Published in final edited form as:
Stat Sin. 2009 ; 19(4): 1641–1663.

An Information Matrix Prior for Bayesian Analysis in Generalized Linear Models with High Dimensional Data

Mayetri Gupta[*] and **Joseph G. Ibrahim**[†]

[*]Department of Biostatistics, Boston University, MA 02118, U.S.A. Email: gupta@bu.edu; Phone: 1.617.414.7946; Fax: 1.617.638.6484

[†]Department of Biostatistics, University of North Carolina at Chapel Hill, NC 27599, U.S.A. Email: ibrahim@bios.unc.edu

Abstract

An important challenge in analyzing high dimensional data in regression settings is that of facing a situation in which the number of covariates p in the model greatly exceeds the sample size n (sometimes termed the "$p > n$" problem). In this article, we develop a novel specification for a general class of prior distributions, called Information Matrix (IM) priors, for high-dimensional generalized linear models. The priors are first developed for settings in which $p < n$, and then extended to the $p > n$ case by defining a ridge parameter in the prior construction, leading to the Information Matrix Ridge (IMR) prior. The IM and IMR priors are based on a broad generalization of Zellner's g-prior for Gaussian linear models. Various theoretical properties of the prior and implied posterior are derived including existence of the prior and posterior moment generating functions, tail behavior, as well as connections to Gaussian priors and Jeffreys' prior. Several simulation studies and an application to a nucleosomal positioning data set demonstrate its advantages over Gaussian, as well as g-priors, in high dimensional settings.

Keywords

Fisher Information; g-prior; Importance sampling; Model identifiability; Prior elicitation

1 Introduction

In the analysis of data arising in many scientific applications, one often faces a scenario in which the number of variables (p) greatly exceeds the sample size (n), often termed the "$p > n$" problem. In these problems, fitting many types of statistical models leads to model nonidentifiability in which parameters cannot be estimated via maximum likelihood. The specification of proper priors can alleviate such a nonidentifiability problem and lead to proper posterior distributions as long as one uses a valid probability density for the data, i.e., $\int f(y|\theta)\, dy < \infty$. Specification of proper priors in the $p > n$ context is not easy since it is desirable that the prior (i) leads to existence of prior or posterior moments, which is not guaranteed and theoretically checking this is not easy; (ii) is relatively non-informative so that the data can essentially drive the inference; (iii) is at least somewhat semi-automatic in nature requiring relatively little or minimal hyper-parameter specification; and (iv) is easy to interpret and computationally feasible.

Properties (i) – (iv) are important to investigate in considering a class of priors for posterior inference. The existence of posterior moments is important since Bayesian inference often relies on the estimation of posterior quantities (e.g. means), via Markov chain Monte Carlo (MCMC) or other sampling procedures. Computing Bayes factors and other model assessment criteria often requires the existence of posterior moments (Chen, Ibrahim, and Yiannoutsos (1999)). Constructing estimates through a black box use of MCMC, without knowing that these estimates are well defined for the class of priors considered, may lead to nonsensical posterior inference. Non-informativeness is desirable in model selection and model assessment settings, or wherever prior information is not available. The specification of semi-automatic priors has been advocated by Spiegelhalter and Smith (1982), Zellner (1986), Mitchell and Beauchamp (1988), Berger and Pericchi (1996), Bedrick, Christensen, and Johnson (1996), Chen, Ibrahim, and Yiannoutsos (1999), and Ibrahim and Chen (2003), and is attractive in high dimensional settings where it is difficult to find contextual interpretations for all parameters.

In the $p > n$ paradigm, there has been little work on the specification of desirable priors and, in particular, priors that attain (i) – (iv) above. West (2003) specifies singular g-priors (Zellner (1986)) for the Gaussian model in the context of Bayesian factor analysis for $p > n$. Liang, Paulo, Molina, Clyde, and Berger (2008) advocate the use of mixtures of g-priors to resolve consistency issues in model selection but their adaptation does not directly extend to cases where $p > n$. When $p > n$, $N_p(0, \gamma I)$ priors are often not desirable– for small to moderate γ, they are typically too informative, and for large γ they often lead to computationally unstable posteriors since the model becomes weakly identified (see Section 4). They also do not capture the a priori correlation in the parameters, and eliciting a prior covariance matrix when $p > n$ is extremely difficult.

The class of priors we consider are called Information Matrix (IM) priors. They can be applied in any parametric regression context but we initially focus on generalized linear models (GLMs). The functional form of the IM prior is obtained once a parametric statistical model is specified for the data, the kernel of the IM prior being specified through the Fisher Information matrix.

Let $(y_1,\ldots, y_n)$ be independent univariate response variables with density $p(y_i|X, \boldsymbol{\beta}, \boldsymbol{\phi})$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, $\boldsymbol{\phi} = (\phi_1,\ldots, \phi_q)$ a vector of dispersion parameters, and X the $n \times p$ matrix of covariates with i-th row $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip})$, where ‘′’ denotes matrix transposition. Let $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\phi})$. Assume for the moment that $\boldsymbol{\phi}$ is known, $p < n$, and rank$(X) = p$. The likelihood function of $\boldsymbol{\beta}$ for a regression model with independent responses is $L(\beta) = \prod_{i=1}^{n} p(y_i|\boldsymbol{x}_i, \beta)$, and the (i, j)-th element of the Fisher information matrix is

$$I_{ij}(\beta) = -E\left(\frac{\partial^2}{\partial\beta_i\partial\beta_j}\log(L(\beta))\right),$$

where the expectation is with respect to $y|\boldsymbol{\beta}$, and β_i is the i-th component of $\boldsymbol{\beta}$, $(i, j = 1, \ldots p)$. Further, assume that the $p \times p$ Fisher information matrix, $I(\boldsymbol{\beta})$, is non-singular, as is often the case when $p < n$. The general IM prior is now defined as

$$\pi_{IM}(\beta) \propto |I(\beta)|^{1/2} \exp\left\{-\frac{1}{2c_0}(\beta - \mu_0)' I(\beta)(\beta - \mu_0)\right\}, \tag{1.1}$$

where $\boldsymbol{\mu}_0$ and $c_0 \geq 0$ are specified location and dispersion hyperparameters. The IM prior (1.1) captures the prior covariance of $\boldsymbol{\beta}$ via the Fisher information matrix, which seems an attractive specification since this matrix plays a major role in the determination of the large sample covariance of $\boldsymbol{\beta}$ in both Bayesian and frequentist inference. The use of the design matrix *X* is attractive since *X* may reveal redundant covariates. The prior (1.1) is semi-automatic, requiring specifications only for $\boldsymbol{\mu}_0$ (which can be taken to be 0), and the scalar c_0. It should be mentioned that some recent work by Wang and George (2007) and Wang (2002) uses a related form of the prior for **β**, with two important differences: (i) the form of the prior for all GLMs is taken to be Gaussian, and (ii) the covariance is dependent on the *observed* information matrix, rather than the *expected* one, leading to a data-dependent prior, unlike our specification here. We now focus on the development of these priors and the resultant posterior estimators in the class of GLMs, where many theoretical and computational properties can be characterized for $p < n$ and $p > n$.

2 IM Priors for Generalized Linear Models

We first consider the IM prior for GLMs when $p < n$, and where the dispersion parameter $\boldsymbol{\phi}$ is known or intrinsically fixed, as for example in the binomial, Poisson, and exponential regression models. For a GLM, $y_i|\boldsymbol{x}_i$, ($i = 1, \ldots, n$) has a conditional density given by

$$p(y_i|\boldsymbol{x}_i,\beta,\phi)=\exp\{a_i^{-1}(\phi)(y_i\theta_i-b(\theta_i))+c(y_i,\phi)\}, i=1,2,\ldots,n, \tag{2.1}$$

where the canonical parameter θ_i satisfies the equations $\theta_i = \theta(\eta_i)$, and $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ determine a particular family in the class. The function θ(.) is the link function for the GLM, often referred to as the θ-link, and $\eta_i = \boldsymbol{x}_i'\boldsymbol{\beta}$. When a canonical link is used, $\theta(\eta_i) = \eta_i$. The function $a_i(\boldsymbol{\phi})$ is commonly of the form $a_i(\phi)=\phi^{-1}w_i^{-1}$, where w_i's are known weights. Without loss of generality and for ease of exposition, let $\boldsymbol{\phi} = 1$ and $w_i = 1$. Then (2.1) can be rewritten as

$$p(y_i|x_i,\beta)=\exp\{y_i\theta_i-b(\theta_i)+c(y_i)\}, i=1,2,\ldots,n. \tag{2.2}$$

Now, let *X* be the $n \times p$ matrix of covariates with *i-th* row $\boldsymbol{x}_i'$, $\theta(X\beta)$ be a component-wise function of $X\boldsymbol{\beta}$ that depends on the link, and $\boldsymbol{J}$ be a $n \times 1$ vector of ones. Then in matrix form, we can write the likelihood function of **β** for the GLM in (2.2) as

$$p(\boldsymbol{y}|\beta)=\exp\{\boldsymbol{y}'\theta(X\beta)-\boldsymbol{J}'b(\theta(X\beta))+c(\boldsymbol{y})\}. \tag{2.3}$$

For the class of GLMs (with $\boldsymbol{\phi} = 1$ and $w_i = 1$), the Fisher information matrix for **β** is

$$I(\beta)=X'\Omega(\beta)X, \tag{2.4}$$

$$\text{where}\ \ \Omega(\beta)=\Delta(\beta)V(\beta)\Delta(\beta), \tag{2.5}$$

where $V(\boldsymbol{\beta})$ is the $n \times n$ diagonal matrix of variance functions with *i-th* diagonal element $\upsilon_i=\upsilon(\boldsymbol{x}_i'\beta)=d^2b(\theta_i)/d\theta_i^2$, and $\Delta(\boldsymbol{\beta})$ is an $n \times n$ diagonal matrix of "link adjustments", with *i-th* diagonal element $\delta_i=\delta(\boldsymbol{x}_i'\beta)=d\theta_i/d\eta_i$. We denote the *i-th* diagonal element of $\Omega(\boldsymbol{\beta})$ by ω_i. We now can write the IM prior for **β** as

$$\pi_{IM}(\beta) \propto |X'\Omega(\beta)X|^{1/2}\exp\left\{-\frac{1}{2c_0}(\beta-\mu_0)'(X'\Omega(\beta)X)(\beta-\mu_0)\right\}. \tag{2.6}$$

The prior in (2.6) can be viewed as a generalization of several types of priors. As $c_0 \to \infty$, (2.6) converges to Jeffreys' prior for GLMs, given by

$$\pi_J(\beta) \propto |X'\Omega(\beta)X|^{1/2}. \tag{2.7}$$

Jeffreys' prior is a popular noninformative prior for Bayesian inference in multiparameter settings involving regression coefficients, due to its invariance and local uniformity properties (Kass (1989, 1990)). Also, as long as $X'\Omega(\boldsymbol{\beta})\, X$ is bounded (for example in the logistic model) as $\boldsymbol{\beta} \to \infty$, the ratio of prior (2.6) to (2.7) is zero, hence showing the IM prior has lighter tails than Jeffreys' in the limit. Jeffreys' prior for GLMs is improper for most models, except for binomial regression models (Ibrahim and Laud (1991)). For the Gaussian linear model, (2.6) reduces to Zellner's g-prior (Zellner (1986)). In this case, $\Omega(\boldsymbol{\beta}) = \sigma^2 I$, where σ^2 denotes the variance of y_i in the Gaussian linear model. For any given GLM, if $\Omega(\boldsymbol{\beta})$ is a constant matrix free of $\boldsymbol{\beta}$, say W, then the IM prior will always be a Gaussian prior, that is $\boldsymbol{\beta} \sim N_p(\boldsymbol{\mu}_0, c_0(X'WX)^{-1})$. For example, for the gamma model with log-link, $\Omega(\boldsymbol{\beta})$ is a constant (identity) matrix. In this sense, the IM prior can be thought of as a "generalized" g-prior.

The tail behavior of the IM prior can be theoretically compared with the Gaussian prior for specific GLMs. To show that its tails are heavier than a Gaussian distribution, we need

$$\lim_{\|\beta\|\to\infty} \frac{\pi_{IM}(\beta)}{\phi(\beta;\nu,\Sigma)} = \infty, \tag{2.8}$$

where $\boldsymbol{\phi}$ $(\boldsymbol{\beta};\nu, \Sigma)$ denotes the p dimensional multivariate normal density with mean ν and covariance matrix Σ. (Showing that the expression in (2.8) goes to zero indicates the tails are lighter than the Gaussian.) For the binomial regression model and the inverse Gaussian model (with canonical link), it can be shown that (2.8) holds, so that the IM prior under these models has heavier tails (Figure 1). For most GLMs, if the IM prior is proper and $\Omega(\beta) \to 0$ elementwise as $\|\boldsymbol{\beta}\| \to \infty$, then the tails of the IM prior will be heavier than those of Gaussian priors. When the elements of $\Omega(\boldsymbol{\beta})$ become infinite as $\|\boldsymbol{\beta}\| \to \infty$, the IM prior has lighter tails than the Gaussian prior, as with the Poisson model with canonical link (Figure 2). Although the IM prior for the Poisson model has lighter tails than a Gaussian prior, the IM prior for this model is much flatter in the "middle" part of the distribution and hence effectively acts as a more noninformative prior.

When the dispersion parameter $\boldsymbol{\phi}$ is unknown, the IM prior based on $\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\boldsymbol{\beta}, \boldsymbol{\phi})$, is typically not proper for GLMs and has properties that are difficult to characterize. In these cases, one can construct the IM prior of $\boldsymbol{\beta}$ conditional on $\boldsymbol{\phi}$, and then specify a proper prior for $\boldsymbol{\phi}$ such as a gamma prior or a product of independent gamma densities (Ibrahim and Laud (1991)). The conditional IM prior of $\boldsymbol{\beta}$ given $\boldsymbol{\phi}$ is defined as

$$\pi_{IM}(\beta|\boldsymbol{\phi}) \propto |I(\beta|\boldsymbol{\phi})|^{1/2}\exp\left\{-\frac{1}{2c_0}(\beta-\mu_0)'I(\beta|\boldsymbol{\phi})(\beta-\mu_0)\right\}, \tag{2.9}$$

$$\text{where} \quad I_{ij}(\beta|\boldsymbol{\phi}) = -\ E\left(\frac{\partial^2}{\partial\beta_i\partial\beta_j}\log(L(\beta,\boldsymbol{\phi}))\right),$$

and $L(\boldsymbol{\beta}, \boldsymbol{\phi})$ is the likelihood function of $(\boldsymbol{\beta}, \boldsymbol{\phi})$. We specify the joint prior of $(\boldsymbol{\beta}, \boldsymbol{\phi})$ as

$$\pi(\beta,\boldsymbol{\phi}) \propto \pi_{IM}(\beta|\boldsymbol{\phi})\pi(\boldsymbol{\phi}), \tag{2.10}$$

where $\pi(\boldsymbol{\phi})$ is a proper prior for $\boldsymbol{\phi}$. When $q = 1$, $\pi(\boldsymbol{\phi})$ can be taken to be a gamma prior and, for a general q, $\pi(\phi_1, \ldots, \phi_q)$ can be assumed to be a product of q independent gamma densities. For GLMs, following the results of Section 3, it can be shown that (2.10) is jointly proper for $(\boldsymbol{\beta}, \boldsymbol{\phi})$, when $\boldsymbol{\phi}$ is distributed as a product of gamma densities.

2.1 Conditions for the existence of the prior MGF when $p < n$

We now present theoretical results establishing conditions for the existence of the prior and posterior moment generating functions (MGFs) using the IM prior for GLMs. The proofs of these results are presented as special cases of more general theorems given in Section 3.

Sufficiency—The sufficient condition for the existence of the prior MGF of $\boldsymbol{\beta}$ for the IM prior in (2.6) is that

$$\int \exp\{\tau\theta^{-1}(r)\}\left(\frac{d^2b(r)}{dr^2}\right)^{\frac{1}{2}}dr<\infty,$$

for τ in some $(-\epsilon, \epsilon)$. This is derived as a special case of Theorem 1, (Corollary 1.2), and matches the condition of MGF existence for Jeffreys' prior when $p < n$ (Ibrahim and Laud (1991)).

Necessity—The necessary condition for existence of the prior MGF of $\boldsymbol{\beta}$ for the IM prior in (2.6) is the finiteness of the p-dimensional integral

$$\int\prod_{i=1}^{p}\left[\upsilon_j(\beta)\delta_j^2(\beta)\right]^{1/2}\exp\left\{-\frac{1}{2c_0}(\beta-\mu_0)'I(\beta)(\beta-\mu_0)+\boldsymbol{t}'\beta\right\}d\beta, \tag{2.11}$$

for any $\boldsymbol{t}$ in some p-dimensional sphere about 0, where $I(\boldsymbol{\beta}) = X'\Omega(\boldsymbol{\beta})X$. In many cases, checking (2.11) reduces to a condition involving a one-dimensional integral (Section 3.1). Note here that (2.11) is less stringent a condition than that for Jeffreys' prior, allowing prior MGFs to exist for models where Jeffreys' prior is improper. This is discussed further in Section 3.1.

2.2 Existence of the posterior MGF when $p < n$

A sufficient condition for the existence of the posterior MGF of $\boldsymbol{\beta}$ for the IM prior in (2.6) is the finiteness of the following one-dimensional integral for $\tau \in (-\epsilon, \epsilon)$, some $\epsilon > 0$:

$$\int\exp\{\tau\theta^{-1}(r)+\phi^{-1}w(yr-b(r))\}\left(\frac{d^2b(r)}{dr^2}\right)^{\frac{1}{2}}dr. \tag{2.12}$$

This is the same condition as for Jeffreys' prior in Ibrahim and Laud (1991). Since the sufficient conditions for the prior and posterior are the same as those for using Jeffreys' prior, the examples which satisfy Sufficiency conditions for Jeffreys' prior in Ibrahim and Laud (1991) also satisfy the sufficient conditions here, e.g., the posterior MGFs exist for binomial, Poisson and Gamma GLMs if none of the observed *y*'s is zero.

3 IM Ridge Priors for GLMs when *p* > *n*

When $p > n$, $\boldsymbol{\beta}$ is not identifiable in the likelihood function and the MLE of $\boldsymbol{\beta}$ does not exist. Specifying a proper prior guarantees a proper posterior in this setting; however, not all proper priors yield the existence of prior and posterior moments which are often the end objectives in Bayesian inference. To generalize the IM priors to a proper prior in the $p > n$ case, we introduce a scalar "ridge" parameter λ in the prior construction, defining the IM Ridge (IMR) prior as

$$\pi_{IMR}(\beta) \propto |X'\Omega(\beta)X+\lambda I|^{1/2}\exp\left\{-\frac{1}{2c_0}(\beta-\mu_0)'(X'\Omega(\beta)X+\lambda I)(\beta-\mu_0)\right\}. \tag{3.1}$$

Here λ can be considered a "ridge" parameter as used in regression models for high dimensional covariates, and to reduce effects of collinearity (Hoerl and Kennard (1970)). With the introduction of λ, the matrix $X'\Omega(\boldsymbol{\beta})X + \lambda I$ is always positive definite regardless of the rank of X and the form of Ω, for any GLM. As $c_0 \to \infty$ in 3.1, the IMR prior converges to a generalized Jeffreys' prior given by $\pi_J^*(\beta) \propto |X'\Omega(\beta)X+\lambda I|^{1/2}$, which is useful in the $p > n$ setting since the usual Jeffreys' prior (2.7) does not exist when $X'\Omega(\boldsymbol{\beta})X$ is singular. This idea is similar in spirit (but considerably different in form) to the introduction of a constant matrix Λ added to the sample covariance matrix S in the construction of a data-dependent prior for the population covariance matrix Σ in multivariate normal settings (Schafer (1997)).

To understand the IMR prior in (3.1), first consider the IMR prior for the linear model. In this case, the IMR prior for $\boldsymbol{\beta}|\sigma^2$ is Gaussian and the posterior distribution is also Gaussian. Specifically, consider the linear model $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 I)$, and $p > n$. The IMR prior in this case becomes $\boldsymbol{\beta}|\sigma^2 \sim N_p(\boldsymbol{\mu}_0, c_0\sigma^2(X'X + \lambda I_p)^{-1})$, and the posterior distribution of $\boldsymbol{\beta}$ given σ^2 is easily shown to be the non-singular Gaussian distribution:

$$\beta|\boldsymbol{y},\sigma^2 \sim N\left(\Sigma\left(\frac{1}{c_0}(X'X+\lambda I)\mu_0+X'\boldsymbol{y}\right),\sigma^2\Sigma\right), \tag{3.2}$$

where $\Sigma = [(1 + 1/c_0)X'X + (\lambda/c_0)I_p]^{-1}$. Theoretical properties of the IMR prior for the linear model are investigated in Section 3.3.

3.1 MGF existence for GLMs using the IMR prior

The prior MGF must exist for the posterior MGF to exist, when $p > n$. We first provide necessary and sufficient conditions for prior MGF existence.

Theorem 1—*A sufficient condition for the existence of the MGF of* $\boldsymbol{\beta}$ *based on the prior (3.1) when p > n is that the one-dimensional integral*

$$\int \exp\left\{-\tfrac{\lambda M}{2c_0}[\theta^{-1}(r)]^2+\tau_i\theta^{-1}(r)\right\}\left[\tfrac{d^2b(r)}{dr^2}\right]^{\frac{1}{2}}dr<\infty, \tag{3.3}$$

for some $\tau_i \in (-\epsilon, \epsilon)$, *where* rank $(X) = n$, *M is such that* $|X^*| \leq M^{-\frac{p}{2}}, X^{*\prime}=[X' \vdots x_0']$, *with* x_0 *as a* $(p-n) \times p$ *matrix selected such that* X^* *is positive definite.*

Proof: Proofs of all theorems are given in the Appendices (in the online Article Supplement).

Corollary 1.1: Let $\omega_k(\beta)$ denote the k-th diagonal element of $\Omega(\boldsymbol{\beta})$ in (2.5). If $\sum_{k=1}^{n}\omega_k(\beta) \leq 1$, the prior MGF of $\boldsymbol{\beta}$ from (3.1) always exists, as the integral is dominated by a Gaussian MGF.

Corollary 1.2: The sufficient condition (3.3) also holds and is identical in the p < n case and, additionally, reduces to the sufficient condition for Jeffreys' prior if $\lambda = 0$.

The IMR prior thus can be useful (and more desirable than the IM prior) even in the $p < n$ case, in the face of high-dimensionality, collinearity, or weak identifiability.

Theorem 2—*If the prior MGF exists when* $p > n$, *the p-dimensional integral*

$$\int a_s(\beta)\exp\left\{-\tfrac{1}{2c_0}(\beta-\mu_0)'(I(\beta)+\lambda I)(\beta-\mu_0)+t'\beta\right\}d\beta \tag{3.4}$$

is finite for some $t \in (-\epsilon, \epsilon)$, *for* $s = 0, 1, \ldots, p$, *where* $a_s(\boldsymbol{\beta}) = |I(\boldsymbol{\beta})^{(i_1, \ldots, i_s)}|$ *is the determinant of the* $(p-s) \times (p-s)$ *sub-matrix of* $I(\boldsymbol{\beta})$ *formed by leaving out the* $(i_1, \ldots, i_s)$*-th rows and columns of* $I(\boldsymbol{\beta}) = X'\,\Omega(\boldsymbol{\beta})X$.

Note here that $a_0(\boldsymbol{\beta}) = |I(\boldsymbol{\beta})|$, $a_{p-1}(\boldsymbol{\beta}) = \text{trace}(I(\boldsymbol{\beta}))$, and $a_p(\boldsymbol{\beta}) = 1$. As a working principle, try to check the Sufficiency condition first: if it does not hold, check the necessity condition. For the necessity condition (Theorem 2) it is easiest to first check the $p = 1$ case. Necessity does not hold if there exists no $t \in (-\epsilon, \epsilon)$ such that

$$\int a_{11}(\beta)^{\frac{1}{2}}\exp\left\{-\frac{1}{2c_0}(\beta-\mu_0)^2(a_{11}(\beta)+\lambda)+t\beta\right\}d\beta$$

is finite, where $a_{11}(\beta)=b''(\theta)\frac{d\theta}{d\eta}$. If this condition is not satisfied, the MGF does not exist. If the necessity condition is satisfied for $p = 1$, we need to check the condition for larger p; if for any p we find that the necessity condition is not satisfied, the prior MGF does not exist.

Corollary 2.1: *The result in Theorem 2 holds when* $n > p$, *and simplifies to (2.11) when* $\lambda = 0$.

Theorem 3—*A sufficient condition for the existence of the posterior MGF for* $\pi_{IMR}(\boldsymbol{\beta})$ *is that*

$$\int \exp\left\{-\frac{M_1\lambda}{2c_0}[\theta^{-1}(r)]^2+\tau\theta^{-1}(r)+\phi^{-1}w(yr-b(r))\right\}\left[\frac{d^2b(r)}{dr^2}\right]^{1/2}dr \tag{3.5}$$

is finite for some τ in an open neighborhood about zero, for j = 1, …, p. For j = n + 1, …, p, the condition is the same as for the existence of the prior MGF.

Corollary 3.1: *When p < n (and λ = 0), a sufficient condition for the existence of the posterior MGF is that*

$$\int \exp\left\{\tau\theta^{-1}(r)+\phi^{-1}w(yr-b(r))\right\}\left[\frac{d^2b(r)}{dr^2}\right]^{1/2}dr \tag{3.6}$$

is finite for some τ in an open neighborhood about zero, and that the MLE exists.

Here we additionally require that the MLE exists (i.e., the likelihood function is bounded above). However, existence of the prior MGF is not necessary.

The next result demonstrates the usage of the necessary and sufficient conditions in some specific examples of GLMs to determine existence of prior and posterior MGFs. Without loss of generality, we assume that $\boldsymbol{\mu}_0 = 0$.

Theorem 4

i. *The sufficient conditions (3.3) and (3.5) guarantee the existence of prior and posterior MGFs for p > n in the Binomial model with canonical and probit link, and in the Poisson model with canonical and identity link.*

ii. According to condition (3.4), prior and posterior MGFs do not exist for the Gamma model with canonical link.

iii. For the Inverse Gaussian model with canonical link, (3.3) is not satisfied, while (3.4) is satisfied, thus it cannot be determined by these conditions alone whether the prior and posterior MGFs exist.

Note that, however, the prior and posterior MGFs do exist for the Gamma model with a log link, due to a result shown in Section 3.2. All derivations are given in the Appendices.

3.2 Connection between IM and g-priors

When the information matrix $\Omega(\boldsymbol{\beta}) = \Delta(\boldsymbol{\beta})V(\boldsymbol{\beta})\Delta(\boldsymbol{\beta})$ is independent of $\boldsymbol{\beta}$, the IMR prior reduces to a “ridge” g-prior for a Gaussian linear model, and is of Gaussian form, proper, and its MGF exists. This provides a quick way to determine existence of the prior and posterior MGF for models for which the Sufficiency conditions do not hold, but the necessary ones do. Examples are the gamma model with log-link and the Gaussian model with canonical link. For the gamma model with log link, note that $b(\theta) = -\log(-\theta)$, so that $b''(\theta) = \upsilon_i(\boldsymbol{\beta}) = 1/\theta^2$; and $\theta = -e^{-\eta}$, thus $\frac{d\theta}{d\eta}=\delta_i(\beta)=\theta$. The diagonal elements of $\Omega(\boldsymbol{\beta})$ are $\upsilon_i(\beta)\delta_i^2(\beta)=1$. Hence $\Omega(\boldsymbol{\beta})$ reduces to an identity matrix, and $\pi_{IMR}(\boldsymbol{\beta})$ is a Gaussian distribution with mean $\boldsymbol{\mu}_0$ and covariance matrix $c_0(X'X + \lambda I)^{-1}$, which is always proper, and for which the MGF exists even if $p > n$.

3.3 Characterization of “frequentist” bias and variance

To determine the influence of λ and c_0 in the IMR prior on the posterior estimates, we first explore the case of Gaussian linear models, where closed forms exist. For simplicity, assume $\boldsymbol{\mu}_0 = 0$.

Theorem 5—*The posterior covariance matrix of* $\boldsymbol{\beta}$ *for the Gaussian linear model,*

$$\sigma^2\Sigma=\sigma^2\left[\left(1+\frac{1}{c_0}\right)X'X+\frac{\lambda}{c_0}I_p\right]^{-1}, \tag{3.7}$$

satisfies

$$\left(\frac{c_0}{\lambda}\right)^{p/2}\frac{\left(1+\frac{c_0+1}{\lambda}x_0^{(p)}\right)^{-n}}{n^{n/2}} \le |\Sigma| \le \left(\frac{c_0}{\lambda}\right)^{p/2},$$

Where $x_0^{(p)}=max\{diag(X'X)\}$, *if* $c_0 > \lambda$.

__Corollary 5.1:__ *If* $p \to \infty$, $c_0 > \lambda$, *and n is finite, then* $|\Sigma| \to \infty$.

Theorem 6—*The posterior bias of* $\boldsymbol{\beta}$, *for the IMR prior, satisfies*

$$\frac{1}{p}\sum_{i=1}^{p}\text{bias}(\beta_p) \le \frac{1}{p}\sum_{i=1}^{p}|\beta_i|.$$

Equivalently, for the determinant of the bias matrix, $D = \Sigma X'X-I$, $\| D \beta \|^2= \boldsymbol{\beta}'D'D\boldsymbol{\beta} \le \boldsymbol{\beta}'\boldsymbol{\beta}$.

It is also of interest to compare the bias and MSE of estimates arising from the use of the IMR prior to those obtained using a Gaussian prior, $N(0, c_0I_p)$. For simplicity, assume $\sigma^2 = 1$, and $\beta_i = 1$, for $i = 1, \ldots, p$. With D_{IM} and D_N denoting the bias matrices of the IMR and Gaussian priors, respectively, it can be shown that the ratio of their determinants is

$$\frac{|D_{IM}|}{|D_N|}=\frac{|\{(1+c_0)X'X+\lambda I\}^{-1}X'Xc_0-I|}{|\{(X'X+c_0I)^{-1}X'X-I\}^{-1}X'Xc_0-I|}\frac{|X'X+\lambda I||X'X+c_0I|}{|(1+c_0)X'X+\lambda I||c_0I|} \tag{3.8}$$

and, similarly, the ratio of the determinants of the mean square error (MSE) matrices is

$$\frac{|MSE_{IM}|}{|MSE_N|}=\frac{|(\Sigma X'X-I)'(\Sigma X'X-I)+\Sigma|}{|[(X'X+c_0I)^{-1}X'X-I]'[(X'X+c_0I)^{-1}X'X-I]+(X'X+c_0I)^{-1}|}, \tag{3.9}$$

where Σ is as given in (3.7). Figure 3 depicts the behavior of these ratios for a set of choices of (n, p). The bias ratio (averaged over 5 data sets) decreases with an increase in c_0, and the IMR prior is uniformly better when c_0 is sufficiently large (> 3) irrespective of whether n is larger or smaller than p. When $n \ge p$, the MSE using the IMR prior is uniformly better when c_0 is moderately large. However, when $p > n$, an increase in c_0 leads to a an increase in the MSE with the IM prior, and a sharper Gaussian prior (with a large bias) is favored in terms of the MSE.

3.4 Elicitation of λ, μ_0 and c_0

The parameter λ plays an important role in the construction of the IMR prior since its introduction makes the prior covariance matrix of **β** nonsingular regardless of *p*. One question is whether to take λ fixed or random in the model. Empirical experience suggests that if λ is random, any prior for λ would have to be quite sharp since there is no information in the data for estimating it. Empirical studies show that taking λ fixed gives similar results with far less computational effort. When taking $0 < \lambda \le 1$, posterior inference about **β** appears quite robust for a wide range of values of λ in (0, 1) (Section 4.2). Note that when c_0 is large, λ and the design matrix *X* have a small impact on the posterior analysis of **β**. Since our main aim is to develop a relatively noninformative IM prior for Bayesian inference, other practical hyperparameter choices include $\boldsymbol{\mu}_0 = \mathbf{0}$ (consistent with Zellner's g-prior) and a moderately large c_0 ($c_0 \ge 1$).

4 Empirical Studies

4.1 Density of the IM prior compared to Jeffreys' and Gaussian priors

We simulated data under a logistic model with sample size $n = 20$, varying the number of covariates in the range $1 \le p \le 500$, to get an idea of the behavior of the IMR prior for $p < n$ and $p > n$. For the prior $\pi_{I\,M\,R}(\boldsymbol{\beta})$, setting $\boldsymbol{\mu}_0 = 0$, the posterior density of **β** is

$$\frac{}{\prod_{i=1}^{n}(1+e^{\boldsymbol{x}_i'\beta})}, \qquad (4.1)$$

where $\Sigma = \Sigma(\boldsymbol{\beta}) = c_0[X'\Omega(\boldsymbol{\beta})X + \lambda I_p]^{-1}$, and $\omega_i(\beta)=\frac{\exp(\boldsymbol{x}_i'\beta)}{[1+\exp(\boldsymbol{x}_i'\beta)]^2}$, (for $i = 1, \ldots, n$). When $p < n$ and $\lambda = 0$, (4.1) reduces to the IM prior. The IM prior was compared to Jeffreys' prior and a Gaussian prior $N(0, c_0(X'X)^{-1}/4)$ (i.e. equivalent to setting $\boldsymbol{\beta} = 0$ in the IM prior). For the $p = 1$ case, we can plot the exact prior densities (normalized computationally to be on the same scale) over a range of values. Figure 1 shows the three priors superimposed on the same plot at three scales. The IMR prior lies between Jeffreys' prior and the Gaussian prior at the center of the range, has heavier tails than do the others for a a wider range than Jeffreys' and, at the tails, converges to a Gaussian prior. We repeated the study with the same *n*, but set $p = 5$ and used the IMR prior. In this case, we cannot analytically derive the marginals, but instead prior and posterior samples are drawn from the respective distributions using Adaptive Rejection Metropolis Sampling (ARMS) (Gilks, Best, and Tan (1995)), using the HI package in the statistical software R (R Development Core Team (2004)). Smoothed kernel density estimates based on a Gaussian kernel are plotted for one of the marginals (Figure 4). This indicates the relative flatness of the IMR prior compared to the posterior density, showing that it is less informative than taking a Gaussian prior. Posterior densities are centered closely around 0.9, the true value of β.

4.2 Robustness of posterior estimates

Effect of c_0 on posterior estimates—We investigated the performance of estimates using the IM prior with different settings of c_0 and for values of *p* from 10 to 500, with a sample size $n = 150$. Data were generated from a logistic model with a design matrix $x = (x_{ij})$, $x_{ij} \sim N(0, 0.1)$, and coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots \beta_p)$, where $\beta_j = 0.9$ ($j = 1, \ldots, p$). When $p > n$, the ARMS algorithm did not work well for generating posterior samples of **β**, the autocorrelations in the samples being quite high. To generate posterior draws, we turned to importance sampling with a multivariate *t* trial density with mean 0, dispersion matrix V_t, and 3 df, denoted as $t_3(0, V_t)$, where $V_t = [(1 + 1/g)X'WX + \lambda/gI]^{-1}$, and *g* is a scalar quantity controlling the dispersion. The performance of the estimates worsened as p increased for a fixed n (Figure 5). With an increase in c_0, the change in the bias and MSE of estimators was small for $p \leq 200$. When $p = 500$, the variance of the estimator overshadowed the bias and hence a moderate c_0 led to a small MSE, with a negligible change in bias. We tested a range of g between 0.1 and 100, for which smaller values appeared to give more accurate results. Results using g between 0.1 and 5 were virtually indistinguishable. The results reported are for a setting of $g = 0.25$, and $\lambda = 0.5$.

Robustness to choice of λ—We tested the effect of λ on posterior estimates when $p > n$. Here we present results from a simulation with data from a logistic model with $p = 100$, $n = 50$, $c_0 = 1$, and λ chosen at approximately equal intervals between 0 to 1 ($\lambda > 0$). Figure 6 shows that very small values of λ, close to zero, led to unstable estimates; however, interestingly, the estimates were remarkably consistent, exhibiting little or no difference over a large range of λ values, between 0.4 and 1. Since the results were robust to this wide range of λ for both the $p < n$ and $p > n$ cases, we chose the value $\lambda = .5$ for all analyses.

Effect of the relationship between *n* and *p* on posterior estimates—We compared the performance of estimates based on the IMR prior to Gaussian priors and a "g-prior" for the logistic model, as the sample size decreased relative to p. We generated sets of data for a fixed $p = 100$, while n varied between $p/2 = 50$ and $2p = 200$. Five data sets were generated for each combination of (n, p), while $\boldsymbol{\beta}$ was generated from a $N(\boldsymbol{\beta}_0, c_0[\boldsymbol{x}'W(\boldsymbol{\beta}_0)\mathbf{x} + \lambda I])$ distribution, with $\boldsymbol{\beta}_0 = (2 \times \boldsymbol{J}_{50}, -2 \times \boldsymbol{J}_{50})'$ where $\boldsymbol{J}_q$ denotes a q-dimensional vector of ones.

Sampling from the posterior distribution was done with a multivariate t trial density. In addition to the bias and MSE, we also computed the theoretical "asymptotic covariance" of the estimates from the inverse of the Hessian matrix of the log-posterior, as $V_{as}=\left[-\frac{\delta^2}{\delta\beta^2}\log\, p(\beta|\boldsymbol{y},\boldsymbol{x})\right]^{-1}$. V_{as} cannot be evaluated in closed form, so we used a numerical computation routine in R to get an approximate estimate. The results from the IM prior (Figure 7) were compared with estimates using (i) $N(0, \gamma I)$ prior distributions for a highly informative prior with $\gamma = 1$ and an almost non-informative prior with $\gamma = 10^6$, the default used in the software BUGS (Gilks, Thomas, and Spiegelhalter (1994)); and (ii) $N(0, \gamma(X'X)^{-1})$ priors for $\gamma = 1, 10^6$ (equivalent to the "g-prior" for a normal linear model). We denote estimates found using the five methods, IMR prior, Gaussian priors $N(0, I)$, $N(0, 10^6 I)$, g-priors $N(0, (X'X)^{-1})$, $N(0, 10^6(X'X)^{-1})$ as IM, NO, NO6, GP, and GP6. Figure 7 shows that (i) when $n > p$, the IM prior had the lowest or comparable bias to NO6; (ii) when $p > n$, the IM prior had almost comparable bias to the NO prior; (iii) the MSE of estimates based on the IM prior was almost uniformly lowest. The GP6 and NO6 performed badly in terms of MSE, especially when $p > n$, whereas for larger n, the NO and GP priors appear too informative and hence led to more biased estimates. Overall, the IMR prior appears to be an attractive choice for computing estimates of the regression coefficients for both cases, whether n is larger or smaller than p.

Comparison to Bayesian model averaging predictions—In high-dimensional regression applications, an alternative method to estimating the full model is Bayesian model averaging (BMA) of the posterior estimates (Hoeting, Madigan, Raftery, and Volinsky (1999)). We compared predictions using the IMR priors in a full model to those obtained using BMA with a g-prior in a scenario where $p > n$ in logistic regression. Since the g-prior is undefined when $p > n$, the model averaging procedure was restricted to involve only $p < n$ models. IMR was compared with (i) BMA using BIC, in a stepwise variable selection algorithm (Yeung, Bumgarner, and Raftery (2005)) implemented as the iBMA routine in the R package “BMA”, and (ii) BMA under Zellner’s g-prior with a marginal likelihood evaluated using the generalized Laplace approximation (Bollen, Ray, and Zavisca (2005)) with model selection through the evolutionary Monte Carlo (EMC) algorithm (Liang and Wong (2000)). The approximation to the marginal likelihood is given by

$p(\boldsymbol{y}) \approx \exp[\sum_{i=1}^{n}\{y_i\boldsymbol{x}_i'\widehat{\beta} - \log(1+e^{\boldsymbol{x}_i'\widehat{\beta}})\}]|c_0(X'X)^{-1}|^{\frac{1}{2}}\phi(\widehat{\beta};\mathbf{0},[X'\Omega(\widehat{\beta})X]^{-1}+c_0(X'X)^{-1})$, where $\boldsymbol{\phi}(\boldsymbol{y};\boldsymbol{\mu},\Sigma)$ denotes the multivariate Gaussian density $N(\boldsymbol{\mu},\Sigma)$ at $\boldsymbol{y}$, $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$, and $\Omega(\cdot)$ is as defined in (2.5).

We first adapted a procedure proposed in Hoeting *et al.* (1999) to compare the predictive performance of the three methods. The data was randomly split into halves, and each model selection method was applied on the first half of the data (“training set”, T). The performance was then measured on the second half (“test set”, t) of the data through an approximation to the *predictive logarithmic score* (Good (1952)) which is given by the sum of the logarithms of the observed ordinates of the predictive density under the model M for each observation in the test set $-\sum_{d\varepsilon t}\log P(d|D^T, M)$, where $d = (y, \boldsymbol{x})$ denotes a data point, and D^T denotes the training data set. For BMA, the predictive score is measured by $-\sum_{d\varepsilon t}\log\{\sum_{M\varepsilon\mathcal{M}}P(d|D_T, M)p(M|D_T)\}$; the smaller the score for a model or model average, the better is the predictive performance. The second criterion used for comparing the performance was obtained through generating the fitted probabilities for the test set using regression coefficients $\hat{\beta}^{(T)}$ estimated from the training set, given by $\widehat{\pi}_i^{(t)}=\exp(X_i^{(t)}\widehat{\beta}^{(T)})/[1+\exp(X_i^{(t)}\widehat{\beta}^{(T)})]$. We then compared the performance of the three methods through their receiver operating characteristic (ROC) curves, based on comparison to the true value of the ordinates. Figure 8 shows the ROC curves for the IMR prior, BMA using BIC, and BMA using the g-prior (gBMA), for a simulated data set under a logistic regression model with test and training data set sizes of 75, and 100 covariates. IMR and gBMA appeared to have a similar performance (with IMR performing better at the two ends of the curve), and both performed much better than the BMA-BIC method at almost all points. The corresponding predictive logarithmic scores were 53.67 (IMR), 77.79 (BMA-BIC), and 51.43 (gBMA). The scores from IMR and gBMA are highly comparable, though BMA requires a much higher computational cost in exploring the high dimensional model space (it sampled a total of 3592 distinct models in 50,000 iterations of EMC). Interestingly, restricting BMA with the g-prior to the top 100 models sampled, the score increased to 197.65, making it extremely inaccurate.

5 Analysis of nucleosomal positioning data

5.1 Data description and background

The misregulation of the chromatin structure in DNA is associated with the progression of cancer, aging, and developmental defects (Johnson (2000)). It is known that the accessibility of genetic information in DNA is dependent on the positioning of histone proteins packaging the chromatin, forming *nucleosomes* (Kornberg and Lorch (1999)), which in turn is dependent upon the underlying DNA sequence. Nucleosomes typically comprise regions of about 147 bp of DNA separated by stretches of “open” DNA (nucleosome-free regions, or NFRs). Nucleosome positioning is known to be influenced by di- and tri-nucleotide repeats (Thastrom, Bingham, and Widom (2004)), but overall, the sequence signals infiuencing positioning are relatively weak and difficult to detect.

To determine how sequence features affect nucleosome positioning, we obtained data from a genome-wide study of chromatin structure in yeast (Hogan, Lee, and Lieb (2006)). The data consist of normalized log-ratios of intensities measured for a tiled array for chromosome III, consisting of 50-mer oligonucleotide probes that overlap every 20 bp. We first fitted a two-state Gaussian hidden Markov model, or HMM (Juang and Rabiner (1991)) to determine the nucleosomal state for each probe. We then refrained from any further use of the probe-level microarray data, as it was our primary interest to determine whether certain sequence features are predictive of nucleosome and nucleosome-free positions in the genome, which would enable us to make predictions for genomic regions for which experimental data is currently unavailable or difficult to obtain.

5.2 Analysis with the sample size *n* > dimension *p*

In order to determine how sequence features might influence the positions of nucleosome-free regions, we concentrated on a region of about 1400 adjacent probes on yeast chromosome III. For each probe, the covariate vector was the set of observed frequencies of nucleotide k-tuples, with $k = 1, 2, 3, 4$. This led to a total of $p = 340$ covariates (without including an intercept in the model). The HMM-based classification gave the observed "state" of each probe, whether corresponding to a nucleosome-free region (NFR) or a nucleosome (N). The results reported here are with $c_0 = 1$, results with $c_0 = 10$ were essentially similar.

We carried out ten-fold cross validation to test the predictive power of the model. The set of probes, with the associated covariates, were divided into ten non-overlapping pairs of training sets (90% of probes) and test sets (10% of probes). Each training set thus had a sample size of $n = 1260$, which is greater than the number of covariates (340). For each training-test set pair, the logistic regression model was first fitted to the training data set (with the three priors: IMR, $N(0, I)$ and $N(0, 10^6 I)$), and the fitted values of $\boldsymbol{\beta}$ used to compute the posterior probabilities of classification into the NFR state, for the corresponding test set. The sensitivity and specificity of cross validation using the three different priors was compared, where any region having an estimated posterior probability of 50% or more with the logistic model was classified as an NFR. The IMR prior showed uniformly higher sensitivity while its specificity was comparable to the other two priors (Table 1). The IMR predicted a slightly higher percentage of NFRs than the true percentage, while the other two methods consistently underestimated the number of NFRs. Overall, using the IMR was about 16% more accurate in predicting NFRs compared to the next best method. We also compared the results with Bayesian model averaging, using a g-prior and using the set of 340 covariates- in which case it performed equally well as the full model with an IMR prior. The computational cost of using BMA was much higher than IMR-5,000 iterations under this setting took about 104 hours on a 1.261 GHz Intel Pentium III compute node running Red Hat Enterprise Linux 4. In comparison, generating 5000 independent samples from the posterior distribution of $\boldsymbol{\beta}$ using the IMR prior required less than an hour.

Out of 340 covariates using the IMR prior, 93 and 91 coefficients had approximate 95% HPD intervals above and below zero. Among the significant dinucleotides, "aa", "at", "tg", and "tt" had a positive effect on the possibility of being an NFR, while "ac", "ag", "cc", "ct", "gc", and "gg" had the opposite effect. It was previously found that "aa" or "tt" repeats have an effect of making DNA rigid, and thus difficult to form nucleosomes, while "gg" and "cc" lead to less rigid DNA for which it is easier to form nucleosomes (Thastrom *et al.* (2004)). Thus these results seem reasonable compared to biological knowledge. On the other hand, we see that dinucleotides alone do not seem to have the strongest power to distinguish NFRs from nucleosomal regions, suggesting that the relationship between sequence factors and nucleosome positioning could be more complex than linear.

5.3 Analysis with *p* > *n*

Next, we increased the number of predictors to test how far the improved model fit, by including the 5-tuple counts, would be offset by the increased covariate dimensionality. We repeated the same process for creating the 10 training-test data set pairs as in the earlier case, except that we included k-tuples for $k = 1, \ldots, 5$, leading to $p = 1364$, with the training set size as 1260. We next carried out the ten-fold cross-validation procedure over each training-test set pair. As seen in Table 1, the IMR prior in this case is the only method that could predict even a proportion of the NFRs correctly. However, the overall predictive power using 5-mers decreased, due to a combination of overfitting with sparse data as well as induced bias due to the massive increase in dimensionality.

The above empirical studies are mainly to illustrate that the use of the IMR prior is beneficial in situations where the number of observations does not significantly exceed, or is even less than the number of observed covariates, and the lack of prior knowledge of the situation prevents selection of a fewer number of covariates in advance of the analysis. We observed some dissimilarities between the sets of covariates found significant between the two situations; however, we suspect the main reason for this is that the covariates are highly collinear (the average correlation between them ranges from −0.7 to 0.9) and the values are highly sparse, especially the counts for k-mers with a large k. The results, however, indicate that the positioning of nucleosomes may indeed depend on a number of underlying sequence factors, and not just a few dinucleotides, as was previously thought. The logistic regression model may be a simplification of the actual relationship between the covariates and response, but is a first step towards modeling a more complex, possibly non-linear relationship with the covariates. Using the logistic model to connect sequence features with nucleosomal state, rather than modeling the probe-level data as an intermediate step, is a direct attempt to determine how sequence features influence positioning. For instance, a model with high predictive power of correct nucleosomal state can provide useful surrogate information for other applications when experimental data, which are expensive to generate, are not available.

6 Discussion

The proposed IM and IMR priors can be viewed as a broad generalization of the “g-prior” (Zellner (1986)) for Gaussian linear models, reducing to Jeffreys’ prior as a limiting case. Although the g-prior was originally conceived (and is still most frequently used) in the context of model selection, the proposed priors provide a desirable alternative to Gaussian or improper priors with high-dimensional data in generalized linear models. The IM and IMR priors appear to produce results similar to a diffuse Gaussian prior, but are computationally more stable with collinear variables. They provide a desirable alternative to Jeffreys’ prior, being proper for most GLMs, but giving more flexibility than Jeffreys’ prior in the choice of tuning parameters, and being less subjective than the choice of an arbitrary Gaussian prior. Theoretical and computational properties of the IM and IMR priors were investigated, demonstrating their effectiveness in a variety of situations. The IM and IMR priors for many GLMs are proper and their moment generating functions (MGFs) exist.

Numerical studies demonstrated that the IMR prior, even with the full model, compared favorably with a more complex Bayesian model averaging procedure with a *g*-prior that involves dimension reduction. With extremely high dimensional data, the BMA procedure becomes computationally infeasible in our examples. The BMA procedure could also be used with an IMR prior– it would be interesting to explore the possibility of improving variable selection methods in GLMs, as in Hans, Dobra, and West (2007) and Liang *et al.* (2008), through the use of an IMR prior instead of conventional priors. This would require the ability to compute accurate approximations for the marginal likelihoods, which is a complex problem outside the Gaussian family of priors. Future work is also needed in developing alternative methods for eliciting λ, such as choosing the λ that maximizes the marginal likelihood. Although our current focus is on GLMs, the IMR prior framework can be easily extended to a variety of other models, for instance, to survival and longitudinal models, and to others used in a multitude of scientific applications.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank two anonymous referees whose valuable insights and suggestions led to major improvements in the overall clarity and presentation of this paper, and to thank Jason Lieb and Greg Hogan for making available the yeast nucleosomal array data. This research was supported in part by the NIH grants GM070335 and HG004946.

Appendix

The appendices are available in the online version of the article at http://www.stat.sinica.tw/statistica.

References

Bedrick EJ, Christensen R, Johnson W. A new perspective on priors for generalized linear models. J. Am. Stat. Assoc 1996;91:1450–1460.

Berger J, Pericchi L. The intrinsic Bayes factor for model selection and prediction. J. Amer. Stat. Assoc 1996;91:109–122.

Bollen, K.; Ray, S.; Zavisca, J. A scaled unit information prior approximation to the Bayes factor; Technical report, SAMSI LVSSS Transition Workshop: Latent Variable Models in the Social Sciences; 2005.

Chen MH, Ibrahim JG, Yiannoutsos C. Prior elicitation, variable selection and Bayesian computation for logistic regression models. J. Roy. Stat. Soc. B 1999;61:223–242.

Gilks WR, Thomas A, Spiegelhalter DJ. A language and program for complex Bayesian modelling. The Statistician 1994;43:169–178.

Gilks WR, Best NG, Tan KKC. Adaptive rejection Metropolis sampling. Applied Statistics 1995;44:455–472.

Good IJ. Rational decisions. J. Roy. Statist. Soc. B 1952;14:107–114.

Hans C, Dobra A, West M. Shotgun stochastic search for “large p” regression. J. Amer. Statist. Assoc 2007;102:507–516.

Hoerl AE, Kennard RW. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 1970;12:55–67.

Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. Statist. Sci 1999;14:382–417.

Hogan GJ, Lee C-K, Lieb JD. Cell cycle-specified fluctuation of nucleosome occupancy at gene promoters. PLoS Genet 2006;2:e158. [PubMed: 17002501]

Ibrahim J, Chen M-H. Conjugate priors for generalized linear models. Statistica Sinica 2003;13:461–476.

Ibrahim JG, Laud PW. On Bayesian analysis of generalized linear models using Jeffreys’ prior. J. Amer. Statist. Assoc 1991;86:981–986.

Johnson CA. Chromatin modification and disease. J Med Genet 2000;37:905–915. [PubMed: 11106353]

Juang B-H, Rabiner LR. Hidden Markov models for speech recognition. Technometrics 1991;33:251–272.

Kass RE. The geometry of asymptotic inference. Statistical Science 1989;4:188–234.

Kass RE. Data-translated likelihood and Jeffreys' rules. Biometrika 1990;77:107–114.

Kornberg RD, Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell 1999;98:285–294. [PubMed: 10458604]

Liang F, Wong WH. Evolutionary Monte Carlo: applications to c_p model sampling and change point problem. Statistica Sinica 2000;10:317–342.

Liang F, Paulo R, Molina G, Clyde MA, Berger JO. Mixtures of g-priors for Bayesian variable selection. J. Am. Stat. Assoc 2008;103:410–423.

Mitchell TJ, Beauchamp JJ. Bayesian variable selection in linear regression. J. Am. Stat. Assoc 1988;83:1023–1032.

R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2004. ISBN 3-900051-07-0.

Schafer, JL. Analysis of Incomplete Multivariate Data. London: Chapman and Hall; 1997.

Spiegelhalter DJ, Smith AFM. Bayes factors for linear and log-linear models with vague prior information. J. Roy. Stat. Soc. B 1982;44:377–387.

Thastrom A, Bingham LM, Widom J. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. J Mol Biol 2004;338:695–709. [PubMed: 15099738]

Wang, X. Ph.D. thesis. Austin: The University of Texas; 2002. Bayesian Variable Selection for GLM.

Wang X, George EI. Adaptive Bayesian criteria in variable selection for generalized linear models. Statistica Sinica 2007;17:667–690.

West M. Bayesian factor regression models in the "large p, small n" paradigm. Bayesian Statistics 2003;7:723–732.

Yeung K, Bumgarner R, Raftery A. Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. Bioinformatics 2005;21:2394–2402. [PubMed: 15713736]

Zellner, A. On assessing prior distributions and Bayesian regression analysis with g-prior distributions, volume Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Goel, PK.; Zellner, A., editors. North-Holland, Amsterdam: 1986. p. 233

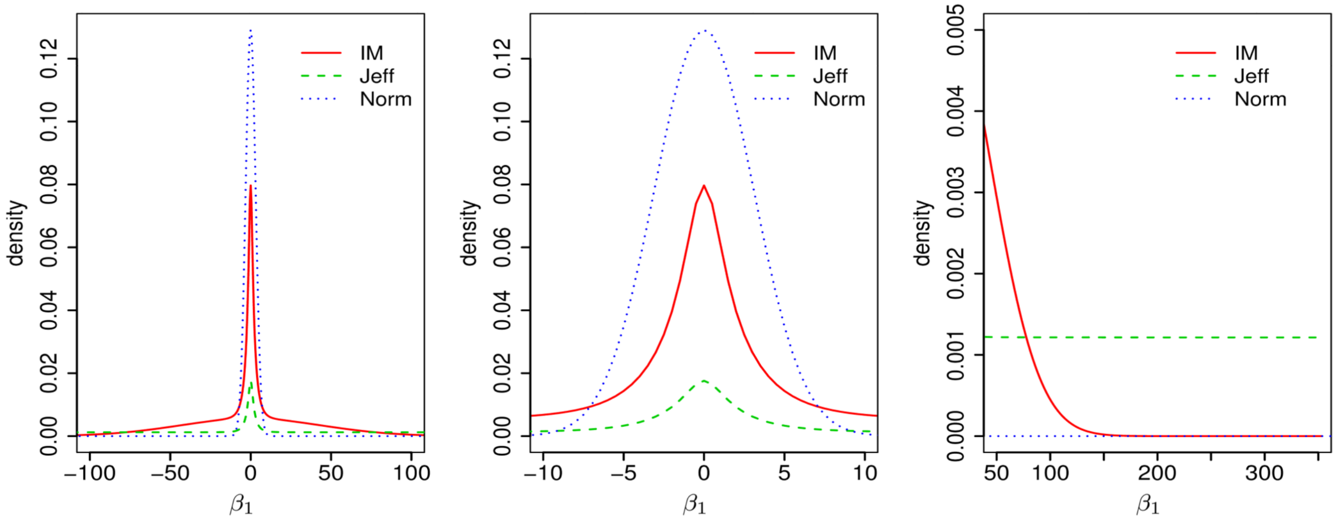

Figure 1.

Density of IM prior under a logistic model with p = 1 compared with Jeffreys’ (“Jeff”) and a Gaussian prior (“Norm”) for a simulated data set with n = 20. The plot is shown at three different scales. The total density under each curve is the same (normalized); however the Gaussian prior has the highest mass near the center while Jeffreys’ prior has much heavier tails (visible in the third panel). The IM prior has lower mass near the center of the distribution compared to the Gaussian prior but thicker tails; while it has thinner tails and more central mass compared to Jeffreys’ prior.

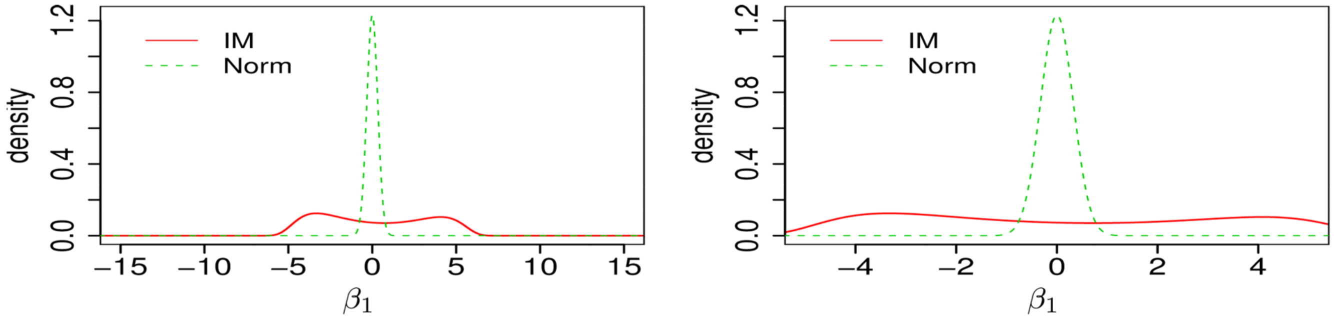

Figure 2.

Density of IM prior under a Poisson model with p = 1 compared with a Gaussian prior for a simulated data set with n = 20, shown at two different scales.

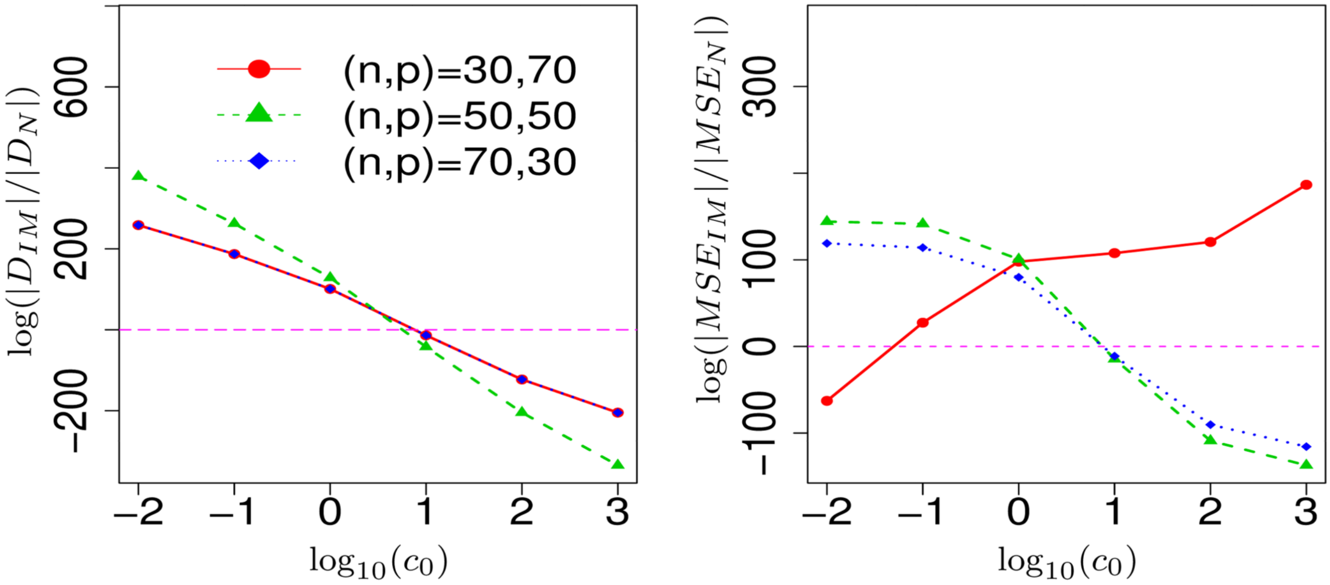

Figure 3.

Ratio of determinants of bias (panel 1) and MSE (panel 2) for the IMR prior compared to a Gaussian N (0, c_0I_p) prior for the Normal linear model.

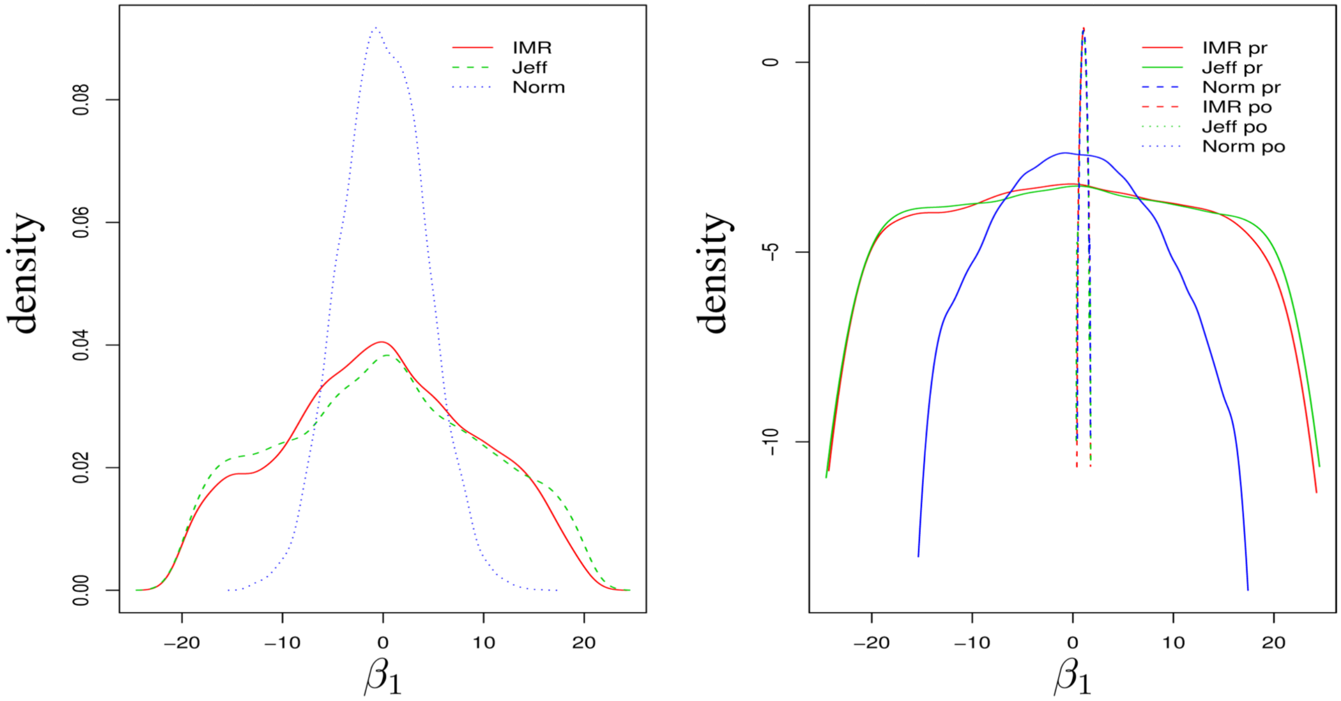

Figure 4.

Panel 1: Density of IMR prior with p = 5 compared with Jeffreys' prior ("Jeff") and a Gaussian prior ("Norm") for a simulated data set with n = 20. Panel 2: Prior ("pr") and posterior ("po") log-densities based on the three priors.

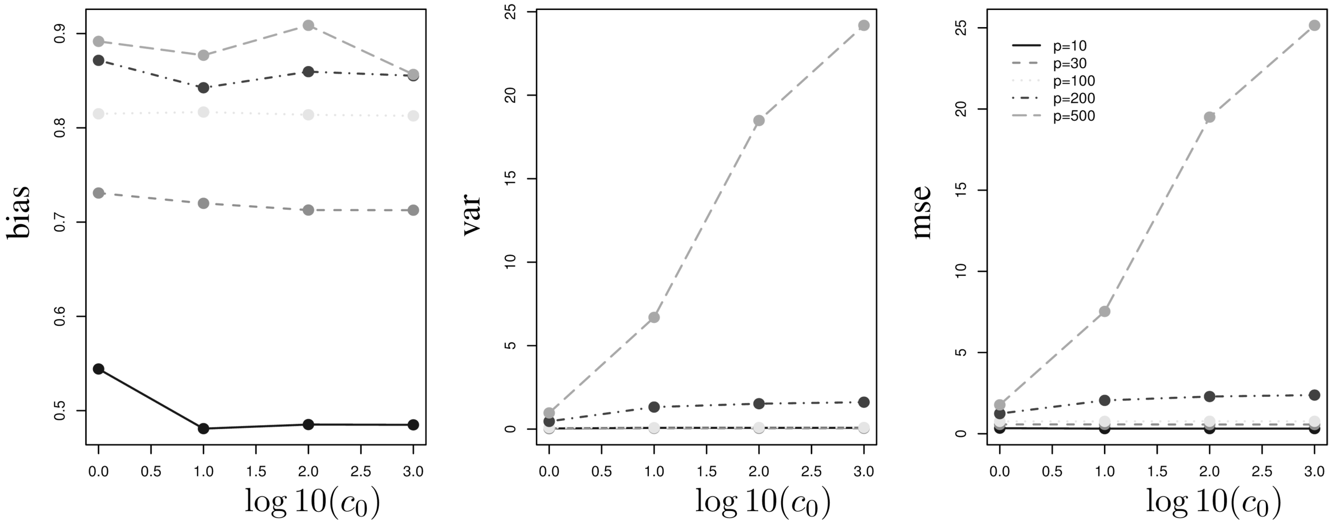

Figure 5.

Performance of IMR prior for different settings of c_0.

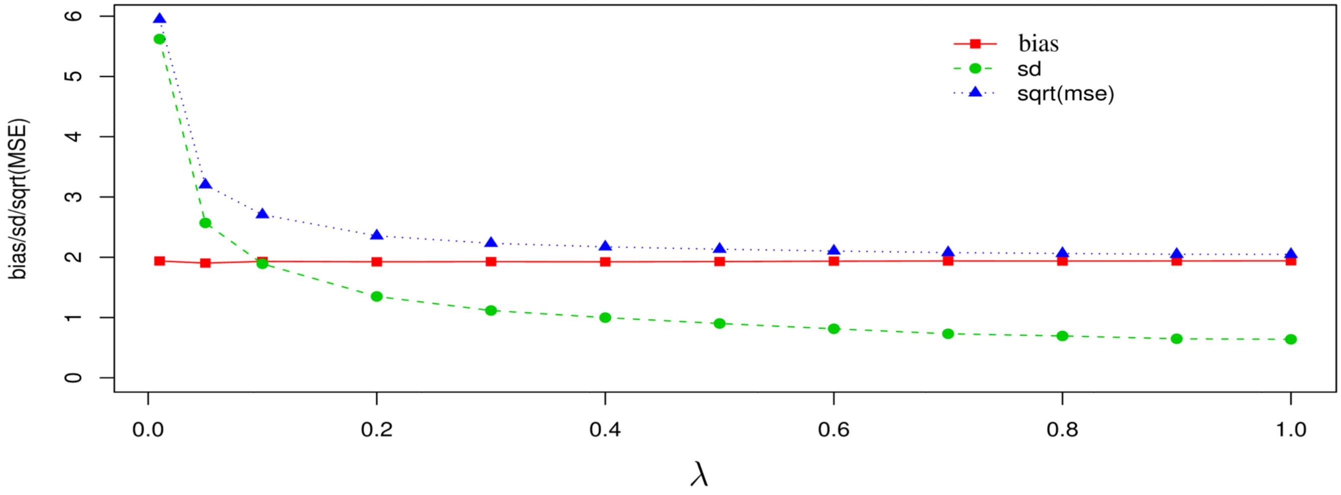

Figure 6.

Performance of estimates using the IMR prior for different settings of λ.

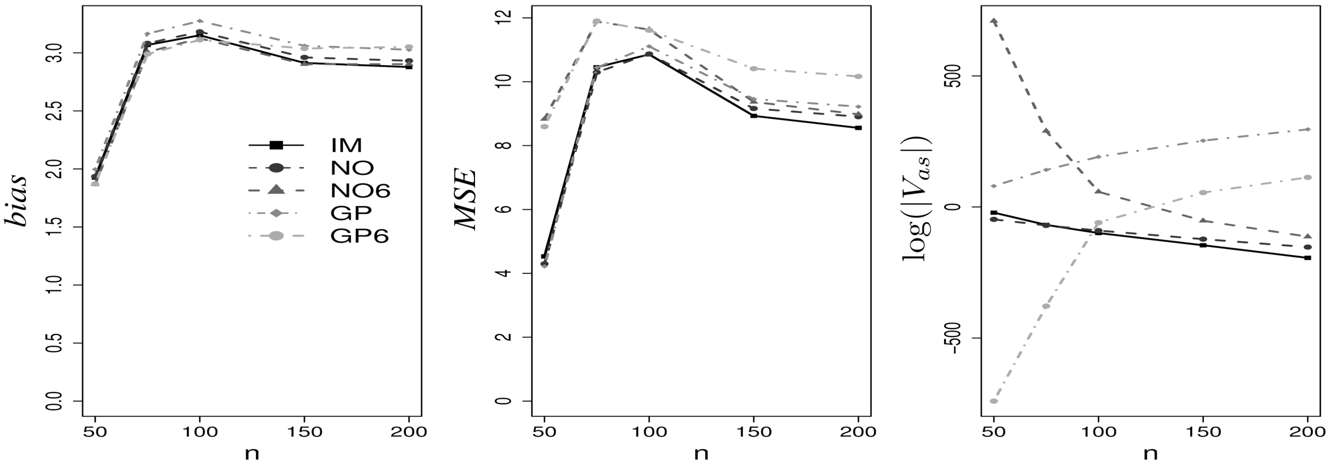

Figure 7.

Comparison of estimates based on the IMR prior with a Gaussian prior N (0, γI), and g-prior N (0, $\gamma(X'X)^{-1}$), with γ = 1 and 10^6 for a data set with p = 100. When n ≥ p, estimates based on the IM prior had the lowest bias and MSE; while when n < p, they performed equally well as the N (0, I) prior.

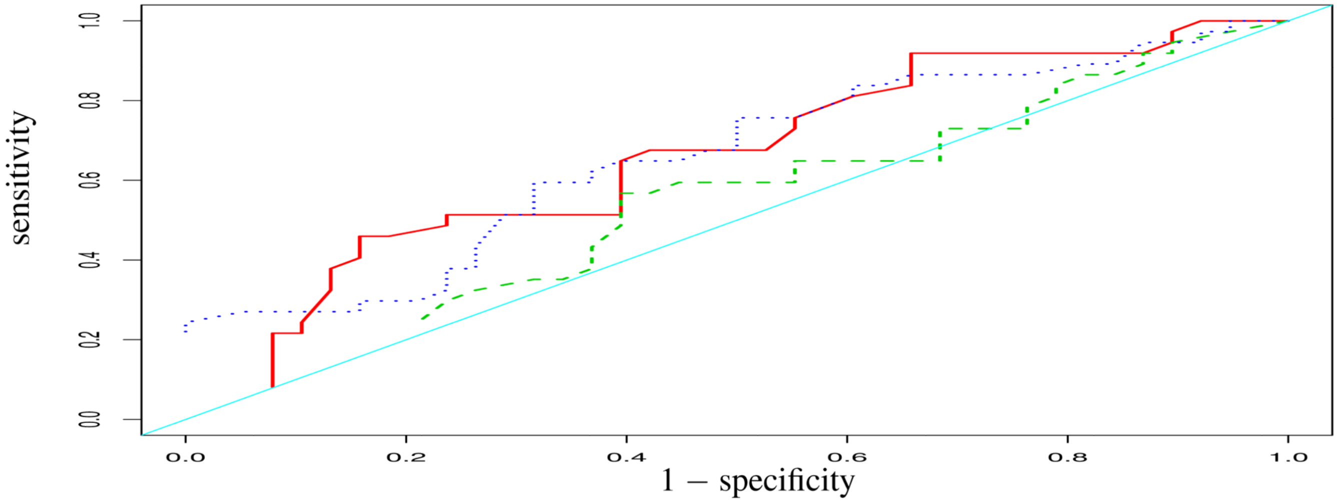

Figure 8.

ROC curves comparing the predictive classification performance between IMR (red solid line), BMA using BIC (green dashed line), and BMA using a g-prior (blue dotted line) for a simulated data set.

Table 1

Overall sensitivity and specificity of methods using three types of priors,under the full model, and BMA using the g-prior (gBMA), by ten-fold cross validation, on set (a): p = 340, n = 1260 and set (b): p = 1364, n = 1260. %pN F R: percentage of predicted nucleosome-free regions by each method; ave.corr: average percent correct classification = (Sens + Spec)/2, averaged over the ten cross-validation data sets. The average percentage of NFRs in the data sets was 43%. Note: It was not possible to use the gBMA procedure for set (b) due to the massive computational cost.

Set	Prior	Range[$E(\beta\|y)$]	%$pNFR$	ave.corr
(a)	*IMR*	(−0.295, 0.443)	0.708	0.664
	$N(0, I)$	(−0.192, 0.161)	0.123	0.509
	$N(0, 10^6 I)$	(−0.188, 0.168)	0.111	0.475
	gBMA	(−8.778, 11.437)	0.569	0.670
(b)	*IMR*	(−0.518, 0.535)	0.387	0.447
	$N(0, I)$	(−1.142, 0.496)	0	–
	$N(0, 10^6 I)$	(−3.424, 2.179)	0	–