$

Stochastic matching pursuit for variable selection

(I) Let A be the number of active variables. With probability p_{add}, go to (II). With probability $p_{\text{delete}} = 1 - p_{\text{add}}$ go to (IV).

(II) With probability $p_{\text{accept}-\text{add}}$ calculated according to Eq. (1), go to (III), and with probability $1 - p_{\text{accept}-\text{add}}$ go back to (I).

(III) Among all the inactive variables i with $\gamma_i = 0$, sample a variable i with probability proportional to z_i, then let $\gamma_i = 1$ and sample β_i as described in (III) of Algorithm 1. Go back to (I).

Stochastic matching pursuit for variable selection

(IV) If $A > 0$, then randomly select an active variable i with $\gamma_i = 1$.

(V) With probability $p_{\text{accept}-\text{delete}}$ calculated according to Eq. (2), accept the proposal of deleting the variable i, i.e., set $\gamma_i = 0$, and $\beta_i = 0$. With probability $1 - p_{\text{accept}-\text{delete}}$, reject the proposal of deleting variable i, and sample β_i as described in (III) of Algorithm 1. Go back to (I).

(VI) After a number of iterations of the above steps, compute the current residual vector, $Res = Y - \sum_i \beta_i X_i$, and then update $\sigma^2 \sim IG(\frac{n+\nu}{2}, \frac{Res' Res + \nu\lambda}{2})$. Go back to (I).

- Combine the strengths of the matching pursuit and the componentwise Gibbs sampler.
  1. Pursue proposing variables.
  2. Don't need to compute the inverse of the large matrix.

Implementation Details

- After a burn-in period, use $\{\gamma^{(i)}, i > T\}$ to estimate $P(\gamma_j = 1|Y)$.
- Selection criteria:
  - The highest posterior probability: $\max P(\gamma_1, \ldots, \gamma_p|Y)$.
  - The median probability criterion in Barbieri and Berger (2004): X_i is included in the model if

$$P(\gamma_i = 1|Y) \geq 1/2.$$

Implementation Details

- Tuning parameters, p_{add} and ρ .

Implementation Details

- Tuning parameters, p_{add} and ρ .
  - Set $p_{add} = 1/2 = p_{delete}$.

Implementation Details

- Tuning parameters, p_{add} and ρ .
  - Set $p_{add} = 1/2 = p_{delete}$.
  - Set $\rho = 1/2$ (George and McCulloch, 1993 and 1997).

Implementation Details

- Tuning parameters, p_{add} and ρ .
  - Set $p_{add} = 1/2 = p_{delete}$.
  - Set $\rho = 1/2$ (George and McCulloch, 1993 and 1997).
- Another tuning parameter, τ

Implementation Details

- Tuning parameters, p_{add} and ρ .
  - Set $p_{add} = 1/2 = p_{delete}$.
  - Set $\rho = 1/2$ (George and McCulloch, 1993 and 1997).
- Another tuning parameter, τ
  - z_i is a decreasing function of τ. Then $P(\gamma_i = 1|\{\beta_k, k \neq i\}, Y)$ is smaller for larger τ.

Implementation Details

- Tuning parameters, p_{add} and ρ .
  - Set $p_{add} = 1/2 = p_{delete}$.
  - Set $\rho = 1/2$ (George and McCulloch, 1993 and 1997).
- Another tuning parameter, τ
  - z_i is a decreasing function of τ. Then $P(\gamma_i = 1|\{\beta_k, k \neq i\}, Y)$ is smaller for larger τ.
  - Cross-Validation approach for selecting τ:
    Use K-fold CV (or Monte Carlo CV) to choose "proper" value of τ. Thus

$$\widehat{\tau} = \arg\min_{\tau} \sum_{k=1}^{K} \sum_{j} (y_{kj} - \widehat{y}_{-kj}(\tau))^2.$$

Implementation Details

- Tuning parameters, p_{add} and ρ .
  - Set $p_{add} = 1/2 = p_{delete}$.
  - Set $\rho = 1/2$ (George and McCulloch, 1993 and 1997).
- Another tuning parameter, τ
  - z_i is a decreasing function of τ. Then $P(\gamma_i = 1|\{\beta_k, k \neq i\}, Y)$ is smaller for larger τ.
  - Cross-Validation approach for selecting τ: Use K-fold CV (or Monte Carlo CV) to choose "proper" value of τ. Thus

$$\widehat{\tau} = \arg\min_{\tau} \sum_{k=1}^{K} \sum_{j} (y_{kj} - \widehat{y}_{-kj}(\tau))^2.$$

- ρ can also be selected by this CV approach.

Large n Small p

Example 3.1: $(n, p) = (60, 5)$

- These five variables, $X_1, \ldots, X_5 \overset{\text{iid}}{\sim} N_{60}(\mathbf{0}, I_{60})$.
- The response variable is generated by

$$Y = X_4 + 1.2X_5 + \epsilon,$$

  where $\epsilon \sim N_{60}(\mathbf{0}, I_{60})$.
- Set $(\rho, \tau) = (0.5, 10)$ for SMP and set $(\nu_0, c) = (0.01, 2500)$ for SSVS.
- Totally there are 1000 replications. Draw 3000 samples from posterior samples.

Large n Small p

Table 1: Variable selection results in Example 3.1

method		Number of selected variables					
		0	1	2	3	4	5
SMP	f_1	0	0	997	3	0	0
	f_2	0	0	997	3	0	0
SSVS	f_1	0	0	997	3	0	0
	f_2	0	0	997	3	0	0

Large n small p

Example 3.2: $(n, p) = (60, 10)$

- 10 variables, $X_1, \ldots, X_{10} \overset{\text{iid}}{\sim} N_{60}(\mathbf{0}, I_{60})$.
- The true model is

$$Y = 2X_1 + 3X_2 + 4X_5 + 5X_6 + 6X_9 + 7X_{10} + \epsilon,$$

  where $\epsilon \sim N_{60}(0, 2.5^2 I_{60})$.
- Set $(\rho, \tau) = (0.5, 15)$ for SMP and set $(\nu_0, c) = (0.01, 2500)$ for SSVS.
- Totally there are 1000 replications. In each replication, draw 3000 samples.

Large n small p

Table 2: Variable selection results in Example 3.2

method		Number of selected variables					
		≤ 3	4	5	6	7	≥ 8
SMP	f_1	0	0	2	961	37	0
	f_2	0	0	0	961	37	0
SSVS	f_1	0	0	1	934	64	1
	f_2	0	0	0	934	64	1

Computational Cost

Table 3: CPU times (in seconds) of 10,000 iterations

	SMP	SSVS
CPU time in Example 3.1 ($p = 5$)	14.6s	9.3s
CPU time in Example 3.2 ($p = 10$)	39.5s	20.8s
CPU time with $p = 100$	2016.6s	7266.4s

Small n Large p

- The gene selection problem in microarray experiments: the number of candidate genes, $p >$ the number of available sample size, n. (Yi et al., 2003, and Lee et al., 2003)
- Overcomplete signal representation: the number of basis functions, $p >$ the size of the signal, n. (Wolf et al., 2004)
- Sparse assumption.

Simulations for Small n Large p Problem

- Shao and Chow (2007) studied the small n large p problem in microarray experiments.
- The ridge regression estimator for β is $\widehat{\beta} = (\mathbf{X}'\mathbf{X} + h_n I_p)^{-1}\mathbf{X}'Y = R_D\mathbf{X}'Y$,
  - I_p is the $p \times p$ identity matrix.
  - h_n is the ridge parameter.
  - $R_D = (\mathbf{X}'\mathbf{X} + h_n I_p)^{-1}$.
- Screen out X_i if $|\widehat{\beta}_i| \leq a_n$, and $a_n \to 0$ as $n \to \infty$.
- Their procedure is asymptotically consistent and their idea is similar to that of the Lasso method (Tibshirani, 1996).

Simulation

- $(n, p) = (50, 200)$ and $(100, 400)$.
- There are 5 true active variables, and

$$\beta = (3, -3.5, 4, -2.8, 3.2, 0, \ldots, 0)'.$$

- The regressor X_i is generated by

$$X_i = G_i + \lambda G,$$

  where G_i and $G \sim N_n(0, I_n)$, and $\lambda = 0$ or 1.
- $\epsilon \sim N_n(0, I_n)$.

Simulation

- Three methods are used here, Shao and Chow (2007), SMP and Lasso + CV.
- The screening method of Shao and Chow (2007): Set $h_n = n^{2/3}$ and $a_n = n^{-1/6}$.
- SMP:
  - τ is selected by 5-fold CV from $\{80, 120, 160, 220\}$ for $(n, p) = (50, 200)$ and from $\{100, 150, 200, 250\}$ for $(n, p) = (100, 400)$.
  - Draw 2000 posterior samples by taking every pth sample.

Simulation

- Lasso + CV: There is a Matlab implementation of the homotopy/LARS-LASSO algorithm for tracing the regularization path of the L1-penalized squared error loss (Rocha, 2006), and this tool-box is available at `http://www.stat.berkeley.edu/twiki/Research/YuGroup/Software`
- Stopping criterion:

$$\|\mathbf{X}^{'}(Y - \mathbf{X}'\widehat{\beta})\|_{\infty} < b.$$

- `lasso_cv`: Fit the parameters of a linear model by using the lasso and k-folds cross validation
- $b \in \{2 \times 10^{-1}, 10^{-1}, 10^{-2}, \ldots, 10^{-5}, 10^{-8}\}$.
- 10-folds CV is used here.
- Identify $\{i||\beta_i| > 0\}$.

Frequencies based on 100 replications with $(n, p) = (50, 200)$

λ	method		Number of selected variables										
			≤ 2	3	4	5	6	7	8	9	10	≥ 11	# of sel.
0	SC	f_1	1	4	6	14	14	20	18	9	10	4	
		f_2	0	0	0	1	1	10	13	9	9	3	46
	Lasso	f_1	0	0	0	17	5	10	14	6	4	44	
		f_2	0	0	0	17	5	10	14	6	4	44	100
	SMP	f_1	0	0	0	94	4	2	0	0	0	0	
		f_2	0	0	0	94	4	2	0	0	0	0	100
1	SC	f_1	1	6	10	21	10	19	13	12	4	4	
		f_2	0	0	0	2	2	9	8	6	4	4	35
	Lasso	f_1	0	0	0	0	1	0	1	4	8	86	
		f_2	0	0	0	0	1	0	1	4	8	86	100
	SMP	f_1	0	0	0	97	3	0	0	0	0	0	
		f_2	0	0	0	97	3	0	0	0	0	0	100

Frequencies based on 100 replications with $(n, p) = (100, 400)$

λ	method		Number of selected variables										
			$\leq$ 2	3	4	5	6	7	8	9	10	$\geq$ 11	# of sel.
0	SC	f_1	0	3	9	49	22	13	3	1	0	0	
		f_2	0	0	0	41	19	13	3	1	0	0	77
	Lasso	f_1	0	0	0	75	6	5	1	0	2	11	
		f_2	0	0	0	75	6	5	1	0	2	11	100
	SMP	f_1	0	0	0	95	5	0	0	0	0	0	
		f_2	0	0	0	95	5	0	0	0	0	0	100
1	SC	f_1	0	1	11	35	28	13	9	2	1	0	
		f_2	0	0	0	33	28	13	8	2	1	0	85
	Lasso	f_1	0	0	0	4	5	3	7	7	7	67	
		f_2	0	0	0	4	5	3	7	7	7	67	100
	SMP	f_1	0	0	0	98	2	0	0	0	0	0	
		f_2	0	0	0	98	2	0	0	0	0	0	100

An illustration in Image Representation

- Gabor regression model (Wolf et al., 2004) is

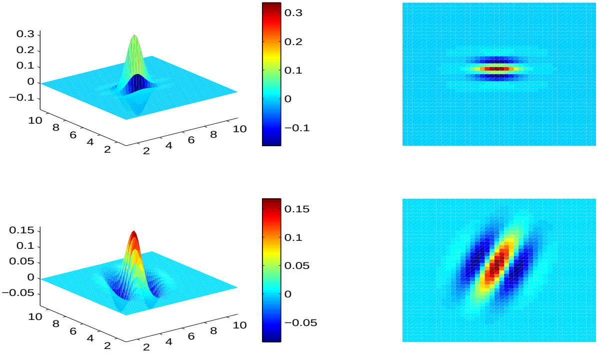
$$f = \sum_i c_i g_i + \varepsilon,$$

where g_i's are the Gabor basis functions.

- The Gabor basis function can be defined as

$$\begin{aligned} g(u,v) &= \exp\left[-\frac{1}{2}(\sigma_u u^2 + \sigma_v v^2)\right]\cos\left[\frac{2\pi u}{\lambda} + \varphi\right], \\ u &= u_0 + x_1\cos\theta - x_2\sin\theta, \\ v &= v_0 + x_1\sin\theta - x_2\cos\theta, \end{aligned}$$

Gabor Regression Function



- Give a grid $\mathcal{X} = \{(x_1, x_2) | x_1 \in \{22, 24, \ldots, 40\}$ and $x_2 \in \{7, 9, \ldots, 25\}\}$.
- Totally we have 200 Gabor basis functions on $\mathcal{X}$ by setting $\varphi = 0$, $\sigma_u = 1$ and $\theta \in \{0, 3/8\pi\}$.
- The response is generated by

$$Y = 7X_{17} - 7X_{71} + 7X_{161} - 7X_{177} + \varepsilon,$$

- SMP:
  - τ is chosen from $\mathcal{A} = \{50, 100, \ldots, 300\}$ by Monte Carlo cross validation with 100 replications.
  - Draw 3000 samples by taking every pth sample.

Selected Bases						
X_{17}	X_{71}	X_{73}	X_{161}	X_{177}	SNR1	SNR2
0.9957	0.5243	0.6317	0.6933	1.0000	0.373	0.080

Conjugate Prior for β

- Smith and Kohn (1996): The prior of β given γ is $N(\mathbf{0}, c\sigma^2(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1})$.
- Obtain $[\gamma|Y]$ by integrating β and σ^2 out.

$$P(\gamma|Y) \propto (1+c)^{-q_\gamma/2} S(\gamma)^{-n/2} \prod_{i=1}^{p} p_i^{\gamma_i}(1-p_i)^{1-\gamma_i},$$

where q_γ is the number of selected variables and

$$S(\gamma) = Y'Y - \frac{c}{1+c} Y'\mathbf{X}_\gamma(\mathbf{X}_\gamma'\mathbf{X}_\gamma)^{-1}\mathbf{X}_\gamma' Y.$$

- Use Gibbs sampler to generate $\gamma_i|Y, \gamma_{-i}$.
- Need to prespecify the prior parameter c. When the norm of X_i is equal to 1, $c \in [10, 1000]$.

- Simulations for the algorithm of Smith and Kohn (1996):
  - Fix $p_i = \rho = 1/2$.
  - Set $n = 50$ and $p = 20, 50, 100, 300$.
  - The variables, $X_1, \ldots, X_p \overset{\text{iid}}{\sim} N_n(\mathbf{0}, I_n)$.
  - The response variable is generated by

$$Y = 3X_1 + 3X_2 + \cdots + 3X_{10} + \epsilon,$$

    where $\epsilon \sim N_n(\mathbf{0}, I_n)$.
  - The median probability criterion
  - Selection results:

c	$p = 20$	$p = 50$	$p = 100$	$p = 300$
10	$\surd$	$\surd$	$\times$	
100	$\surd$	$\surd$	$\surd$	$\times$
1000			$\surd$	$\times$

  - SMP with $\tau = 250$ works.

Summarization

- Stochastic matching pursuit + median probability criterion works for both the cases of large n small p and small n large p.
- Tune the parameters, ρ and τ, via CV approach.
- "Full" Bayesian procedure
- CPU times: Componentwise Gibbs sampler $<$ SMP $<$ SSVS
- Window (or block) version
- Selection criterion
- Theoretical Properties
- Other applications

Analysis of Supersaturated Design

- Supersaturate design:
  - Investigates p factors in only $n(< p+1)$ experimental runs.
  - Particularly useful in factor screening.
- Analysis methods:
  - Lin (1993): Stepwise regression approach.
  - Chipman (1996) and Chipman et al. (1997): Propose different priors for SSVS.
  - Beattie et al. (2002): A two-stage method via SSVS.
  - Phoa et al. (2009): Dantzing selection method.

Analysis Approach

- Use componentwise Gibbs sampler:
  - The sample correlations between the factors are not so high.
  - The variance would not be too small.
- Follow the pre-process in Phoa et al. (2009), standardize Y and X_i's are unit norm.
- Use leave-two-out cross-validation approach to choose the proper parameters, ρ and τ.
- Selection criterion: the median probability criterion and the highest posterior probability criterion.

Example 1. Cast Fatigue Experiment

Run	A	B	C	D	E	F	G
1	+	+	−	+	+	+	−
2	+	−	+	+	+	−	−
3	−	+	+	+	−	−	−
4	+	+	+	−	−	−	+
5	+	+	−	−	−	+	−
6	+	−	−	−	+	−	+
7	−	−	−	+	−	+	+
8	−	−	+	−	+	+	−
9	−	+	−	+	+	−	+
10	+	−	+	+	−	+	+
11	−	+	+	−	+	+	+
12	−	−	−	−	−	−	−

Example 1. Cast Fatigue Experiment

- Consider main effect model. i.e. $n = 12$ and $p = 7$.
- Fix $\rho = 1/2$.
- Iterate $10000 \times p$ times and get 1000 samples from last $5000 \times p$ iterations.
- τ is select from $\mathcal{A} = \{1, 2, 3, 4, 5\}$. $\widehat{\tau} = 2$.
- The marginal posterior probabilities

Variable	A	B	C	D	E	F	G
Prob.	0.350	0.353	0.341	0.553	0.292	0.899	0.279

- Wu and Hamada (2000) and Phoa et al. (2009): [F (D)]

Example 1. Cast Fatigue Experiment

- Consider main effects + two-factor interactions. i.e. $n = 12$ and $p = 28$.
- τ is select from $\mathcal{A} = \{40, 80, 120, 160, 200\}$. $\widehat{\tau} = 120$.
- The marginal posterior probabilities

Variable	F	FG	AE	AC	BD	BC	AB
Prob.	0.763	0.759	0.129	0.015	0.014	0.014	0.014

- The highest posterior probability criterion: [F FG].
- Same as Phoa et al. (2009) by mAIC.

Example 2. Blood Glucose Experiment

- Sample size, $n = 18$.
- $p = 15$: 1 two-level factors, A, 7 three-level factors, $B, \ldots, H$ and 7 quadratic contrasts of these seven three-level factors, $B^2, \ldots, H^2$.
- τ is select from $\mathcal{A} = \{3, 4, 5, 6, 7, 8\}$. $\widehat{\tau} = 4$.
- The marginal posterior probabilities

Variable	F^2	E^2	C	B	G	F	A
Prob.	0.596	0.538	0.384	0.383	0.378	0.364	0.322

- Same as Wu and Hamada (2000) and Phoa et al. (2009)

Example 2. Blood Glucose Experiment

- Include two-factor interaction terms, $p = 113$.
- τ is select from $\mathcal{A} = \{40, 80, 120, 160, 200\}$. $\widehat{\tau} = 80$.
- The marginal posterior probabilities

Variable	BH^2	B^2H^2	EG	AH^2	DE	BC	DE^2
Prob.	0.821	0.748	0.578	0.496	0.154	0.147	0.145

- The highest posterior probability criterion:

Model	Post. Prob.	R^2
$AH^2\ BH^2\ EG\ B^2H^2$	0.116	0.9568
$BH^2\ B^2H^2$	0.027	0.7696
$BH^2\ EG\ B^2H^2$	0.018	0.8737
$AH^2\ BH^2\ EG\ B^2H^2\ E^2G^2$	0.017	0.9766

Example 3. An Example in Lin (1993)

- A supersaturated design with $n = 14$ and $p = 23$.
- Fix $\rho = 1/2$.
- Iterate 10000 times and get 1000 samples from last 5000 iterations.
- τ is select from $\mathcal{A} = \{20, 40, 60, 80, 100\}$. $\widehat{\tau} = 20$.
- The marginal posterior probabilities

Variable	14	12	19	4	10	11	15
Prob.	0.967	0.574	0.561	0.444	0.099	0.069	0.063

Example 3. An Example in Lin (1993)

- The highest posterior probability criterion:

Model	Post. Prob.	R^2
4 12 14 19	0.206	0.9548
14	0.133	0.6317
12 14 19	0.034	0.8706
12 14	0.031	0.7401
14 19	0.023	0.7225

- Li and Lin (2003): [4 12 14 19]
- Phoa et al. (2009): [14]

Future Works

- Apply SMP when p is large.

Future Works

- Apply SMP when p is large.
- Select (ρ, τ) via CV approach.

Future Works

- Apply SMP when p is large.
- Select (ρ, τ) via CV approach.
- Two-stage procedure via CGS: First screen out useless factors and then select the important factors.

Future Works

- Apply SMP when p is large.
- Select (ρ, τ) via CV approach.
- Two-stage procedure via CGS: First screen out useless factors and then select the important factors.
- Other examples

Future Works

- Apply SMP when p is large.
- Select (ρ, τ) via CV approach.
- Two-stage procedure via CGS: First screen out useless factors and then select the important factors.
- Other examples
- One-stage method or two-stage method?

Future Works

- Apply SMP when p is large.
- Select (ρ, τ) via CV approach.
- Two-stage procedure via CGS: First screen out useless factors and then select the important factors.
- Other examples
- One-stage method or two-stage method?
- The idea of Chipman (1996) and Chipman et al. (1997)

Example 3. An Example in Lin (1993)

- Main effects + Two-factor interaction effects
- Totally 252 variables (23 + 229)
- Fix $\rho = 1/2$.
- Iterate 10000 times and get 1000 samples from last 5000 iterations.
- τ is select from $\mathcal{A} = \{150, 170, 190, 210, 230\}$. $\widehat{\tau} = 170$.
- The marginal posterior probabilities

Var.	14	7×15	13×20	6×10	3×5	7×19	9×22
Prob.	0.367	0.133	0.116	0.059	0.057	0.056	0.053

Example 3. An Example in Lin (1993)

- Select the variables whose marginal probabilities > 0.04.
- Totally 21 variables.
- Fix $\rho = 1/2$.
- Iterate 10000 times and get 1000 samples from last 5000 iterations.
- τ is select from $\mathcal{A} = \{5, 10, 15, 20, 25\}$. $\widehat{\tau} = 5$.
- The marginal posterior probabilities

Var.	5×20	23	14	6×10	11	9×21	7×15
Prob.	0.593	0.551	0.548	0.542	0.47	0.45	0.314