Approximate belief updating via semiparametric variational Bayes

Cristian Castiglione & Mauro Bernardi

Statistical methods and models for complex data — Poster Session: 21 September 2022

Motivation

We propose a **variational algorithm** for performing approximate Bayesian inference and Bayesian belief updating for mixed regression and classification models. Specifically, we combine mean field and parametric variational approximations to handle both **non–conjugate** and **non–regular** models within a unified algorithmic approach.

Following Bissiri et al. (2016), we consider models defined by a **minimum risk criterion** for which a proper likelihood function may not be available. Then, the generalized posterior, or belief update, we try to approximate is given by

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}) \exp\{-nR(\boldsymbol{\theta};\mathbf{y})\}, \qquad (1)$$

where $R(\boldsymbol{\theta};\mathbf{y})$ is a risk function linking the parameter $\boldsymbol{\theta} \in \Theta$ and the data $\mathbf{y} \in \mathcal{Y}$.

Model specification

Empirical risk function

We consider regression and classification models which attempt to predict the response y_i using the linear predictor η_i. We measure the misfit between y_i and η_i through the (negative) empirical risk function, i.e. pseudo–likelihood,

$$-nR(\boldsymbol{\theta};\mathbf{y}) = -\frac{n}{\alpha}\log\sigma_\varepsilon^2 - \frac{1}{\alpha\sigma_\varepsilon^2}\sum_{i=1}^n \psi(y_i,\eta_i), \qquad (2)$$

where $\psi(y,\eta)$ is a loss function, σ_ε^2 is a dispersion parameter and α is a non–stochastic constant.

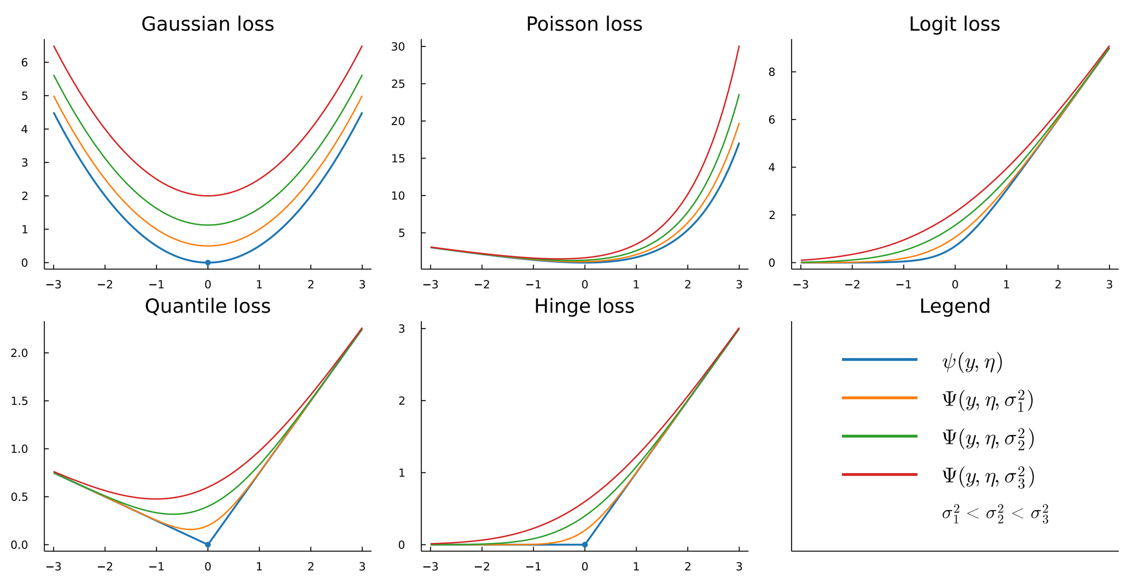


Figure 1. Examples of variational loss functions as defined in Equation (9).

Mixed and additive linear model

We assume an additive model specification for the linear predictor, that is

$$\eta_i = (\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u})_i, \qquad \mathbf{Z}\boldsymbol{u} = \sum_{h=1}^{\mathrm{H}} \mathbf{Z}_h\boldsymbol{u}_h, \qquad (3)$$

where $\mathbf{C} = (\mathbf{X}, \mathbf{Z}_1, \dots, \mathbf{Z}_{\mathrm{H}})$ and $\boldsymbol{u} = (\boldsymbol{u}_1^\top, \dots, \boldsymbol{u}_{\mathrm{H}}^\top)^\top$. The term $\mathbf{X}\boldsymbol{\beta}$ is the fixed effect component, while $\mathbf{Z}_h\boldsymbol{u}_h$ is the h–th random effect component.

Prior distributions

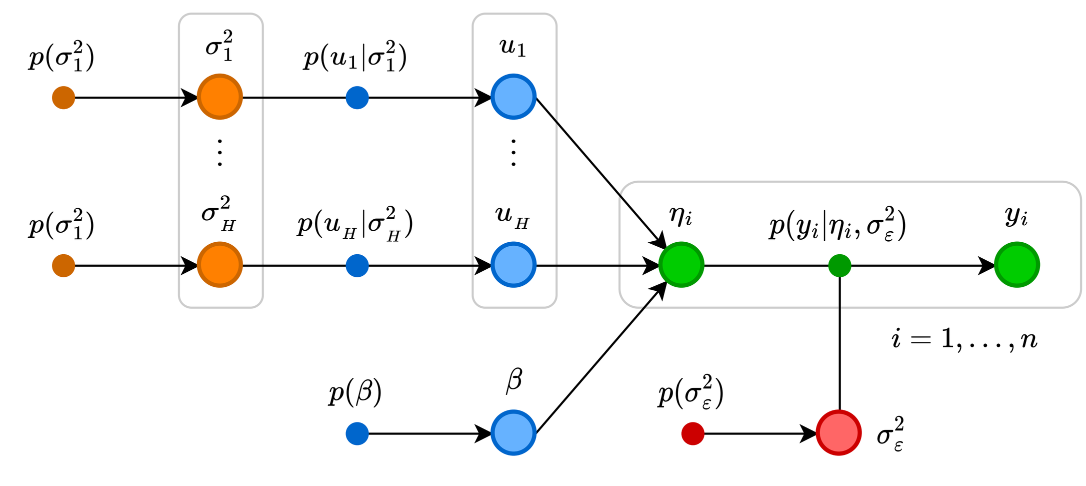
We assume the following set of prior distributions:

$$\begin{aligned} \boldsymbol{u}_h|\sigma_h^2 &\sim \mathsf{N}_{d_h}(\mathbf{0}_{d_h}, \sigma_h^2\mathbf{Q}_h^{-1}), & \sigma_h^2 &\sim \mathsf{IG}(A_h, B_h), \\ \boldsymbol{\beta} &\sim \mathsf{N}_p(\mathbf{0}_p, \sigma_\beta^2\mathbf{I}_p), & \sigma_\varepsilon^2 &\sim \mathsf{IG}(A_\varepsilon, B_\varepsilon), \end{aligned} \qquad (4)$$

where $\sigma_\beta^2, A_\varepsilon, B_\varepsilon, A_h, B_h > 0$ and $\mathbf{Q}_h \succeq 0$, $h = 1, \dots, \mathrm{H}$, are fixed prior parameters, while $\mathrm{K} = p + d_1 + \dots + d_{\mathrm{H}}$ is the total number of regression parameters in the model.

Remark 1. *We do not assume conditional* ***conjugacy*** *or the existence of equivalent* ***data–augmented*** *conjugate models.*

Directed acyclic graph representation



Variational inference

We perform the posterior inference by substituting the true posterior law $p(\boldsymbol{\theta}|\mathbf{y})$ with a variational density $q(\boldsymbol{\theta}) \in \mathcal{Q}$. According to the **mean field** approach (Ormerod and Wand, 2010), we assume that the variational posterior $q(\boldsymbol{\theta})$ factorizes as

$$p(\boldsymbol{\theta}|\mathbf{y}) \approx q(\boldsymbol{\theta}) = q(\boldsymbol{\beta},\boldsymbol{u})\, q(\sigma_1^2) \,\dots\, q(\sigma_{\mathrm{H}}^2)\, q(\sigma_\varepsilon^2). \qquad (5)$$

Moreover, we impose the **parametric restriction**

$$q(\boldsymbol{\beta},\boldsymbol{u};\boldsymbol{\mu},\boldsymbol{\Omega}) \sim \mathsf{N}_{\mathrm{K}}(\boldsymbol{\mu},\boldsymbol{\Omega}). \qquad (6)$$

Then, we select to optimal approximation by maximizing the **evidence lower bound**

$$q^*(\boldsymbol{\theta}) = \underset{q\in\mathcal{Q}}{\operatorname{argmax}}\; L\{\mathbf{y}; q(\boldsymbol{\theta})\}, \qquad (7)$$

where $L\{\mathbf{y}; q(\boldsymbol{\theta})\} = \log p(\mathbf{y}) - \mathrm{KL}\{q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta}|\mathbf{y})\}$.

The optimal coordinatewise solution for $q^*(\sigma_\varepsilon^2)$ and $q^*(\sigma_h^2)$ are available in closed form as Inverse–Gamma densities. For the parametric solution of $q^*(\boldsymbol{\beta},\boldsymbol{u})$ we rely on the **fully simplified multivariate Gaussian update** by Knowles and Minka (2011) and Wand (2014):

$$\begin{aligned} &\text{(update)} & \hat{\boldsymbol{\mu}} &\leftarrow \hat{\boldsymbol{\mu}} - \mathbf{H}^{-1}\mathbf{g}, \quad \hat{\boldsymbol{\Omega}} \leftarrow -\mathbf{H}^{-1}, \\ &\text{(gradient)} & \mathbf{g} &\leftarrow -\mathbf{R}\hat{\boldsymbol{\mu}} - \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top\boldsymbol{\Psi}^{(1)}/\alpha, \\ &\text{(Hessian)} & \mathbf{H} &\leftarrow -\mathbf{R} - \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top\mathrm{diag}\big[\boldsymbol{\Psi}^{(2)}\big]\,\mathbf{C}/\alpha, \end{aligned} \qquad (8)$$

where $\mathbf{R} \leftarrow \mathrm{blockdiag}\big[\sigma_\beta^{-2}\mathbf{I}_p, \mu_{q(1/\sigma_1^2)}\mathbf{Q}_1, \dots, \mu_{q(1/\sigma_{\mathrm{H}}^2)}\mathbf{Q}_{\mathrm{H}}\big]$.

Variational loss derivatives

We define $\boldsymbol{\Psi}^{(r)} = (\Psi_1^{(r)}, \dots, \Psi_n^{(r)})^\top$ and

$$\Psi^{(r)}(y_i, \hat{\eta}_i, \hat{\sigma}_i^2) = \int_{-\infty}^{+\infty} \psi^{(r)}(y_i, x)\, \phi(x; \hat{\eta}_i, \hat{\sigma}_i^2)\, \mathrm{d}x \qquad (9)$$

with $r = 0,1,2$ and $i = 1,\dots,n$. Here, $\psi^{(r)}$ is the r–th weak derivative of ψ calculated wrt η, while

$$\hat{\eta}_i = \mathbf{c}_i^\top\hat{\boldsymbol{\mu}} \quad \text{and} \quad \hat{\sigma}_i^2 = \mathbf{c}_i^\top\hat{\boldsymbol{\Omega}}\,\mathbf{c}_i. \qquad (10)$$

Theorem 1. *Let $\psi(y,\eta)$ be a continuous, convex function wrt η with r–th order weak derivative $\psi^{(r)}$. Then, we have:*

1. *$\Psi^{(r)}(y,\eta,\sigma^2)$ is infinitely* ***differentiable*** *wrt η and σ^2;*
2. *$\Psi^{(0)}(y,\eta,\sigma^2)$ is jointly* ***convex*** *wrt η and σ^2;*
3. *$\Psi^{(0)}(y,\eta,\sigma^2) \geq \psi(y,\eta)$ for any η and σ^2;*
4. *$\Psi^{(0)}(y,\eta,\sigma^2) \to \psi(y,\eta)$ as $\sigma^2 \to 0$.*

Remark 2. *Because of Theorem 1, $L\{\mathbf{y}; q(\boldsymbol{\theta})\}$ is concave and differentiable wrt $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$. Therefore, all the solutions of (7) are* ***global maximizers*** *and belong in a* ***closed convex set****.*

Algorithm

We end up with a semiparametric variational Bayes routine which can be viewed as a variational implementation of the **penalized iterated reweighted least squares** algorithm (Wood, 2017).

Semiparametric variational Bayes algorithm	
Initialize $\hat{A}_\varepsilon, \hat{B}_\varepsilon, \hat{A}_h, \hat{B}_h, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}$;	
While *convergence is not reached* **do**:	
Evaluate $\boldsymbol{\Psi}^{(0)}, \boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}$;	$\mathcal{O}(n\mathrm{K}^2)$
$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \{A_\varepsilon + n/\alpha\}/\{B_\varepsilon + \mathbf{1}_n^\top\boldsymbol{\Psi}^{(0)}/\alpha\}$;	$\mathcal{O}(n)$
$\mu_{q(1/\sigma_1^2)} \leftarrow \{A_1 + d_1/2\}/\{B_1 + \frac{1}{2}[\hat{\boldsymbol{\mu}}_1^\top\mathbf{Q}_1\hat{\boldsymbol{\mu}}_1 + \mathrm{trace}(\mathbf{Q}_1\hat{\boldsymbol{\Sigma}}_{11})]\}$;	$\mathcal{O}(d_1^3)$
$\dots \leftarrow \dots$	$\dots$
$\mu_{q(1/\sigma_{\mathrm{H}}^2)} \leftarrow \{A_{\mathrm{H}} + d_{\mathrm{H}}/2\}/\{B_{\mathrm{H}} + \frac{1}{2}[\hat{\boldsymbol{\mu}}_{\mathrm{H}}^\top\mathbf{Q}_{\mathrm{H}}\hat{\boldsymbol{\mu}}_{\mathrm{H}} + \mathrm{trace}(\mathbf{Q}_{\mathrm{H}}\hat{\boldsymbol{\Sigma}}_{\mathrm{HH}})]\}$;	$\mathcal{O}(d_{\mathrm{H}}^3)$
$\mathbf{R} \leftarrow \mathrm{blockdiag}[\sigma_\beta^{-2}\mathbf{I}_p, \mu_{q(1/\sigma_1^2)}\mathbf{Q}_1, \dots, \mu_{q(1/\sigma_{\mathrm{H}}^2)}\mathbf{Q}_{\mathrm{H}}]$;	
$\mathbf{g} \leftarrow -\mathbf{R}\hat{\boldsymbol{\mu}} - \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top\boldsymbol{\Psi}^{(1)}/\alpha$;	$\mathcal{O}(n\mathrm{K}^2)$
$\mathbf{H} \leftarrow -\mathbf{R} - \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top\mathrm{diag}[\boldsymbol{\Psi}^{(2)}]\mathbf{C}/\alpha$;	$\mathcal{O}(n\mathrm{K}^2)$
$\rho \leftarrow \mathrm{LineSearch}(\mathbf{g},\mathbf{H})$; $\quad \hat{\boldsymbol{\Sigma}} \leftarrow -\mathbf{H}^{-1}$; $\quad \hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} - \rho\,\mathbf{H}^{-1}\mathbf{g}$;	$\mathcal{O}(\mathrm{K}^3)$
End of while	

Total computational complexity: $\mathcal{O}(n\mathrm{K}^2 + \mathrm{K}^3)$

Extensions

- Streamlined algorithms for **cross-random effects**, **DLM**, **GMRF**;
- Inducing **shrinkage** and **sparsity** prior distributions;
- **Skew normal** variational approximations;
- Frequentist mixed models with **non–regular likelihood**.

Simulation results

We simulated 100 datasets having 500 observations each from a non–linear heteroscedastic model. We estimated the 90% conditional quantile of the data by using a Bayesian semiparametric quantile regression model. The posterior inference is performed via **Markov chain Monte Carlo** (MCMC), **conjugate mean field variational Bayes** (MFVB) and **semiparametric variational Bayes** (SVB).

Method	Accuracy	RMSE	Iterations	Exe. Time
MCMC		0.764 (0.054)	10000	3.944 (0.041)
MFVB	0.776 (0.021)	0.763 (0.051)	41.919 (13.502)	0.084 (0.033)
SVB	**0.951** (0.011)	**0.763** (0.050)	**44.694** (14.632)	**0.097** (0.053)

Table 1. Performance measures (standard errors).

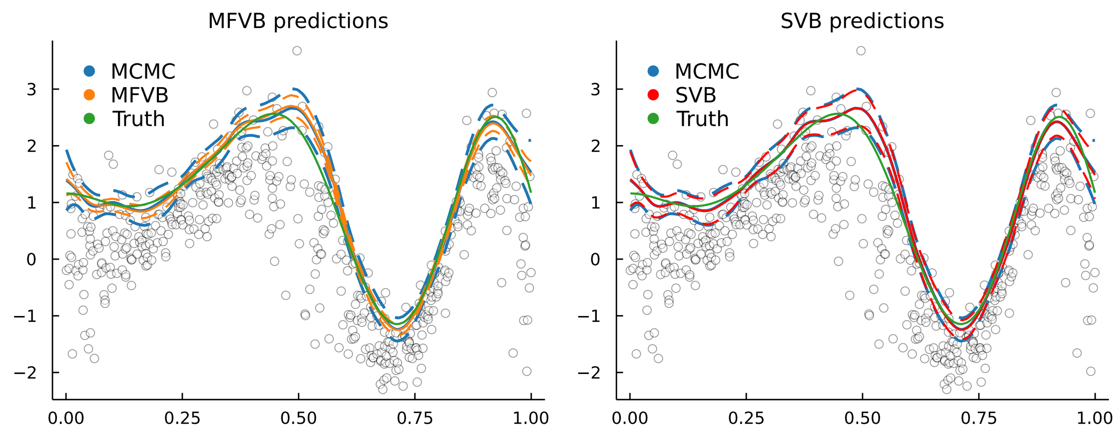


Figure 2. Posterior predictive distributions.

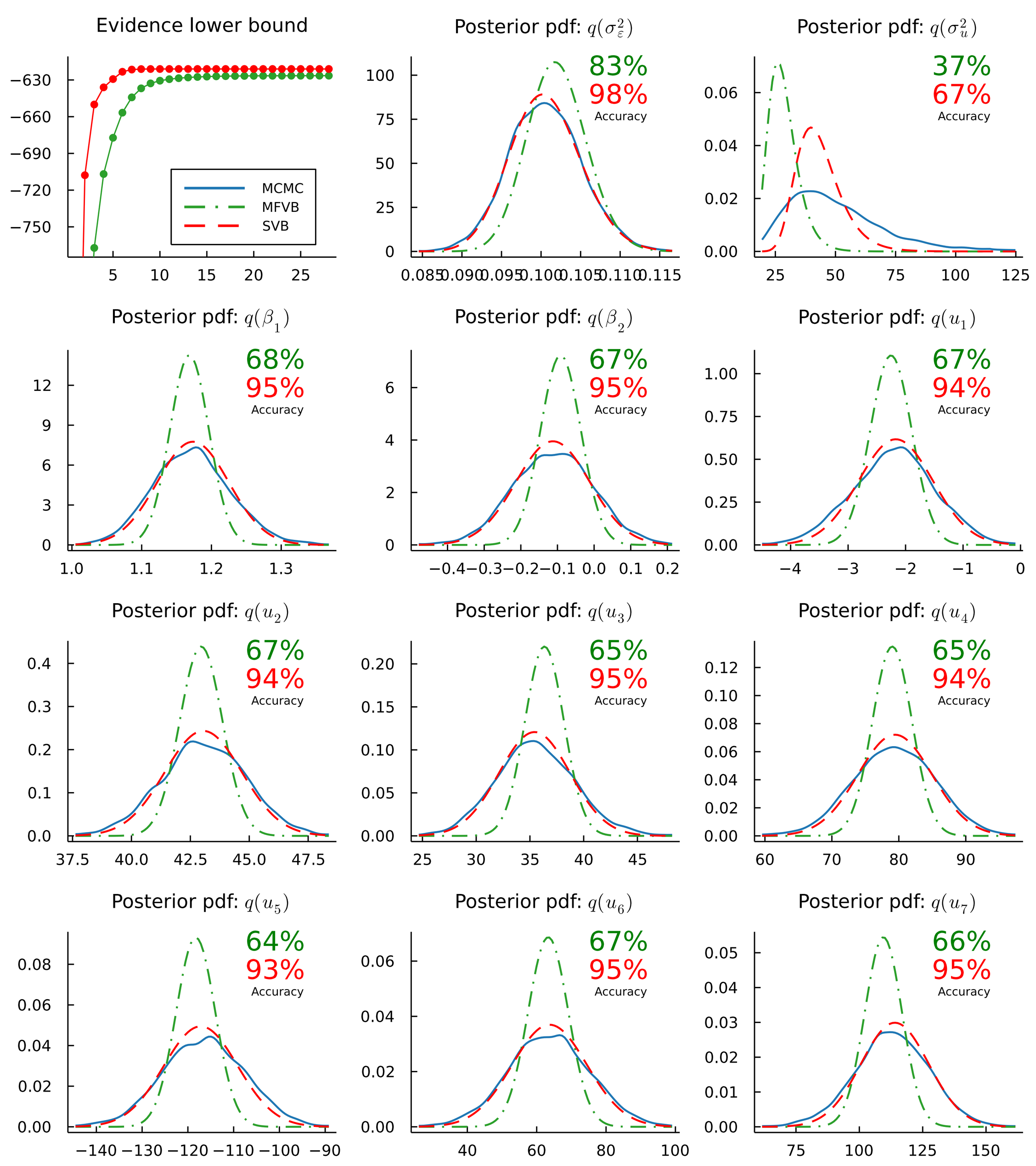


Figure 3. Marginal posterior density functions.

References

Bissiri, P.G., Holmes, C.C., and Walker, S.G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, **78**(5), 1103 – 1130.

Castiglione, C., Bernardi, M. (2022). Bayesian non–conjugate regression via variational belief updating. *arXiv preprint, arXiv:2206.09444*.

Knowles, D., Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. *Advances in Neural Information Processing Systems*, **24**, 1701 – 1709.

Ormerod, J.T., Wand, M.P. (2010). Explaining variational approximations. *The American Statistician*, **64**(2), 140 – 153.

Wand, M.P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research*, **15**, 1351 – 1369.

Wood, S. N. (2017). *Generalized additive models. An introduction with R, Second edition*. CRC Press, Boca Raton, FL.

Contact information

Cristian Castiglione, Ph.D. student
Department of Statistical Sciences, University of Padova
cristian.castiglione@phd.unipd.it
CristianCastiglione