doi:10.20944/preprints202102.0551.v1

Variational Bayesian Learning and Semiparametric Models on the Double Exponential Family

Abstract

In this paper, we focus on variational Bayesian learning deterministic optimization methods for inference in biparametric exponential models where the parameters follow semiparametric regression structures. This combination of data models and algorithms contributes to solving real-world problems and reduces the computation time. This allows both the rapid exploration of many data models and the accurate estimation of the mean and variance functions through the connection between generalized linear models and graph theory. A simulation study was carried out to assess the performance of the deterministic approximation. Finally, herein, we present an application using macroeconomic data to emphasize the benefits of the proposed approach.

keywords: Variational learning Bayes; semiparametric heterocedastic models; calculus of variations; optimization; biparametric exponential models

1 Introduction

Variational Bayesian learning is an optimization-based framework that provides computationally tractable posterior distributions over hidden or latent variables to approximate the true joint distribution. This complexity of statistical and probabilistic models has continuously increased due to the demand for new applications as a result of the advances in information technology. This expansion includes methods such as hidden Markov models, probabilistic graphical models, and mixed and latent variable models. Furthermore, empirical applications are linked to various analyses in the statistics and machine learning cultures, for example, estimating uncertainty, handling non-*IID* data, and hyper-parameter tuning, among others. Summaries of variational approximation research may be found in [2], [11], and [12]. A recent introduction from the statistical perspective was provided in [13]. The implementation of these models requires the solution of high-dimensional integrals in likelihood functions and posterior probability densities, which emerge in each setting. Variational Bayesian learning (*VBL*), which relies on the calculus of variation optimization techniques, approximates the parameters' values by minimizing various objective functions. This optimization determines a sufficient approximation to the posterior distribution in order to avoid overfitting problems, with the benefit of faster and easier scaling to large datasets through the use of stochastic and distributed optimization techniques.

The general *VBL* setup can be divided into three main stages. First, a family of approximating densities $\mathcal{Q}$ is proposed for the latent variables set. Then, the member of that family, which minimizes a functional, shaped with a finite number of unknown parameters, is found. This *functional* is related to the posterior distribution and is known as the *free energy* to the exact posterior. Finally, the posterior is approximated with the optimized member of that family, $r*$. In other words, the goal of *VBL* is to minimize the *upper bound*, which is known as the *Evidence lower bound (ELBO)*of the surprise or *free energy*. This optimization strategy makes use of the *conditional conjugacy* assumption that allows practical models to be built by combining basic distributions.

Even though complex Bayesian modeling has benefited from Markov Chain Monte Carlo (*MCMC*) and its variants[1], there are cases where this framework could face difficulties as a result of large datasets and the complex geometry of the posterior distributions. Thus, these approximations need massive computing resources, which can lead to slow convergence and incorrect conditional and predictive posterior distributions. Therefore, *VBL* can be seen as a complementary alternative to Markov chain Monte Carlo methods.

Recent studies including generalized linear models and semiparametric regression have con-

[1] The *MCMC* simulation-based framework has provided outstanding sampling tools and specifically has facilitated the approximation in semiparametric heteroscedastic Bayesian models.

sidered the presence of *nonconjugate models* arising when one or more conditional posteriors are unknown. Models of this type were studied by [6] who relied on the mixed model formulation of penalized spline regression. The variational methodology for heteroscedastic semiparametric regression models in the *GLM* context was initially developed in [13]. On the other hand, [10] achieved simultaneously nonparametric function estimation using a Gaussian process approach. Next, [11] developed a *mean field* variational approximation where both the mean and variance are smooth functions of the predictors in a nonparametric sense and the nonconjugate structure was addressed using a fixed-form Gaussian update. Furthermore, *semiparametric mean-field variational Bayes* were introduced in [7] as a combination of the *kullback–Leibler* divergence and the mean-field restriction that include parametric and nonparametric densities. For the heterocedastic regression model, [3] implemented a mean-field variational approximation with an embedded *Laplace* approximation to account for the nonconjugate structure for semiparametric regression in the presence of errors with nonconstant variance. From the applied point of view, variational approximations have been implemented, for example, in latent variable models, low-rank clustering, and time series *nowcasting* involving mixture models.

The main novelty of this paper is the connection of variational Bayesian learning theory, developed in [12], with the *biparametric exponential family* to model regression structures for mean and variance parameters. This strategy was developed under the framework of generalized linear models (*GLIM*) with penalized splines. In this context, we designed an algorithm to simultaneously estimate the mean and variance functions for heteroscedastic regression.

This paper contains five sections apart from the introduction and proceeds as follows. In Section 2, we describe the main issues behind the variational Bayesian learning methodology. In Section 3, the biparametric exponential family applied to the heteroscedastic semiparametric model and its approximation methodology are explained. In Section 4, we provide a simulation study and an application to real data. Finally, in Section 5, we present our conclusions and propose future lines of research. The technical results are presented in the Appendix.

2 Variational Bayesian Learning

In this section, we describe the components of variational Bayesian learning and the optimization framework required for *free-energy* minimization. Thus, variational Bayesian learning (*VBL*) can be seen as an optimization setup that approximates the posterior density by solving a constrained maximization problem through the application of the calculus of variations. This procedure relies on the ideas of both *free energy* and *conditional conjugacy*.

2.1 Free energy

The *free energy*, $F(r)$, is a functional that depends on the path of an arbitrary distribution r. This concept is related to the upper bound of the *Bayes free energy*, which is the negative logarithm of the marginal likelihood, $-\log p(\mathcal{D})$, where $\mathcal{D}$ represents the observed data. Thus, the free-energy minimization is summarized in the following three steps:

Firstly, consider the *Kullback–Leibler* (KL) divergence from an arbitrary trial distribution $r(\omega)$ on the parameter ω to the posterior $p(\omega|\mathcal{D})$ given by this measure:

$$\mathcal{KL}\left(r\left(\omega\right)||P\left(\omega|\mathcal{D}\right)\right)=\int r\left(w\right)\log\frac{r(\omega)}{p\left(\omega|\mathcal{D}\right)}d\omega=\left\langle \log\frac{r(\omega)}{p\left(\omega|\mathcal{D}\right)}\right\rangle_{r(\omega)}$$

On the basis of the work in [12] and [2], which depends on applying the KL properties, the minimizer of this integral corresponds to the Bayes posterior given by

$$P\left(\omega|\mathcal{D}\right)=\min_{r}\mathcal{KL}\left(r\left(w\right)||p\left(\omega|\mathcal{D}\right)\right)$$

Secondly, according to [12], an equivalent problem for this optimization can be issued as

$$P\left(\omega|\mathcal{D}\right)=\min_{r}F\left(r\right),$$

where the functional of r, $F(r)$ is known as the *free-energy functional* [4]. This functional is obtained by replacing the posterior distribution $p(w/\mathcal{D})$ with the proportional joint distribution $p(\omega,\mathcal{D})$ from the $\mathcal{KL}$ divergence, according to Bayes theorem:

$$F\left(r\right)=\int r\left(\omega\right)\log\frac{r(\omega)}{p\left(\omega,\mathcal{D}\right)}d\omega=\left\langle \log\frac{r(\omega)}{p\left(\omega,\mathcal{D}\right)}\right\rangle_{r(\omega)} \tag{1}$$

$$=\mathcal{KL}\left(r\left(\omega\right)||p\left(\omega|\mathcal{D}\right)\right)-\log p\left(\mathcal{D}\right),$$

where the normalizing factor $p(\mathcal{D})$ does not depend on ω. The functional $F\left(r\right)$ is known as an upper bound of the traditional *Bayes free energy* $-log(p(\mathcal{D}))$ and $-F\left(r\right)$ is called the *evidence lower-bound* $(ELBO)$. Finally, to guarantee that the evaluation of the functional is tractable for optimal $r\left(w\right)$, the following optimization problem is solved by restricting the search space to $\mathcal{G}$:

$$\min_{r}F\left(r\right)\qquad \textit{subject to}\qquad r\epsilon(\mathcal{G}),$$

where G is a chosen tractable distribution. However, thanks to *conditional conjugacy*, a weaker constraint restricts the optimal distribution to being in a tractable class.

2.2 Conditional conjugacy

In various empirical applications, the model likelihood has no conjugate prior distribution, which has turned Bayesian analysis to numerical approximation algorithms, such as *MCMC* and its variants. However, the notion of *conditional conjugacy* appears to be crucial to assess the inference problem in terms of deterministic optimization algorithms.

The objective of conditional conjugacy is to divide the set of unknown parameters into two parts: $\omega = (\omega_1, \omega_2)$. For instance, if the posterior distribution of ω_1 is in the same distribution family as the prior $p(w_1)$, where ω_2 is a given constant, then the conditional posterior is

$$p(\omega_1|\omega_2, \mathcal{D}) \propto p(\mathcal{D}|\omega_1, \omega_2)\, p(\omega_1)\,,$$

where the prior $p(\omega_1)$ is called the *conditional conjugate prior* of the likelihood $p(\mathcal{D}|\omega)$ with respect to the the parameter ω_1, given the fixed parameter ω_2.

2.3 Constraint design

The design of a tractable *VBL* algorithm is straightforward once the conditional conjugacy for every unknown parameters is found. The procedure considers the set of the parameteres $\omega = (\omega_1, ..., \omega_s)$, such that for each $s = 1, .., S$, the model likelihood $p(D|\omega) = p(D|\omega_{s,s',s'\neq s})$ has a conditionally conjugate prior $p(\omega_s)$ with respect to ω_s, given $\omega_{s'\neq s}$ as fixed constant. Then, if the prior distribution is set as

$$p(\omega) = \prod_{s=1}^{S} p(\omega_s),$$

the posterior distribution

$$p(\omega|\mathcal{D}) \propto p(\mathcal{D}|\omega)\, p(\omega)$$

is, as a function of ω_s, in the same family distribution as the prior, $p(\omega_s)$. Furthermore, the moments and characteristics of the posterior distribution are tractable if the other parameters, $\{\omega_{s'}\}_{s'\neq s}$, are given.

Consequently, in order to take into account this *conditional conjugacy* property, the independence constraint between the parameter groups should be imposed on the approximate posterior.

doi:10.20944/preprints202102.0551.v1

$$r\left(\omega\right)=\prod_{s=1}^{S}r_s\left(\omega_s\right).$$

The constraint allows us to compute the moments with respect to w_s and to optimize each factor $\{r_s\}_{s=1}^{s}$ separately. This factorized form of variational inference corresponds to a framework known as *mean field theory*. See [?] for details.

As a result, the *VB* posterior is defined as

$$\widehat{r}=\min_{r}F\left(r\right)\qquad s.t\qquad r\left(\omega\right)=\prod_{s=1}^{S}r_s\left(\omega_s\right)$$

Optimizing each factor requires the use of the theory of *calculus of variations*. From this, the free energy, $F(r)$, is expressed as an explicit function with a finite number of unknown parameters. In other words, this framework provides stationary conditions of the function with respect to the posterior.

2.4 Calculus of variations

According to [14], the introduction of *variational principles* leads to a more clear and compact modeling strategy. For instance, quantities such as entropy or energy that depends on probability distributions as inputs can be found as a result of the optimization setup. Moreover, *VBL* provides a structure to solve the problem using variational methods focused on minimizing integral functionals.

Hence, *calculus of variations* is a method related to finding maxima and minima functions that occurs when the derivatives vanish and the gradient of f equals zero at the stationary points (x_0, y_0). Moreover, it allows the extremes of *functionals* to be found, which are defined as functions that depend on the path of one or more functions and their domain is a set of admissible functions. Therefore, determining the stationary points of functionals is the fundamental problem of the calculus of variations, from where different solutions can be implemented. For instance, reducing the variational problem to differential calculus to obtain the *Euler* equations is popular among economists in order to predict future patterns of key variables.

2.5 Variational Bayesian Learning in action

In this section, we summarize the algorithm and its components to solve the *VBL* optimization problem. First, to approximate the posterior, according to the mean field theory, the *free-energy* functional is minimized, conditional on the independence constraint between groups of parameters. This setup is given by

$$\widehat{r} = \min_{r} F\left(r\right) \qquad s.t \qquad r\left(\omega\right) = \prod_{s=1}^{S} r_s\left(\omega_s\right)$$

Then, the conditionally conjugate prior, which can be decomposed, is assumed:

$$p\left(\omega\right) = \prod_{s=1}^{S} p\left(\omega_s\right).$$

Furthermore, on the basis of calculus of variations, the stationary conditions are computed. The *free energy* is stated as

$$F\left(r\right) = \int \prod_{s=1}^{S} r_s\left(\omega_s\right) \left[\log \frac{\prod_{s=1}^{S} r_s\left(\omega_s\right)}{p\left(\mathcal{D}|\omega\right) \prod_{s=1}^{S} p\left(\omega_s\right)}\right] \mathrm{d}\omega,$$

where the derivatives of the free energy with respect to $r_s(\omega_s)$ equal zero, i.e., $\frac{\partial F}{\partial r_s} = 0$. From this first-order condition, the following stationary conditions as a function of ω_s are obtained:

$$F\left(r\right) = \int \prod_{s=1}^{S} r_s\left(\omega_s\right) \left[\log \frac{\prod_{s=1}^{S} r_s\left(\omega_s\right)}{p\left(\mathcal{D}|\omega\right) \prod_{s=1}^{S} p\left(\omega_s\right)}\right] \mathrm{d}\omega$$

The VBL posterior, $r_s\left(W_s\right)$, is in a parametric form depending on the *variational parameters*. The stationary conditions found before are used to update these variational parameters through an iterative algorithm that ends when the free energy is minimized. In other words, this computational design provides a local minimizer, $\widehat{r}$, for the free-energy functional, which is considered as the *VBL- posterior*. From this posterior distribution, the moments and marginal, conditional, and predictive functions can be computed. For example, the mean *VBL* estimator can be defined as $\widehat{W} = \langle W \rangle_{\widehat{r}(W)}$.

Even though there are cases where a model does not have a conditionally conjugate prior, Bayesian variational learning can be adapted to consider such intractable functions. Various methods have been proposed according to the analyzed model.[2] In this paper, we follow the work in [3], where the nonconjugate posterior was approximated by the Laplace method. This method is briefly described below.

[2] The most frequently used are Black box variational inference, which is suitable as data size increases, and Local variation approximation and expect propagation, which are mostly used in the Statistics community.

doi:10.20944/preprints202102.0551.v1

2.5.1 Laplace approximation

In the *laplace approximation*, the posterior is approximated by a Gaussian:

$$r_\omega = Gauss_D(\omega, \widehat{\omega}, \widehat{\Sigma})$$

VBL locates the variational parameters $\lambda = (\omega, \Sigma)$ by minimizing the free energy $F(r)$. Then, the *MAP* mean and covariance estimators are located.

$$\widehat{\omega}^{LA} = \widehat{\omega}^{MAP} = \min_{w}\ p(\mathcal{D}/\omega)p(\omega)$$

$$\widehat{\omega}^{LA} = \widehat{F}^{-1}$$

2.5.2 Empirical Variational Bayesian Learning

When the model involves hyperparameters in the likelihood and/or the priors, the joint distribution is specified as

$$p\left(\mathcal{D}, \omega | hyper\right) = p\left(\omega | hyper\right) p\left(\mathcal{D} | \omega, hyper\right)$$

Then, the free energy is written as

$$F\left(r, hyper\right) = \int r\left(\omega\right) \log \frac{r\left(\omega\right)}{p\left(D, \omega | hyper\right)} d\omega = \left\langle \log \frac{r\left(\omega\right)}{p\left(\mathcal{D}, \omega | hyper\right)} \right\rangle_{r(\omega)}$$

$$= \mathcal{KL}\left(r(\omega) || p(\omega | \mathcal{D}, hyper\right) - \log p\left(\mathcal{D} | hyper\right)$$

Under this framework, the hyperparameters can be estimated from observation by simultaneously minimizing the following:

$$\left(\widehat{r}, \widehat{hyper}\right) = \min_{r, hyper} F\left(r, hyper\right)$$

3 Biparametric exponential model

Frequently, in empirical applications, the implicit mean-variance relation, in which the variance is specified as a function of the mean, is not validated by data. Thus, extending this condition to a larger collection of models has been proposed as a result of the advantages of the exponential family framework [5]. In this section, we outline the *VBL* framework and derive the variational posterior distribution for the semiparametric heterocedastic model under the *GLM* architecture. Following [15], the Doubly semiparametric stochastic generalized linear models with splines as random effects can be stated as

$$y_i \sim DE(\mu_i, \tau_i), \tag{2}$$

where $DE(\mu_i, \tau_i)$ denotes the double exponential distribution with mean μ_i and variance τ_i. Under the semiparametric framework, the mean and variance functions depend on both the predictors of the model and the basis functions from the smoothing technique.

$$\mu_i = \beta_0 + \beta_1 x + \sum u_k z_k$$

$$\log(\tau_i) = \{\gamma_o + \gamma_1 x + \sum v_k z_k\}$$

As a result, the entire model could be stated as

$$y/u, v \sim DE\,(X\beta + Z_u, Diag\{exp(X\gamma + Z_v\})$$

with the random effect being doubled

$$\begin{bmatrix} u \\ v \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2 I & 0 \\ 0 & \sigma_v^2 I \end{bmatrix}\right)$$

Then, by combining the fixed and random coefficients for the mean and variance functions into single vectors, these vectors can be expressed as

$$\theta = \begin{bmatrix} \beta \\ u \end{bmatrix} \quad \theta_v = \begin{bmatrix} \gamma \\ v \end{bmatrix}$$

On the other hand, concatenating the fixed and random effects for the mean and variance functions, the design matrices can be written as $C_u = [X, u]$ $C_v = [Z, v]$.

The equations can be simplified as

$$X\beta + Z_u = C_u\theta \quad and \quad X\gamma + Z_v = C_v\theta_v$$

Taking advantage of the connection between GLM models and graphical models, the correspondent directed acyclic graph (DAG) that identifies the causality among parameters for semiparametric heterocedastic model is presented below:

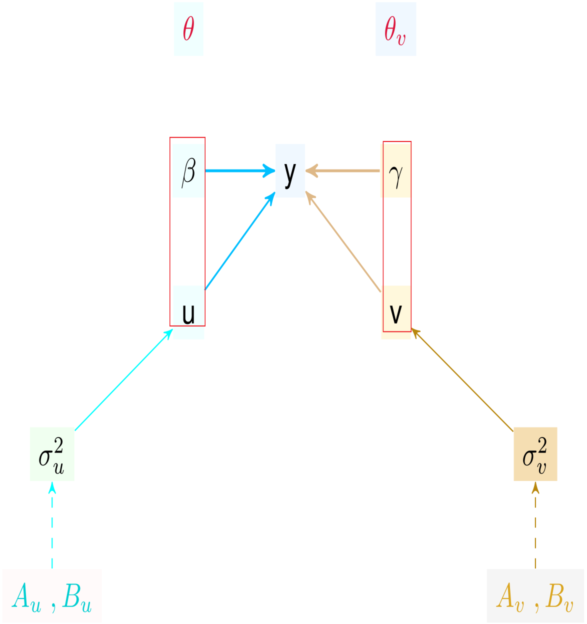


Figure 1: Directed acyclic graph for the model. The node y corresponds to the observed data vector. The set of parameters $\theta = (\beta, U)$ and $\theta_v = (\gamma, V)$.

3.1 Prior specification

The prior beliefs for the parameters of the model are described in this section. First, The fixed effects parameters β_i and γ_i are a priori independent. That is,

$$\beta \sim N(0, \sigma_\beta^2 I_p) \quad \gamma \sim N(0, \sigma_\gamma^2 I_r) \tag{3}$$

where σ_β^2 and σ_γ^2 are the *hyperparameters* with very large values. In the applications, we set 10^6 as was the case in [3] and [6].

On the other hand, the priors for the dispersion parameters are specified as

$$\sigma_u^2 \sim \mathcal{IG}\,(A_u, B_u) \quad \sigma_v^2 \sim \mathcal{IG}\,(A_v, B_v) \tag{4}$$

with *hyperparameters* A_u, A_v, B_u, and B_v, respectively. The covariance matrices for fixed and random effects are formed by

$$\Sigma_\theta = blockdiag(\sigma^2_\beta, \sigma^2_\gamma) \quad and \quad \Sigma_{\theta_v} = blockdiag(\sigma^2_u, \sigma^2_v)$$

3.2 Conditional posterior distributions

We now turn to set the conditional posterior distributions for each unknown parameter in the semiparametric Bayesian context defined by the *DAG* graphical model. Under the *VBL* framework, the majority of these distributions are in a closed form, but some of them may be intractable. These distributions could be approximated by causal learning ideas that link the directed acyclic graph (*DAG*) with the following concept: the *markov blanket* of a node is specified by the relevant variables for a node, which are the parents, children, children's parents, and spouses of that node. In other words, the *markov blanket* contains all variables that carry information about the node that cannot be obtained from any other variable, for instance, $p(\delta_i \mid \cdot)= p(\delta_i \,|\, markov\, blanket)$. This framework is explained in detail in [8].

Furthermore, it is worth noting that for the semiparametric model, the work of [6] provided the conditional posteriors with closed and tractable standard forms for the mean regression fixed parameters $p(\theta|\cdot)$, and the variance parameters for both the mean $p(\sigma^2_u|\cdot)$ and the variance equation $p(\sigma^2_v|\cdot)$. In contrast, the parameter distribution for the variance equation $p(\theta_v|\cdot)$ is *nonconjugate* and tractability is not easy to achieve. It can be accomplished by imposing the *Laplace* approximation as in [3]. In summary, the conditional posteriors are stated as follows:

Fixed effects in the mean equation:

$$\sigma^2_u \mid \cdot \sim \mathcal{IG}\,(A_u + \frac{K}{2}, B_u + \frac{||u||^2}{2}) \tag{5}$$

Variance of random effects in the mean equation:

$$\sigma^2_u \mid \cdot \sim \mathcal{IG}\,(A_u + \frac{K}{2}, B_u + \frac{||u||^2}{2}) \tag{6}$$

Variance of random effects in the variance equation:

$$\sigma^2_v \mid \cdot \sim \mathcal{IG}\,(A_v + \frac{K_v}{2}, B_u + \frac{||v||^2}{2}) \tag{7}$$

Fixed effects in the variance equation:

$$\sigma^2_v \mid \cdot \sim \mathcal{IG}\,(A_v + \frac{K_v}{2}, B_u + \frac{||v||^2}{2}) \tag{8}$$

3.3 Variational Bayesian learning of heterocedastic model

The *VBL* implementation is detailed in this section. The work of [3] described a variational approximation for the heterocedastic semiparametric regression *via* spline basis for both the mean and variance functions. The *mean field* factorizes the joint distribution in the following way:

$$r(\theta\,,\theta_v\,,\sigma_v^2\,,\sigma_u^2) = r_\theta(\theta)\,r_{\theta_v}(\theta_v)\,r_{\sigma_u^2}(\sigma_u^2)\,r_{\sigma_v^2}(\sigma_v^2)$$

According to the *VBL* framework, the solution can be derived form the following optimization problem:

$$F\left(r\right) = \int \prod_{s=1}^{S} r_s\left(\omega_s\right)\left[\log \frac{\prod_{s=1}^{S} r_s\left(\omega_s\right)}{p\left(\mathcal{D}|\omega\right)\prod_{s=1}^{S} p\left(\omega_s\right)}\right]d\omega$$

Under the *mean field* constraint, the free energy is written as

$$F\left(r\right) = \int \prod_{s=1}^{S} r_s\left(\omega_s\right)\left[\log \frac{\prod_{s=1}^{S} r_s\left(\omega_s\right)}{p\left(\mathcal{D}|\omega\right)\prod_{s=1}^{S} p\left(\omega_s\right)}\right]d\omega$$

where the parameters set $\omega_s = [\theta, \theta_v, \sigma_u^2, \sigma_v^2]$.

The following stationary conditions can be obtained from the method described above.

$\mathbf{r}_{\theta\,(\theta)}$: optimal distribution for the fixed effects in the mean equation

$$r_\theta\left(\theta\right) \propto exp\left\langle \log p\left(\theta|\omega_{s'},\mathcal{D}\right)\right\rangle_{\Pi_{s'\neq s} r_{s'}(w_{s'})}$$

$$\propto exp\left[\mathbb{E}_{-\theta}\left\{\log\, p(\theta/.)\right\}\right]$$

By substituting the conditional posterior into Eq

$$r_\theta\left(\theta\right) \propto exp\left[\mathbb{E}_{-\theta}\{-\frac{1}{2}(\theta - MC^T\Sigma^{-1}y)^T M^{-1}(\theta - MC^T\Sigma^{-1}y)\right]$$

$$r_\theta\,(\theta) \propto exp\left[\mathbb{E}_{-\theta}\,\{-\frac{1}{2}(\theta - MC^T\Sigma^{-1}y)^T M^{-1}(\theta - MC^T\Sigma^{-1}y)\right]$$

Thus, completing the square as follows:

$$r_\theta\,(\theta) \propto exp\left[-\frac{1}{2}(\theta - \Sigma_{r(\theta)}\,C^T\Gamma\,y)^T\,\Sigma^{-1}_{r(\theta)}(\theta - \Sigma^{-1}_{r(\theta)}C^T\Gamma\,y)\right]$$

This implies that the posterior is *Gaussian*. This, more specifically, can be written as

$$r_\theta(\theta) = \mathcal{N}_{\mathcal{M}}\,(\mu_{r(\theta)}\,, \Sigma_{r(\theta)})\ where\ \mu_{r(\theta)} = \Sigma_{r(\theta)}\,C^T\Gamma\,y$$

$\mathbf{r}_{\sigma^2_{\mathbf{u}}}(\sigma^2_{\mathbf{u}})$: Optimal distribution for the variance of random effects in the mean equation

$$r_{\sigma^2_u}(\sigma^2_u) \propto exp\,[\mathbb{E}_{-\sigma^2_u}\,\{\log p(\sigma^2_u|\cdot)]$$

$$\propto exp\,[\mathbb{E}_{-\sigma^2_u}\,\{-(A_u + \frac{K}{2})\,\log(\sigma^2_u) - \frac{1}{\sigma^2_u}(B_u + ||u||^2)]$$

$$r_{\sigma^2_u} \sim \mathcal{IG}\,(A_u + \frac{K_V}{2}, B_{r(\sigma^2_u)})$$

$\mathbf{r}_{\sigma^2_{\mathbf{v}}}(\sigma^2_{\mathbf{v}})$: Optimal distribution for the variance of random effects in the variance equation

$$r_{\sigma^2_v}(\sigma^2_v) \propto exp\,[\mathbb{E}_{-\sigma^2_v}\,\{\log p(\sigma^2_v|\cdot)\}]$$

$$\propto exp\,[\mathbb{E}_{-\sigma^2_v}\,\{-(A_v + \frac{K}{2})\,\log(\sigma^2_v) - \frac{1}{\sigma^2_v}(B_v + ||v||^2)]$$

$$r_{\sigma^2_v} \sim \mathcal{IG}\,(A_v + \frac{K_V}{2}, B_{r(\sigma^2_v)})$$

$\mathbf{r}_{\theta_{\mathbf{v}}}(\theta_{\mathbf{v}})$: Optimal distribution for the variance of fixed effects in the variance equation

$$r_{\theta_v}(\theta_v) \propto exp\left[\mathbb{E}_{-\theta_v}\{\log\ p(\theta_v|\cdot)\right]$$

Therefore, the *Free energy* as a function of the unknown variational parameters can be obtained by substituting the proposal densities into *free energy*.

$$F\,(r) = \int r_\theta\,(\theta) r_{\sigma_u^2}(\sigma_u^2) r_{\sigma_v^2}(\sigma_v^2) r_{\theta_v}(\theta_v) \log \frac{r_\theta(\theta) r_{\sigma_u^2}(\sigma_u^2) r_{\sigma_v^2}(\sigma_v^2) r_{\theta_v}(\theta_v)}{p\,(y|\theta\,,\sigma_u^2,\sigma_v^2,\theta_v)\,p(\theta)p(\sigma_u^2)p(\sigma_v^2)p(\theta_v)} d\theta\, d\sigma_u^2\, d\sigma_v^2\, d\theta_v$$

$$= \left\langle \log \frac{r_\theta(\theta) r_{\sigma_u^2}(\sigma_u^2) r_{\sigma_v^2}(\sigma_v^2) r_{\theta_v}(\theta_v)}{p\,(y|\theta\,,\sigma_u^2,\sigma_v^2,\theta_v)\,p(\theta)p(\sigma_u^2)p(\sigma_v^2)p(\theta_v)} \right\rangle_{r_\theta(\theta)\, r_{\sigma_u^2}(\sigma_u^2)\, r_{\sigma_v^2}(\sigma_v^2)\, r_{\theta_v}(\theta_v)}$$

Additionally, the empirical variational Bayesian procedure is performed by minimizing the free energy with respect to the hyperparameters. Thus, the derivative of the free energy with respect to the hyperparameters is $(\frac{\partial F(r)}{\partial A_u}\,, \frac{\partial F(r)}{\partial B_u}\,, \frac{\partial F(r)}{\partial A_v}\,, \frac{\partial F(r)}{\partial B_v})$. Technical details are presented in the Appendix.

Using the r densities described above, the variational learning Bayesian algorithm consists of updating the parameters associated with each $r - density$ until r has approximated the *posterior*. In each step of the algorithm, the most recent parameter value is used. This is called the coordinate ascent (CA) method.

Algorithm 1 Learning for heterocedastic semiparametric regression

```
1: Initialize the variational parameters $(\mu_{r(\theta)}\,, \Sigma_{r(\theta)}\,, B_{r(\sigma_u^2)}\,, B_{r(\sigma_v^2))}$ and the hyperparameters
   $(A_u\,, B_u\,, A_v\,, B_v)$ in the $DAG$.
2: Apply (substitute the right-hand side into the left-hand side to update
   $\mu_{r(\theta)}\,, \Sigma_{r(\theta)}\,, B_{r(\sigma_u^2)}\,, B_{r(\sigma_v^2)}\,, \mu_{r(\theta_V}\,, \Sigma_{r(\theta_V)}$ , respectively.
3: Update $A_u\,, B_u\,, A_v\,, B_v$, respectively.
4: Evaluate the free energy.
5: Iterate steps 2 through 4 until convergence (until the energy decrease becomes smaller than
   a threshold $\epsilon = 10^{-4}$)
6: Construct parameter estimates using means of variational approximations
```

4 Simulation

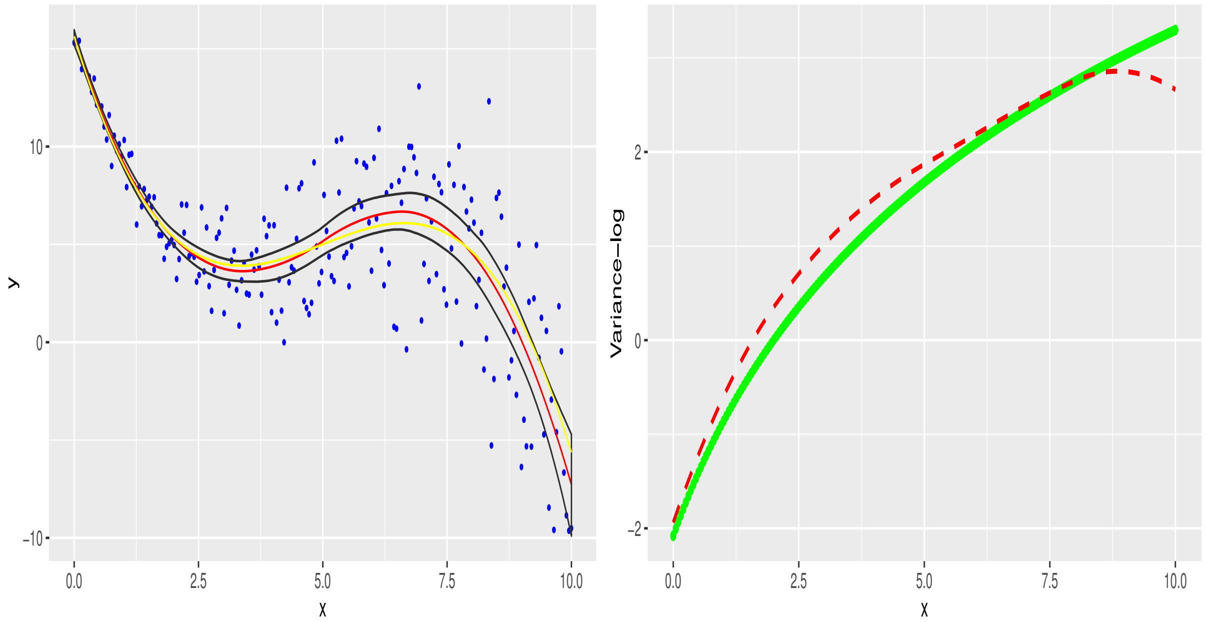
To illustrate the methodology, a particular case from a simulation experiment, adapted from [3], is presented for different sample size scenarios. The covariate vector $x = (x_1, x_2)$, where x_1

is an intercept and x_2 is a continuous variable. The coefficient values for the mean function were $(\beta_1, \beta_2) = (1, 0.12)$ and the variance function was $(\gamma_0, \gamma_1, \gamma_2)$=$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

$$\begin{cases} y_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \\ \mu_i = x - \frac{1}{8}(x-5)^3 \\ \sigma_i^2 = (\frac{1}{2} + \frac{1}{4}x)^2 \end{cases}$$

where $x \sim \mathcal{U}\,[1, 10]$ and $n = 50, 200, 500$ points.

The model was fitted using penalized splines for both the mean and variance levels and the mean squared errors $(RMSE)$. The hyperparameters are set The variance hyperparameters related with the random effects are.



(a) (b)

Figure 2: (a) Simulated observations are depicted by the yellow line. The red line represents the approximation with the learning Bayesian algorithm. (b) The green line represents the true log variance function. The dashed line is the variational learning approximation.

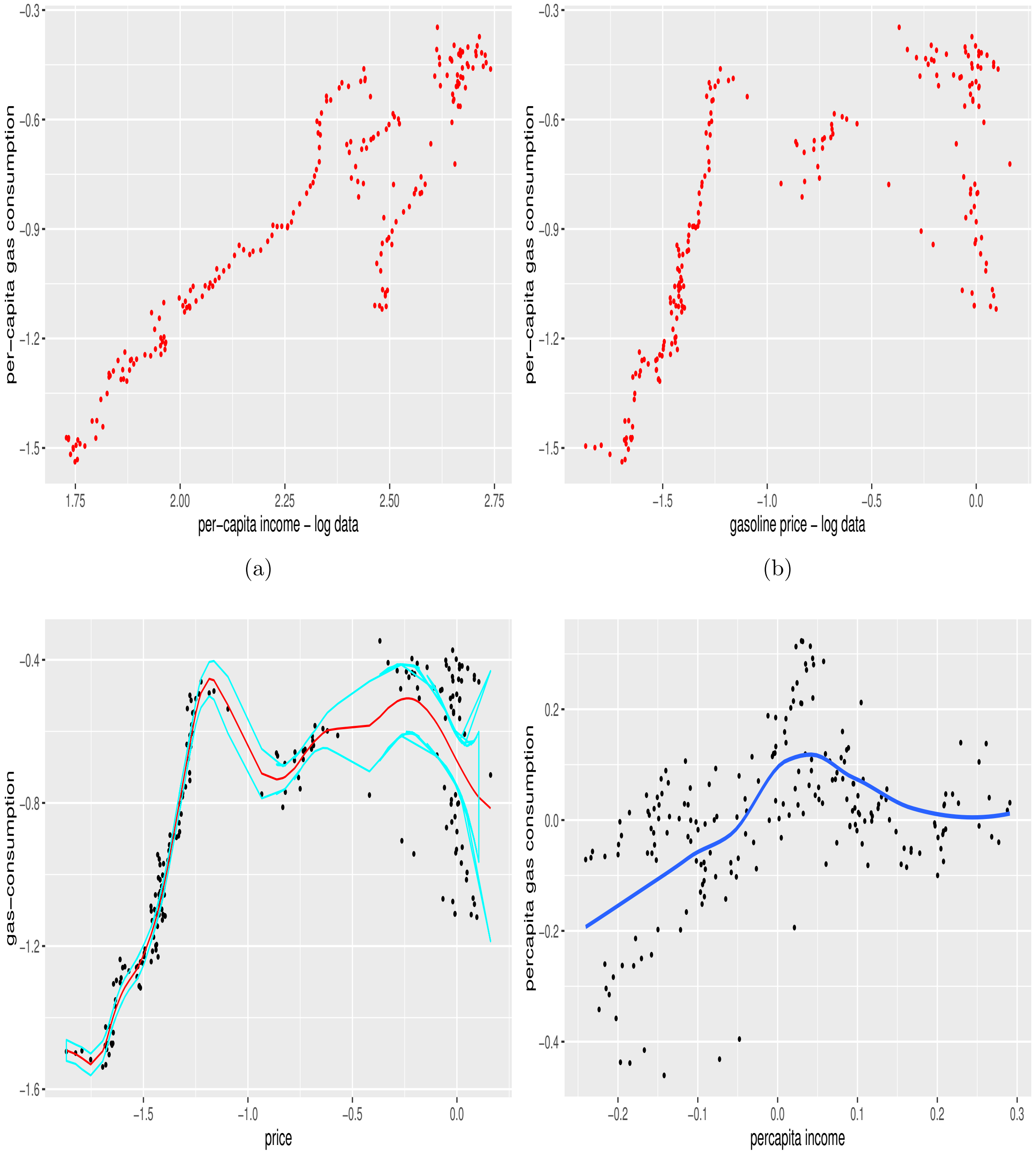
5 Real application

We illustrate the learning methodology with a quarterly dataset of gas consumption determinants from 1947.01 to 1997.04, which is used to explain US economic growth. In this type of macroeconomic study, the goal is to estimate the effects of key variables (price, income) on the dependent variable (gasoline demand). These types of *growth regressions* are described in [1]. The *scatterplots* of per capita *US* gasoline demand vs. price and per capita income can be seen in panels a) and b) of Figure 1. From these, we inferred that a linear relationship is not easy to justify. In panel c), we implemented the variational Bayesian algorithm to smooth the scatter plot between price and gasoline demand. Finally, the residual plot indicates the effect that per capita income has on gas demand, suggesting a nonlinear relationship.

The proposed learning Bayesian model was applied to fit this relation with the heterocedastic semiparametric regression methodology. More specifically, we assume

$$\begin{cases} Gas\ demand_t \sim \mathcal{N}(\mu_t, \sigma_t^2) \\ \mu_t = \beta_0 + \beta_1\, price_t + \beta_2\, income_t + \sum u_k z_k \\ \log(\sigma_t^2) = \gamma_o\ + \gamma_1\, price_t + \sum v_k z_k \end{cases}$$

The *partial residual plot*, which was introduced by [9], is a device that is commonly used by econometricians to apply multivariate regression. It assesses the importance of an independent variable in the presence of all other independent variables in predicting the dependent variable. Thus, it represents the multivariate regression results as bivariate scatterplots by previously *netting-out* the effect of the *other* variables. According to the theorem that was proved by *Gauss–Frisch–Waugh*, the same coefficient and standard deviation can be obtained for a given covariate by using either the partial residual regression or the multivariate regression. Additionally, this partial residual plot asseses the importance of nonlinearity. We adapted this tool to our model and found a curvilinear relationship, which is different from the literature. This suggests a much more precise and accurate assessment of the effects.



(a)

(b)

(c)

(d)

Figure 3: Comparison of Markov Chain Monte Carlo (*MCMC*) and *PROC MIXED* with *P-splines* and *RKHS* based on each of the four simulations settings.

6 Conclusions

In this article, we connect the variational Bayesian learning theory with the machine learning literature to tackle inference in heterocedastic semiparametric regression models. The setup was based on the generalized linear model (*GLM*) framework with penalized splines. In this context, we designed an algorithm for simultaneously estimating the mean and variance functions for heteroscedastic regression. Thus, on the basis of *Bayesian variational learning*, a deterministic approximation of the posterior density in heterocedastic semiparametric regression models is provided. Furthermore, closed formulae for estimating the mean and variance function in the *GLM* model architecture were derived. It is our belief that with the increasing availability of structured and unstructured data and advances in computing, the field of statistics needs to adapt to the new demands. From the Bayesian point of view, *MCMC* and its variants are excellent tools; however, faster approximations are required for various real-world applications. To this end, the Bayesian learning methodology provides an accurate solution.

References

[1] X Barro S, Sala i Martin. *Economic growth.* MIT, 2004.

[2] C. Bishop. *Pattern recognition and machine learning.* Springer, 2016.

[3] B. Bugbee, J. Bredit, and M. Van der Woerd. Laplace Variational Approximation for Semiparametric Regression in Presence of Heterocedastic Errors. *Journal of Computational and Graphical Statistics*, 25:225–245, 2016.

[4] Potgieter C. and Genton M. *Bayesian analysis of two-piece normal regression models.* 2003.

[5] E. Cepeda. *Variability modeling in Generalized Linear models.* PhD thesis, Unpublished Ph.D thesis, Matematics Institute Universidade Federal do Rio de Janeiro, 2001.

[6] C. Crainiceanu. Spatially Adaptative Bayesian Penalized Splines with Heteroscedastic Errors. *Journal of Computational and Graphical Statistics*, (2):265–288, 2007.

[7] C. Faes and M.P. Wand. Semiparametric mean field variational bayes: General principles and numerical issues. *Journal of Machine Learning Research*, (17):1–47, 2016.

[8] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* The MIT Press, 2010.

[9] W. Larsen and S. McCleary. The use of partial residual plots in regression analysis. *Technometrics*, (14):781–790, 1970.

[10] M Lazaro. Bayesian warped Gaussian processes. *Advances in Neural Information Processing Systems*, 26:225–245, 2013.

[11] M. Mencitas and M. Wand. Variational inference for heteroscedastic semiparametric regression. *School of mathematical sciences, University of Technology. Sydney, Australia*, 2014.

[12] Watanabe K Nakajima S and M Sugiyama. *Variational Bayesian Learning Theory*. Cambridge University press, 2019.

[13] M. Nott, D. Tran and A. Kuk. Efficient variational inference for generalized linear mixed models with large datasets. *arXiv preprint*, 2013.

[14] F. Rindler. *Calculus of Variations*. Springer-Verlag, 2016.

[15] H. Zárate and E. Cepeda. Semiparametric Smoothing Spline in Joint Mean and Dispersion Models with Responses from the Biparametric Exponential Family: A Bayesian Perspective. *Statistics, Optimization and Information Computing*, (forthcoming), 2020.

doi:10.20944/preprints202102.0551.v1

7 appendix

7.1 Free energy as a function of variational parameters

Substituting the equations into the free energy is explicit as a function of the variational parameters. $[\theta\,, \sigma_u^2\,, \sigma_v^2\,, \theta_v]$

$$F\,(r) = \int r_\theta\,(\theta) r_{\sigma_u^2}(\sigma_u^2) r_{\sigma_v^2}(\sigma_v^2) r_{\theta_v}(\theta_v) \log \frac{r_\theta(\theta) r_{\sigma_u^2}(\sigma_u^2) r_{\sigma_v^2}(\sigma_v^2) r_{\theta_v}(\theta_v)}{p\,(y|\theta\,, \sigma_u^2, \sigma_v^2, \theta_v)\; p(\theta) p(\sigma_u^2) p(\sigma_v^2) p(\theta_v)} d\theta\, d\sigma_u^2\, d\sigma_v^2\, d\theta_v$$

$$= \int r_\theta\,(\theta) r_{\sigma_u^2}(\sigma_u^2) r_{\sigma_v^2}(\sigma_v^2) r_{\theta_v}(\theta_v) \log[[r_\theta(\theta) r_{\sigma_u^2}(\sigma_u^2) r_{\sigma_v^2}(\sigma_v^2) r_{\theta_v}(\theta_v)] - \log[p\,\big(y|\theta\,, \sigma_u^2, \sigma_v^2, \theta_v)\; p(\theta) p(\sigma_u^2) p(\sigma_v^2) p(\theta_v)\big)]] d\theta\, d\sigma_u^2\, d\sigma_v^2\, d\theta_v$$

$$2F = 2\Big\langle \log \frac{r_\theta(\theta) r_{\sigma_u^2}(\sigma_u^2) r_{\sigma_v^2}(\sigma_v^2) r_{\theta_v}(\theta_v)}{p\,(y|\theta\,, \sigma_u^2, \sigma_v^2, \theta_v)\; p(\theta) p(\sigma_u^2) p(\sigma_v^2) p(\theta_v)} \Big\rangle_{r_\theta(\theta)\, r_{\sigma_u^2}(\sigma_u^2)\, r_{\sigma_v^2}(\sigma_v^2)\, r_{\theta_v}(\theta_v)}$$

$$= 2\Big\langle \log \frac{r_\theta(\theta) r_{\sigma_u^2}(\sigma_u^2) r_{\sigma_v^2}(\sigma_v^2) r_{\theta_v}(\theta_v)}{p(\theta) p(\sigma_u^2) p(\sigma_v^2) p(\theta_v)} \Big\rangle_{r_\theta(\theta)\, r_{\sigma_u^2}(\sigma_u^2)\, r_{\sigma_v^2}(\sigma_v^2)\, r_{\theta_v}(\theta_v)} -2\big\langle \log p\,\big(y|\theta\,, \sigma_u^2, \sigma_v^2, \theta_v\big)\big\rangle_{r_\theta(\theta)\, r_{\sigma_u^2}(\sigma_u^2)\, r_{\sigma_v^2}(\sigma_v^2)\, r_{\theta_v}(\theta_v)}$$

Replacing the equation's stationary conditions, the prior distributions and the likelihood and using the properties of log

$$\begin{aligned} &= p\log(\sigma_\beta^2) + r\log(\sigma_\gamma^2) - A_b \log(B_b) + \log(\Gamma(A_b)) - A_c\log(B_c) + \log(\Gamma(A_c)) - \log(\Sigma_{q(\theta)}) \\ &-\log(\Sigma_{q(\theta_V)}) + (A_b + \frac{k}{2})\log(B_q(\sigma_b^2)) + (A_c + \frac{k}{2})\log(B_q(\sigma_c^2)) - \log(\Gamma(A_c + \frac{K_V}{2})) - \log(\Gamma(A_b + \frac{K}{2})) \\ &- \sigma_\beta^{-2}(\mu_{q(\beta)})^2 - trace(\Sigma_{q(\beta)}) - \sigma_\gamma^{-2}(\mu_{q(\gamma)})^2 - trace(\Sigma_{q(\gamma)}) - (p+k) - (r+K_V) \\ &+ \sum((y_i - C_i^T\mu_{q(\theta)})^2 + C_i^T\Sigma_{q(\theta)}C_i)exp(-C_i^T\mu_{q(\theta)} + \frac{1}{2}C_i^T\Sigma_{q(\theta)}C_i) \end{aligned} \quad (9)$$

7.2 The empirical variational Bayesian

The algorithm is completed by minimizing the free energy with respect to the *hyperparameters*

$$\frac{\partial F(r)}{\partial A_u} = -\log(B_u) + \log(A_u) + \frac{1}{2A_u} + log(B_q(\sigma_u^2)) + \log(A_u + \frac{K}{2}) + \frac{1}{(2A_u + \frac{K}{2})} \tag{10}$$

$$\frac{\partial F(r)}{\partial B_u} = -\frac{A_b}{B_b} \tag{11}$$

$$\frac{\partial F(r)}{\partial A_v} = -\log(B_v) + \log(A_v) + \frac{1}{2A_v} + log(B_q(\sigma_v^2)) + \log(A_v + \frac{K}{2}) + \frac{1}{(2A_v + \frac{K}{2})} \tag{12}$$

$$\frac{\partial F(r)}{\partial B_v} = -\frac{A_b}{B_b} \tag{13}$$