UNIVERSITÀ DEGLI STUDI DI PADOVA

Efficient estimation of mixed effect models via variational message passing

Cristian Castiglione & Mauro Bernardi

Department of Statistical Sciences, University of Padova

Motivation

Classical mean field variational inference relies on conjugate families of distributions, eventually obtained via **data-augmentation** (Wand et al., 2011). However, convenient stochastic representations of the likelihood are not always available (e.g., Poisson and Gamma regression cases).

As an alternative to conjugate approximations and stochastic variational inference, we here propose an efficient marginal **variational message passing** algorithm with (almost) closed-form updates to estimate non-linear mixed models. Remarkably, the proposed approach applies to both **non-conjugate** and **non-regular** models and, moreover, it does not require model-specific transformations of the likelihood.

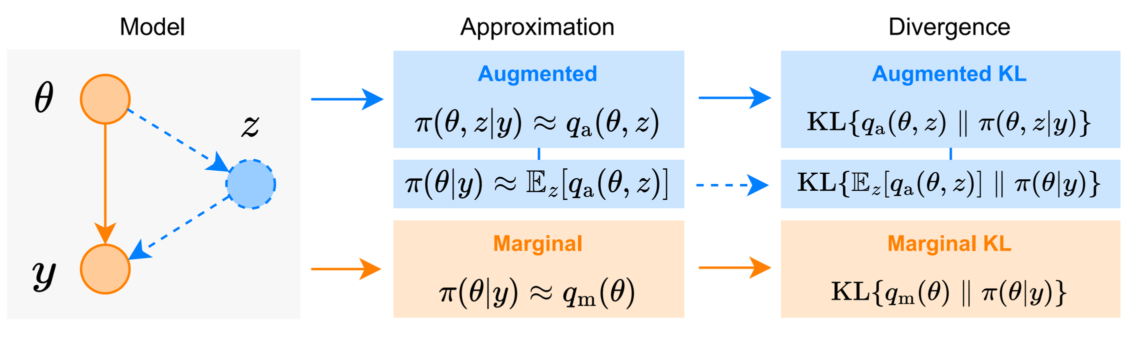


Figure 1. Graphical illustration of **augmented** and **marginal** variational inference.

Model specification

Empirical risk function

We consider generalized Bayesian models (Bissiri et al., 2016) having posterior belief update distribution

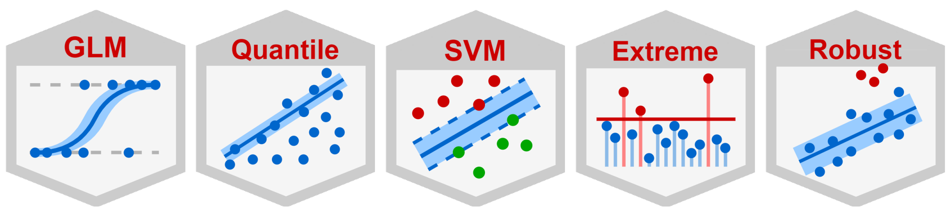
$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\boldsymbol{\theta}) \exp\{-nR(\boldsymbol{\theta}; \mathbf{y})\}, \tag{1}$$

where $R(\boldsymbol{\theta}; \mathbf{y})$ is an **empirical risk function**.

In a GLM fashion, we model the response y_i through a linear predictor η_i, eventually transformed using a bijective link function g. Within this class of models, we define the following empirical risk measure

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

where ψ is a loss function, σ_ε^2 is a dispersion parameter and ϕ is a non-stochastic calibration constant.



Additive specification

We assume an additive model specification

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

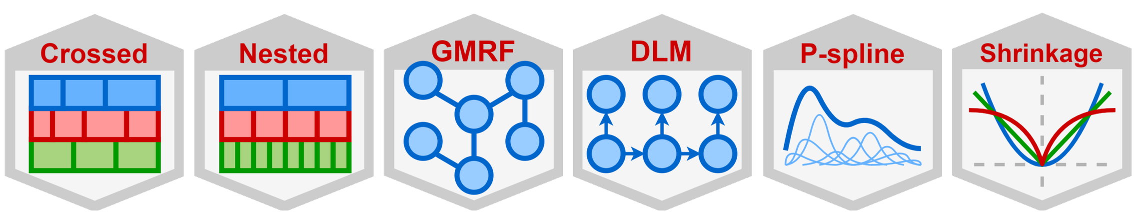
where $\mathbf{C} = (\mathbf{X}, \mathbf{Z}_1, \ldots, \mathbf{Z}_{\mathrm{H}})$ and $\boldsymbol{u} = (\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_{\mathrm{H}}^\top)^\top$. The term $\mathbf{X}\boldsymbol{\beta}$ is the **fixed effect** component, while $\mathbf{Z}_h \boldsymbol{u}_h$ is the h–th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

Prior distributions

We assume the following set of prior distributions:

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

where $\sigma_\beta^2, A_\varepsilon, B_\varepsilon, A_h, B_h > 0$ and $\mathbf{Q}_h \succeq 0$ are known prior hyperparameters.



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.

Wand, M.P., Ormerod, J.T., Padoan, S.A., Frühwirth, 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.

Variational inference

We perform posterior inference via the variational approximation $\pi(\boldsymbol{\theta}|\mathbf{y}) \approx q(\boldsymbol{\theta}) \in \mathcal{Q}$. We then seek the best variational density $q^*(\boldsymbol{\theta}) \in \mathcal{Q}$ by minimizing the **Kullback-Leibler divergence**

$$\mathrm{KL}\{q(\boldsymbol{\theta}) \,\|\, \pi(\boldsymbol{\theta}|\mathbf{y})\} = \int_\Theta q(\boldsymbol{\theta}) \log \left\{ \frac{\pi(\boldsymbol{\theta}|\mathbf{y})}{q(\boldsymbol{\theta})} \right\} \mathrm{d}\boldsymbol{\theta} \tag{5}$$

under the following restrictions on $\mathcal{Q}$:

$$\begin{aligned} &\text{(factorization)} & q(\boldsymbol{\theta}) &= q(\boldsymbol{\beta}, \boldsymbol{u})\, q(\sigma_1^2) \ldots q(\sigma_{\mathrm{H}}^2)\, q(\sigma_\varepsilon^2), \\ &\text{(Gaussianity)} & q(\boldsymbol{\beta}, \boldsymbol{u}) &= q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\mu}, \boldsymbol{\Omega}) \sim \mathrm{N}_{\mathrm{K}}(\boldsymbol{\mu}, \boldsymbol{\Omega}). \end{aligned} \tag{6}$$

Variational message passing

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):

$$\begin{aligned} &\text{(update)} & \boldsymbol{\mu} &\leftarrow \boldsymbol{\mu} - \mathbf{H}^{-1}\mathbf{g}, \quad \boldsymbol{\Omega} \leftarrow -\mathbf{H}^{-1}, \\ &\text{(gradient)} & \mathbf{g} &\leftarrow -\mathbf{R}\boldsymbol{\mu} - \mu_{q(1/\sigma_\varepsilon^2)} \mathbf{C}^\top \boldsymbol{\Psi}_1/\phi, \\ &\text{(Hessian)} & \mathbf{H} &\leftarrow -\mathbf{R} - \mu_{q(1/\sigma_\varepsilon^2)} \mathbf{C}^\top \mathrm{diag}(\boldsymbol{\Psi}_2)\, \mathbf{C}/\phi, \end{aligned} \tag{7}$$

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

$$\Psi_{r,i} = \Psi_r(y_i, \bar{\eta}_i, \bar{\nu}_i^2) = \int_{-\infty}^{+\infty} \psi_r(y_i, x)\, \phi(x; \bar{\eta}_i, \bar{\nu}_i^2)\, \mathrm{d}x, \tag{8}$$

with $r = 0, 1, 2$, $\bar{\eta}_i = \mathbf{c}_i^\top \boldsymbol{\mu}$ and $\bar{\nu}_i^2 = \mathbf{c}_i^\top \boldsymbol{\Omega}\, \mathbf{c}_i$.

Theorem 1. *Let $\psi_0(y, \eta) = \psi(y, g(\eta))$ be a continuous, convex function wrt η with rth order weak derivative $\psi_r(y, \eta) = D_\eta^r \psi_0(y, \eta)$. Then, we have:*

1. *$\Psi_r(y, \eta, \nu)$ is infinitely **differentiable** wrt η and ν;*
2. *$\Psi_0(y, \eta, \nu)$ is jointly **convex** wrt η and ν;*
3. *$\Psi_0(y, \eta, \nu) \geq \psi_0(y, \eta)$ for any η and ν;*
4. *$\Psi_0(y, \eta, \nu) \to \psi_0(y, \eta)$ as $\nu \to 0$.*

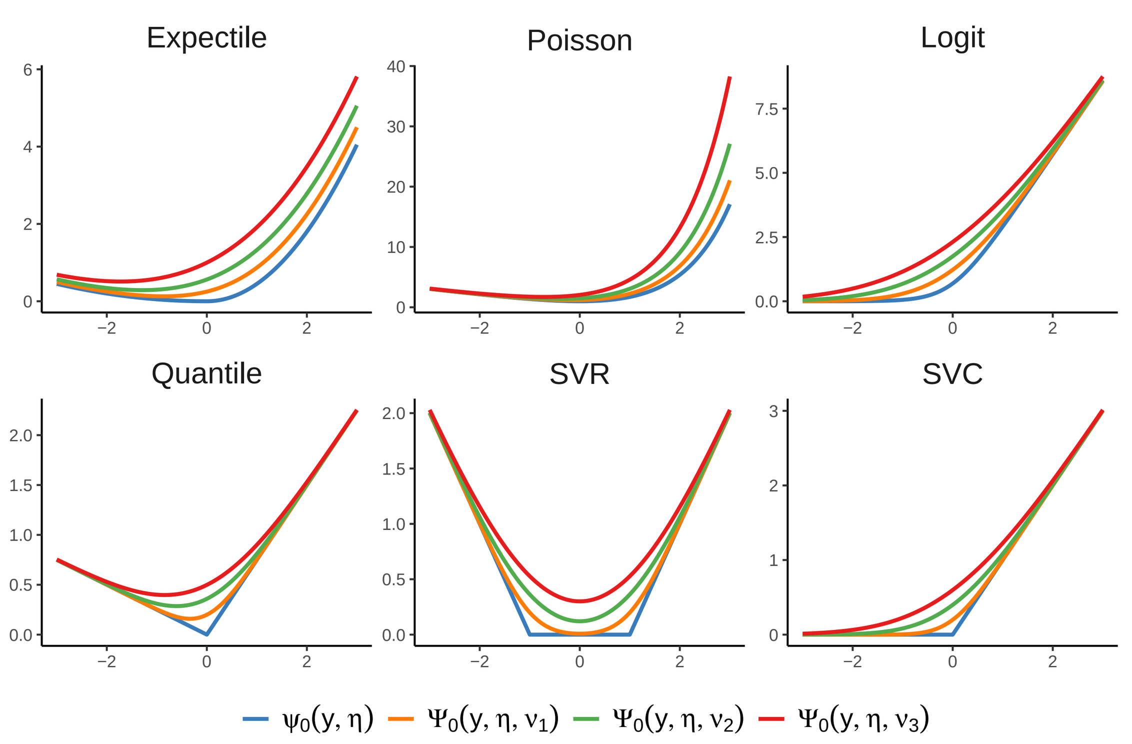


Figure 2. Comparison between $\psi_0(y, \eta)$ and $\Psi_0(y, \eta, \nu)$ for $\nu_1 < \nu_2 < \nu_3$.

Comparison with data-augmentation

Theorem 2. *Let $q_{\mathrm{m}}^*(\boldsymbol{\theta}) = \mathrm{argmin}\ \mathrm{KL}\{q(\boldsymbol{\theta}) \,\|\, \pi(\boldsymbol{\theta}|\mathbf{y})\}$ and $q_{\mathrm{a}}^*(\boldsymbol{\theta}, \boldsymbol{z}) = \mathrm{argmin}\ \mathrm{KL}\{q(\boldsymbol{\theta}, \boldsymbol{z}) \,\|\, \pi(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})\}$, then, under mild regularity conditions, we have*

$$\begin{aligned} \mathrm{KL}\{q_{\mathrm{m}}^*(\boldsymbol{\theta}) \,\|\, \pi(\boldsymbol{\theta}|\mathbf{y})\} &\leq \mathrm{KL}\{\mathbb{E}_{\boldsymbol{z}}[q_{\mathrm{a}}^*(\boldsymbol{\theta}, \boldsymbol{z})] \,\|\, \pi(\boldsymbol{\theta}|\mathbf{y})\} \\ &\leq \mathrm{KL}\{q_{\mathrm{a}}^*(\boldsymbol{\theta}, \boldsymbol{z}) \,\|\, \pi(\boldsymbol{\theta}, \boldsymbol{z}|\mathbf{y})\} \end{aligned} \tag{9}$$

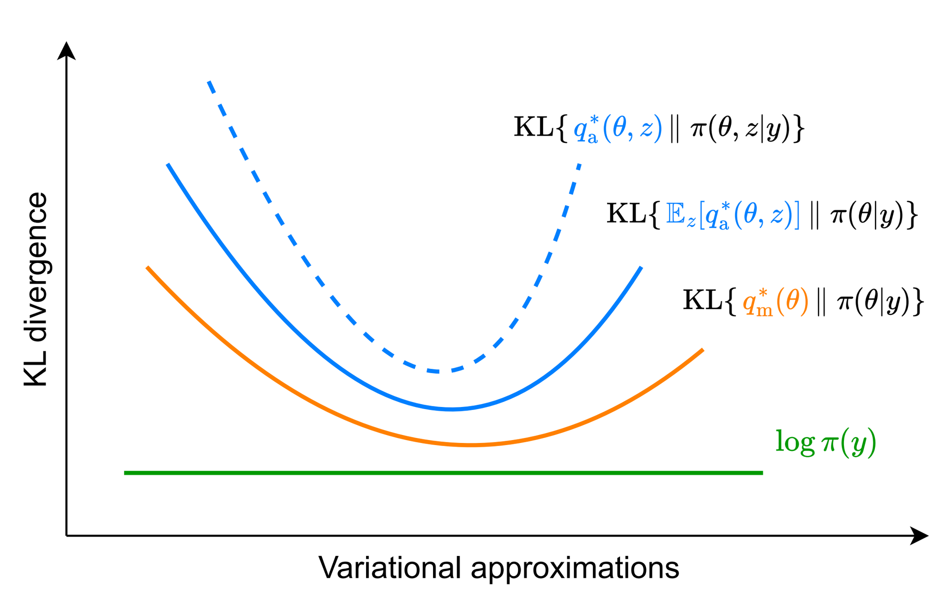


Figure 3. Graphical representation of inequality (9).

Contact information

Cristian Castiglione
Department of Statistical Sciences, University of Padova
cristian.castiglione@unipd.it

Mauro Bernardi
Department of Statistical Sciences, University of Padova
mauro.bernardi@unipd.it

Simulation study

- Setting A: $n \in \{250, 500, 1000, 2500, 5000\}$, $p = 2$, $d = 10$
- Setting B: $n = 500$, $p = 2$, $d \in \{5, 10, 25, 50, 100\}$
- 100 replications for each combination of $\{n, p, d\}$
- 5 prediction models (3 regression, 2 classification)
- Random intercept model: $\eta_{ij} = \beta_0 + \beta_1 x_{ij} + u_j$
- Algorithms for approximate posterior inference:
  - Markov chain Monte Carlo (MCMC)
  - conjugate mean field variational Bayes (MFVB)
  - non-conjugate variational message passing (VMP)

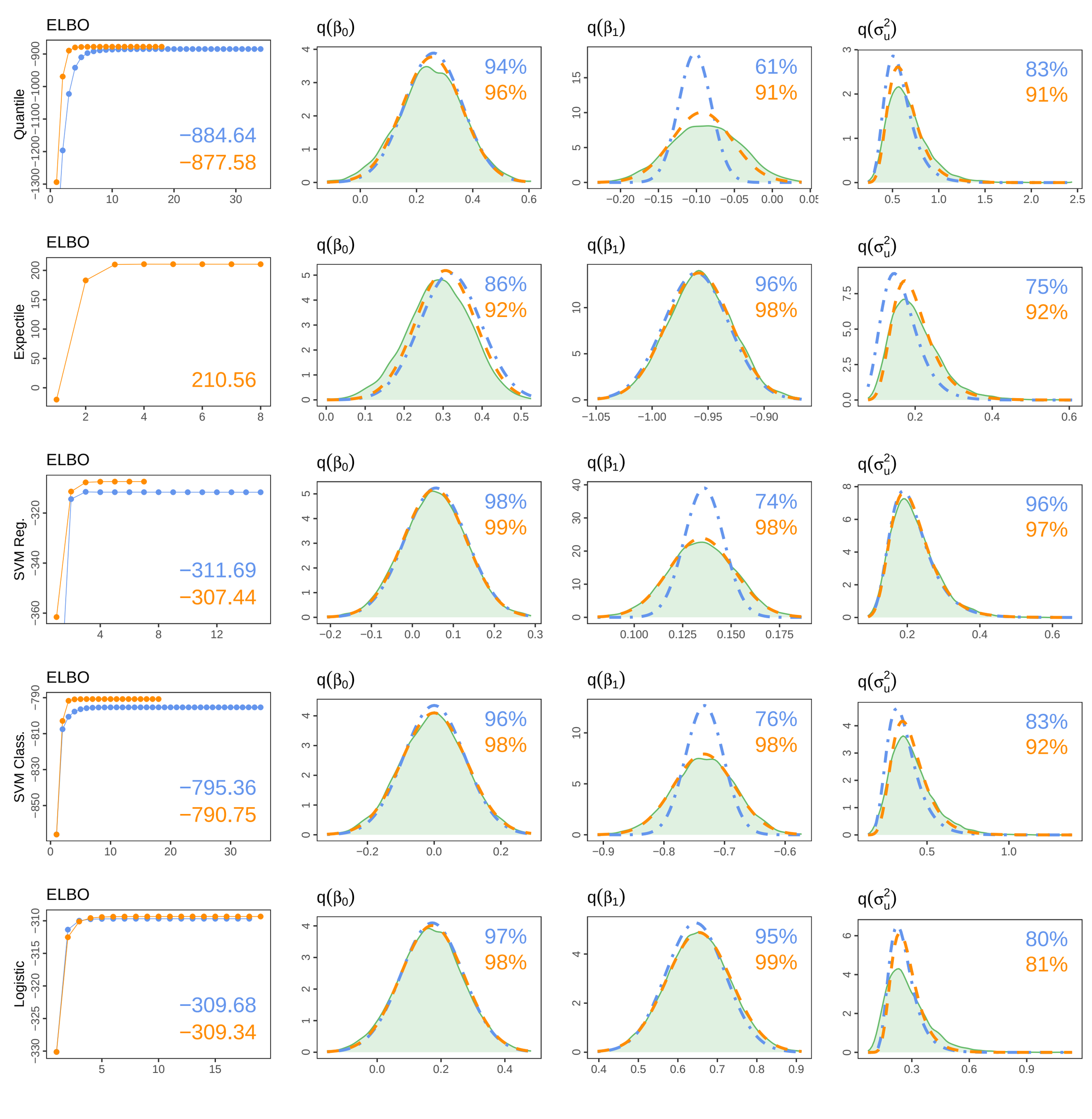


Figure 4. Marginal posterior density functions (setting B, $n = 500$, $p = 2$, $d = 50$).

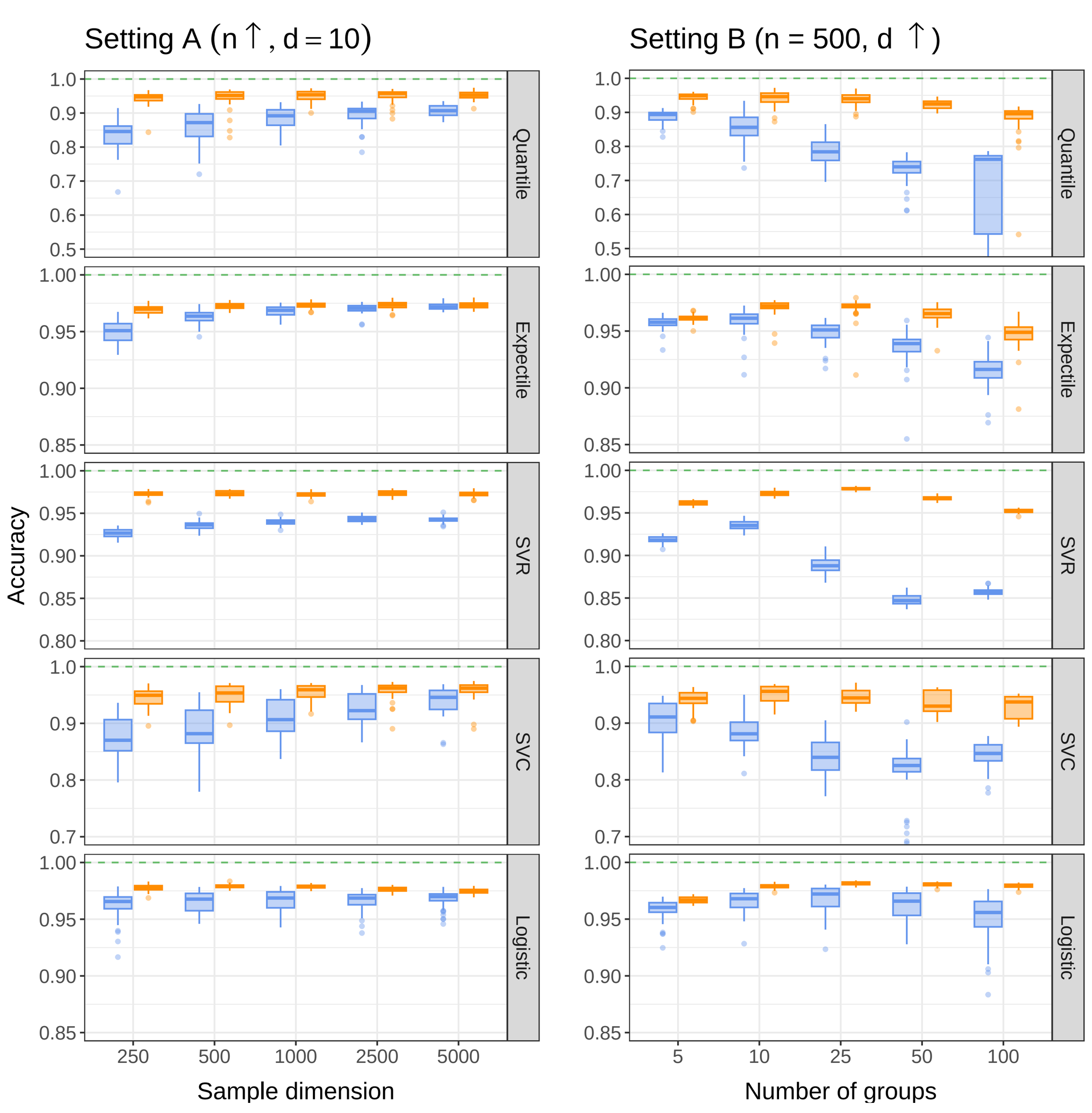


Figure 5. Boxplot of the marginal accuracy scores.

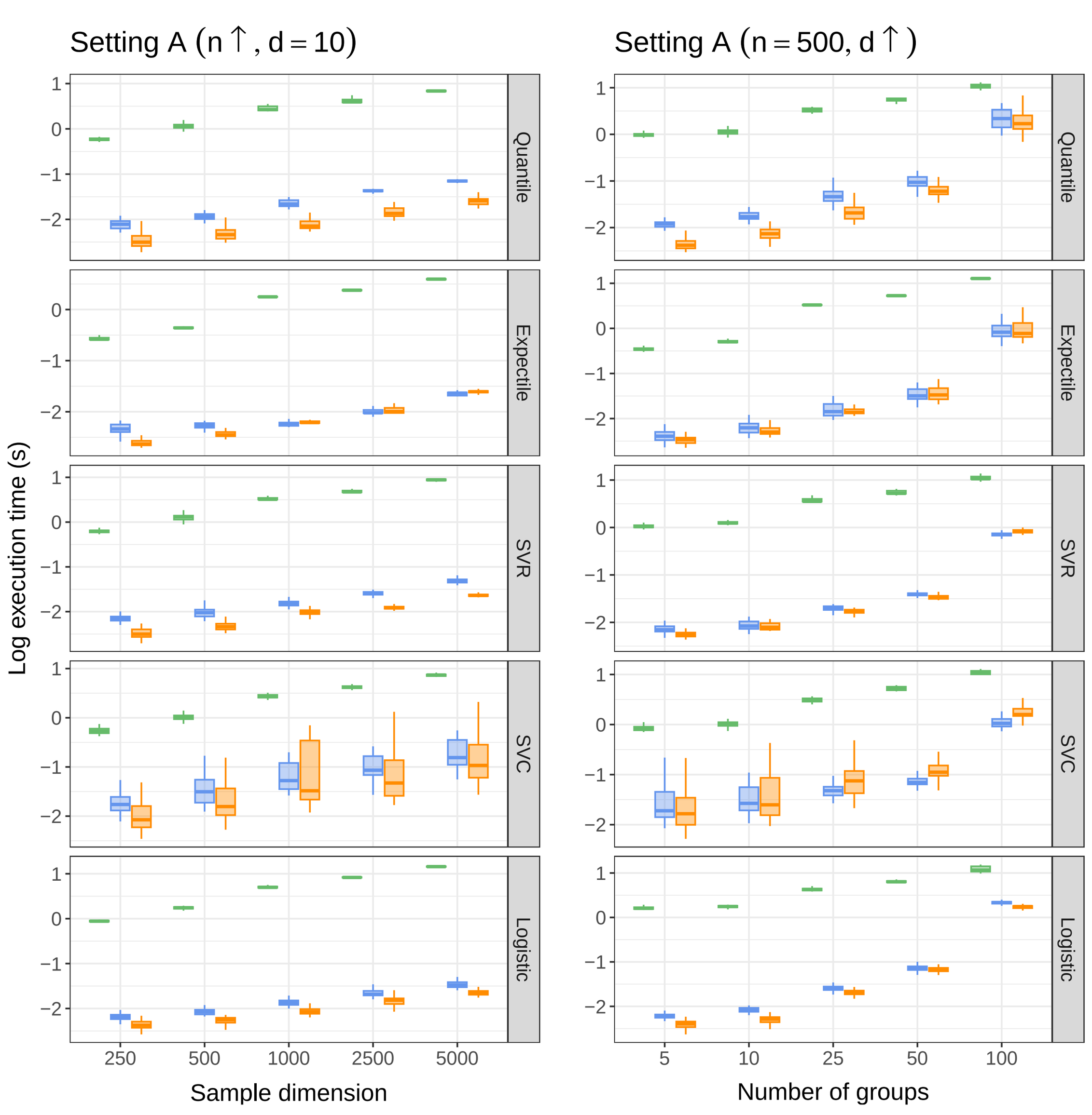


Figure 6. Boxplot of the elapsed execution times.