Grace Yoon[1] / Wenxin Jiang[2] / Lei Liu[3] / Ya-Chen Tina Shih[4]

Simple Quasi-Bayes Approach for Modeling Mean Medical Costs

[1] Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX, USA, E-mail: gyoon@stat.tamu.edu. https://orcid.org/0000-0003-3263-1352.
[2] Department of Statistics, Northwestern University, Evanston, IL, USA
[3] Department of Biostatistics, Washington University in St. Louis, St. Louis, MO, USA
[4] Department of Health Services Research, MD Anderson Cancer Center, Houston, TX, USA

Abstract:
Several statistical issues associated with health care costs, such as heteroscedasticity and severe skewness, make it challenging to estimate or predict medical costs. When the interest is modeling the mean cost, it is desirable to make no assumption on the density function or higher order moments. Another challenge in developing cost prediction models is the presence of many covariates, making it necessary to apply variable selection methods to achieve a balance of prediction accuracy and model simplicity. We propose Spike-or-Slab priors for Bayesian variable selection based on asymptotic normal estimates of the full model parameters that are consistent as long as the assumption on the mean cost is satisfied. In addition, the scope of model searching can be reduced by ranking the Z-statistics. This method possesses four advantages simultaneously: *robust* (due to avoiding assumptions on the density function or higher order moments), *parsimonious* (feature of variable selection), *informative* (due to its Bayesian flavor, which can compare posterior probabilities of candidate models) and *efficient* (by reducing model searching scope with the use of Z-ranking). We apply this method to the Medical Expenditure Panel Survey dataset.

Keywords: Spike-or-Slab prior, variable selection, sandwich variance estimator, health econometrics

DOI: 10.1515/ijb-2018-0122

Received: January 5, 2018; **Accepted:** April 26, 2019

1 Introduction

Health care spending has grown faster than other sectors of the United States economy for decades. Total healthcare expenditures accounted for 17 % of Gross Domestic Product (GDP) in 2015, a percentage far exceeding all other Organization for Economic Co-operation and Development (OECD) countries. This percentage is projected to continue rising over the next decade [1].

Given the prominence of healthcare in the U.S. economy, predictive models of medical costs, especially mean medical costs, are of great interest to policy makers. But medical costs are often right-skewed and heteroscedastic, making statistical analysis challenging. Logarithmic transformation has been often proposed, but requires modeling on the transformed scale, and heteroscedasticity may bias estimates derived from logged response variable models [2, 3]. It is therefore desirable to estimate $\log E(Y)$ instead of $E(\log Y)$ to obtain the mean medical cost easily, and to consider robust statistical methods by avoiding unnecessary assumptions. For example, Chen, Liu, Zhang, and Shih [4] and Chen, Liu, Zhang, Shih, and Severini [5] developed models using log link and a fixed set of predictors, under assumptions only on the mean cost without additional restrictive assumptions on higher order moments.

Variable selection is critically important for cost prediction models because policy makers wish to know which variables "drive" medical costs, and favor simple models that are easy to interpret and estimate - usually models with fewer predictors.

Statistical methods of variable selections can be performed using either Bayesian or frequentist approaches. Bayesian approaches are more informative than frequentist counterparts, as they can provide posterior probability ratios for selected models, stating that the top model A is five times more likely to be the true model than the next model B, etc., which cannot be done by frequentist approaches. While the Bayesian approach of variable selection is appealing, it is computationally challenging. Specifically, Bayesian model selection methods such as BIC or Bayes Factor require assumptions on the true likelihood function. Since the posterior in

Grace Yoon is the corresponding author.

a Bayesian approach is proportional to the product of the likelihood and the prior, it is seemingly impossible to conduct Bayesian analysis for robust methods without likelihood functions. To resolve this problem, we use an asymptotic likelihood based on the asymptotic normality of a parameter estimate instead of the original data and apply the robust sandwich estimate of the asymptotic variance, so that Bayesian inference based on this asymptotic normal likelihood function is valid under the assumption of the mean model. We employ a "quasi-Bayesian" approach from the econometrics literature (see, e. g. Chernozhukov and Hong [6], Inoue and Shintani [7]), that is less known in statistical applications. When multiplied by a prior density that either proposes a "spike" or a "slab" for the parameter components, we can form such an asymptotic posterior and perform Bayesian variable selection.

The method developed in this paper has four goals: (i) robustness to skewness and heteroscedasticity with only one assumption on the mean model; (ii) parsimonious variable selection; (iii) generation of posterior probabilities of candidate models; (iv) computational efficiency by avoiding either numerical or Monte Carlo integration, and limiting model search efforts. Our proposed method simplifies search for the model with the highest posterior probability and achieve computing efficiency based on ranking the Z-statistics, following Jiang and Liu [8]. In addition, with the Spike-or-Slab priors, the integration of the posterior computation is exact and no iteration is needed. Although recent work by Li and Jiang [9] addressed these three goals (i), (ii) and (iii) simultaneously, their method is computationally cumbersome because it involves a much more complicated posterior function that needs to be simulated by Markov chain Monte Carlo (MCMC) methods, which is no longer necessary in our model. Jiang and Liu [8] used the same idea of posterior inference based on parameter estimates as in this paper, but they did not do exact integration for the posterior probability of each candidate model as what we will do here. In addition, they did not use spike or slab prior densities to describe the prior belief of either small or large regression coefficients.

LASSO (Least Absolute Shrinkage and Selection Operator) and sslasso other commonly used variable selection methods, but neither achieves all our goals. LASSO is a frequentist approach and thus does not provide posterior probabilities of the candidate models. Ročková and George [10] proposed the spike-and-slab lasso and Tang, Shen, Zhang, and Yi [11] extended it to generalized linear models, termed sslasso, which could be applied to model the medical costs nonlinearly against the regressors. Their method is implemented in R package `BhGLM`. sslasso, as our method, is a Bayesian method with spike and slab priors. However, it does not use a robust estimate of the asymptotic variance, and therefore cannot guarantee that the reported posterior probabilities are asymptotically valid from the Bayesian perspective, when only the mean model is correctly specified. We will compare our method empirically to these alternative variable selection methods.

Our motivating example uses the Medical Expenditure Panel Survey (MEPS) dataset. We include 3,376 individuals aged over 65 from the MEPS 2014 Full Year Consolidated Data File. The mean medical cost in 2014 is \$10,321 and the median is \$4,394.50, illustrative of the right skewness. As a large scale survey of health, health services use and associated costs of noninstitutionalized civilians in the United States, the MEPS included many variables, and is thus well suited example on which to compare the abilities of variable selection methods to develop parsimonious cost prediction models.

The paper is organized as follows. Details about our proposed method are provided in Section 2. We compare our method's estimation performance to other alternatives in Section 3. In Section 4, we demonstrate our method on the MEPS dataset. We discuss the results and draw conclusions in Section 5.

2 Methods

2.1 Model

Consider a log-linear mean model without any distribution assumption for a response variable. Let Y_i denote the medical cost and X_i a p-dimensional covariate vector for subject i, we assume a log linear model for the mean response:

$$\begin{aligned} E(Y_i|X_i) = \mu(X_i, \boldsymbol{\beta}) &= e^{X_i^T\boldsymbol{\beta}} \qquad \text{for } i = 1, \dots, n, \\ X_i &= (1, x_{i1}, x_{i2}, \dots, x_{ip})^T, \\ \boldsymbol{\beta} &= (\beta_0, \beta_1, \beta_2, \dots, \beta_p)^T, \end{aligned} \tag{1}$$

where β_j's are unknown coefficients. We are interested in selecting important covariates to construct a predictive model for Y with the highest posterior probability. If the coefficient β_j is not negligible, then the jth variable is included in the model. Thus, the variable selection problem corresponds to finding which β_j's are nonzero.

We propose a Spike-or-Slab (SorS) prior for $\boldsymbol{\beta}$ as follows:

$$\boldsymbol{\beta} \sim N(0, U) \text{ where } U = \text{diag}(u_0, u_1, \ldots, u_p), \tag{2}$$

where u_j is the prior variance of β_j, taking either a prespecified small value (spike variance), or a prespecified large value (slab variance). The specification of these small or large variance values will be described later. This strategy can approximately select component β_j if u_j is large, and neglect β_j if u_j is small. Let u_0 be large so we always have the intercept in the model. Since U contains the information about which variables are included in the model, we simply treat U as a model index by examining which components u_j's take the larger values.

Now suppose that we have an asymptotic normal parameter estimate $\hat{\boldsymbol{\beta}}$ of the parameter $\boldsymbol{\beta}$, so that $\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, U \sim N(\boldsymbol{\beta}, V)$. This, together with a normal prior $\boldsymbol{\beta}|U \sim N(0, U)$ from (2), implies that $\hat{\boldsymbol{\beta}}|U \sim N(0, U+V)$. Then the marginal likelihood $p(\hat{\boldsymbol{\beta}}|U)$ of the model index U has an explicit expression:

$$p(\hat{\boldsymbol{\beta}}|U) = \frac{1}{\sqrt{\det(2\pi(U+V))}} \exp\{-\frac{1}{2}\hat{\boldsymbol{\beta}}^T (U+V)^{-1}\hat{\boldsymbol{\beta}}\}. \tag{3}$$

The posterior distribution for model U can be obtained from

$$p(U|\hat{\boldsymbol{\beta}}) \propto p(\hat{\boldsymbol{\beta}}|U)p(U).$$

Here $p(U)$ is the prior for model U, and for simplicity, we first suppose that all models have the equal prior probabilities. Then we can find the model with the highest posterior probability.

Next, we choose the asymptotic normal estimate $\hat{\boldsymbol{\beta}}$ to be the maximizer of a *naive* likelihood based on Poisson-like regression of the Y_i's against the full model of X_i's. Note that $\hat{\boldsymbol{\beta}}$ can be obtained by taking the derivative of *naive* Poisson log likelihood and setting it to $\mathbf{0}$:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} X_i \left(Y_i - e^{X_i^T\boldsymbol{\beta}}\right) = \mathbf{0}.$$

This is an unbiased estimating equation, i. e. $E\left[S(\boldsymbol{\beta})|\boldsymbol{\beta}\right] = 0$, as long as the mean model (1) is correct. Therefore, $\hat{\boldsymbol{\beta}}$ is asymptotic normal for the true parameter $\boldsymbol{\beta}$ of interest, even if the true model of Y_i is NOT Poisson, and we can use this $\hat{\boldsymbol{\beta}}$ to compute the posterior probability of any candidate model U explicitly, using (3). Obviously, such asymptotic normal estimate $\hat{\boldsymbol{\beta}}$ could be chosen by other unbiased estimating equations. However, a major advantage of the Poisson estimating equation is that the *naive* Poisson likelihood is typically globally concave, and so the maximizer $\hat{\boldsymbol{\beta}}$ is unique and easy to compute.

Of note, since the Poisson model is a *naive* probability model, the asymptotic variance V of $\hat{\boldsymbol{\beta}}$ should be estimated by a robust sandwich formula (see, e. g. White [12]).

Now we describe how to specify the spike or slab variance for the prior variance $u_j = \text{var}(\beta_j)$. We wish to do it in a scale invariant way, e. g. to stipulate that the variance ratio u_j/V_{jj} takes either a small value a (spike variance) or a large value A (slab variance), where V_{jj} is the jth diagonal element of $V = \text{var}(\hat{\boldsymbol{\beta}})$. This can approximately select components β_j if $u_j/V_{jj} = A$, and neglect β_j if $u_j/V_{jj} = a$. a and A work as tuning parameters in regularized methods, such as ridge or LASSO. We fix A to be relatively large to avoid unnecessary penalization on selected coefficients. Setting a to be a small but positive value, rather than zero, helps to absorb negligible nonzero coefficients into the spike distribution. The values a and A are not sensitive from our numerical experience. Appendix A heuristically verifies that this method is consistent in selecting true variables under a wide range of choices of a and A. In practice, we can choose a using grid search from the cross-validation method to minimize Root Mean Square Error (RMSE) and A as n which is the sample size. In simulation and real data application to the MEPS data, we performed 10-fold cross-validation to select an optimal tuning parameter a based on the RMSE.

2.2 Model selection procedure

For fixed values of a and A, we now describe how to do variable selection, i. e. how to assign them to the p components of the prior variance U, in order to maximize the posterior probability. We do the following:

Step 1. Calculate Z-statistics for each variable using $\hat{\boldsymbol{\beta}}$ and the sandwich variance estimate from a Poisson unbiased estimating equation based on the full model. For $j = 1, \ldots, p$, we define Z-statistics as

$$Z_j = \hat{\beta}_j / \sqrt{\hat{V}_{jj}}$$

where $\hat{\beta}_j$ is an MLE of the *naive* Poisson regression coefficient β_j and $\hat{V}_{jj}$ is the j-th diagonal element of the sandwich formula variance estimate $\hat{V}$:

$$\hat{V} = \left(\sum_{i=1}^{n} X_i X_i^T e^{X_i^T \hat{\boldsymbol{\beta}}}\right)^{-1} \left(\sum_{i=1}^{n} X_i \left(Y_i - e^{X_i^T \hat{\boldsymbol{\beta}}}\right)\left(Y_i - e^{X_i^T \hat{\boldsymbol{\beta}}}\right)^T X_i^T\right)\left(\sum_{i=1}^{n} X_i X_i^T e^{X_i^T \hat{\boldsymbol{\beta}}}\right)^{-1}.$$

Step 2. Rank all p variables by the absolute values of Z-statistics.

Step 3. Compare the posterior probability of p different candidate models in Z-scope. The Z-scope can be denoted as

$$\Phi_Z = \left\{\mathcal{M}_{(1)}, \mathcal{M}_{(2)}, \ldots, \mathcal{M}_{(p)}\right\},$$

where $\mathcal{M}_{(j)} = \left\{k : k \in \{1, \ldots, p\}, |Z_k| \geq |Z_{(j)}|\right\}$ for each $j = 1,\ldots,p$, Z_k is the usual Z-statistic of the variable X_k, and $Z_{(j)}$ is the jth largest of the Z_k's. Among 2^p candidate models, ranking Z-statistics can reduce the scope of model searching to p candidate models, without losing the true model asymptotically, see Zheng and Loh [13] and Jiang and Liu [8]. Then, each candidate model $\mathcal{M}_{(j)}$ assigns $u_k = A\hat{V}_{kk}$ to the top j variables of $|Z_k| = \left|\hat{\beta}_k / \sqrt{\hat{V}_{kk}}\right|$ and $u_k = a\hat{V}_{kk}$ to the other $p - j$ variables. As an example, consider the case that we have three variables X_1, X_2, X_3 and assume that the ranks are X_3, X_1, X_2 based on the size of Z-statistics. Then, $\mathcal{M}_{(1)} = \{X_3\}$, $\mathcal{M}_{(2)} = \{X_1, X_3\}$, $\mathcal{M}_{(3)} = \{X_1, X_2, X_3\}$ and we assign the prior variance as below.

$$\begin{aligned}
&\text{For } \mathcal{M}_{(1)}, \quad u_0 = AV_{00},\ u_1 = aV_{11},\ u_2 = aV_{22} \text{ and } u_3 = AV_{33}.\\
&\text{For } \mathcal{M}_{(2)}, \quad u_0 = AV_{00},\ u_1 = AV_{11},\ u_2 = aV_{22} \text{ and } u_3 = AV_{33}.\\
&\text{For } \mathcal{M}_{(3)}, \quad u_0 = AV_{00},\ u_1 = AV_{11},\ u_2 = AV_{22} \text{ and } u_3 = AV_{33}.
\end{aligned}$$

Then $p(U|\hat{\boldsymbol{\beta}})$ is computed and compared for each of the p candidate models using (2), we can identify the best model with the highest posterior probability $p(U|\hat{\boldsymbol{\beta}})$. Since we have an explicit formula for posterior probability, calculating posterior probability for each $\mathcal{M}_{(1)}, \ldots, \mathcal{M}_{(p)}$ is fast and straightforward.

3 Simulation study

We conduct simulation studies to assess the performance of our proposed method (SorS). The response Y are generated from Gamma distribution where Gamma(a,b) density function is $f(y) = \left[b^a/\Gamma(a)\right] y^{a-1}e^{-by}$. In our application study, we have many binary variables as covariates, so we only consider binary variables for X in the simulation. Correlated binary variables are generated using R package `bindata`. In all cases, we generated $R = 100$ datasets, each with sample size $n = 1000$ and $p = 50$ where $\mu_i = e^{X_i\boldsymbol{\beta}}$. The true parameter coefficient has an intercept and four nonzero components ($p_0 = 4$) as below:

$$\boldsymbol{\beta} = (2, 2, -2, 2, -2, 0, \ldots, 0).$$

Case 1. $Y_i \sim \text{Gamma}(\mu_i, 1)$, that is, $E(Y_i) = \mu_i$, $V(Y_i) = \mu_i$.

(1) (Independent predictors) $\mathbf{x}_1, \ldots, \mathbf{x}_p$'s are iid from Bernoulli(0.5).

(2) (Correlated predictors) $\mathbf{x}_1, \ldots, \mathbf{x}_p \sim$ Bernoulli(0.5) with $\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = 0.5^{|j-k|}$ for $1 \leq j,k \leq p$.

Case 2. $Y_i \sim \text{Gamma}(1, 1/\mu_i)$, that is, $E(Y_i) = \mu_i$, $V(Y_i) = \mu_i^2$.

(1) (Independent predictors) $\mathbf{x}_1, \ldots, \mathbf{x}_p$'s are iid from Bernoulli(0.5).

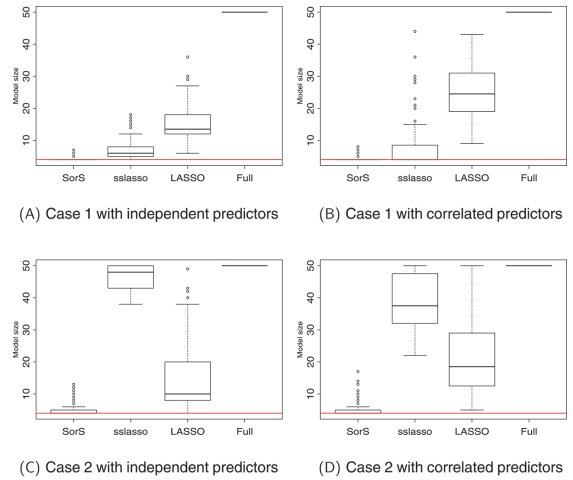
(2) (Correlated predictors) $\mathbf{x}_1, \ldots, \mathbf{x}_p \sim$ Bernoulli(0.5) with $\text{corr}(\mathbf{x}_j, \mathbf{x}_k) = 0.5^{|j-k|}$ for $1 \leq j,k \leq p$.

We compare SorS with three different methods: sslasso, LASSO and full model without variable selection. For estimation, we would like to compare RMSE $= \sqrt{\sum_{i=1}^{n}(Y_i^* - \hat{Y}(X_i^*))^2/n}$ where $\hat{Y}(X_i^*) = e^{X_i^*\hat{\boldsymbol{\beta}}}$ and the test data (X_i^*, Y_i^*), $i = 1,\ldots,n = 1000$ are generated independently from the same distribution as the training set. In addition, we will also present the following criteria to evaluate the performance of these methods. The comparison results are shown in the following figures and table.

- RMSE of coefficient estimates (bRMSE) $= \sqrt{\sum_{j=0}^{p}\left(\beta_j - \hat{\beta}_j\right)^2/(p+1)}$.
- Selected model size.
- In the table,
  - Coverage probability (Cov)= $\sum_{r=1}^{R} I(\mathscr{M}^* \subset \hat{\mathscr{M}}^{(r)})/R$,
  - True negative rate (TNR)= $\sum_{r=1}^{R}\sum_{j=1}^{p} I(\hat{\beta}_j^{(r)} = 0, \beta_j = 0)/[R(p-p_0)]$,
  - False negative rate (FNR)= $\sum_{r=1}^{R}\sum_{j=1}^{p} I(\hat{\beta}_j^{(r)} = 0, \beta_j \neq 0)/[Rp_0]$,
  - Exact selection probability (Ext)= $\sum_{r=1}^{R} I(\mathscr{M}^* = \hat{\mathscr{M}}^{(r)})/R$,

  where $\mathscr{M}^*$ and $\hat{\mathscr{M}}^{(r)}$ denote a true model and a selected model at r th generated dataset, respectively.
- Average accuracy rate of variable selection (Acc) $= \sum_{j=1}^{p} I(\hat{\gamma}_j = \gamma_j)/p$, where $\gamma_j = I(\beta_j \neq 0)$ and $\hat{\gamma}_j = I(\hat{\beta}_j \neq 0)$.

In Case 1 with moderate heteroscedasticity, variance is the same as mean. The median model size selected by LASSO is 15 as shown in Figure 1A, which implies a lot of false positives. For the case without heteroscedasticity (not shown here), LASSO works well. As is well known, LASSO is very fast and works better with independent predictors. The estimated coefficient $\hat{\boldsymbol{\beta}}$ from LASSO is re-fitted without shrinkage using only selected variables. Although sslasso is fast and still works well in the correlated setting, the weakness of sslasso is its incapability to deal with heteroscedasticity. As heteroscedasticity becomes more severe in Case 2 where variance is the square of the mean, SorS shows its advantage. Not only has SorS the smallest RMSE in Figure 2, but also SorS is the most parsimonious method and includes only a few false positives as seen in Figure 1. In Case 2 when heteroscedasticity is severe, its average accuracy is the highest as 0.97 and still can identify the true model more than half times in both independent and correlated covariate settings as shown in Table 1. In summary, SorS performs better than the other comparable methods in terms of selection (Figure 1), prediction (Figure 2) and estimation performance (see Figure 3). All method is implemented in R and the source code is available in https://github.com/GraceYoon/SorS.



(A) Case 1 with independent predictors (B) Case 1 with correlated predictors

(C) Case 2 with independent predictors (D) Case 2 with correlated predictors

Figure 1: Boxplots of selected model sizes over 100 replications. The horizontal lines indicate the true model size, which is 4.

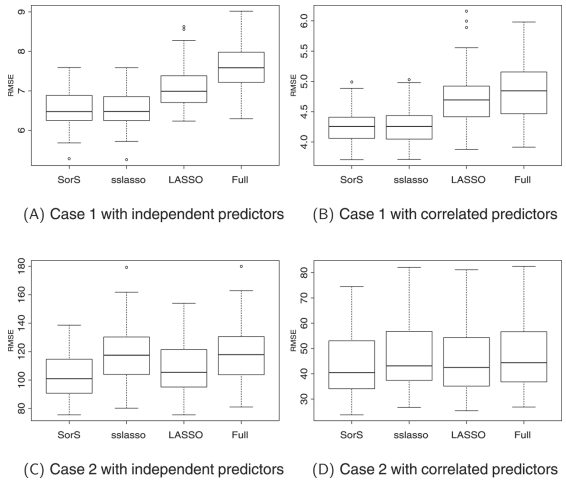


Figure 2: Boxplots illustrating RMSE for test data in all cases.

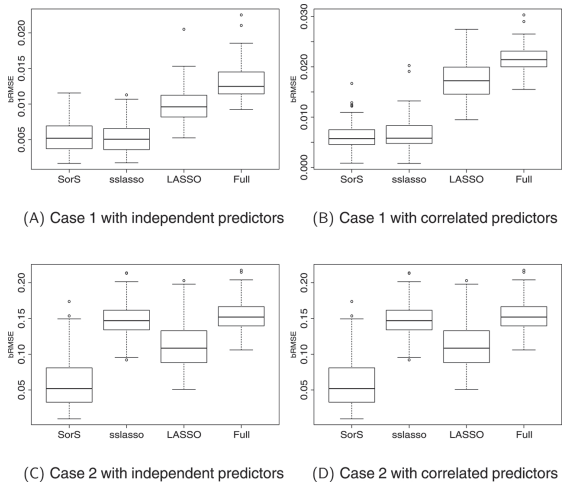


Figure 3: RMSE of coefficient estimates in each case from 100 replications.

Table 1: Simulation results.

		Oracle	Cov 1	TNR 1	FNR 0	Ext 1	Acc 1
Case 1 $\text{var}(Y_i) = \mu_i$	independent predictors	Full	1	0	0	0	0.08
		LASSO	1	0.7630	0	0.01	0.7820
		sslasso	1	0.9283	0	0.14	0.9340
		SorS	1	0.9939	0	0.81	0.9944
	correlated predictors	Full	1	0	0	0	0.08
		LASSO	1	0.5667	0	0	0.6014
		sslasso	1	0.9139	0	0.62	0.9208
		SorS	1	0.9959	0	0.87	0.9962
Case 2 $\text{var}(Y_i) = \mu_i^2$	independent predictors	Full	1	0	0	0	0.08
		LASSO	1	0.7835	0	0.01	0.8008
		sslasso	1	0.0541	0	0	0.1298
		SorS	1	0.9824	0	0.68	0.9838
	correlated predictors	Full	1	0	0	0	0.08
		LASSO	1	0.6404	0	0.01	0.6692
		sslasso	1	0.1333	0	0	0.2026
		SorS	1	0.9741	0	0.61	0.9762

Our proposed method is based on estimating a Poisson estimating equation with the log-linear mean specification. Our method is robust against violation of the Poisson-type variance assumption, as is shown in our simulation that allows the variance to be unequal to the mean. As is typical in the biostatistical applications of the estimating equations, the proposed method is NOT intended to be robust against mean specification. For

example, for a true model $Y = \exp(X^2)$ on $X \sim \text{Unif}(-1, 1)$, fitting a Poisson model with log-linear mean in X would obtain a constant fit, suggesting (incorrectly) that $E(Y|X) =$ constant, independent of X.

Specifically, for medical cost data, generalized linear models (GLM) with log link have been used most often to describe the skewed and heteroscedastic medical cost data in the original scale [3, 14–18]. In this paper, we have not considered other type of link functions. However, more flexible link functions, e. g. Box-Cox transformation [19], could be considered in the future work. Another option is to add nonlinear functions for covariate effects, e. g. through polynomial functions or splines [4], to accommodate the curvature and account for mean misspecification.

4 Application to MEPS data

We construct an analytical sample from the 2014 MEPS Full Year Consolidated Data File, including every subject ≥ 65 years old in the survey year, and extract a total of 33 variables to account for the impact of each individual's disease(s), symptom(s) or any related conditions on medical costs. These variables are listed in Appendix B. Most are self-explanatory. EDUCAT, the highest degree of education in 2014, is coded as an integer from 1 to 3, for < 13 years, 13–16 years, or > 16 years, respectively. POVCAT is coded as: 1: poor, 2: nearly poor, 3: low income, 4: middle income, 5: high income. Other continuous variables are standardized in our analyses. Correlations of covariates range from -0.55 to 0.6. The final study sample includes 3,376 individuals. As noted, the response variable in our model, annual medical costs in U.S. dollars, is highly right-skewed; 3.6 % of subjects had no medical cost, which our method accommodates since the *naive* Poisson likelihood supports 0 values.

We apply Spike-or-Slab normal priors for the parameters a uniform prior for model selection. Posterior probabilities are obtained from the exact integration as described in Section 2.2. To assess the performance of various variable selection methods, we randomly sample half the observations to form a training set and use the remaining as test data to compute the RMSE, repeating this 100 times. The results are summarized in Figure 4. RMSE values for SorS method are close to those from other methods, but selected model sizes are less than half of others.

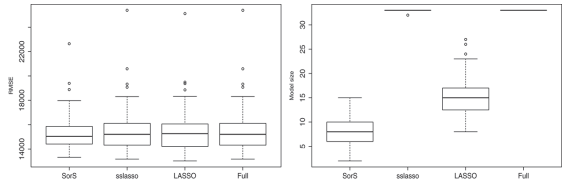


Figure 4: Prediction performance and selected model size by different methods in MEPS data analysis.

Applying the SorS method to all the data, the model with 10 variables has the largest posterior probability. The candidate model with 11 variables has the second highest posterior probability and is 21.21 % as likely as the model with 10 variables. Using 10 variables in the highest posterior probability model, the estimated coefficients are shown in Table 2.

Table 2: Estimates of coefficients and sandwich standard errors for MEPS data.

	Estimate	s.e.	p-value
(Intercept)	5.9330	0.2520	<0.0001
HOSPEXP	1.2325	0.0581	<0.0001
INSCOV	2.1102	0.2474	<0.0001
DIABETES	0.2707	0.0501	<0.0001
PCS	–0.1561	0.0270	<0.0001
ANYLMT	0.3280	0.0636	<0.0001
EMERG	0.2353	0.0559	<0.0001
CANCER	0.2060	0.0536	0.0001
STRK	0.2127	0.0654	0.0011
CORHRT	0.1519	0.0538	0.0048
EDUCAT	0.1487	0.0352	<0.0001

Results from the best performing model (with 10 variables) show that hospitalization (HOSPEXP) and emergency room visit (EMERG) increase the medical cost by a large percentage (3.4 times and 1.3 times, respectively). Heart and blood vessel disease (CORHRT and STRK), body movement disorder (ANYLMT), cancer and dia-

betes are all significantly associated with annual medical costs. Gender and race variables are not selected in our model.

Expected medical costs are higher for the higher educated: more educated individuals may be more likely to have regular checkups and better able and prone to spend more for treatment. Physical Composite Scores (PCS), a quality of life measure, is inversely associated with medical costs.

5 Discussion

In this work, we propose a new quasi-Bayesian variable selection method with Spike-or-Slab (SorS) priors to simultaneously accommodate severe skewness and heteroscedasticity of the medical cost data without any transformation. We integrate four features: log-linear mean model; spike and slab prior; ranking by Z-statistics; and posterior computation without MCMC. The novelty of our work lies in the innovative integration and application to skewed and heteroscedastic data. Our proposed method is applicable to many situations where the likelihood function is not available. In addition, our method uses the robust sandwich formula for constructing the posterior, and is therefore more robust in the sense that the asymptotic validity of the reported posterior probabilities depends only on the correct specification of the mean model without any additional assumptions on the variance structure or density function.

This method however is limited to low-dimensional data. Our asymptotic likelihood is based on asymptotic normality of the parameter estimates which relies on the standard large sample theory and ranking based on Z-statistics method cannot be implemented when the dimension of covariates is higher than the sample size. In high-dimensional case, sslasso works and is computationally feasible, but it performs poorly when the heteroscedasticity level is severe. It will be an interesting area of future research to consider robust Bayes variable selection for higher dimensional data with $p > n$.

Funding

This work was supported by the Agency for Healthcare Research and Quality (Grant Number: R01 HS 020263, Funder Id: http://dx.doi.org/10.13039/100000133) and National Cancer Institute (Grant Number: T32-CA090301, Funder Id: http://dx.doi.org/10.13039/100000054).

Appendix

A Heuristic verification for model selection consistency in a simple example

We have $\boldsymbol{\beta} \sim N(0, U)$, $\hat{\boldsymbol{\beta}}|\boldsymbol{\beta}, U \sim N(\boldsymbol{\beta}, V)$ and $\hat{\boldsymbol{\beta}}|U \sim N(0, U+V)$. Define $U = \mathrm{diag}(u_1, \ldots, u_p)$ where $u_j = V_{jj}\rho_j$, $\rho_j \in \{a, A\}$. We will consider a simple special case and verify that there is a wide range of choices of a small a and a large A, so that the proposed method of variable selection will be consistent in the frequentist sense. In the special case, V is a diagonal matrix, which typically has elements of order $\frac{1}{n}$, since V is an asymptotic variance based on n iid observations. We would like to choose the model U which maximizes the posterior probability:

$$\begin{aligned} p(U|\hat{\boldsymbol{\beta}}) \propto & p(\hat{\boldsymbol{\beta}}|U)p(U) \\ \propto & \prod_{j=1}^{p} \exp\left[-\frac{1}{2}\left(\frac{\hat{\beta}_j^2}{V_{jj}(1+\rho_j)} + \log\left(2\pi V_{jj}(1+\rho_j)\right)\right)\right] \equiv \prod_{j=1}^{p} \pi_j(\rho_j). \end{aligned}$$

The optimization can be done for each j separately due to the product form. For each j, we find $\rho_j \in \{a, A\}$ to maximize the jth factor $\pi_j(\rho_j)$ in the product. Suppose $\rho_j = A$ gives a bigger jth factor $\pi_j(\rho_j)$ than $\rho_j = a$, then the posterior maximization implies that the prior variance parameter is large, i. e. $\mathrm{var}(\beta_j) = u_j = V_{jj}\rho_j = V_{jj}A$. In other words, β_j comes from a "slab" prior and we select the variable X_j.

Let $z_j = \dfrac{\hat{\beta}_j^2}{V_{jj}}$. Then $\pi_j(\rho_j = A) > \pi_j(\rho_j = a)$ (and we select the variable X_j) if and only if

$$\begin{aligned}\frac{z_j^2}{1+A} + \log(1+A) &< \frac{z_j^2}{1+a} + \log(1+a)\\ \frac{z_j^2}{1+a} - \frac{z_j^2}{1+A} &> \log\frac{1+A}{1+a}\\ z_j^2 &> \frac{\log(1+a)^{-1} - \log(1+A)^{-1}}{(1+a)^{-1} - (1+A)^{-1}} \equiv \Delta.\end{aligned}$$

For two stochastic sequences $\{a_n\}$ and $\{b_n\}$, let $a_n \prec b_n$, $a_n \succ b_n$ and $a_n \sim b_n$ respectively denote $a_n = o(b_n)$, $b_n = o(a_n)$ and a_n, b_n having the same order as $n \to \infty$. Since $V_{jj} \sim \dfrac{1}{n}$,

$$z_j^2 = \frac{\hat{\beta}_j^2}{V_{jj}} = \begin{cases} \sim n & \text{if the true } \beta_j \neq 0 \\ \sim \chi_1^2 = O_p(1) & \text{if the true } \beta_j = 0. \end{cases}$$

Therefore, if a and A are set to make the order of the threshold Δ to be between 1 and n, then our model selection procedure would work well. For example, it is easy to verify that $\Delta \sim n$, as long as a is of $O(1)$ and A is of order n. For this wide range of choices of a and A, $z_j^2 \sim n \succ \log n \sim \Delta$ and X_j will be selected, whenever the true $\beta_j \neq 0$; and $z_j^2 \sim 1 \prec \log n \sim \Delta$ and X_j will not be selected, whenever the true $\beta_j = 0$. So the proposed variable selection method is consistent in the frequentist sense.

B Description of 33 variables used in the analysis

Table 3 summarizes simple information of the variables used in the MEPS data analysis in Section 4. ANYLMT is the summary of variables related to any limitation: whether a person has any Instrumental Activities of Daily Living (IADL), Activities of Daily Living (ADL), activity limitation (ACTLIM) or functional limitations. IADL/ADL indicates if they need help or supervision with using the telephone, paying bills, taking medications, preparing light meals, doing laundry, or going shopping. ACTLIM implies any limitation in work (WRKLIM), housework (HSELIM), or school (SCHLIM). Functional limitations mean having difficulties walking (WLKDIF), climbing stairs (STPDIF), grasping objects (FNGRDF), reaching overhead (RCHDIF), lifting (LFTDIF), bending or stooping (BENDIF) or standing for long periods of time (STNDIF). More information can be found in the webpage https://meps.ahrq.gov/data_stats/download_data/pufs/h171/h171doc.shtml.

Table 3: Description of 33 variables used in the analysis.

Variable name	Description	Type
DECEASED	Deceased during 2014	Binary (Yes=1, No=0)
AGE	Age as of 12/31/14	Continuous
MALE	Sex (Male=1, Female=0)	Binary
WHITE	Race (White=1, else=0)	Binary
HISP	Hispanic ethnicity (Hispanic=1, else=0)	Binary
EDUCAT	Category of highest degree of education	Categorical (1,2,3)
HBP	High blood pressure diagnosis	Binary
CORHRT	Coronary heart disease diagnosis	Binary
STRK	Stroke diagnosis	Binary
EMPHY	Emphysema diagnosis	Binary
CHBRON	Chronic bronchitis diagnosis	Binary
CHOLES	High cholesterol diagnosis	Binary
CANCER	Cancer diagnosis	Binary
DIABETES	Diabetes diagnosis	Binary
JTPAIN	Join pain last 12 months	Binary
ARTH	Arthritis diagnosis	Binary
ASTH	Asthma diagnosis	Binary
ANYLMT	Any limitation during 2014	Binary
SMK	Current smoking status	Binary

PCS	Self-Administered Questionnaire: physical component summary	Continuous
MCS	Self-Administered Questionnaire: mental component summary	Continuous
USABORN	Born in USA	Binary
HAVEUSC	Have USC provider	Binary
POVCAT	Category of poverty level	Categorical (1, 2, 3, 4, 5)
INSCOV	Had health insurance coverage during 2014	Binary
MDCARE	Covered by Medicare	Binary
DTCARE	Dental insurance	Binary
PMED	Prescription drug private insurance	Binary
CALLSDEPT	The number of calls with office and outpatient departments	Continuous
CALLSPHY	The number of calls with office and outpatient physicians	Continuous
HOSPEXP	Having any hospitalization during 2014	Binary
EMERG	Having any emergency room visit during 2014	Binary
HMHLTH	Informal home health provider days	Continuous

Physical and Mental Health Composite Scores (PCS and MCS) are summary scores computed based on the Short-Form 12 Version 2 (SF-12v2). SF-12v2 is a shorter version of the SF-36v2 Health Survey that uses just 12 questions. Twelve questions are combined, scored, and weighted to create two scales, PCS and MCS range from 0 (lowest) to 100 (highest level of health). More details also can be found in the same webpage above.

References

[1] Keehan S, Stone D, Poisal J, Cuckler G, Sisko A, Smith S, Madison A, Wolfe C, Lizonitz J. National health expenditure projections, 2016–25: Price increases, aging push sector to 20 percent of economy. Health Affairs. 2017;36:553–63.
[2] Duan N. Smearing estimate: a nonparametric retransformation method. J Am Stat Assoc. 1983;78:605–10.
[3] Manning W. The logged dependent variable, heteroscedasticity, and the retransformation problem. J Health Econ. 1998;17:283–95.
[4] Chen J, Liu L, Zhang D, Shih Y-C. A flexible model for the mean and variance functions, with application to medical cost data. Stat Med. 2013;32:4306–18.
[5] Chen J, Liu L, Zhang D, Shih Y-C, Severini T. A flexible model for correlated medical costs, with application to medical expenditure panel survey data. Stat. Med. 2016;35:883–894.
[6] Chernozhukov V, Hong H. An MCMC approach to classical estimation. J Econ. 2003;115:293–346.
[7] Inoue A, Shintani M. Quasi-Bayesian model selection, 2014. https://my.vanderbilt.edu/inoue/files/2014/08/submitted_version.pdf, technical report, Vanderbilt University.
[8] Jiang W, Liu X. Consistent model selection based on parameter estimates. J Stat Plann Infer. 2004;121:265–83.
[9] Li C, Jiang W. On oracle property and asymptotic validity of Bayesian generalized method of moments. J Multivariate Anal. 2016;145:132–47.
[10] Ročková V, George E. The spike-and-slab LASSO. J Am Stat Assoc. 2017; in press. DOI: 10.1080/01621459.2016.1260469.
[11] Tang Z, Shen Y, Zhang X, Yi N. The spike-and-slab lasso generalized linear models for prediction and associated genes detection. Genetics. 2017;205:77–88.
[12] White H. Maximum likelihood estimation of misspecified models. Econometrica. 1982;50:1–25.
[13] Zheng X, Loh W-Y. Consistent variable selection in linear models. J Am Stat Assoc. 1995;90:151–6.
[14] Blough DK, Madden CW, Hornbrook MC. Modeling risk using generalized linear models. J Health Econ. 1999;18:153–71.
[15] Buntin MB, Zaslavsky AM. Too much ado about two-part models and transformation? comparing methods of modeling medicare expenditures. J Health Econ. 2004;23:525–42.
[16] Manning WG, Basu A, Mullahy J. Generalized modeling approaches to risk adjustment of skewed outcomes data. J Health Econ. 2005;24:465–88.
[17] Manning WG, Mullahy J. Estimating log models: to transform or not to transform? J Health Econ. 2001;20:461–94.
[18] Mullahy J. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. J Health Econ. 1998;17:247–81.
[19] Basu A, Rathouz PJ. Estimating marginal and incremental effects on health outcomes using flexible link and variance function models. Biostatistics. 2005;6:93–109.