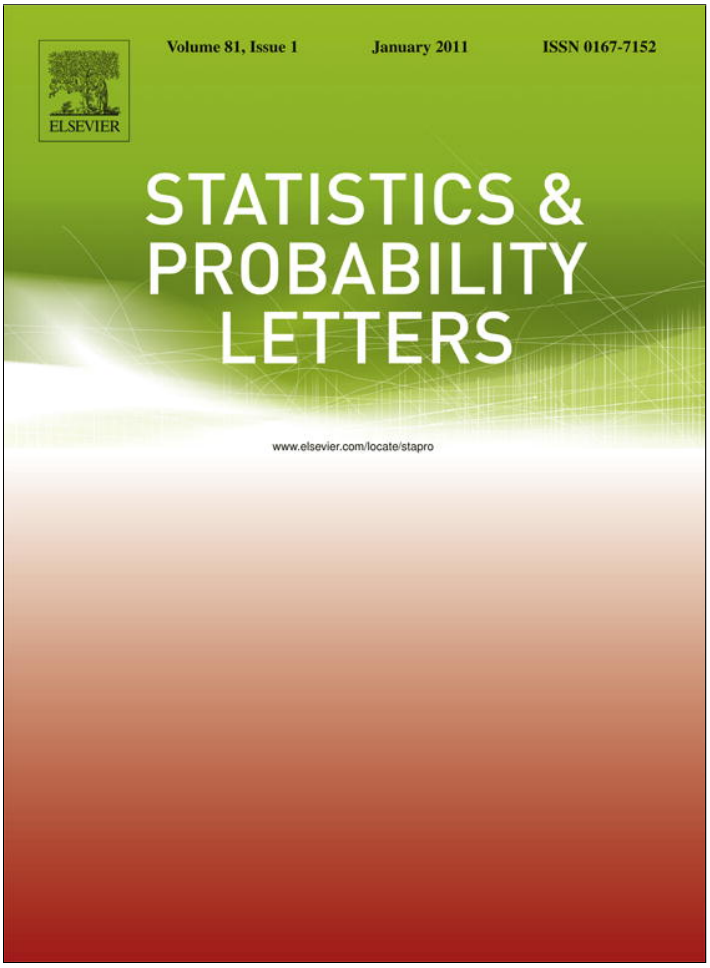

(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

Statistics and Probability Letters 82 (2012) 438–445

Contents lists available at SciVerse ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Prior influence in linear regression when the number of covariates increases to infinity

Luis Leon-Novelo[a], George Casella[b,*]

[a] Department of Statistics, University of Florida, 102 Griffin-Floyd Hall, Gainesville, FL 32611, United States
[b] Department of Statistics, University of Florida, Gainesville, FL 32611, United States

ARTICLE INFO

Article history:
Received 29 March 2011
Received in revised form 19 October 2011
Accepted 26 October 2011
Available online 11 November 2011

Keywords:
Linear models
Regression
Bayes
Model selection

ABSTRACT

It is becoming more typical in regression problems today to have the situation where "$p > n$", that is, where the number of covariates is greater than the number of observations. Approaches to this problem include such strategies as model selection and dimension reduction, and, of course, a Bayesian approach. However, the discrepancy between p and n can be so large, especially in genomic data, that examining the limiting case where $p \to \infty$ can be a relevant calculation. Here we look at the effect of a prior distribution on the coefficients, and in particular characterize the conditions under which, as $p \to \infty$, the prior does not overwhelm the data. Specifically, we find that the prior variance on the growing number of covariates must approach zero at rate $1/p$, otherwise the prior will overwhelm the data and the posterior distribution of the regression coefficient will equal the prior distribution.

1. Introduction

In clinical data it is now often the case that the number of different measurements per patient is larger than the number of patients in the study. This becomes more dramatic in genomic studies, when many different gene expressions are measured for each subject. For example, there may be thousands of genes with fewer than one hundred patients. Lee et al. (2003) cite studies where the number of genes selected as having an impact on a response variable (for example, a disease) is significantly larger than the number of sample points.

Here we are interested in the relationship between the covariates and the outcome when considering a regression model with p covariates and n observations. Classical (frequentist) regression requires $p < n$ to have unequivocal inferences. When $p > n$, in order to obtain inferences, classical techniques require some data reduction technique (for example, principal components) to identify an effective set of covariates that have dimension less than n. However, in such cases, the interpretation of the results in terms of the original covariates becomes complicated.

In contrast, Bayesian regression with proper priors can be directly applied when $p > n$. If the credible interval of a regression coefficient does not include zero then we assume that this covariate has an impact in the response. As long as we use a proper prior for the regression coefficients we can keep adding covariates, that is, we can let $p \to \infty$. Here we explore the behavior of the posterior distribution of the regression coefficients when this is the case.

[*] Corresponding author.
E-mail addresses: luis@stat.ufl.edu (L. Leon-Novelo), casella@ufl.edu (G. Casella).

0167-7152/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2011.10.018

1.1. Background

The literature relating to this problem is quite small. Jiang (2007), working in the context of Bayesian Variable Selection (BVS) in generalized linear models, gives an example with $p \gg n$ where using the full model (the one including all the covariates available) for the inference yields undesirable results. However, he shows that BVS performs well when the priors for the parameters of each model satisfy certain conditions, in that the predictive model implied by the selected model, $p(y \mid D_n)$ is close to the true model $p^*(y)$ where D_n is the observed data.

Gupta and Ibrahim (2009) propose a family of prior distributions when the sampling distribution belongs to the exponential family. Their prior is based on the Fisher information matrix of the coefficients of the linear element of the GLM, and reduces to the g-prior of Zellner (1971, Section 10.4) in the normal linear regression case. Using ridge regression ideas, they handle the problem of $p \gg n$ by adding $\gamma\mathbf{I}$ to the Fisher information matrix, which adds a new tuning parameter to the prior specification. This is the ridge g-prior.

1.2. Summary

Here we restrict our study to the linear regression case and, in contrast to Jiang (2007) we do not perform variable selection but rather focus on the inference on the parameters of the model. To simplify the analysis, throughout the paper, we also assume the error variance in the regression is known. We first consider the Bayesian regression setting:

$$\begin{aligned}&\mathbf{Y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim MVN_n(0, \Sigma_\epsilon)\\ &\beta \sim MVN_r(0, \Sigma_\beta),\end{aligned} \tag{1}$$

where $MVN_r(m, B)$ represents the *r-variate* normal distribution with mean m and covariance matrix B. In Section 2 we examine, as a function of the prior variances, the behavior of the posterior distribution of β when covariates are added to the model. We focus on the case Σ_ϵ proportional to the identity matrix and Σ_β either proportional to the identity or specified according to the ridge *g-prior*. We find that, as the number of additional covariates, p, grows, the prior variance must decrease like $1/p$, otherwise the data may have no influence on the posterior distribution. This is the case both for an independence prior, and the ridge g-prior. In Section 3 we give conditions on how the covariates themselves must grow in order for the prior not to overwhelm the data. Section 4 contains a short discussion, and there is a small technical Appendix.

2. The influence of the prior as $p \to \infty$

The posterior distribution of the regression parameter β in model (1) is then $MVN_p(\delta_\beta, \Lambda_\beta)$ where,

$$\Lambda_\beta = \left(\mathbf{X}'\Sigma_\epsilon^{-1}\mathbf{X} + \Sigma_\beta^{-1}\right)^{-1}, \quad \text{and} \quad \delta_\beta = \Lambda_\beta \mathbf{X}'\Sigma_\epsilon^{-1}\mathbf{Y}. \tag{2}$$

To examine the behavior of the coefficients as $p \to \infty$, we augment the model with

$$\mathbf{Y} = (\mathbf{X}{:}\, \mathbf{C})\beta_{r+p} + \epsilon, \quad \epsilon \sim MVN(0, \sigma^2\mathbf{I}) \tag{3}$$

where we augment the $n \times r$ matrix $\mathbf{X}$ with an $n \times p$ matrix $\mathbf{C}$, column bound to the design matrix $\mathbf{X}$. Here we consider r fixed, which may reflect coefficients of major interest, and we look at the effect on these coefficient estimates as $p \to \infty$. As the number of columns in $\mathbf{C}$ grows, we must be concerned with the limiting behavior of $(1/p)\mathbf{CC}'$. In this section we make the common assumption that

$$(1/p)\mathbf{CC}' \to \mathbf{S}, \quad \text{a positive semi-definite matrix} \tag{4}$$

and in Section 3 we relax this assumption.

2.1. The posterior distribution from independent coefficients

Now we let the regression coefficients, β, be *a priori* independent and distributed according to a normal law with known variance, in particular, we partition $\beta_{r+p} = (\beta_r, \beta_p)$, giving it a normal prior distribution with mean and covariance matrix

$$\mathrm{E}\beta_{r+p} = 0 \quad \text{and} \quad \mathrm{Var}\beta_{r+p} = \begin{pmatrix} \tau_r^2\mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \tau_p^2\mathbf{I}_p \end{pmatrix}, \tag{5}$$

where $\mathbf{I}_m$ is an $m \times m$ identity matrix. We summarize the posterior distribution of β_r in the following lemma.

Lemma 1. *For the model* (3) *with prior* (5), *the posterior distribution of* β_r *is normal with mean and variance*

$$\delta_{\beta_r} = \mathrm{E}(\beta_r \mid \mathbf{Y}) = \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\mathbf{X}'\Gamma\mathbf{X} + \frac{1}{\tau_r^2}\mathbf{I}_r\right)^{-1}\mathbf{X}'\Gamma\mathbf{Y}$$

$$V_{\beta_r} = \mathrm{Var}(\beta_r \mid \mathbf{Y}) = \left(\frac{1}{\sigma^2}\mathbf{X}'\Gamma\mathbf{X} + \frac{1}{\tau_r^2}\mathbf{I}_r\right)^{-1},$$

where

$$\Gamma = \mathbf{I}_n - \mathbf{C}\left(\mathbf{C}'\mathbf{C} + \frac{\sigma^2}{\tau_p^2}\mathbf{I}_p\right)^{-1}\mathbf{C}' = \left(\frac{\tau_p^2}{\sigma^2}\mathbf{C}\mathbf{C}' + \mathbf{I}_n\right)^{-1}. \tag{6}$$

Proof. The posterior variance of β_{r+p} is

$$\begin{pmatrix} A = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} + \frac{1}{\tau_r^2}\mathbf{I}_r & B = \frac{1}{\sigma^2}\mathbf{X}'\mathbf{C} \\ B' = \frac{1}{\sigma^2}\mathbf{C}'\mathbf{X} & D = \frac{1}{\sigma^2}\mathbf{C}'\mathbf{C} + \frac{1}{\tau_p^2}\mathbf{I}_p \end{pmatrix}^{-1} = \begin{pmatrix} A^- & B^- \\ B'^- & D^- \end{pmatrix}$$

where

$$A^- = (A - BD^{-1}B')^{-1} = \left(\frac{1}{\sigma^2}\mathbf{X}^t\ \Gamma\mathbf{X} + \frac{1}{\tau^2}\mathbf{I}_p\right)^{-1}$$

with Γ given in (6). Clearly, the matrix A^- is the covariance matrix of β_r. The calculation of $E(\beta_r \mid \mathbf{Y})$ is straightforward. □

Note that the second expression for Γ follows from Woodbury's Theorem (Woodbury, 1950; see also Hager 1989), and this matrix plays a fundamental role. The effect of the augmented coefficients on β_r is only through this matrix, which will govern it limiting behavior. In particular, if $\Gamma \to \mathbf{0}$ (or equivalently, since Γ is positive definite and $\mathbf{X} \neq \mathbf{0}$ fixed, $\mathbf{X}'\Gamma\mathbf{X} \to \mathbf{0}$) the posterior mean goes to the prior mean, 0, and the posterior covariance matrix goes to the prior covariance $\tau_r^2\mathbf{I}_p$. Thus, the prior overwhelms the data and inference is solely dependent on the prior. However, if Γ converges to a finite value then, even if p is infinite, the data still have an effect on the posterior distribution.

Using the assumption that $(1/p)\mathbf{C}\mathbf{C}' \to \mathbf{S}$, we have

$$\Gamma = \left(p\frac{\tau_p^2}{\sigma^2}[(1/p)\mathbf{C}\mathbf{C}'] + \mathbf{I}_n\right)^{-1} \approx \left(p\frac{\tau_p^2}{\sigma^2}\mathbf{S} + \mathbf{I}_n\right)^{-1},$$

and thus the limiting behavior of $p\frac{\tau_p^2}{\sigma^2}$ controls the posterior distribution of β_r. We summarize this in the following theorem.

Theorem 2. *For the model situation of* Lemma 1

$$\lim_{p\to\infty}\Gamma = \begin{cases} \mathbf{I} & \text{if } p\frac{\tau_p^2}{\sigma^2} \to 0 \\ (\mathbf{I} + c\mathbf{S})^{-1} & \text{if } p\frac{\tau_p^2}{\sigma^2} \to c \\ \mathbf{0} & \text{if } p\frac{\tau_p^2}{\sigma^2} \to \infty \end{cases}$$

where $c > 0$ *is a constant. Thus,*

(a) *If* $\Gamma \to \mathbf{I}$ *the posterior density of* β_r *converges to the usual posterior ignoring the augmented matrix* $\mathbf{C}$,
(b) *If* $\Gamma \to \mathbf{0}$, *the posterior density converges exactly to the prior density.*[1]
(c) *If* $\lim \Gamma = (\mathbf{I} + c\mathbf{S})^{-1}$, *then the limiting posterior distribution of* β_r *is from the model*

$$\mathbf{Y} = \mathbf{X}\beta_r + \epsilon, \quad \epsilon \sim N(0, \sigma^2(\mathbf{I} + c\mathbf{S})), \beta_r \sim N(0, \tau_r^2\mathbf{I}).$$

[1] The convergence of the densities implies convergence in total variation or in distribution.

We summarize the details in the following table:

$\lim p\frac{\tau_p^2}{\sigma^2}$	Limiting posterior distribution	
0	$\pi^{(0)}(\beta_r \mid \mathbf{Y}) = N\left(\frac{1}{\sigma^2}V_{\beta_r}\mathbf{X}'\mathbf{Y}, V_{\beta_r}\right)$,	$V_{\beta_r} = \left(\frac{1}{\sigma^2}\mathbf{X}'\mathbf{X} + \frac{1}{\tau_r^2}\mathbf{I}_r\right)^{-1}$
$0 < c < \infty$	$\pi^{(c)}(\beta_r \mid \mathbf{Y}) = N\left(\frac{1}{\sigma^2}V_{\beta_r}\mathbf{X}'(\mathbf{I}+c\mathbf{S})^{-1}\mathbf{Y}, V_{\beta_r}\right)$,	$V_{\beta_r} = \left(\frac{1}{\sigma^2}\mathbf{X}'(\mathbf{I}+c\mathbf{S})^{-1}\mathbf{X} + \frac{1}{\tau_r^2}\mathbf{I}_r\right)^{-1}$
∞	$\pi^{(\infty)}(\beta_r \mid \mathbf{Y}) = N(0, \tau_r^2\mathbf{I}_r)$	

So the limiting behavior of $p\frac{\tau_p^2}{\sigma^2}$ is key in understanding the effect of the infinite augmentation. In particular, we need to be concerned when this quantity goes to infinity, as then the prior overwhelms that data, which is never a good idea! This quantity remains finite only if $p\frac{\tau_p^2}{\sigma^2} \to c$, which means that $\frac{\tau_p^2}{\sigma^2} = O(1/p)$ and, in particular, goes to zero as $p \to \infty$. Since the prior mean is also zero, this means that the coefficients of β_p are all converging *a priori* to zero in probability, and thus have little or no effect on the estimation of β_r.

The prior probability that the prior on β_p gives to any ball centered at 0 of radius c is $P(\chi_p^2 \le c/\tau_p^2)$ with χ_p^2 denoting a chi-square r.v. with p degrees of freedom. Since $\chi_p^2 \to \infty$ a.s. as $p \to \infty$, if the limiting value of τ_p^2 is either a positive constant or infinity, then mass will be placed infinitely far away from any fixed point in the parameter space of β_ps (i.e., outside any ball with fixed radius) and hence will influence the estimation greatly. In the latter case it completely wipes out any influence of the data on the estimation of β_r. If the limiting value of $p\tau_p^2$ is constant, this effect does not overwhelm the data.

2.2. Ridge g-prior

Now consider the model (1) with a ridge g-prior on β defined by Gupta and Ibrahim (2009). They assume that $\Sigma_\epsilon = \sigma^2\mathbf{I}_n$, $\Sigma_\beta = g\sigma^2(\mathbf{X}'\mathbf{X} + \gamma\mathbf{I})^{-1}$ and $g, \gamma > 0$ are chosen by the user, where we assume that σ^2 is known. This is a modification of the original g-prior of Zellner (1971), which assumes $\Sigma_\beta = g\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Sabanés Bové and Held (2011) notice that the original g-prior can be interpreted as the posterior after observing an imaginary sample $\mathbf{Y}_0 = 0$ of size n from the regression model with known variance, $\mathbf{Y}_0 \mid \beta \sim N_n(\mathbf{X}_0\beta, g\sigma^2\mathbf{I}_n)$, with prior $p(\beta) \propto 1$ and $\mathbf{X}_0$ any design matrix such that $\mathbf{X}_0'\mathbf{X}_0 = \mathbf{X}'\mathbf{X}$.

In order to include the "$p > n$" case Gupta and Ibrahim (2009) add a ridge regression parameter γ to the covariance considered by Zellner. However, as we know from the results in the previous section, we cannot put the same prior on β_r and β_p, so we modify the g-prior, and let the prior β_{r+p} have prior mean zero and prior variance

$$g\sigma^2\begin{pmatrix}\mathbf{X}'\mathbf{X} + \gamma_r\mathbf{I}_r & \mathbf{X}'\mathbf{C} \\ \mathbf{C}'\mathbf{X} & \mathbf{C}\mathbf{C}' + \gamma_p\mathbf{I}_p\end{pmatrix}^{-1}$$

with $\gamma_r, \gamma_p > 0$ chosen by the user. Similar to the calculations in the previous section, the posterior distribution of β_r is normal with mean and variance

$$\delta_{\beta_r} = \mathrm{E}(\beta_r \mid \mathbf{Y}) = \left((g+1)\mathbf{X}'\Gamma_R\mathbf{X} + \gamma_r\mathbf{I}\right)^{-1}\mathbf{X}'\Gamma_R\mathbf{Y}$$
$$V_{\beta_r} = \mathrm{Var}(\beta_r \mid \mathbf{Y}) = g\sigma^2\left((g+1)\mathbf{X}'\Gamma_R\mathbf{X} + \gamma_r\mathbf{I}\right)^{-1},$$

where

$$\Gamma_R = \mathbf{I} - (g+1)\mathbf{C}[(g+1)\mathbf{C}'\mathbf{C} + \gamma_p\mathbf{I}_p]^{-1}\mathbf{C}' = \left(\frac{(g+1)p}{\gamma_p}\frac{1}{p}\mathbf{C}\mathbf{C}' + \mathbf{I}_n\right)^{-1}.$$

Noting the similarity to (6), we see that here p/γ_p will play the same role as $p\frac{\tau_p^2}{\sigma^2}$ in Section 2.1, that is, γ_p must grow like p. We can summarize the conclusions in the following table:

$\lim \frac{p}{\gamma_p}$	Limiting posterior distribution	
∞	$\pi^{(0)}(\beta_r \mid \mathbf{Y}) = N\left(\frac{1}{\sigma^2}V_{\beta_r}\mathbf{X}'\mathbf{Y}, V_{\beta_r}\right)$,	$V_{\beta_r} = g\sigma^2\left((g+1)\mathbf{X}'\mathbf{X} + \gamma_r\mathbf{I}\right)^{-1}$
$0 < c < \infty$	$\pi^{(c)}(\beta_r \mid \mathbf{Y}) = N\left(\frac{1}{\sigma^2}V_{\beta_r}\mathbf{X}'(\mathbf{I}+c(g+1)\mathbf{S})^{-1}\mathbf{Y}, V_{\beta_r}\right)$,	$V_{\beta_r} = \left(\frac{1}{\sigma^2}\mathbf{X}'(\mathbf{I}+c(g+1)\mathbf{S})^{-1}\mathbf{X} + \gamma_r\mathbf{I}_r\right)^{-1}$
0	$\pi^{(\infty)}(\beta_r \mid \mathbf{Y}) = N\left(0, g\sigma^2\left(\mathbf{X}'\mathbf{X} + \gamma_r\mathbf{I}\right)^{-1}\right)$	

3. Limiting behavior of the augmented matrix

Thus far we have assumed (4), that the covariates $\mathbf{C}$ result in a convergent $\frac{1}{p}\mathbf{C}\mathbf{C}'$ matrix. We now relax that assumption, and look more closely at conditions on $\mathbf{C}$ under which Γ remains finite or converges to $\mathbf{0}$ for the case considered in Section 2.1.

3.1. Conditions on eigenvalues

We work with the matrix $\mathbf{CC}' = \Delta' D \Delta$ with $D = \text{diag}(\lambda_1, \ldots, \lambda_n)$ and $\Delta'\Delta = \mathbf{I}_n$, the spectral decomposition of $\mathbf{CC}'$, and assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. From (6) we have

$$\Gamma = \left(\frac{\tau_p^2}{\sigma^2}\mathbf{CC}' + \mathbf{I}_n\right)^{-1} = \left(\Delta'\left(\frac{\tau_p^2}{\sigma^2}D + \mathbf{I}_n\right)\Delta\right)^{-1} = \Delta'\left(\frac{\tau_p^2}{\sigma^2}D + \mathbf{I}_n\right)^{-1}\Delta$$

$$= \Delta'\begin{pmatrix} 1/\left(1+\frac{\tau_p^2}{\sigma^2}\lambda_1\right) & & \\ & \ddots & \\ & & 1/\left(1+\frac{\tau_p^2}{\sigma^2}\lambda_n\right)\end{pmatrix}\Delta. \tag{7}$$

Clearly if $\frac{\tau_p^2}{\sigma^2}\lambda_n \to \infty$, we have $\Gamma \to \mathbf{0}$ and the posterior of β_r converges to the prior. We now look at two cases in which the posterior may be different from the prior. We write $\lambda_i(p)$ to make explicit the dependence of λ_i on p.

1. If $\lambda_n(p) \to \infty$ as $p \to \infty$.
   We require $\lambda_n(p)\tau_p^2/\sigma^2 \to c < \infty$ or equivalently (when considering σ^2 fixed), *a priori* $\beta_p \to 0$ in probability, so that the posterior density does not tend to the prior density. If this is the case, since the sample size is fixed, applying Lemma 1 to β_p (instead of β_r), it is easy to see that each entry of $\beta_p \to 0$ in probability *a posteriori*.
   We note that a necessary condition for $\lambda_n(p) \to \infty$ is that the Euclidean norm of each of the rows of $\mathbf{C}$ goes to infinity (see Theorem 3 below).
2. If $\lim \tau_p/\sigma^2 \to c > 0$, where c is finite.
   We require that $\lim \lambda_i(p) < \infty$, for some $1 \leq i \leq n$ so that $\Gamma \nrightarrow \mathbf{0}$. If $i > 1$ then $\Gamma \to \Gamma^*$, a nonzero matrix that is not of full rank. In this case there is a posterior different from the prior but the coefficients of β_r are not estimable (remember, a parameter is estimable if distinct values of the parameter always correspond to distinct values of the likelihood function). If $\lim \lambda_1(p) < \infty$, then $\Gamma \to (I + c_\infty S_\infty)^{-1}$ for some finite constant c_∞ and finite matrix S_∞. This puts us in the same situation as when $\lim p\frac{\tau_p^2}{\sigma^2}$ is finite in Section 2.1. Here β_r has a limiting posterior distribution that depends on the data, and all coefficients are estimable. The Appendix establishes the convergence result.

   We note that similar results hold for the ridge g-prior, but will omit the details.

3.2. Conditions on the rows

Here we examine conditions on the rows of $\mathbf{C}$ that lead to different limiting behavior, and characterize how the rows of the augmented matrix (the values of the covariates) must behave in order for the posterior of β_r to converge to the prior. The next theorem gives conditions under which the eigenvalues of a matrix $\mathbf{ZZ}'$, with $\mathbf{Z}$ a matrix of dimension $n \times p$, tend to infinity as the number of columns, p, of $\mathbf{Z}$ goes to infinity. Notice that this is the case if each new row of Z (i) stays orthogonal to the others, and (ii) its Euclidean norm diverges. The following theorem is a relaxation of these two conditions. Let z_{ij} be the ij-entry of $\mathbf{Z}$, let Z_i denote its i-th row, and define $\langle Z_i, Z_k\rangle = \sum_{l=1}^p z_{il}z_{kl}$ as the standard Euclidean norm, that is, the dot product.

Theorem 3. *Let* $\mathbf{Z}$ *be an* $n \times p$ *matrix. If*

$$\lim_{p\to\infty}\left(\langle Z_i, Z_i\rangle - \sum_{k\in\{1,\ldots,n\}\setminus\{i\}}\langle Z_i, Z_k\rangle\right) = \infty, \quad \textit{for all } i = 1, \ldots, n, \tag{8}$$

then the minimum singular value of $\mathbf{ZZ}'$, $\lambda_n(\mathbf{ZZ}')$, *tends to infinity as* $p \to \infty$.

Proof. The theorem is an application of Eq. (1) in Johnson and Szulc (1998) stated here: Let $\mathbf{A} = (a_{ij})$ be an $n \times n$ square complex matrix and define, for $k = 1, \ldots, n$,

$$P_k(\mathbf{A}) = \sum_{j\neq k}|a_{kj}| \quad \text{and} \quad Q_k(\mathbf{A}) = \sum_{j\neq k}|a_{jk}|$$

then the smallest singular value of $\mathbf{A}$ satisfies

$$\lambda_n(A) \geq \min_{1\leq k\leq n}\left\{\frac{1}{2}\left(|a_{kk}| - \frac{1}{2}[P_k(\mathbf{A}) + Q_k(\mathbf{A})]\right)\right\}. \tag{9}$$

Setting $\mathbf{A} = \mathbf{ZZ}'$ and noticing that $P_i(\mathbf{ZZ}') = Q_i(\mathbf{ZZ}')$ and $(\mathbf{ZZ}')_{ik} = \langle Z_i, Z_k \rangle$, we obtain the left side of (8) (without the $1/2$) and prove the theorem. □

The next lemma approximates the situation of standardized independent covariates. It says that, no matter how large the coefficients β_r are in the true model generating the data, if we keep adding noisy independent covariates, our estimates of $\beta_r \to 0$ in probability, leading to an incorrect inference.

Lemma 4. *Consider the model in (3) with the prior in (5) with τ_r^2 and τ_p^2 fixed. Also assume that the entries of $\mathbf{C}$, c_{ij}, are iid with mean 0 and variance 1. Then the posterior for β_r converges to the prior as the number of covariates $p \to \infty$.*

Proof. We just need to prove that $\Gamma \to \mathbf{0}$ or equivalently that $\mathbf{C}$ satisfies (8). Starting from

$$\mathrm{E}\left(c_{il}^2 - \sum_{k\neq i} c_{il}c_{kl}\right) = \mathrm{E}(c_{il}^2) - \mathrm{E}(c_{il})\mathrm{E}\left(\sum_{k\neq i} c_{kl}\right) = 1,$$

and applying the "law of total variance", we obtain

$$\begin{aligned}\mathrm{Var}\left(c_{il}^2 - \sum_{k\neq i} c_{il}c_{kl}\right) &= \mathrm{E}\left(\mathrm{Var}\left(c_{il}^2 - \sum_{k\neq i} c_{il}c_{kl} \mid c_{il}\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(c_{il}^2 - \sum_{k\neq i} c_{il}c_{kl} \mid c_{il}\right)\right)\\ &= \mathrm{E}\left(c_{il}^2\mathrm{Var}\left(\sum_{k\neq i} c_{kl}\right)\right) + \mathrm{Var}\left(c_{il}^2 - c_{il}\mathrm{E}\left(\sum_{k\neq i} c_{kl}\right)\right)\\ &= (n-1)\mathrm{E}\left(c_{il}^2\right) + \mathrm{Var}\left(c_{il}^2\right)\\ &= (n-1) + 1 = n.\end{aligned}$$

Therefore, $c_{il}^2 - \sum_{k\neq i} c_{il}c_{kl} - 1$ has mean 0 and variance n, and the law of large numbers yields $(1/p)\sum_{l=1}^{p}\left(c_{il}^2 - \sum_{k\neq i} c_{il}c_{kl}\right) - 1 \to 0$ a.s. as $p \to \infty$. □

From a frequentist point of view, when $\Gamma \to \mathbf{0}$ and τ_p^2/σ^2 converges to a finite number, regardless of the data generating distribution g, $\lim_{p\to\infty} \mathrm{E}_g\,[E(\beta_r|\mathbf{Y})] = 0$ and $\lim_{p\to\infty} \mathrm{Var}_g\,(E(\beta_r|\mathbf{Y})) = \mathbf{0}$. This is easy to see from the expression of $E(\beta_r \mid \mathbf{Y})$.

4. Discussion

Bayesian regression easily provides a solution to the case "$p > n$" in the sense that, if a proper prior is used, the resulting matrices are nonsingular and coefficients estimates can be computed. However, care must be taken because, unless the prior is carefully specified, it can overwhelm the data and result in useless inferences. From the spectral decomposition (7), we see that the quantities $\frac{\tau_p^2}{\sigma^2}\lambda_i(p)$, $i = 1, \ldots, n$ govern the limiting behavior of the posterior distribution of β_r. Only when the limits of these quantities are finite do we get a posterior distribution for β_r that depends on the data. In Section 2 we looked at the case when $\lambda_i(p)/p$ remained finite, as reflected in the assumption (4). In such a case we then need $\frac{\tau_p^2}{\sigma^2} \propto 1/p$ in order to have a data-dependent posterior.

The results in Section 3.2, about the limiting behavior of the augmented rows, are more in a frequentist vein (although the overall framework is still Bayesian). There we see that the augmented rows must grow at the correct rate in order to insure that the posterior will reflect the data.

There is actually an interplay between the eigenvalues of $\mathbf{CC}'$ and τ_p^2/σ^2, as can be seen from (7). To have a nondegenerate posterior for β_r requires only that $\frac{\tau_p^2}{\sigma^2}\lambda_i(p)$ have a finite limit for some $1 \le i \le n$. This suggests choosing τ_p^2 proportional to $[\lambda_1(p)]^{-1}$.

Lastly, we make two further observations. First, there is actually a stronger convergence result that can be established in that, if $\lambda_n(p) \to \infty$, the Euclidean distance between the posterior and prior expectations go to zero. And, we note that, although we have worked in a special case of linear regression, it seems reasonable to expect that some version of these results will apply to more complex linear and generalized linear models.

Acknowledgments

Thanks to the Editor and reviewers for their careful reading of our paper and suggestions that enhanced the presentation. The first author was supported by NIH 1R01GM081704. The second author was supported by National Science Foundation Grants DMS-0631632 and SES-0631588, and NIH 1R01GM081704, United States.

Appendix. Convergence Proof

We first establish convergence when the largest eigenvalue converges, that is, when $\lim_{p\to\infty} \lambda_1(p) < \infty$. If this limit is infinity, but $\lim_{p\to\infty} \lambda_i(p) < \infty$ for some $1 \le i \le n$, the situation is slightly more complicated, but is summarized in Corollary 6.

Theorem 5. *In the notation of Section 3, let $\lambda_i(p)$, $i = 1, \dots, n$ be the eigenvalues of $\mathbf{CC}'$. If $\lim_{p\to\infty} \lambda_i(p) < \infty$, for $1 \le i \le n$, then Γ converges to a finite limit.*

Proof. In the proof write $\mathbf{C} = \mathbf{C}_p$, the $n \times p$ matrix, to clarify the dependence on p.

We note the following:

1. The n largest singular values of $\mathbf{C}_p\mathbf{C}'_p$ are the same as those of $\mathbf{C}'_p\mathbf{C}_p$. It is also the case that for this latter matrix, $\lambda_i(p) = 0$ for $i > n$.
2. The sequence $\{\lambda_i(p)\}_{p=1}^{\infty}$ is nondecreasing for each $i = 1, \dots, n$. The Interlacing Property (Bhatia, 1997, page 59) establishes that if $\mathbf{A}$ is an $n \times n$ symmetric matrix and $\mathbf{A}_i$ is a matrix obtained from $\mathbf{A}$ by removing the i-th row and column, then the eigenvalues of $\mathbf{A}$ satisfy
$$\lambda_1(\mathbf{A}) \ge \lambda_1(\mathbf{A}_i) \ge \lambda_2(\mathbf{A}) \ge \lambda_2(\mathbf{A}_i) \ge \cdots \ge \lambda_{n-1}(\mathbf{A}_i) \ge \lambda_n(\mathbf{A}).$$
If $\mathbf{C}_{p+1} = (\mathbf{C}_p \colon \mathbf{c})$, that is, we augment $\mathbf{C}_p$ with one more column, then applying the Interlacing Property to $\mathbf{C}'_{p+1}\mathbf{C}_{p+1}$ for $p > n$, we get,
$$\lambda_1(p+1) \ge \lambda_1(p) \ge \lambda_2(p+1) \ge \lambda_2(p) \ge \cdots \ge \lambda_n(p+1) \ge \lambda_n(p) \ge \lambda_{n+1}(p+1) = 0.$$
3. Using 1., 2., and the hypothesis that $\lim_{p\to\infty} \lambda_1(p) < \infty$ imply that, as $p \to \infty$, $\lambda_i(p) \uparrow \lambda_i^*$ with λ_i^* finite, for $i \ge 1$.
4. Γ converges in the Hilbert space of $n \times n$ matrices with Frobenius norm.
The Frobenius norm of an $n \times n$ matrix $\mathbf{A}$ is
$$\|\mathbf{A}\|_F^2 = \sum_{1\le i,j\le n} (\mathbf{A})_{ij}^2 = trace(\mathbf{A}^*\mathbf{A}) = \sum_{i=1}^{n} \nu_i(\mathbf{A})^2,$$
where $\mathbf{A}^*$ is the conjugate transpose of $\mathbf{A}$ and $\nu_i(\mathbf{A})$ denotes the i-th largest eigenvalue of the matrix $\mathbf{A}$.
To prove statement 4, by (7) it is enough to prove that $\mathbf{C}_p\mathbf{C}'_p$ converges. In order to do so, we prove that the sequence $\{\mathbf{C}_p\mathbf{C}'_p\}_{p=1}^{\infty}$ is Cauchy. From the Interlacing Property it follows that $\{\sum_{i=1}^{n} \lambda_i(p)\}_{p=1}^{\infty}$ is Cauchy. Thus, given $\epsilon > 0$, there exists $p_0 > 0$ such that for all $p > p_0$ and $m > 0$, $\sum_{i=1}^{n} \lambda_i(p+m) - \sum_{i=1}^{n} \lambda_i(p) < \epsilon$. Moreover, it also follows that
$$\mathbf{C}_p\mathbf{C}'_p - \mathbf{C}_{p+m}\mathbf{C}'_{p+m} = \left[\mathbf{C}_p\mathbf{C}'_p - (\mathbf{C}_p \colon \mathbf{C}_m)(\mathbf{C}_p : \mathbf{C}_m)'\right] = -\mathbf{C}_m\mathbf{C}'_m,$$
and therefore
$$\|\mathbf{C}_p\mathbf{C}'_p - \mathbf{C}_{p+m}\mathbf{C}'_{p+m}\|_F^2 = \text{tr}\left(\mathbf{C}_m\mathbf{C}'_m\right)^2 = \sum_{i=1}^{n} \lambda_i'^2(m),$$
where $\lambda'_1(m) \ge \lambda'_2(m) \ge \cdots \ge \lambda'_n(m)$ are the eigenvalues of $\mathbf{C}_m\mathbf{C}'_m$. We then have
$$\sum_{i=1}^{n} \lambda_i(p+m) = tr(\mathbf{C}_p\mathbf{C}'_p) + tr(\mathbf{C}_m\mathbf{C}'_m) = \sum_{i=1}^{n} \lambda_i(p) + \sum_{i=1}^{n} \lambda'_i(m)$$
implying,
$$\sum_{i=1}^{n} \lambda'_i(m) = \sum_{i=1}^{n} \lambda_i(p+m) - \sum_{i=1}^{n} \lambda_i(p) < \epsilon \quad \text{for all } p > p_0 \text{ and } m > 0.$$
Then, if $0 < \epsilon < 1$, since $0 < \lambda'_i(m) < \epsilon < 1$ for $i = 1, \dots, n$, we have
$$\|\mathbf{C}_p\mathbf{C}'_p - \mathbf{C}_{p+m}\mathbf{C}'_{p+m}\|_F^2 = \sum_{i=1}^{n} (\lambda'_i(m))^2 < n\epsilon^2. \quad \square$$

Corollary 6. *In the notation of Section 3, let $\lambda_i(p)$, $i = 1, \dots, n$ be the eigenvalues of $\mathbf{CC}'$. If $\lim_{p\to\infty} \lambda_i(p) < \infty$, for some $1 \le i \le n$, then Γ does not degenerate to the zero matrix.*

Proof. It should be clear from (7) that if some, but not all, of eigenvalues of $\mathbf{CC}'$ go to infinity, the limiting Γ matrix will be different from the zero matrix. However, this limit may not be unique. □

References

Bhatia, R., 1997. Matrix analysis. In: Graduate Texts in Mathematics. Springer.

Gupta, M., Ibrahim, J.G., 2009. An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data. Statistica Sinica 19 (4), 1641–1663.

Hager, W.W., 1989. Updating the inverse of a matrix. SIAM Review 31, 221–239.
Jiang, W., 2007. Bayesian variable selection for high dimensional generalized linear models: convergence rates of the fitted densities. Annals of Statistics 35 (4), 1487–1511.
Johnson, C.R., Szulc, T., 1998. Further lower bounds for the smallest singular value. Linear Algebra and its Applications 272 (1–3), 169–179.
Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M., Mallick, B.K., 2003. Gene selection: a Bayesian variable selection approach. Bioinformatics 19 (1), 90–97.
Sabanés, Bové, D., Held, L., 2011. Hyper-g priors for generalized linear models. Bayesian Analysis 6 (1), 1–24.
Woodbury, M.A., 1950, Inverting modified matrices. Technical Report 42, Statistical research group, Princeton, NJ: Princeton University.
Zellner, A., 1971. An Introduction to Bayesian Inference in Econometrics. Wiley.