Testing for collinearity using Bayesian Analysis

Abstract

When faced with the problem of multicollinearity most tourism researchers recommend mean-centering the variables. This procedure however does not work. It is actually one of the biggest misconceptions we have in the field. We propose instead using Bayesian ridge regression and treat the biasing constant as a parameter about which inferences are to be made. It is well known that many estimates of the biasing constant have been proposed in the literature. When the coefficients in ridge regression have a conjugate prior distribution, formal selection can be based on the marginal likelihood. In the non-conjugate case we propose a conditionally conjugate prior for the biasing constant, and show that Gibbs sampling can be employed to make inferences about the ridge regression parameters as well as the biasing constant itself. We examine posterior sensitivity and apply the techniques to a tourism data set.

Keywords: Multicollinearity; Bayesian analysis; ridge regression; Gibbs sampling.

Introduction

The problem of multicollinearity is highly common in tourism research. One particular example is the regression model with moderators. Such model is usually highly prone to having collinearity problems because the interaction term is created by multiplying two exogenous variables to create another exogenous variable. To "alleviate" the potential problems of collinearity, tourism researchers routinely mean center the variables by subtracting the item value from the mean value of the item. This simply does not fix the problem. Mean centering does not really help or harm (Echambadi and Hess, 2007; and Dalal and Zickar, 2011). While the mean-centered coefficients have different interpretations than the original coefficients, we rarely see them being compared against each other in the tourism literature. In fact, anytime an interaction is included in the model, the original coefficients should not be used directly to assess the impact of X on Y. Instead, one needs to use the marginal effect which is actually what we obtain when we mean center the variables.

Assuming that the data for the dependent variable are arranged in the $n \times 1$ vector $\mathbf{y}$ and the data for the explanatory variables are in the $n \times p$ matrix $\mathbf{X}$, so that we have n observations and p regressors, it is well established that the least squares (LS) estimator $\mathbf{b} = \left(\mathbf{X'X}\right)^{-1}\mathbf{X'y}$, under the stated assumptions about the error term is the best linear unbiased estimator (BLUE). However, multicollinearity can result in ill conditioning of the matrix $\mathbf{X'X}$ rendering the LS estimator undesirable. For example when this matrix is nearly non-invertible, the covariance matrix will have large elements in the diagonal, implying that standard errors of LS estimators will be quite large. Effectively, in specific samples, it is quite likely that we may end up with LS coefficients having the wrong sign, being non-significant *etc*.

A regularization method that has been proposed is the use of the ridge regression estimator (Hoerl and Kennard, 1970), with a biasing constant k, usually small. Effectively, "the procedure can be used to portray the sensitivity of the estimates to the particular set of data being used, and it can be used to obtain a point estimate with a smaller mean square error" (Hoerl and Kennard, 1970, p.55). As a matter of fact, Hoerl and Kennard (1970) discussed the Bayesian foundation of their approach (p.64) and also proposed a more general ridge regression.

A main challenge in the literature has been finding the appropriate value of k, as different procedures (Dorugade and Kashid, 2010; Uslu, Egrioglu and Bas, 2014) have been used for that purpose. Hoerl and Kennard (1970) suggested using the ridge trace to find the appropriate value of k for which the regression coefficients have been stabilized. Hoerl and Kennard (1976) proposed an iterative approach for selecting k. However, their procedure does not necessarily converge. As there is no consensus on what is a reasonable procedure to select the value of k, we propose here a Bayesian approach to address this issue. Our aim is to provide tourism researchers with more flexibility in estimating ridge regressions. The Bayesian approach is appealing because it treats k as a parameter which is to be selected in light of the data. In fact, we do not select a single value of k, but we produce the whole marginal posterior of this parameter given the data. This, in turn, is one attractive way to address the uncertainty about k.

The push for Bayesian estimation is taking place across several disciplines such as management (Zyphur & Oswald, 2015; Cabantous and Gond, 2015; McKee and Miller, 2015), marketing (Rossi and Allenby, 2003; Rossi et al. 2012), psychology (Van De Schoot, et al., 2017) and tourism (Assaf and Tsionas, 2018 a, b). Over the last decade, we have seen a strong increase in the use of the Bayesian methodology in tourism and other related fields (Wong et al. 2006; Wang et al. 2011; Assaf, 2012; Barros, 2014; Assaf et al. 2017; Assaf et al., 2018). A recent special issue in the Journal of Management is a clear indication on the growing popularity of this method (Zyphur & Oswald, 2015).

Recent papers has provided comprehensive introductions on the advantages of the Bayesian approach (Muthen, 2010, Zyphur and Oswald, 2015). The Bayesian approach is not simply about fitting more advanced models with MCMC (Markov chain Monte Carlo) but is a completely different paradigm and philosophy in statistics. It offers several advantages in the estimation of regression models including " rich diagnostic information about parameters and models; controlling for multiple comparisons as a function of the data; handling low-frequency, unbalanced, missing data; and exploration of prior assumptions about model parameters" (Zyphur and Oswald, 2013, p.7). Probably, on the most known advantages of the method is its ability to incorporate prior information about a parameter and form a prior distribution. For instance, the Bayes' theorem can be expressed as: $p(\theta \mid y) \propto p(\mathrm{y} \mid \theta) p(\theta)$, where $\propto$ is the proportionality symbol. Here, $p(\theta \mid y)$ is the posterior distribution which is used to carry out all inferences, and is proportional to the product of the prior $p(\theta)$ and the likelihood function $p(y|\theta)$[1]. Difference choices of priors can be used such as conjugate vs. non-conjugate priors. The prior is said to be conjugate if it belongs to the family of distribution as the posterior distribution (i.e. the posterior has posterior has the same distributional form as the prior distribution). For example, in the context when the likelihood function is binomial $y \sim Bin(n,\theta)$, a conjugate prior in the form of a beta distribution on θ will also lead to a posterior distribution that follows a beta distribution. A prior distribution which is not conjugate is called a non-conjugate prior.

We illustrate below the flexibility of the Bayesian approach and prior information within the context of ridge regression. In particular, we introduce a Bayesian ridge estimator for both conjugate and non-conjugate priors though we rely more on the non-conjugate prior as the conjugate priors are restrictive and have certain problems, for example they have the same tails with the likelihood and they are rarely used in practice. A singular advantage of the Bayesian approach is that ridge regression can be interpreted as Bayes posterior mean when the prior on the regression parameters is multivariate normal with zero mean and covariance matrix a diagonal matrix whose diagonal elements have the same variance / precision. Moreover, the significance of the Bayesian approach to regression is that the celebrated James-Stein estimator has a direct empirical Bayes estimator. The James-Stein estimator is well-known to improve on maximum likelihood / OLS estimator in terms of risk and MSE across all values of the parameter space.

In this paper we proceed as follows: In section 2 we provide an introduction to ridge regression. Sections 3 and 4 present the Bayesian ridge regression approach with conjugate and non-conjugate setting in comparison with the diffuse prior assumptions. We conduct a Monte Carlo study in section 5 to illustrate the issue diagnosing and correcting the effect multicollinearity. We then present illustration on the Bayesian ridge regression using a tourism application.

2. How to Proceed?

So, if mean centering does not work, how to proceed from here? One of the most common approaches is to use ridge regression to analyze regression data that is subject to multicollinearity. As mentioned, with OLS the regression parameters can be estimated using the following formula:

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$$

The ridge regression differentiates by adding a biased constant $k > 0$ to the diagonal elements of the correlations matrix:

$$\mathbf{b}_k = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathbf{y},$$

[1] The likelihood function summarizes the information from the data.

This is where the term "ridge regression" comes from as the diagonal of one in the correlation matrix are thought of as a ridge). What we know from Hoerl and Kennard (1970) is that there is always a $k \in \left(0, \bar{k}\right)$ for which ridge regression dominates OLS in terms of mean squared error (MSE), and $\bar{k} = \frac{\sigma^2}{\alpha_{\max}^2}$, where $\mathbf{X}'\mathbf{X} = \mathbf{P}'\mathbf{\Lambda}\mathbf{P}$, and $\boldsymbol{\alpha} = \mathbf{P}\boldsymbol{\beta}$. Here, $\mathbf{P}$ is the orthonormal matrix of eigenvectors of $\mathbf{X}'\mathbf{X}$, and $\Lambda = diag\left(\lambda_1, .., .\lambda_p\right)$, where $\lambda_1, .., \lambda_p$ represent the distinct eigenvalues of $\mathbf{X}'\mathbf{X}$. Another result of Hoerl and Kennard (1970) was that the total MSE of the ridge estimator is[2]:

$$MSE\left(\mathbf{b}_k\right) = \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{\left(\lambda_i + k\right)^2} + k^2 \boldsymbol{\beta}' \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p\right)^{-2} \boldsymbol{\beta}$$

Minimizing the MSE, unfortunately, depends on the ratio of β / σ. Depending on this result several settings for the parameter k have been proposed. See for example Khalaf and Shukur (2005), Lawless and Wang (1976), Nomura (1988) and Maruyama and Strawderman (2005). A similar idea is the Bayesian lasso regression (Park and Casella, 2008, Hans, 2009).

The goal of this paper is to propose a more flexible way to estimate k using the Bayesian approach. As mentioned, one of the advantages is that with Bayesian approach we do not (necessarily) select a single value of k but we produce the whole marginal posterior of this parameter given the data. We aim here to diagnose and correct the effects of multicollineatity through a full non-conjugate Bayesian approach to Ridge Regression. In particular we take up Bayesian inference in conjugate and non-conjugate ridge regression models by using the fact that a prior can be placed on the ridge parameter(s) k and proceed with posterior analysis on all parameters using MCMC techniques. We run different simulations to illustrate the performance of the method. We also provided evidence based on a real dataset from the hotel industry. Our goal is to show that collinearity can be simultaneously diagnosed and corrected using priors on all parameters. Our techniques detect and correct the adverse effects of collinearity in a transparent way.

Specifically, given the general regression model, we consider first ridge regression from the Bayesian point of view treating the biasing constant (k) as a parameter about which inferences are to be made to avoid selecting a particular value of k. For the conjugate case we have derived the marginal likelihoods and showed how selection of the k parameter can be performed to choose the appropriate value. It is important to notice that the original ridge regression estimators depend crucially on a conjugacy assumption, namely that the regression coefficients, $\boldsymbol{\beta} \mid \sigma, k \sim N_p\left(0, \frac{\sigma^2}{k}\mathbf{I}_p\right)$. Conjugate priors have certain problems, for example they have the same tails with the likelihood and they are rarely used in practice.

[2] The notation A^{-2} for a matrix A, means $A^{-2} = A^{-1}A^{-1}$.

The reader can refer to Leamer (1969) and Judge et al (1985) regarding this point. As they mention, despite the fact that the natural conjugate setting is a convenient approach (since it provides an analytical solution to the integrations involved), it has been criticized because it employs the prior information as a previous imaginary sample from the same process: When we set the degrees of freedom and the precision matrix equal to zero to obtain the limiting distribution of the normal-(inverse) gamma prior, the resulting ignorant prior is different to the usual diffuse prior and the posterior distribution has different degrees of freedom. Therefore, a non - conjugate prior can be adopted instead, and numerical posterior inference can rely on the Gibbs sampler. In the Gibbs sampler k is treated as a parameter and, therefore, formal statistical inferences can be made about this parameter thus solving a long-standing problem in the literature. Moreover, a formal test for collinearity can be developed if we compare the marginal posterior of k with its value at $k=0$ (corresponding to OLS or Bayes with diffuse prior). An equivalent test is to compare the marginal likelihood at the optimal k with its value when $k=0$.

We discuss below a Bayesian ridge estimator for both conjugate and non-conjugate priors though we rely more on the non-conjugate prior ad the conjugate priors are restrictive and have certain problems, for example they have the same tails with the likelihood and they are rarely used in practice.

3. Bayesian ridge regression

For the Bayesian interpretation of the ridge regression estimator, the model is given by:

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{u},\ \mathbf{u}\sim N_n\left(\mathbf{0},\ \sigma^2\mathbf{I}_n\right)$$

The prior on the unknown parameters is

$$\boldsymbol{\beta}\,|\,\sigma\sim N_p\left(\mathbf{0},\frac{\sigma^2}{k}\mathbf{I}_p\right), p(\sigma)\propto\sigma^{-1},$$

where $k>0$ is prior precision *relative* to the error variance, σ^2. It is not difficult to show that, under these conditions, the posterior mean is given by: $E\left(\boldsymbol{\beta}\,|\,\mathbf{y},\mathbf{X}\right)=\mathbf{b}_k$, that is the ridge regression estimator. As mentioned, since the choice of parameter k has been an active area of research for many years, and many choices have been proposed, it is natural to investigate the implications of a fully Bayesian approach to the problem. To this effect, we consider both a conjugate and non-conjugate prior on the regression parameters, $\boldsymbol{\beta}$.

The conjugate prior provides explicitly in analytical form the ridge regression estimator so there is much in favor of it. However, the non-conjugate case is also interesting and can be considered as an alternative.

3.1 Optimal biasing parameter through conjugacy

Suppose $\boldsymbol{\beta}\,|\,\sigma\sim N_p\left(0,\ \frac{\sigma^2}{k}\mathbf{I}_p\right)$, and $\frac{\underline{\nu}\underline{s}^2}{\sigma^2}\sim\chi^2\left(\underline{\nu}\right)$, where $\underline{\nu},\underline{s}^2$ are prior hyperparameters. This prior is conjugate because it depends on σ and matches exactly the likelihood to provide as posterior mean $\mathbf{b}_k$ below.

The marginal likelihood (or "evidence"), for a given value of k, can be derived analytically in this case[3]:

$$p_k(\mathbf{y}) \propto \left(\frac{|\mathbf{V}|}{|\underline{\mathbf{V}}|}\right)^{1/2} \left(\overline{\nu}\overline{s}^2\right)^{-\overline{\nu}/2}, \text{where}$$

$$\mathbf{V} = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p\right)^{-1}, \underline{\mathbf{V}} = k^{-1}\mathbf{I}_p,$$

$$\overline{\nu}\overline{s}^2 = \underline{\nu}\underline{s}^2 + \mathbf{y}'\left(\mathbf{I}_p - \mathbf{X}\left(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p\right)^{-1}\mathbf{X}'\right)\mathbf{y}$$

$$\mathbf{b} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ and } \overline{\nu} = \underline{\nu} + n.$$

Moreover, $\frac{|\mathbf{V}|}{|\underline{\mathbf{V}}|} = \frac{k^p}{\prod_{i=1}^{p}(k+\lambda_i)}$, where $\lambda_1,\dots,\lambda_p$ are the eigenvalues of $\mathbf{X}'\mathbf{X}$, and $\overline{\nu}\overline{s}^2 = \underline{\nu}\underline{s}^2 + RSS + \mathbf{b}'\left(\mathbf{X}'\mathbf{X}\right)\left(\mathbf{b} - \mathbf{b}_k\right)$, $RSS = \left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)'\left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)$. By (2.3) in Hoerl and Kennard (1970), we have $\left[\mathbf{I}_p + k\left(\mathbf{X}'\mathbf{X}\right)^{-1}\right]^{-1}\mathbf{b} = \mathbf{b}_k$, where $\mathbf{b}_k = \left(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p\right)^{-1}\mathbf{X}'\mathbf{y}$ is the ridge estimate. So the log marginal likelihood simplifies to the expression:

$$\log p_k(\mathbf{y}) \doteq 0.5p\log k - 0.5\sum_{i=1}^{p}\log\left(\lambda_i + k\right) - 0.5\overline{\nu}\log\left(\underline{\nu}\underline{s}^2 + RSS + \mathbf{b}'\left(\mathbf{X}'\mathbf{X}\right)\left(\mathbf{b} - \mathbf{b}_k\right)\right) \quad (1)$$

This expression involves only the eigenvalues of $\mathbf{X}'\mathbf{X}$, standard LS quantities and the ridge estimates.

3.2. Bayesian ridge regression in the non-conjugate case

In the Bayesian context it is reasonable to treat k as unknown parameter whose prior is $p(k)$ *independently* of $\boldsymbol{\beta}$ and σ. Therefore, it is useful to depart from the conjugate case which involves the unpleasant feature that the tails of the posterior and the prior are the same. Then we have:

$$\boldsymbol{\beta} \mid \sigma, k \sim N_p\left(\mathbf{0}, \tfrac{1}{k}\mathbf{I}_p\right),\ p(\sigma \mid k) \propto \sigma^{-1},$$

and the prior of k is proportional to $p(k)$. It is not necessary for this prior to be proper. The joint posterior is as follows:

$$p\left(\boldsymbol{\beta}, \sigma, k \mid \mathbf{y}, \mathbf{X}\right) \propto \sigma^{-(n+1)} k^{p/2} p(k) \exp\left[-\frac{\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) + k\sigma^2\boldsymbol{\beta}'\boldsymbol{\beta}}{2\sigma^2}\right] \quad (2)$$

[3]See Zellner (1971), p.309.

Completing the square $Q=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+k\sigma^2\boldsymbol{\beta}'\boldsymbol{\beta}$, we obtain the expression:

$$Q=(\boldsymbol{\beta}-\mathbf{b}_k)'(\mathbf{X}'\mathbf{X}+k\sigma^2\mathbf{I}_p)(\boldsymbol{\beta}-\mathbf{b}_k)+\mathbf{y}'\mathbf{M}_k\mathbf{y},$$

where $\mathbf{M}_k=\mathbf{I}_p-\mathbf{X}\mathbf{V}_k\mathbf{X}'$ and $\mathbf{V}_k=(\mathbf{X}'\mathbf{X}+k\sigma^2\mathbf{I}_p)^{-1}$. Therefore, the posterior distribution is:

$$p(\boldsymbol{\beta},\sigma,k\mid\mathbf{y},\mathbf{X})\propto\sigma^{-(n+1)}k^{p/2}p(k)\exp\left[-\frac{(\boldsymbol{\beta}-\mathbf{b}_k)'(\mathbf{X}'\mathbf{X}+k\sigma^2\mathbf{I}_p)(\boldsymbol{\beta}-\mathbf{b}_k)+\mathbf{y}'M_k\mathbf{y}}{2\sigma^2}\right] \quad (3)$$

In this expression, $\mathbf{b}_k=(\mathbf{X}'\mathbf{X}+k\sigma^2\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y}$ is the ridge regression estimate. From the expression in (3), we can extract the following posterior conditional distributions:

$$\boldsymbol{\beta}\mid\sigma,k,\mathbf{y},\mathbf{X}\sim N_p\left(\mathbf{b}_k,\sigma^2(\mathbf{X}'\mathbf{X}+k\sigma^2\mathbf{I}_p)^{-1}\right),$$

$$\left.\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma^2}\right|\boldsymbol{\beta},k,\mathbf{y},\mathbf{X}\sim\chi^2(n),$$

where $\chi^2(n)$ denotes the chi-square distribution with n degrees of freedom. Finally, the posterior conditional distribution of the biasing parameter can be derived from (2) as:

$$p(k\mid\boldsymbol{\beta},\sigma,\mathbf{y},\mathbf{X})\propto k^{p/2}p(k)\exp\left(-\frac{k\boldsymbol{\beta}'\boldsymbol{\beta}}{2}\right) \quad (4)$$

The conditionally conjugate prior for the biasing parameter, k, is clearly a $Gamma(\bar{A}/2,\ \bar{B}/2)$[4] distribution whose density is of the following form: $p(k)=\frac{(\bar{B}/2)^{\bar{A}/2}}{\Gamma(\bar{A}/2)}k^{\bar{A}/2-1}\exp(-\frac{\bar{B}}{2}k)$, where $\bar{A}\geq 0$ and $\bar{B}\geq 0$ are hyperparameters. Then we obtain:

$$p(k\mid\boldsymbol{\beta},\sigma,\mathbf{y},\mathbf{X})\propto k^{(p+\bar{A})/2-1}\exp\left(-\frac{\bar{B}+\boldsymbol{\beta}'\boldsymbol{\beta}}{2}k\right)$$

Therefore, the posterior conditional distribution of the biasing parameter is:

$$k\mid\boldsymbol{\beta},\sigma,\mathbf{y},\mathbf{X}\sim Gamma\left(\frac{n+\bar{A}}{2},\frac{\bar{B}+\boldsymbol{\beta}'\boldsymbol{\beta}}{2}\right) \quad (5)$$

Prior elicitation of the hyperparameters $\bar{A}$ and $\bar{B}$ is facilitated by the fact that, in the prior, $E(k)=\frac{\bar{A}}{\bar{B}}$ and $Var(k)=2\frac{\bar{A}}{\bar{B}^2}$. If we believe that $E(k)=0.1$ and the standard deviation of the biasing parameter is σ_k, then $\bar{B}=\frac{0.2}{\sigma_k^2}$. If σ_k^2 is 0.1, 0.5, 1 or 5 then we obtain respectively that $\bar{B}$ is 2.0, 0.40, 0.20 or 0.04. Therefore $\bar{A}$ must be, respectively, 0.2, 0.004, 0.02 or 0.004. In what follows

[4]Notice that $Gamma(\bar{A}/2,\bar{B}/2)$ reduce to an exponential prior for k, by setting $\bar{A}=2$.

we adopt the reference prior $p(\sigma) \propto \sigma^{-1}$.

4. Sampling properties of diagnosing and correcting multicollinearity

4.1. Diagnosing Collinearity

Two reasonable questions: how we diagnose for collinearity using the Bayesian approach, and how the Bayesian ridge model in section 3.2[5] behaves compared to ordinary least square (OLS)[6].

Specifically, the question is whether the Bayesian approach can be useful in both diagnosing and correcting the possibly harmful effects of multicollinearity in circumstances that are encountered in practice. To illustrate this, we choose a design with $n = 500$ observations and $p = 10$ regressors. The first regressor, say X_{t1}, is generated from a standard normal distribution. The remaining regressors are $X_{tj} = \alpha X_{t,j-1} + \omega Z_{tj}$, $Z_{tj} \sim iidN(0,1)$, $j = 2,..., p$, and ω is set to 1/500 with $\alpha = 1$ for collinear data, and $\alpha = 0$, $\omega = 1$ for independent data. The data generating process is $y_t = \sum_{j=1}^{p} \beta_j x_{tj} + u_t$, where all regression coefficients are $\beta_j = 1\ (j = 1,..., p)$, and $u_i \sim iidN(0, \sigma^2)$, with $\sigma = 1$[7]. Setting all coefficients equal to one is done only for simplicity and the results in no way depend on the exact true values of the coefficients.

As our prior on the biasing constant, k, we choose an exponential with parameter $\bar{B} = 10^3$ implying a prior average value of k equal to $E(k) = 10^{-3}$ which seems reasonable in view of experience with collinear data. Holding the matrix of regressors, X, fixed we generate D=10,000 different data sets. For each data set, the Gibbs sampling technique presented in section 4.2 is applied using 11,000 draws, omitting the first 1,000 and taking only every other tenth draw (for a total for 1,000 draws). From the 1,000 available, approximately independent, draws we compute the posterior means, $\bar{\boldsymbol{\beta}}^{(d)} = E(\boldsymbol{\beta} | \mathbf{y})$, $d = 1,..., D$. Our objective is to compare the sampling distribution resulting from $\bar{\boldsymbol{\beta}}^{(d)}$s (as an approximation to the actual sampling distribution) with the sampling distribution of posterior means resulting from a diffuse prior, which is based on OLS quantities that are readily available for each different data set.

With a diffuse prior (i.e. OLS), the sampling distribution of posterior means should be more dispersed compared to the sampling distribution of $E(\boldsymbol{\beta} | \mathbf{y})$ under the stated prior on the biasing constant, k. The sampling results are presented in the three panels of Figure 1. For orthogonal data

---

[5] As mentioned, we recommend relying on the non-conjugate prior for the reasons mentioned in Section 4.2.

[6] This is similar to Bayesian analysis using diffuse priors of the form $p(\boldsymbol{\beta}, \sigma) \propto \sigma^{-1}$, from the sampling-theory point of view

[7] All computations were performed using the WinGauss software. The codes can be provided by the authors upon request.

(i.e. no collinearity) ridge and diffuse posteriors are extremely close (Figure 1a), but this is not the case when we have collinear data (Figures 1b and 1c). In other words, for collinear data, one would observe significant difference between the diffuse posterior (i.e. OLS) and the ridge regression results.

This, in turn, would provide a useful way to detect whether there is harmful multicollinearity and, at the same time, correct it based on the Bayesian ridge estimator. The test can be made more formal by using a Kolmogorov-Smirnov test for testing the equality of the two distributions. We believe, however, that visual presentation is much more informative. We can also use a formal test for collinearity in Bayesian analysis. In OLS settings such tests are not possible. For example, variance inflation factors (VIF) commonly used are diagnostics of collinearity, not statistical tests.

In our case, one can formally test for collinearity using the Bayes factor. Given the marginal posterior $p(k \mid \mathbf{y})$ the Bayes factor in favor of ridge regression and against OLS can be approximated using

$$BF \simeq \frac{p(\hat{k} \mid \mathbf{y})}{p(k=0 \mid \mathbf{y})} \tag{6}$$

where $\hat{k}$ is the modal value of the marginal posterior. In our case the denominator is practically zero, so the BF diverges to a very large value, indicating that the ridge regression model fits the data best. For this approach see Berger (1980, p. 156).

4.2. Correcting for Collinearity

To correct for collinearity we propose the Bayesian ridge estimator proposed in Section 4.2. While we saw in Figure 1.c how the Bayesian diffuse prior (i.e. OLS) can seriously affect the regression results in the presence of collinear data, the important issue that remains is how reliable our Bayesian estimator is in the presence of collinearity.

We run a similar experiment to the above where we compared the performance of the Bayesian diffuse prior against the Bayesian ridge regression. We also included in the comparison a traditional ridge regression model (non-Bayesian) with a pre-specified value for k =0.001[8]. For the diffuse prior context, we followed a common practice in the literature and tried to drop the collinear variables from the model.

We use 10,000 Monte Carlo replications to compare the above models. We tried three versions of the diffuse prior model by dropping one, two and three variables at a time. The true model for the simulation is

$y_i = x_{i1} + x_{i2} + x_{i3} + x_{i4} + x_{i5} + 0.1u_i$ where $u_i \sim N(0,1)$ and the regressors are generated as follows.

[8] This is a standard value for k used in the literature. Of course, our proposed approach focuses on optimizing the value of k and not on pre-specifying the value of k, as discussed previously.

$x_{i1} \sim N(0,1)$

$x_{ij} = x_{i1} + 0.1v_{ij}, v_{ij} \sim N(0,1), j = 2,...,5$.

For testing purposes we set all $\beta_j \left(j = 1,..., p \right)$ as equal to 1. Again, the results do not depend on the exact values of these coefficients.

The results are presented in Figure 2. We can see that the Bayesian ridge regression based on the optimal prior seems to performs best and is the one most centered around the true value of β. Contrary to common belief, the practice of dropping variables from the models, on the other hand, does not seem to be a good choice for correcting the results of the regression model. The closest to our model is the traditional ridge regression, but this has the problem of pre-specifying the value of k in advance. Further illustration with real data is presented next.

Figure 1. Sampling distributions of different estimators of β_2.

Figure 1a. Sampling Distributions of Posterior Mean from Bayesian diffuse analysis and Bayesian ridge analysis ($\bar{B} = 1$). Orthogonal data.

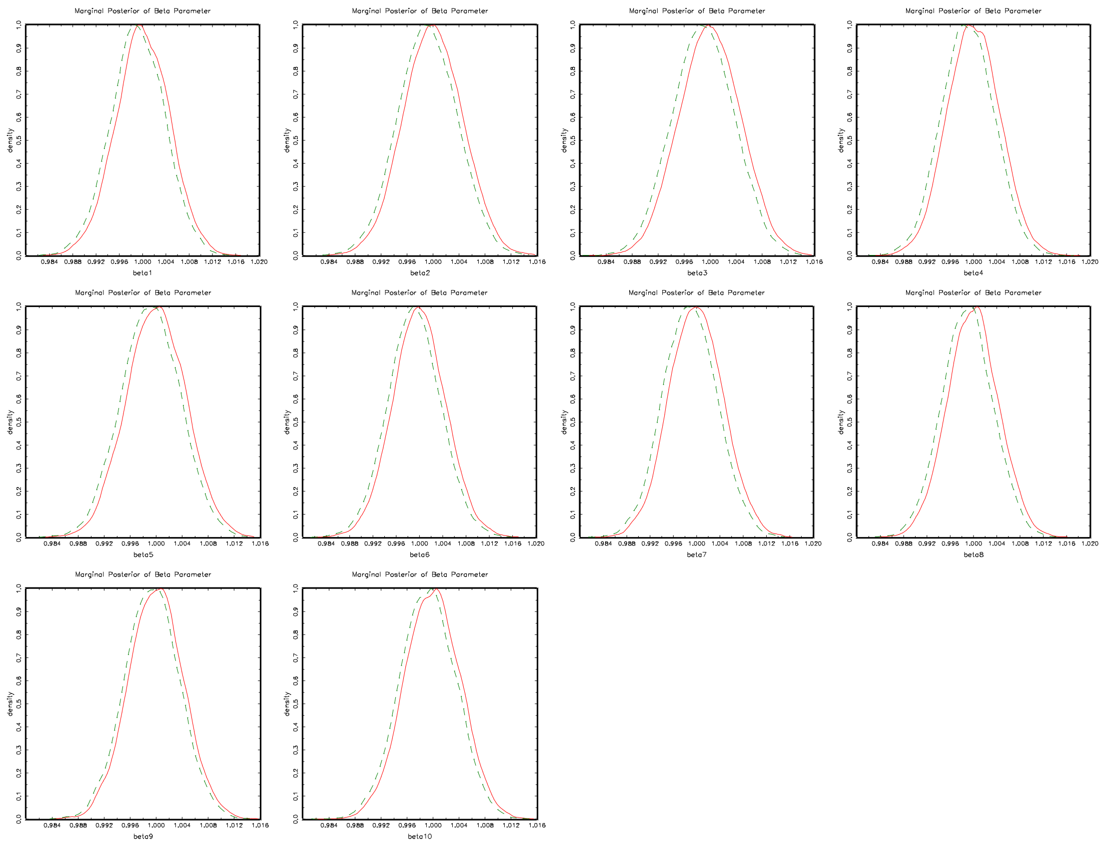


Figure 1b. Sampling Distributions of Posterior Mean from Bayesian diffuse analysis and Bayesian ridge analysis ($\bar{B} = 10^3$). Collinear data.

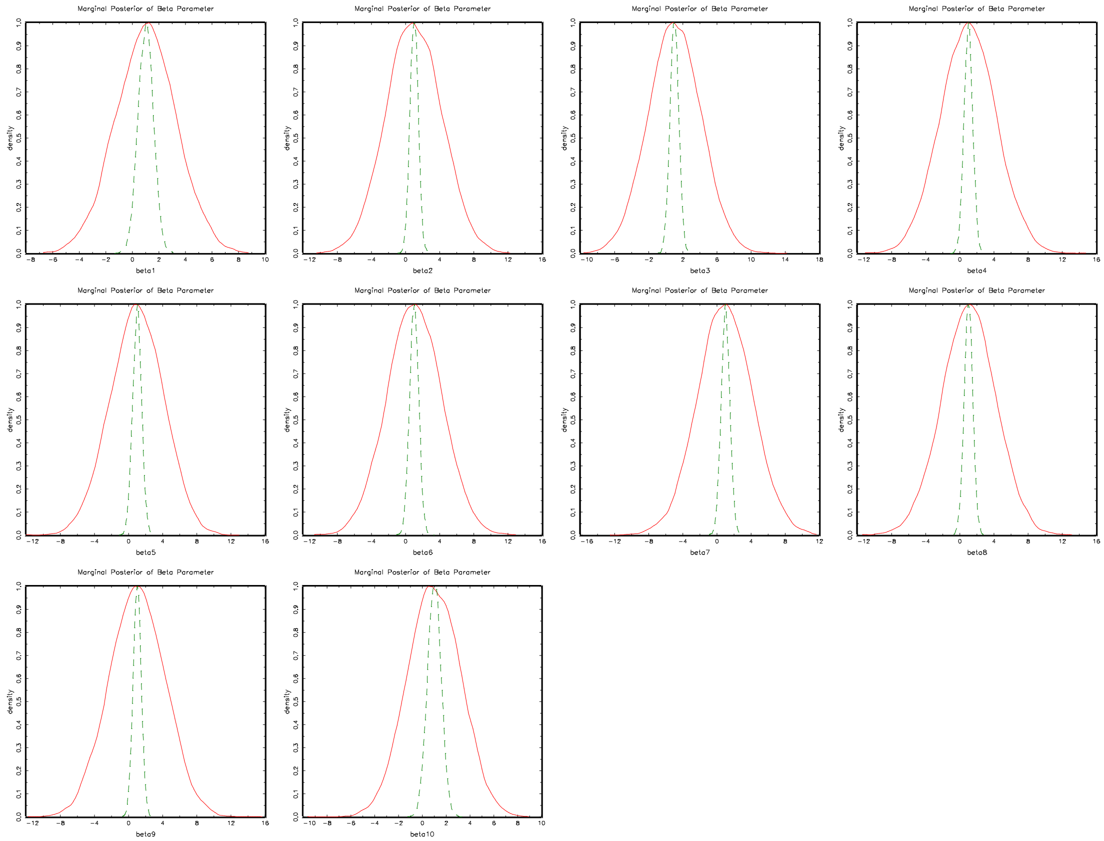


Figure 1c. Sampling Distributions of Posterior Mean from Bayesian diffuse analysis and Bayesian ridge analysis ($\bar{B}=1$). Collinear data.

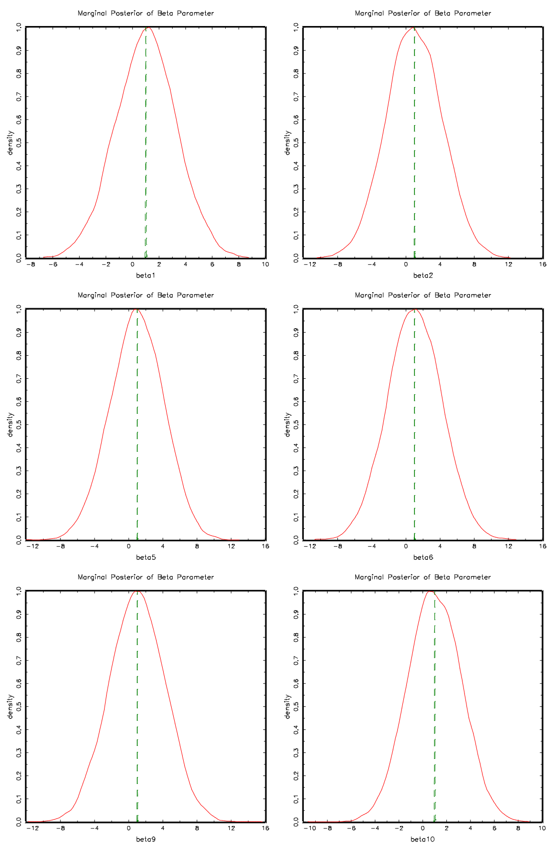
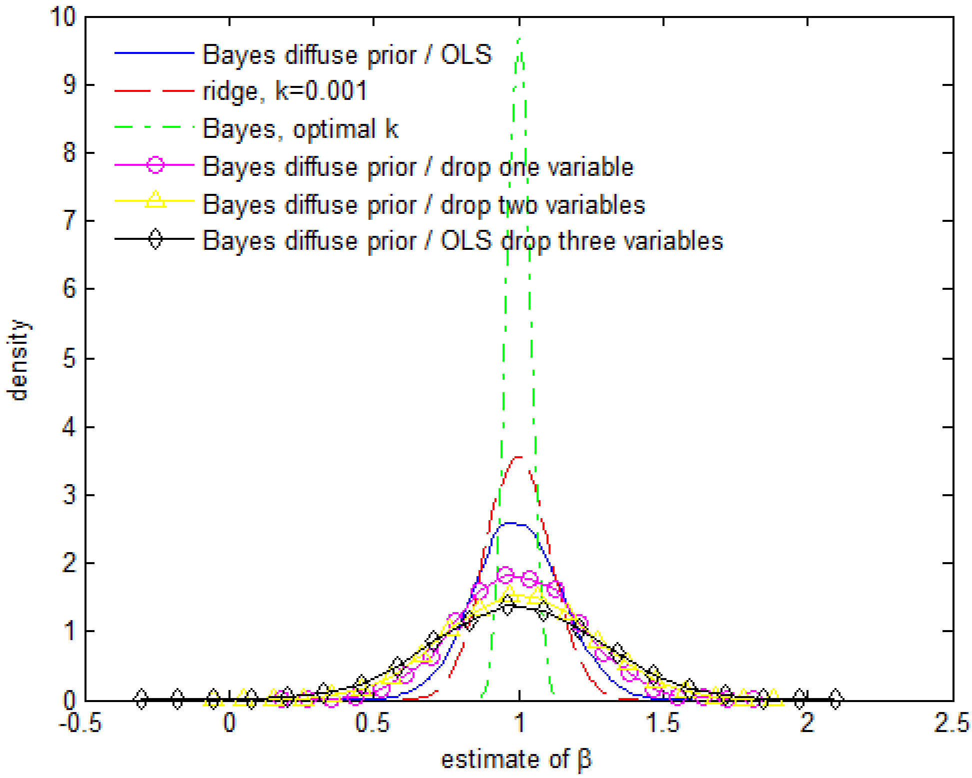


Figure 2. Performance of the Bayesian ridge regression against other alternatives



5. Illustration using Real Data

We also test our Bayesian ridge regression using a real application the hotel industry. The model focuses on the relationship between room revenue and the following covariates: room expenses, food and beverage (F&B) expenses, utility expenses, marketing expenses, property and maintenance (POM) expenses and number of rooms. All these variables are expected to be positively correlated with room revenue as higher revenue usually results in higher expenses in these categories.

The dataset for this study was obtained from Smith Travel Research, an independent company that tracks lodging supply and demand data for most major hotels in the US and internationally. The STR's data are highly comprehensive, reliable and mostly commonly by hotels to track their performance[9].

We use here a unique panel sample of 78 US hotels (for the years 2012-2016). So, in total we have 390 observations. The correlation matrix for all variables included in the model (Table 4) clearly illustrates the high collinearity problem. Further evidence on the collinearity problem in this dataset is illustrated in Table 5 where we can see that the variance inflation factors (VIFs) for five of the six

[9] At least in the United States.

covariates are >10. Our Bayes factor (equation 6) also diverges to a very large value, indicating that the ridge regression model fits the data best.

We report in Table 6 the results from Bayesian ridge regression and linear regression (i.e. OLS). For the Bayesian estimation, we used the non-conjugate prior described in Section 4.2. As mentioned, one of the advantages is that with the Bayesian approach, we do not pre-set or select a single value of k but we produce the whole marginal posterior of this parameter given the data. For example, we report in Figure 3 the overall posterior density of k.

The posterior mean of k is also included in Table 6. We can clearly the differences between the results obtained from the Bayesian ridge regression vs OLS. For instance, despite the high positive correlation between the various covariates and the dependent variable several coefficients from OLS have a negative sign and only three of them are significant. The Bayesian ridge regression however indicates that all coefficients are positive and significant. This confirms our earlier results from the simulation that when collinearity exits, ridge and least square results can be very different.

Of course, we are not implying that collinearity is of less concern than is often implied in the literature. While our method seems to be more tolerant to collinearity, the results should not encourage tourism researchers to throw any variable into the model and expect the results to come out perfectly. The selection of variables should still be based on an educated theoretical approach. In contexts when collinearity cannot be avoided, the practices of mean centering, or dropping variables do not seem to be good choices for correcting the results of the regression model. Rather, the regression estimation should be conducted using more robust approaches such as the one we propose in this study.

Table 4. Correlation Matrix

	1	2	3	4	5	6	7
Room revenue (1)	1						
Room Expenses (2)	0.9867	1					
F&B expenses (3)	0.9640	0.9744	1				
Utility Expenses (4)	0.9815	0.9741	0.9543	1			

Marketing Expenses(5)	0.9463	0.9279	0.9379	0.9523	1		
POM Expenses (6)	0.9615	0.9717	0.9741	0.9602	0.9514	1	
Number of Rooms (7)	0.8773	0.8839	0.8685	0.9075	0.9242	0.8967	1

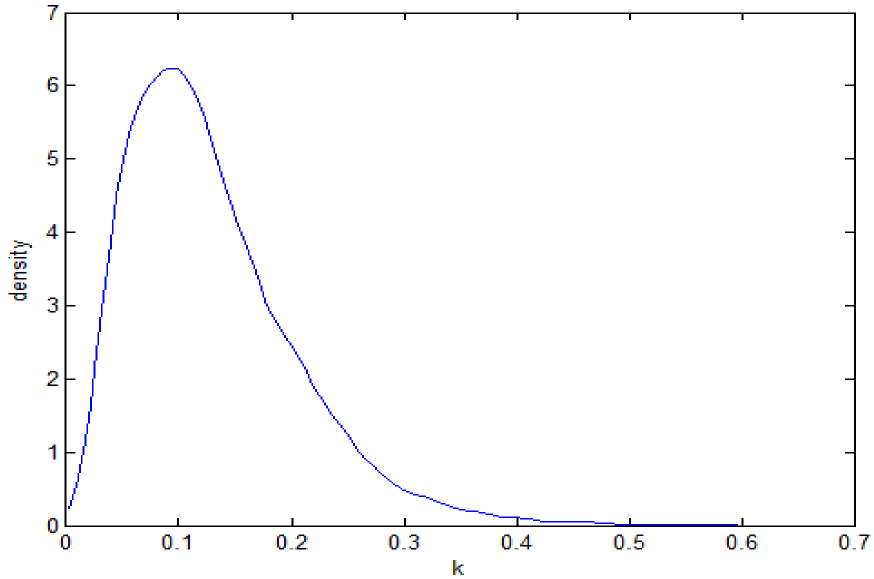
Table 5. Multicollinearity Diagnostic Criteria

	Eigen Values	VIF	1/VIF
Room Expenses	5.7033	34.3711	0.0291
F&B expenses	0.157	25.0432	0.0399
Utility Expenses	0.0566	28.6689	0.0349
Marketing Expenses	0.0423	19.6543	0.0509
POM Expenses	0.0242	33.7577	0.0296
Number of Rooms	0.0165	8.3613	0.1196

Table 6. Bayesian Ridge Regression vs. OLS

	Bayesian Ridge		OLS	
Variable	Posterior Mean	Posterior t-stat	Estimate	t-stat
Room Expenses	5.871	11.068	10.442	16.202
F&B expenses	0.803	3.902	-0.504	-1.127
Utility Expenses	1.414	11.542	0.957	1.844
Marketing Expenses	3.913	38.391	4.628	12.172
POM Expenses	0.960	3.357	-1.378	-2.018
Number of Rooms	1.101	4.871	-0.911	-1.953
sigma	0.211	0.010		
k	0.129	0.075		

Figure 3. Posterior Distribution of the Bayesian ridge parameter (k)



6. Concluding Remarks

In this paper we have taken up Bayesian inference in conjugate and non-conjugate ridge regression models by using the fact that a prior can be placed on the ridge parameter(s) k proceed with posterior analysis on all parameters using standard MCMC techniques. For the conjugate case we have derived the marginal likelihoods and showed how selection of the k or g parameter can be based in an empirical Bayes context to choose the appropriate value. It is important to notice that the original ridge regression estimators depend crucially on a conjugacy assumption, namely that the regression coefficients, $\boldsymbol{\beta} \mid \sigma, k \sim N_p\left(0, \frac{\sigma^2}{k}\mathbf{I}_p\right)$. In the absence of σ the prior of $\boldsymbol{\beta}$ is no longer in the normal-gamma prior form which is necessary for ordinary ridge regression to emerge. A non-conjugate prior of the form $\boldsymbol{\beta} \mid k \sim N_p\left(0, \frac{1}{k}\mathbf{I}_p\right)$ can be adopted instead, and numerical posterior inference can rely on the Gibbs sampler. Conjugate priors have certain problems, for example they have the same tails with the likelihood and they are rarely used in practice.

We have applied these ideas to show that collinearity can be simultaneously diagnosed and corrected using priors on all parameters. We also illustrated that the Bayesian ridge regression performs better than a Bayesian regression with diffuse prior (i.e. OLS). Contrary to common belief, the practice of dropping variables from the models, does not also seem to be a good choice for correcting the results of the regression model.

One limitation of the paper is that we focus on a single k. Different ks can be used for each regressor easily, although at the cost of computing several values of the marginal likelihood depending on the value of such k coefficients. Our methods illustrate, in the context of tourism studies, that biased estimators yielding lower mean squared error (MSE) are clearly desirable and, thus, future research could focus more on generalizations of our procedure. Another limitation is, of course, the assumption of normality of errors which, however, can be relaxed to consider more general models including other elliptical distributions, distributions with fat tails and / or asymmetry, etc.

References

Assaf, A. G. (2012). Benchmarking the Asia Pacific tourism industry: A Bayesian combination of DEA and stochastic frontier. *Tourism Management*, 33(5), 1122-1127.

Assaf, A. G., & Tsionas, M. (2018a). Bayes factors vs. P-values. *Tourism Management*, 67, 17-31.

Assaf, A. G., & Tsionas, M. (2018b). The estimation and decomposition of tourism productivity. *Tourism Management*, 65, 131-142.

Assaf, A. G., Tsionas, M., & Oh, H. (2018). The time has come: Toward Bayesian SEM estimation in tourism research. *Tourism Management*, 64, 98-109.

Assaf, A. G., Oh, H., & Tsionas, M. (2017). Bayesian Approach for the Measurement of Tourism Performance: A Case of Stochastic Frontier Models. *Journal of Travel Research*, 56(2), 172-186.

Barros, C. P. (2014). Airports and tourism in Mozambique. *Tourism Management*, 41, 76-82.

Berger, J. O. (1980). Statistical decision theory and Bayesian analysis, Springer, Berlin.

Cabantous, L., & Gond, J. P. (2015). The resistible rise of Bayesian thinking in management: Historical lessons from Decision Analysis. *Journal of Management*, 41(2), 441-470.

Dalal, D. K., & Zickar, M. J. (2012). Some common myths about centering predictor variables in moderated multiple regression and polynomial regression. *Organizational Research Methods*, 15(3), 339-362.

Dorugade, A. V., & Kashid, D. N. (2010). Alternative method for choosing ridge parameter for regression. *Applied Mathematical Sciences*, 4(9), 447-456.

Echambadi, R., & Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science*, 26(3), 438-445.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96(4), 835-845.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67.

Judge, O. G., W. E. Griffith, R. C. Hill, C. H. Lee, and H. Lutkepohl.(1985). The theory and practice of econometrics, second edition, Wiley, New York.

Khalaf, G., & Shukur, G. (2005). Choosing Ridge Parameter for Regression Problems. *Communication in Statistics: Theory and Methods*, 34, 1177-1182.

Lawless, J. F., & Wang, P. (1976). Simulation study of ridge and other regression estimators. *Communication in Statistics: Theory and Methods*, 4, 307-323.

Leamer, E. E., (1978) Specification Searches, Wiley, New York.

Maruyama, Y., & Strawderman, W. E. (2005). A new class of generalized Bayes minimax ridge regression estimators. *The Annals of Statistics*, *33*(4), 1753-1770.

McKee, R. A., & Miller, C. C. (2015). Institutionalizing Bayesianism within the organizational sciences: A practical guide featuring comments from eminent scholars. *Journal of Management*, *41*(2), 471-490.

Muthén, B. (2010). Bayesian analysis in Mplus: A brief introduction. *Unpublished manuscript. www. statmodel. com/download/IntroBayesVersion*, *203*.

Nomura, M. (1988). On the almost unbiased ridge regression estimator. *Communications in Statistics-Simulation and Computation*, 17(3), 729-743.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482), 681-686.

Rossi, P. E., & Allenby, G. (2003). Bayesian statistics and marketing. *Marketing Science* 22(3), 304-328.

Rossi, P. E., Allenby, G, and McCulloch, R. (2012). *Bayesian statistics and marketing*. John Wiley & Sons.

Uslu, V. R., Egrioglu, E., & Bas, E. (2014). Finding optimal value for the shrinkage parameter in ridge regression via particle swarm optimization. *American Journal of Intelligent Systems*, 4(4), 142-147.

Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, 22(2), 217.

Wang, W., Zeng, G., & Tang, D. (2011). Bayesian intelligent semantic mashup for tourism. *Concurrency and Computation: Practice and Experience*, 23(8), 850-862.

Wong, K. K., Song, H., & Chon, K. S. (2006). Bayesian models for tourism demand forecasting. *Tourism Management*, 27(5), 773-780.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics* (No. 519.54 Z4).

Zyphur, M. J., & Oswald, F. L. (2013). Bayesian probability and statistics in management research: A new horizon. *Journal of Management*, 39(1), 5-13.

Zyphur, M. J., & Oswald, F. L. (2015). Bayesian estimation and inference: A user's guide. *Journal of Management*, 41(2), 390-420.