Approximate General Bayesian Inference via Semiparametric Variational Bayes

Cristian Castiglione Mauro Bernardi

University of Padova, Department of Statistical Sciences

cristian.castiglione@phd.unipd.it *mauro.bernardi@unipd.it*

June 29, 2022

Overview

Some background

Risk–based parameter definition

$$
\begin{aligned}
\text{Data distribution}: &\quad y \sim (\mathcal{Y}, \mathcal{F}, \mathbb{P}) \\
\text{Parameter of interest}: &\quad \boldsymbol{\theta} = \operatorname{argmin}\ R(\,\cdot\,; \mathbb{P}) \\
\text{Theoretical risk function}: &\quad R(\boldsymbol{\theta}; \mathbb{P}) = \mathbb{E}\{L(y, \boldsymbol{\theta})\}
\end{aligned}
$$

General belief updating (Bissiri et al., 2016)

$$
\begin{aligned}
\text{Subjective prior belief}: &\quad \boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \\
\text{Empirical risk function}: &\quad R(\boldsymbol{\theta}; \mathbf{y}) = \textstyle\sum_{i=1}^{n} L(y_i, \boldsymbol{\theta})/n \\
\text{Bayesian belief updating}: &\quad p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}) \exp\{-nR(\boldsymbol{\theta}; \mathbf{y})\}
\end{aligned}
$$

Some background

Risk–based parameter definition

Data distribution : $y \sim (\mathcal{Y}, \mathcal{F}, \mathbb{P})$

Parameter of interest : $\boldsymbol{\theta} = \mathrm{argmin}\ R(\,\cdot; \mathbb{P})$

Theoretical risk function : $R(\boldsymbol{\theta}; \mathbb{P}) = \mathbb{E}\{L(y, \boldsymbol{\theta})\}$

General belief updating (Bissiri et al., 2016)

Subjective prior belief : $\boldsymbol{\theta} \sim p(\boldsymbol{\theta})$

Empirical risk function : $R(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i=1}^{n} L(y_i, \boldsymbol{\theta})/n$

Bayesian belief updating : $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}) \exp\{-nR(\boldsymbol{\theta}; \mathbf{y})\}$

Overview

Bayesian mixed model

Empirical risk function

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

- $\{y_i, \mathbf{x}_i, \mathbf{z}_i\} \in \mathcal{Y} \times \mathbb{R}^p \times \mathbb{R}^d$
- $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}$
- $\psi(y, \eta)$: loss function
- σ_ε^2: dispersion parameter
- α: calibrating parameter

Bayesian mixed model

Empirical risk function

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

- $\{y_i, \mathbf{x}_i, \mathbf{z}_i\} \in \mathcal{Y} \times \mathbb{R}^p \times \mathbb{R}^d$
- $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}$
- $\psi(y, \eta)$: loss function
- σ_ε^2: dispersion parameter
- α: calibrating parameter

Bayesian mixed model

Empirical risk function

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

- $\{y_i, \mathbf{x}_i, \mathbf{z}_i\} \in \mathcal{Y} \times \mathbb{R}^p \times \mathbb{R}^d$
- $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}$
- $\psi(y, \eta)$: loss function
- σ_ε^2: dispersion parameter
- α: calibrating parameter

Bayesian mixed model

Empirical risk function

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

- $\{y_i, \mathbf{x}_i, \mathbf{z}_i\} \in \mathcal{Y} \times \mathbb{R}^p \times \mathbb{R}^d$
- $\eta_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \mathbf{z}_i^\top \boldsymbol{u}$
- $\psi(y,\eta)$: loss function
- σ_ε^2: dispersion parameter
- α: calibrating parameter

Prior distributions

$$\boldsymbol{u}|\sigma_u^2 \sim \mathsf{N}_d(\mathbf{0}_d, \sigma_u^2\mathbf{Q}^{-1}), \qquad \sigma_u^2 \sim \mathsf{IG}(A_u, B_u),$$
$$\boldsymbol{\beta} \sim \mathsf{N}_p(\mathbf{0}_p, \sigma_\beta^2\mathbf{I}_p), \qquad \sigma_\varepsilon^2 \sim \mathsf{IG}(A_\varepsilon, B_\varepsilon),$$

Here, $\sigma_\beta^2, A_\varepsilon, B_\varepsilon, A_u, B_u > 0$ and $\mathbf{Q} \succeq 0$ are fixed prior parameters.

Generalized posterior distribution

$$\underbrace{p(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2, \sigma_\varepsilon^2 \mid \mathbf{y})}_{\text{Generalized posterior}} \propto \underbrace{p(\sigma_\varepsilon^2)\, p(\boldsymbol{\beta})\, p(\sigma_u^2)\, p(\boldsymbol{u}|\sigma_u^2)}_{\text{Prior beliefs}}\ \underbrace{\exp\{-nR(\mathbf{y}, \boldsymbol{\theta})\}}_{\text{Pseudo–likelihood}}$$

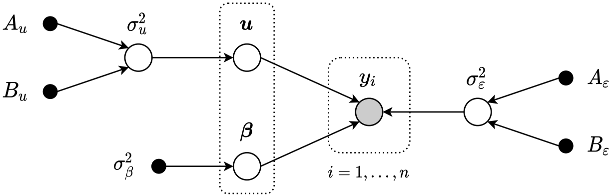


Figure: Direct acyclic graph representing the Bayesian model.

Overview

Semiparametric variational inference

- Variational approximation: $p(\boldsymbol{\theta}|\mathbf{y}) \approx q(\boldsymbol{\theta};\boldsymbol{\xi}) = q_1(\boldsymbol{\theta}_1)\cdots q_K(\boldsymbol{\theta}_K)\, q_\phi(\phi;\boldsymbol{\xi})$
- Variational problem: $(q_1^*, \ldots, q_K^*, q_\phi^*) = \operatorname{argmin}\ \mathsf{KL}(q(\boldsymbol{\theta};\boldsymbol{\xi}) \,\|\, p(\boldsymbol{\theta}|\mathbf{y}))$

Mean field variational Bayes

$$q_k^*(\boldsymbol{\theta}_k) \propto \exp\left[\mathbb{E}_{-k}\{\log p(\boldsymbol{\theta}_k \mid \mathsf{rest})\}\right]$$

Knowles–Minka–Wand recursion

- $q_\phi(\phi;\boldsymbol{\mu},\Sigma) \sim \mathsf{N}(\boldsymbol{\mu},\Sigma)$
- $\hat{\mu} \leftarrow \hat{\mu} - \rho\,\mathbf{H}^{-1}\mathbf{g}$
- $\hat{\Sigma} \leftarrow -\mathbf{H}^{-1}$

- $f(\boldsymbol{\mu},\Sigma) = \mathbb{E}_q\{\log p(\mathbf{y},\boldsymbol{\theta})\}$
- $\mathbf{g}(\boldsymbol{\mu},\Sigma) = \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu},\Sigma)$
- $\mathbf{H}(\boldsymbol{\mu},\Sigma) = \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu},\Sigma)$

See Ormerod and Wand (2010), Blei et al. (2017) for MFVB.
See Knowless and Minka (2011), Wand (2014), Rohde and Wand (2016) for KMW.

Semiparametric variational inference

- Variational approximation: $p(\boldsymbol{\theta}|\mathbf{y}) \approx q(\boldsymbol{\theta};\boldsymbol{\xi}) = q_1(\boldsymbol{\theta}_1)\cdots q_K(\boldsymbol{\theta}_K)\, q_\phi(\phi;\boldsymbol{\xi})$
- Variational problem: $(q_1^*,\dots,q_K^*,q_\phi^*) = \operatorname{argmin}\ \mathsf{KL}(q(\boldsymbol{\theta};\boldsymbol{\xi}) \,\|\, p(\boldsymbol{\theta}|\mathbf{y}))$

Mean field variational Bayes

$$q_k^*(\boldsymbol{\theta}_k) \propto \exp\left[\mathbb{E}_{-k}\{\log p(\boldsymbol{\theta}_k \mid \text{rest})\}\right]$$

Knowles–Minka–Wand recursion

- $q_\phi(\phi;\boldsymbol{\mu},\Sigma) \sim \mathrm{N}(\boldsymbol{\mu},\Sigma)$
- $\hat{\mu} \leftarrow \hat{\mu} - \rho\,\mathbf{H}^{-1}\mathrm{g}$
- $\hat{\Sigma} \leftarrow -\mathbf{H}^{-1}$

- $f(\boldsymbol{\mu},\Sigma) = \mathbb{E}_q\{\log p(\mathbf{y},\boldsymbol{\theta})\}$
- $\mathrm{g}(\boldsymbol{\mu},\Sigma) = \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu},\Sigma)$
- $\mathbf{H}(\boldsymbol{\mu},\Sigma) = \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu},\Sigma)$

See Ormerod and Wand (2010), Blei et al. (2017) for MFVB.
See Knowless and Minka (2011), Wand (2014), Rohde and Wand (2016) for KMW.

Semiparametric variational inference

- Variational approximation: $p(\boldsymbol{\theta}|\mathbf{y}) \approx q(\boldsymbol{\theta};\boldsymbol{\xi}) = q_1(\boldsymbol{\theta}_1)\cdots q_K(\boldsymbol{\theta}_K)\, q_\phi(\boldsymbol{\phi};\boldsymbol{\xi})$
- Variational problem: $(q_1^*,\dots,q_K^*,q_\phi^*) = \operatorname{argmin}\, \mathsf{KL}(q(\boldsymbol{\theta};\boldsymbol{\xi}) \,\|\, p(\boldsymbol{\theta}|\mathbf{y}))$

Mean field variational Bayes

$$q_k^*(\boldsymbol{\theta}_k) \propto \exp\left[\mathbb{E}_{-k}\{\log p(\boldsymbol{\theta}_k \mid \text{rest})\}\right]$$

Knowles–Minka–Wand recursion

- $q_\phi(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma}) \sim \mathsf{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$
- $\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} - \rho\,\mathbf{H}^{-1}\mathbf{g}$
- $\hat{\boldsymbol{\Sigma}} \leftarrow -\mathbf{H}^{-1}$

- $f(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathbb{E}_q\{\log p(\mathbf{y},\boldsymbol{\theta})\}$
- $\mathbf{g}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu},\boldsymbol{\Sigma})$
- $\mathbf{H}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu},\boldsymbol{\Sigma})$

See Ormerod and Wand (2010), Blei et al. (2017) for MFVB.
See Knowless and Minka (2011), Wand (2014), Rohde and Wand (2016) for KMW.

Semiparametric variational inference

- Variational approximation: $p(\boldsymbol{\theta}|\mathbf{y}) \approx q(\boldsymbol{\theta};\boldsymbol{\xi}) = q_1(\boldsymbol{\theta}_1)\cdots q_K(\boldsymbol{\theta}_K)\, q_\phi(\phi;\boldsymbol{\xi})$
- Variational problem: $(q_1^*,\ldots,q_K^*,q_\phi^*) = \operatorname{argmin}\ \mathsf{KL}(q(\boldsymbol{\theta};\boldsymbol{\xi})\,\|\,p(\boldsymbol{\theta}|\mathbf{y}))$

Mean field variational Bayes

$$q_k^*(\boldsymbol{\theta}_k) \propto \exp\left[\mathbb{E}_{-k}\{\log p(\boldsymbol{\theta}_k \mid \text{rest})\}\right]$$

Knowles–Minka–Wand recursion

- $q_\phi(\phi;\boldsymbol{\mu},\boldsymbol{\Sigma}) \sim \mathsf{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$
- $\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} - \rho\,\mathbf{H}^{-1}\mathbf{g}$
- $\hat{\boldsymbol{\Sigma}} \leftarrow -\mathbf{H}^{-1}$

- $f(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathbb{E}_q\{\log p(\mathbf{y},\boldsymbol{\theta})\}$
- $\mathbf{g}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu},\boldsymbol{\Sigma})$
- $\mathbf{H}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu},\boldsymbol{\Sigma})$

See Ormerod and Wand (2010), Blei et al. (2017) for MFVB.
See Knowless and Minka (2011), Wand (2014), Rohde and Wand (2016) for KMW.

Semiparametric variational inference

- Variational approximation: $p(\boldsymbol{\theta}|\mathbf{y}) \approx q(\boldsymbol{\theta};\boldsymbol{\xi}) = q_1(\boldsymbol{\theta}_1)\cdots q_K(\boldsymbol{\theta}_K)\, q_\phi(\boldsymbol{\phi};\boldsymbol{\xi})$
- Variational problem: $(q_1^*, \dots, q_K^*, q_\phi^*) = \operatorname{argmin} \mathsf{KL}(q(\boldsymbol{\theta};\boldsymbol{\xi}) \,\|\, p(\boldsymbol{\theta}|\mathbf{y}))$

Mean field variational Bayes

$$q_k^*(\boldsymbol{\theta}_k) \propto \exp\left[\mathbb{E}_{-k}\{\log p(\boldsymbol{\theta}_k \mid \text{rest})\}\right]$$

Knowles–Minka–Wand recursion

- $q_\phi(\boldsymbol{\phi};\boldsymbol{\mu},\boldsymbol{\Sigma}) \sim \mathsf{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$
- $\hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} - \rho\,\mathbf{H}^{-1}\mathbf{g}$
- $\hat{\boldsymbol{\Sigma}} \leftarrow -\mathbf{H}^{-1}$

- $f(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathbb{E}_q\{\log p(\mathbf{y},\boldsymbol{\theta})\}$
- $\mathbf{g}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu},\boldsymbol{\Sigma})$
- $\mathbf{H}(\boldsymbol{\mu},\boldsymbol{\Sigma}) = \nabla_{\boldsymbol{\mu}} f(\boldsymbol{\mu},\boldsymbol{\Sigma})$

See Ormerod and Wand (2010), Blei et al. (2017) for MFVB.
See Knowless and Minka (2011), Wand (2014), Rohde and Wand (2016) for KMW.

Optimal variational distributions

Assumptions:

- mean field factorization: $q(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2, \sigma_\varepsilon^2) = q(\sigma_\varepsilon^2)\, q(\sigma_u^2)\, q(\boldsymbol{\beta}, \boldsymbol{u})$
- parametric restriction: $q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathsf{N}\left(\begin{bmatrix} \boldsymbol{\mu}_\beta \\ \boldsymbol{\mu}_u \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\beta\beta} & \boldsymbol{\Sigma}_{\beta u} \\ \boldsymbol{\Sigma}_{u\beta} & \boldsymbol{\Sigma}_{uu} \end{bmatrix} \right)$

Closed form solutions:

- $q^*(\sigma_\varepsilon^2) \sim \mathsf{IG}(\hat{A}_\varepsilon, \hat{B}_\varepsilon)$ where $\hat{A}_\varepsilon \leftarrow A_\varepsilon + n/\alpha$ and $\hat{B}_\varepsilon \leftarrow B_\varepsilon + \mathbf{1}_n^\top \boldsymbol{\Psi}^{(0)}/\alpha$
- $q^*(\sigma_u^2) \sim \mathsf{IG}(\hat{A}_u, \hat{B}_u)$ where $\hat{A}_u \leftarrow A_u + d/2$ and $\hat{B}_u \leftarrow B_u + \frac{1}{2}\hat{\boldsymbol{\mu}}_u^\top \mathbf{Q}\, \hat{\boldsymbol{\mu}}_u + \frac{1}{2}\mathrm{trace}\left[\mathbf{Q}\hat{\boldsymbol{\Sigma}}_{uu}\right]$

Optimal variational distributions

Assumptions:

- mean field factorization: $q(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2, \sigma_\varepsilon^2) = q(\sigma_\varepsilon^2)\, q(\sigma_u^2)\, q(\boldsymbol{\beta}, \boldsymbol{u})$
- parametric restriction: $q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathsf{N}\left(\begin{bmatrix} \boldsymbol{\mu}_\beta \\ \boldsymbol{\mu}_u \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\beta\beta} & \boldsymbol{\Sigma}_{\beta u} \\ \boldsymbol{\Sigma}_{u\beta} & \boldsymbol{\Sigma}_{uu} \end{bmatrix} \right)$

Closed form solutions:

- $q^*(\sigma_\varepsilon^2) \sim \mathsf{IG}(\hat{A}_\varepsilon, \hat{B}_\varepsilon)$ where $\hat{A}_\varepsilon \leftarrow A_\varepsilon + n/\alpha$ and $\hat{B}_\varepsilon \leftarrow B_\varepsilon + \mathbf{1}_n^\top \boldsymbol{\Psi}^{(0)}/\alpha$
- $q^*(\sigma_u^2) \sim \mathsf{IG}(\hat{A}_u, \hat{B}_u)$ where $\hat{A}_u \leftarrow A_u + d/2$ and $\hat{B}_u \leftarrow B_u + \frac{1}{2}\hat{\boldsymbol{\mu}}_u^\top \mathbf{Q}\, \hat{\boldsymbol{\mu}}_u + \frac{1}{2}\mathrm{trace}\big[\mathbf{Q}\hat{\boldsymbol{\Sigma}}_{uu}\big]$

Knowles–Minka–Wand recursion

Parametric solution:

$$\text{(update)} \qquad \boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \rho\,\mathbf{H}^{-1}\mathbf{g}, \quad \boldsymbol{\Sigma} \leftarrow -\mathbf{H}^{-1},$$

$$\text{(gradient)} \qquad \mathbf{g}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{\mu}_\beta \\ \mu_{q(1/\sigma_u^2)}\mathbf{Q}\,\boldsymbol{\mu}_u \end{bmatrix} - \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top \boldsymbol{\Psi}^{(1)}/\alpha,$$

$$\text{(Hessian)} \qquad \mathbf{H}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\begin{bmatrix} \sigma_\beta^{-2}\mathbf{I}_p & \mathbf{O} \\ \mathbf{O} & \mu_{q(1/\sigma_u^2)}\mathbf{Q} \end{bmatrix} - \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top \mathrm{diag}\big[\boldsymbol{\Psi}^{(2)}\big]\,\mathbf{C}/\alpha,$$

where $\mathbf{C} = \big[\mathbf{X},\, \mathbf{Z}\big]$, $\mu_{q(1/\sigma_u^2)} = \mathbb{E}_q(1/\sigma_u^2)$, $\mu_{q(1/\sigma_\varepsilon^2)} = \mathbb{E}_q(1/\sigma_\varepsilon^2)$ and

$$\Psi_i^{(r)} = \Psi^{(r)}(y_i, \mathbf{c}_i^\top\boldsymbol{\mu}, \mathbf{c}_i^\top\boldsymbol{\Sigma}\,\mathbf{c}_i) = \mathbb{E}_q\left\{\frac{\partial^r}{\partial\eta^r}\psi(y_i, \eta_i)\right\}, \qquad r = 0, 1, 2.$$

Knowles–Minka–Wand recursion

Parametric solution:

$$\text{(update)} \qquad \boldsymbol{\mu} \leftarrow \boldsymbol{\mu} - \rho\,\mathbf{H}^{-1}\mathbf{g}, \quad \boldsymbol{\Sigma} \leftarrow -\mathbf{H}^{-1},$$

$$\text{(gradient)} \qquad \mathbf{g}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\begin{bmatrix} \sigma_\beta^{-2}\boldsymbol{\mu}_\beta \\ \mu_{q(1/\sigma_u^2)}\mathbf{Q}\,\boldsymbol{\mu}_u \end{bmatrix} - \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top\boldsymbol{\Psi}^{(1)}/\alpha,$$

$$\text{(Hessian)} \qquad \mathbf{H}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\begin{bmatrix} \sigma_\beta^{-2}\mathbf{I}_p & \mathbf{O} \\ \mathbf{O} & \mu_{q(1/\sigma_u^2)}\mathbf{Q} \end{bmatrix} - \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top\mathrm{diag}\big[\boldsymbol{\Psi}^{(2)}\big]\,\mathbf{C}/\alpha,$$

where $\mathbf{C} = \big[\mathbf{X},\,\mathbf{Z}\big]$, $\mu_{q(1/\sigma_u^2)} = \mathbb{E}_q(1/\sigma_u^2)$, $\mu_{q(1/\sigma_\varepsilon^2)} = \mathbb{E}_q(1/\sigma_\varepsilon^2)$ and

$$\Psi_i^{(r)} = \Psi^{(r)}(y_i, \mathbf{c}_i^\top\boldsymbol{\mu}, \mathbf{c}_i^\top\boldsymbol{\Sigma}\,\mathbf{c}_i) = \mathbb{E}_q\left\{\frac{\partial^r}{\partial\eta^r}\psi(y_i, \eta_i)\right\}, \qquad r = 0, 1, 2.$$

Ψ–functions

Proposition

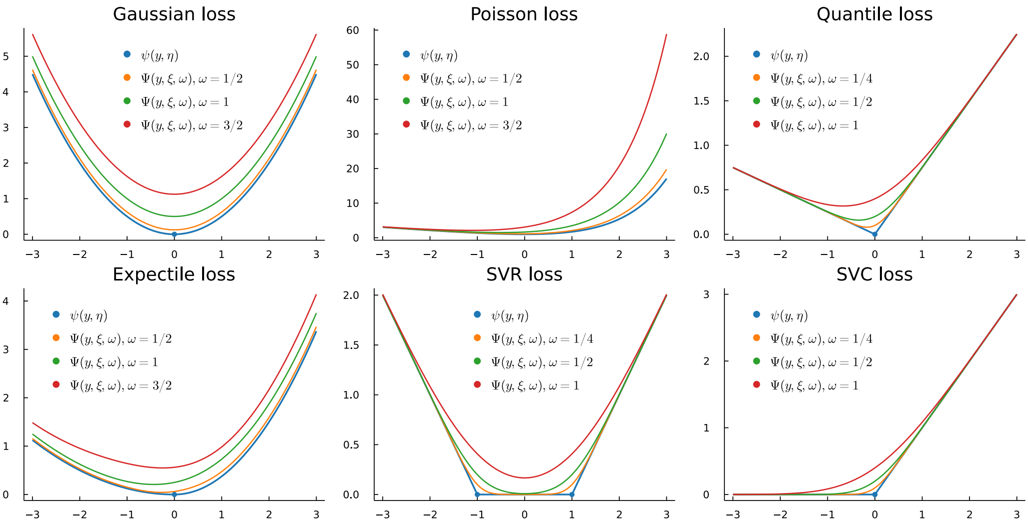
Let $\psi(y, \eta)$ be a **convex**, 2–times **weakly differentiable** function wrt to η, then:

1. $\Psi^{(r)}(y, \mathbf{c}^\top \boldsymbol{\mu}, \mathbf{c}^\top \boldsymbol{\Sigma}\, \mathbf{c})$ has infinitely many derivatives wrt $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$;
2. $\Psi^{(0)}(y, \mathbf{c}^\top \boldsymbol{\mu}, \mathbf{c}^\top \boldsymbol{\Sigma}\, \mathbf{c})$ is jointly convex wrt $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$;
3. $\Psi^{(0)}(y, \mathbf{c}^\top \boldsymbol{\mu}, \mathbf{c}^\top \boldsymbol{\Sigma}\, \mathbf{c}) \geq \psi(y, \mathbf{c}^\top \boldsymbol{\mu})$ for any y and $\mathbf{c}$.

Then

- the optimum of the KL–divergence is unique
- the KMW recursion converges to the optimum

Models


Gaussian loss
$\psi(y,\eta)$
$\Psi(y,\xi,\omega), \omega = 1/2$
$\Psi(y,\xi,\omega), \omega = 1$
$\Psi(y,\xi,\omega), \omega = 3/2$
Poisson loss
$\psi(y,\eta)$
$\Psi(y,\xi,\omega), \omega = 1/2$
$\Psi(y,\xi,\omega), \omega = 1$
$\Psi(y,\xi,\omega), \omega = 3/2$
Quantile loss
$\psi(y,\eta)$
$\Psi(y,\xi,\omega), \omega = 1/4$
$\Psi(y,\xi,\omega), \omega = 1/2$
$\Psi(y,\xi,\omega), \omega = 1$
Expectile loss
$\psi(y,\eta)$
$\Psi(y,\xi,\omega), \omega = 1/2$
$\Psi(y,\xi,\omega), \omega = 1$
$\Psi(y,\xi,\omega), \omega = 3/2$
SVR loss
$\psi(y,\eta)$
$\Psi(y,\xi,\omega), \omega = 1/4$
$\Psi(y,\xi,\omega), \omega = 1/2$
$\Psi(y,\xi,\omega), \omega = 1$
SVC loss
$\psi(y,\eta)$
$\Psi(y,\xi,\omega), \omega = 1/4$
$\Psi(y,\xi,\omega), \omega = 1/2$
$\Psi(y,\xi,\omega), \omega = 1$

Algorithm

Semiparametric variational Bayes algorithm for approximate Bayesian inference

---

Initialize $\hat{A}_\varepsilon$, $\hat{B}_\varepsilon$, $\hat{A}_u$, $\hat{B}_u$, $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$;

While *convergence is not reached* **do**:

Evaluate $\boldsymbol{\Psi}^{(0)}$, $\boldsymbol{\Psi}^{(1)}$, $\boldsymbol{\Psi}^{(2)}$;

$\hat{A}_u \leftarrow A_u + d/2; \quad \hat{B}_u \leftarrow B_u + \frac{1}{2}\hat{\boldsymbol{\mu}}_u^\top \mathbf{Q}\,\hat{\boldsymbol{\mu}}_u + \frac{1}{2}\mathrm{trace}\big[\mathbf{Q}\,\hat{\boldsymbol{\Sigma}}_{uu}\big];$

$\hat{A}_\varepsilon \leftarrow A_\varepsilon + n/\alpha; \quad \hat{B}_\varepsilon \leftarrow B_\varepsilon + \mathbf{1}_n^\top \boldsymbol{\Psi}^{(0)}/\alpha;$

$\mu_{q(1/\sigma_u^2)} \leftarrow \hat{A}_u/\hat{B}_u; \quad \mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \hat{A}_\varepsilon/\hat{B}_\varepsilon;$

$\mathbf{g} \leftarrow -\,\mathrm{stack}\big[\sigma_\beta^{-2}\hat{\boldsymbol{\mu}}_\beta, \mu_{q(1/\sigma_u^2)}\mathbf{Q}\,\hat{\boldsymbol{\mu}}_u\big] - \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top\boldsymbol{\Psi}^{(1)}/\alpha;$

$\mathbf{H} \leftarrow -\,\mathrm{blockdiag}\big[\sigma_\beta^{-2}\mathbf{I}_p, \mu_{q(1/\sigma_u^2)}\mathbf{Q}\big] - \mu_{q(1/\sigma_\varepsilon^2)}\mathbf{C}^\top\mathrm{diag}\big[\boldsymbol{\Psi}^{(2)}\big]\mathbf{C}/\alpha;$

$\rho \leftarrow \mathrm{LineSearch}(f, \mathbf{g}, \mathbf{H}); \quad \hat{\boldsymbol{\Sigma}} \leftarrow -\mathbf{H}^{-1}; \quad \hat{\boldsymbol{\mu}} \leftarrow \hat{\boldsymbol{\mu}} - \rho\,\mathbf{H}^{-1}\mathbf{g};$

End of while

Overview

Simulation setup

1. **Data generating process**:
   - $y_i | x_i \overset{\text{ind.}}{\sim} \begin{cases} \mathsf{N}(f(x_i), \{g(x_i)\}^2) & \text{for regression} \\ \mathsf{Be}(\text{expit}\{f(x_i)\}) & \text{for classification} \end{cases}$
   - mean function: $f(x) = 1.6 \sin(3\pi x^2)$
   - variance function: $g(x) = \exp\{-0.6 + 0.5 \cos(4\pi x)\}$
2. **Linear predictor**:
   - Bayesian penalized semiparametric regression
3. **Loss functions**:
   - quantile regression (MCMC: Kozumi and Kobayashi, 2011; MFVB: Wand et al., 2011)
   - expectile regression (MCMC: Waldmann et al., 2017; Laplace approximation)
   - support vector regression (MCMC: Polson and Scott, 2011; MFVB: Luts and Ormerod, 2014)
   - support vector classification (MCMC: Polson and Scott, 2011; MFVB: Luts and Ormerod, 2014)

Simulation setup

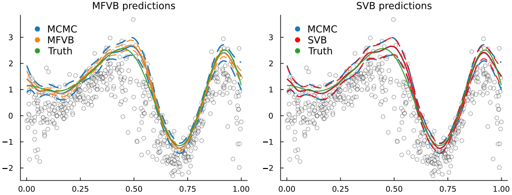
1. **Data generating process**:
   - $y_i | x_i \stackrel{\text{ind.}}{\sim} \begin{cases} \mathsf{N}(f(x_i), \{g(x_i)\}^2) & \text{for regression} \\ \mathrm{Be}(\mathrm{expit}\{f(x_i)\}) & \text{for classification} \end{cases}$
   - mean function: $f(x) = 1.6\sin(3\pi x^2)$
   - variance function: $g(x) = \exp\{-0.6 + 0.5\cos(4\pi x)\}$
2. **Linear predictor**:
   - Bayesian penalized semiparametric regression
3. **Loss functions**:
   - quantile regression (MCMC: Kozumi and Kobayashi, 2011; MFVB: Wand et al., 2011)
   - expectile regression (MCMC: Waldmann et al., 2017; Laplace approximation)
   - support vector regression (MCMC: Polson and Scott, 2011; MFVB: Luts and Ormerod, 2014)
   - support vector classification (MCMC: Polson and Scott, 2011; MFVB: Luts and Ormerod, 2014)

Simulation setup

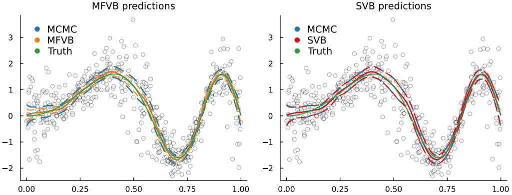
1. **Data generating process**:
   - $y_i | x_i \stackrel{\text{ind.}}{\sim} \begin{cases} \mathsf{N}(f(x_i), \{g(x_i)\}^2) & \text{for regression} \\ \mathsf{Be}(\text{expit}\{f(x_i)\}) & \text{for classification} \end{cases}$
   - mean function: $f(x) = 1.6 \sin(3\pi x^2)$
   - variance function: $g(x) = \exp\{-0.6 + 0.5 \cos(4\pi x)\}$
2. **Linear predictor**:
   - Bayesian penalized semiparametric regression
3. **Loss functions**:
   - quantile regression (MCMC: Kozumi and Kobayashi, 2011; MFVB: Wand et al., 2011)
   - expectile regression (MCMC: Waldmann et al., 2017; Laplace approximation)
   - support vector regression (MCMC: Polson and Scott, 2011; MFVB: Luts and Ormerod, 2014)
   - support vector classification (MCMC: Polson and Scott, 2011; MFVB: Luts and Ormerod, 2014)

Predictive distributions

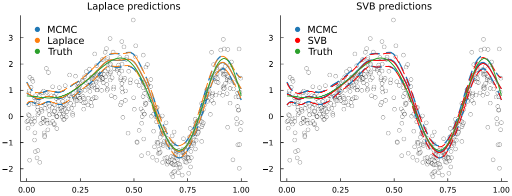
Quantile regression ($\tau = 0.9$)



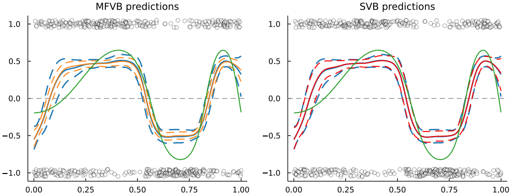
Support vector regression ($\epsilon = 0.01$)



Expectile regression ($\tau = 0.9$)



Support vector classification



Marginal approximations

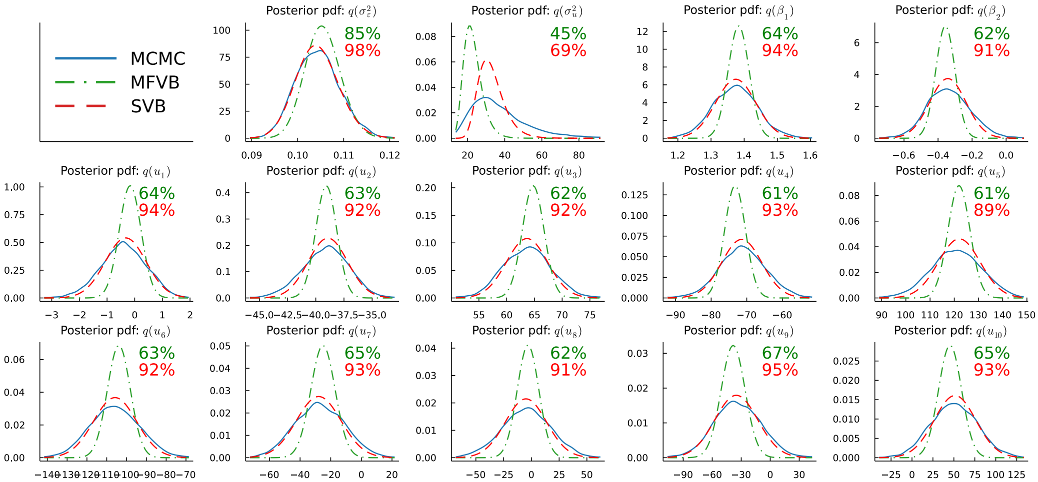


Figure: Marginal posterior density functions for the quantile regression model.

Results

Model	Method	ELBO	Accuracy	Iter.Count	Exe.Time (S.E.)
QReg	MCMC	–	–	10000	4170.000 (32.482) ms
	MFVB	-713.9727	0.8613	47	81.697 (3.463) ms
	SVB	-710.1009	0.9196	27	80.252 (3.761) ms
EReg	MCMC	–	–	10000	3710.000 (33.565) ms
	Laplace	–	0.9669	26	46.734 (3.014) ms
	SVB	153.8963	0.9687	35	62.979 (3.113) ms
SVR	MCMC	–	–	10000	4867.000 (30.114) ms
	MFVB	-637.3713	0.9090	33	61.805 (3.477) ms
	SVB	-634.9547	0.9521	20	75.474 (3.256) ms
SVC	MCMC	–	–	10000	4396.000 (35.974) ms
	MFVB	-537.9229	0.8579	69	118.710 (2.885) ms
	SVB	-536.1095	0.9070	37	129.785 (4.563) ms

Results

Model	Method	ELBO	Accuracy	Iter.Count	Exe.Time (S.E.)
QReg	MCMC	–	–	10000	4170.000 (32.482) ms
	MFVB	-713.9727	0.8613	47	81.697 (3.463) ms
	SVB	-710.1009	0.9196	27	80.252 (3.761) ms
EReg	MCMC	–	–	10000	3710.000 (33.565) ms
	Laplace	–	0.9669	26	46.734 (3.014) ms
	SVB	153.8963	0.9687	35	62.979 (3.113) ms
SVR	MCMC	–	–	10000	4867.000 (30.114) ms
	MFVB	-637.3713	0.9090	33	61.805 (3.477) ms
	SVB	-634.9547	0.9521	20	75.474 (3.256) ms
SVC	MCMC	–	–	10000	4396.000 (35.974) ms
	MFVB	-537.9229	0.8579	69	118.710 (2.885) ms
	SVB	-536.1095	0.9070	37	129.785 (4.563) ms

Results

Model	Method	ELBO	Accuracy	Iter.Count	Exe.Time (S.E.)
QReg	MCMC	–	–	10000	4170.000 (32.482) ms
	MFVB	-713.9727	0.8613	47	81.697 (3.463) ms
	SVB	-710.1009	0.9196	27	80.252 (3.761) ms
EReg	MCMC	–	–	10000	3710.000 (33.565) ms
	Laplace	–	0.9669	26	46.734 (3.014) ms
	SVB	153.8963	0.9687	35	62.979 (3.113) ms
SVR	MCMC	–	–	10000	4867.000 (30.114) ms
	MFVB	-637.3713	0.9090	33	61.805 (3.477) ms
	SVB	-634.9547	0.9521	20	75.474 (3.256) ms
SVC	MCMC	–	–	10000	4396.000 (35.974) ms
	MFVB	-537.9229	0.8579	69	118.710 (2.885) ms
	SVB	-536.1095	0.9070	37	129.785 (4.563) ms

Results

Model	Method	ELBO	Accuracy	Iter.Count	Exe.Time (S.E.)
QReg	MCMC	–	–	10000	4170.000 (32.482) ms
	MFVB	-713.9727	0.8613	47	81.697 (3.463) ms
	SVB	-710.1009	0.9196	27	80.252 (3.761) ms
EReg	MCMC	–	–	10000	3710.000 (33.565) ms
	Laplace	–	0.9669	26	46.734 (3.014) ms
	SVB	153.8963	0.9687	35	62.979 (3.113) ms
SVR	MCMC	–	–	10000	4867.000 (30.114) ms
	MFVB	-637.3713	0.9090	33	61.805 (3.477) ms
	SVB	-634.9547	0.9521	20	75.474 (3.256) ms
SVC	MCMC	–	–	10000	4396.000 (35.974) ms
	MFVB	-537.9229	0.8579	69	118.710 (2.885) ms
	SVB	-536.1095	0.9070	37	129.785 (4.563) ms

Overview

Conclusions

Methodological innovations:

- we introduced a simplified general algorithm for the estimating risk–based mixed models
- existing algorithms are included into our framework (GLMs)
- new algorithms for Quantile, Expectile and SVM models

Empirical evidences:

- improvement over existing data–augmented MFVB approximations
- good–to–excellent performance in posterior approximation

Possible extensions:

- streamlined algorithms for structured prior distributions (cross–random effects, GMRF)
- hierarchical prior for inducing sparsity and shrinkage on the estimates
- application to frequentist mixed models with non-regular likelihood

Conclusions

Methodological innovations:

- we introduced a simplified general algorithm for the estimating risk–based mixed models
- existing algorithms are included into our framework (GLMs)
- new algorithms for Quantile, Expectile and SVM models

Empirical evidences:

- improvement over existing data–augmented MFVB approximations
- good–to–excellent performance in posterior approximation

Possible extensions:

- streamlined algorithms for structured prior distributions (cross–random effects, GMRF)
- hierarchical prior for inducing sparsity and shrinkage on the estimates
- application to frequentist mixed models with non-regular likelihood

Conclusions

Methodological innovations:

- we introduced a simplified general algorithm for the estimating risk–based mixed models
- existing algorithms are included into our framework (GLMs)
- new algorithms for Quantile, Expectile and SVM models

Empirical evidences:

- improvement over existing data–augmented MFVB approximations
- good–to–excellent performance in posterior approximation

Possible extensions:

- streamlined algorithms for structured prior distributions (cross–random effects, GMRF)
- hierarchical prior for inducing sparsity and shrinkage on the estimates
- application to frequentist mixed models with non-regular likelihood

Conclusions

Methodological innovations:

- we introduced a simplified general algorithm for the estimating risk–based mixed models
- existing algorithms are included into our framework (GLMs)
- new algorithms for Quantile, Expectile and SVM models

Empirical evidences:

- improvement over existing data–augmented MFVB approximations
- good–to–excellent performance in posterior approximation

Possible extensions:

- streamlined algorithms for structured prior distributions (cross–random effects, GMRF)
- hierarchical prior for inducing sparsity and shrinkage on the estimates
- application to frequentist mixed models with non-regular likelihood

Overview

Reference

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

References I

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.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D. (2017)
Variational inference: A review for statisticians
Journal of the American Statistical Association, 112(518), 859–877.

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

Kozumi, H., Kobayashi, G. (2011)
Gibbs sampling methods for Bayesian quantile regression
Journal of statistical computation and simulation, 81(11), 1565–1578.

Luts, J., Ormerod, J.T. (2014)
Mean field variational Bayesian inference for support vector machine classification
Computational Statistics and Data Analysis, 73, 163–176.

McLean, M.W., Wand, M.P. (2019)
Variational message passing for elaborate response regression models
Bayesian Analysis, 14(2), 371–398.

References II

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

Polson, N. G., Scott, S. L. (2011)
Data augmentation for support vector machines
Bayesian Analysis, 6(1), 1–23.

Rohde, David and Wand, Matt P. (2016)
Semiparametric mean field variational Bayes: general principles and numerical issues
Journal of Machine Learning Research, 17(1), 5975–6021.

Wand, M.P., Ormerod, J.T., Padoan, S.A., Frührwirth, R. (2011)
Mean field variational Bayes for elaborate distributions
Bayesian Analysis, 6(4), 847–900.

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

Waldmann, E., Sobotka, F., and Kneib, T. (2017)
Bayesian regularisation in geoadditive expectile regression
Statistics and Computing, 268(6), 1539–1553.

Thank you for your attention!