Summary and discussion of: Covariances, Robustness, and Variational Bayes

Xinyi YU

1 Summary

Variational Bayes (VB) is a method for approximating Bayesian posteriors, which aims to minimize the divergence among a sub-class of tractable distributions from the exact posterior. Mean-field Variational Bayes (MFVB) is one simple and widely-used flavor of VB, where the tractable sub-class posteriors are assumed to be factorizable. Having no general accuracy guarantees, MFVB usually produces accurate enough posterior mean estimates of certain parameters. However, MFVB typically underestimates marginal variances. Also, techniques for assessing Bayesian robustness have not been developed for MFVB before this paper. In short, this paper provides both improved covariance estimates and local robustness measures for MFVB.

The main idea is utilizing the direct correspondence between derivatives and covariance. The proposed linear response variational Bayes (LRVB) covariances by using the sensitivity of MFVB posterior expectations to the priors better approximate the exact posteriors than the posterior covariances. Also, LRVB provides a good estimate of local sensitivities, under the assumption that the posterior means are well estimated by MFVB for all the perturbations of interest. Experiments collaborate that LRVB covariances are superior to MFVB and Laplace posterior covariances and have significant computational advantages over MCMC.

1.1 Bayesian Covariances and Sensitivity

Suppose we are interested in the posterior expectation of some function $g(\theta)$: $\mathbb{E}_{p_\alpha}[g(\theta)]$, where $p_\alpha(\theta)$ is the prior of θ parameterized by $\alpha \in \mathcal{A} \subset \mathbb{R}^D$. Since it is impractical to calculate $\mathbb{E}_{p_\alpha}[g(\theta)]$ for all $\alpha \in \mathcal{A}$, we can examine the changes of $\mathbb{E}_{p_\alpha}[g(\theta)]$ over values near $\alpha_0 \in \mathcal{A}$. The local sensitivity at α_0 is defined as $\boldsymbol{S}_{\alpha_0} = \frac{d\mathbb{E}_{p_\alpha}[g(\theta)]}{d\alpha}\Big|_{\alpha_0}$. $\boldsymbol{S}_{\alpha_0}$ measures sensitivity to hyperparameters with in a small region near $\alpha = \alpha_0$ and it can be estimated by a posterior covariance which is one contribution of this paper. The important general result relating sensitivity and covariance is (Theorem 1):

$$\frac{d\mathbb{E}_{p_\alpha}[g(\theta)]}{d\alpha}\Big|_{\alpha_0}^{\mathrm{T}} = \mathrm{Cov}_{p_0}\left(g(\theta), \frac{\partial \rho(\theta, \alpha)}{\partial \alpha}\Big|_{\alpha_0}\right) \tag{1}$$

where $\rho(\theta, \alpha)$ is the log perturbation of posterior $p_0(\theta)$ and $p_\alpha(\theta) \propto p_0(\theta)\exp(\rho(\theta, \alpha))$. This theorem allows us to calculate $\boldsymbol{S}_{\alpha_0}$ as a covariance:

$$\boldsymbol{S}_{\alpha_0} = \mathrm{Cov}_{p_0}\left(g(\theta), \frac{\partial \rho(\theta, \alpha)}{\partial \alpha}\Big|_{\alpha_0}\right) \tag{2}$$

An estimate of the covariance can be easily caculated by MCMC.

1.2 Variational Bayesian Covariances and Sensitivity

The variational approximation $q_\alpha(\theta) \in \mathcal{Q}$ to $p_\alpha(\theta)$ is defined by the minimizer of KL divergence:

$$q_\alpha(\theta) \coloneqq argmin_{q\in\mathcal{Q}}\{KL(q(\theta;\eta)||p_\alpha(\theta))\} \tag{3}$$

Here we focus on the mean-field family:

$$\mathcal{Q}_{mf} \coloneqq \left\{q(\theta) : q(\theta) = \prod_k q(\theta_k;\eta_k)\right\} \tag{4}$$

Theorem 1 provides a way to estimate the sensitivity of exact posterior means to generic perturbations. Theorem 2 derives a VB analogue of Theorem 1: under some regularity conditions,

$$\frac{d\mathbb{E}_{q_\alpha}[g(\theta)]}{d\alpha^{\mathrm{T}}}\Big|_{\alpha_0} = \boldsymbol{g}_\eta \boldsymbol{H}_{\eta\eta}^{-1}\boldsymbol{f}_{\alpha\eta}^{\mathrm{T}} \tag{5}$$

where

$$\boldsymbol{H}_{\eta\eta} \coloneqq \frac{\partial^2 KL(q(\theta;\eta)||p_0(\theta)}{\partial\eta\partial\eta^{\mathrm{T}}}\Big|_{\eta_0^*} \quad \boldsymbol{f}_{\alpha\eta} \coloneqq \frac{\partial^2 \mathbb{E}_{q(\theta;\eta)}[\rho(\theta,\alpha)]}{\partial\alpha\partial\eta^{\mathrm{T}}}\Big|_{\eta_0^*,\alpha_0} \quad \boldsymbol{g}_\eta \coloneqq \frac{\partial \mathbb{E}_{q(\theta;\eta)}[g(\theta)]}{\partial\eta^{\mathrm{T}}}\Big|_{\eta_0^*} \tag{6}$$

When Condition 1:

$$\mathbb{E}_{q_0}[g(\theta)] \approx \mathbb{E}_{q_\alpha}[g(\theta)] \quad and \quad \frac{d\mathbb{E}_{q_\alpha}[g(\theta)}{d\alpha^{\mathrm{T}}}\Big|_{\alpha_0} \approx \frac{d\mathbb{E}_{p_\alpha}[g(\theta)}{d\alpha^{\mathrm{T}}}\Big|_{\alpha_0} \tag{7}$$

holds, variational approximations and their sensitivity measures will be useful meaning that we can approximate $\frac{d\mathbb{E}_{p_\alpha}[g(\theta)}{d\alpha^{\mathrm{T}}}\Big|_{\alpha_0}$ using Theorem 2.

Note that we do not expect Condition 1 to hold for all $g(\theta)$ but only for some parameters of interest. For example, a mean-field approximation to a multivariate normal posterior can have a bad KL divergence from p_α, but Condition 1 holds exactly for the location parameters. The accuracy of approximated posterior mean allows us to better estimate posterior variance using the local sensitivity.

1.3 LRVB and Local Prior Sensitivity for MFVB

By Theorem 1 and Theorem 2, we can calculate improved covariance estimates and prior sensitivity measures for MFVB.

First, for covariances of variational Bayes, it is known that MFVB typically underestimates posterior covariances. This is determined by the nature of mean field approximating

family as the covariances between sub-components of θ are zero. However, the relation between covariance and sensitivity and the accuracy of MFVB posterior means together hints us to estimate covariances using sensitivity.

Specifically, by taking linear log perturbation $\rho(\theta, \alpha) = \alpha^{\mathrm{T}} g(\theta)$ and $\alpha_0 = 0$, under Condition 1, we have:

$$\mathrm{Cov}_{q_0}^{LR}(g(\theta)) = \frac{d\mathbb{E}_{q_\alpha}[g(\theta)]}{d\alpha^{\mathrm{T}}}\Big|_{\alpha=0} \approx \frac{d\mathbb{E}_{p_\alpha}[g(\theta)]}{d\alpha^{\mathrm{T}}}\Big|_{\alpha=0} = \mathrm{Cov}_{p_0}(g(\theta)) \tag{8}$$

where $\mathrm{Cov}_{q_0}^{LR} \coloneqq \boldsymbol{g}_\eta \boldsymbol{H}_{\eta\eta}^{-1} \boldsymbol{g}_\eta^{\mathrm{T}}$ is the linear response variational Bayes (LRVB) approximation. The intuition that LRVB covariances can better estimate exact posterior covariances than MFVB covariances lies in the fact that the optimal value of every component of $\mathbb{E}_{q_\alpha}[g(\theta)]$ will be affected by other components via the log perturbation $\alpha^{\mathrm{T}} g(\theta)$. Thus, $\mathrm{Cov}_{q_0}^{LR}(g(\theta)$ are able to estimate covariances between elements of $g(\theta)$ on the contrary to MFVB covariances.

$\mathrm{Cov}_{q_0}^{LR}(g(\theta)$ will improve $\mathrm{Cov}_{q_0}(g(\theta)$ when q_α fail to estimate the second moment of $g(\theta)$ accurately. In other words, when the variational approximation for the first and second moments of $g(\theta)$ are both adequate, there will be:

$$\mathrm{Cov}_{q_0}^{LR}(g(\theta)) \approx \mathrm{Cov}_{q_0}(g(\theta)) \approx \mathrm{Cov}_{p_0}(g(\theta)) \tag{9}$$

An illustrative example showing the improvement of $\mathrm{Cov}_{q_0}^{LR}(g(\theta)$ is the bivariate normal distribution. The exact posterior of a bivariate normal $p_0(\theta) = \mathcal{N}(\theta|\mu, \boldsymbol{\Sigma})$, $\boldsymbol{\Lambda} \coloneqq \boldsymbol{\Sigma}^{-1}$ is approximated by $q(\theta)$ in the mean-field family:

$$\mathcal{Q}_{mf} = \{q(\theta) : q(\theta) = q(\theta_1)q(\theta_2)\}$$

The optimal MFVB approximation q_α is given by:

$$q_0(\theta_1) = \mathcal{N}(\theta_1|\mu_1, \boldsymbol{\Lambda}_{11}^{-1}), \quad q_0(\theta_2) = \mathcal{N}(\theta_2|\mu_2, \boldsymbol{\Lambda}_{22}^{-1})$$

The posterior mean of θ_1 and θ_2 are exactly estimated by MFVB:

$$\mathbb{E}_{q_0}[\theta_1] = \mu_1 = \mathbb{E}_{p_0}[\theta_1], \quad \mathbb{E}_{q_0}[\theta_2] = \mu_2 = \mathbb{E}_{p_0}[\theta_2]$$

While as long as $\boldsymbol{\Sigma}_{12} \neq 0$, the second moments and covariance are underestimated:

$$\mathbb{E}_{q_\alpha}[\theta_1^2] = \mu_1^2 + \boldsymbol{\Gamma}_{11}^{-1} < \mu_1^2 + \boldsymbol{\Sigma}_{11} = \mathrm{E}_{p_\alpha}[\theta_1^2]$$

$$\mathbb{E}_{q_\alpha}[\theta_2^2] = \mu_2^2 + \boldsymbol{\Gamma}_{22}^{-1} < \mu_2^2 + \boldsymbol{\Sigma}_{22} = \mathrm{E}_{p_\alpha}[\theta_2^2]$$

and

$$\mathrm{Cov}_{q_0}(\theta_1, \theta_2) = 0 \neq \boldsymbol{\Sigma}_{12} = \mathrm{Cov}_{p_0}(\theta_1, \theta_2)$$

The exactness of MFVB means is utilized by LRVB to provide exact posterior covariances. Consider the log perturbation $\rho(\theta, \alpha) = \theta_1\alpha_1 + \theta_2\alpha_2$. Since the perturbation is actually of the exponential family form, the perturbed posterior distribution remains bivariate normal. Thus, we still have

$$\mathbb{E}_{q_\alpha}[\theta_1] = \mathbb{E}_{p_\alpha}[\theta_1], \quad \mathbb{E}_{q_\alpha}[\theta_2] = \mathbb{E}_{p_\alpha}[\theta_2]$$

Since the $\mathbb{E}_{q_\alpha}[\theta]$ are exact for all α, the derivative $\frac{d\mathbb{E}_{q_\alpha}[g(\theta)]}{d\alpha^{\mathrm{T}}}\Big|_{\alpha=0}$ is also exact, giving the exact posterior covariances. Therefore, in this example, Condition 1 holds with equality for $g(\theta) = \theta$.

Besides MFVB, LRVB can also be used for Laplace posterior approximation if MAP estimator is viewed as a variational approximation.

The Laplace approximation to $p_0(\theta)$ takes the form of multivariate normal distribution whose mean is the MAP estimate and covariance is inverse of Fisher information matrix at MAP estimate:

$$\hat{\theta}_{Lap} \coloneqq \operatorname*{argmax}_{\theta} p_0(\theta) \tag{10}$$

$$\mathrm{Cov}^{Lap}_{q_{Lap}}(\theta) \coloneqq \boldsymbol{H}^{-1}_{Lap}, \quad \boldsymbol{H}_{Lap} \coloneqq -\frac{\partial^2 p_0(\theta)}{\partial\theta\partial\theta^{\mathrm{T}}}\Big|_{\hat{\theta}} \tag{11}$$

If MAP estimator is viewed as a special variational approximation where the variational family $\mathcal{Q}_{Lap}$ is defined to to include point masses at θ_0 with constant entropy. Thus, $\mathrm{Cov}_{q_{Lap}} = 0$, an apparent underestimation of $\mathrm{Cov}_{q_0}(\theta)$. Consider the linear response covariance:

$$\boldsymbol{H}_{\eta\eta} = -\frac{\partial^2 KL(q(\theta;\theta_0)||p_0(\theta)}{\partial\theta_0\partial\theta_0^{\mathrm{T}}}\Big|_{\theta_0^*}$$

So $\mathrm{Cov}^{LR}_{q_0}(\theta)$ is equivalent to the Laplace approximation covariance $\mathrm{Cov}^{Lap}_{q_{Lap}}(\theta)$.

Although $\mathrm{Cov}^{Lap}_{q_{Lap}}(\theta) = \mathrm{Cov}^{LR}_{q_0}(\theta)$ for $\mathcal{Q}_{Lap}$, since variational Bayes usually use a $\mathcal{Q}$ more expressive than $\mathcal{Q}_{Lap}$ which only uses the mode of posterior distribution thus more accurate mean estimates, we can expect LRVB to provide more accurate covariance estimates for general $\mathcal{Q}$.

Second, for prior sensitivity of MFVB, we use the defined local sensitivity of $\mathbb{E}[g(\theta)]$:

$$\boldsymbol{S}^q_{\alpha_0} \coloneqq \frac{\partial\mathbb{E}_{q_\alpha}[g(\theta)]}{d\alpha}\Big|_{\alpha_0}$$

to approximate $\boldsymbol{S}_{\alpha_0}$. Specifically, under Condition 1, we have $\boldsymbol{S}^q_{\alpha_0} \approx \boldsymbol{S}_{\alpha_0}$.

Furthermore, by Theorem 2, the estimated prior sensitivity can be calculated as:

$$\boldsymbol{S}^q_{\alpha_0} = \boldsymbol{g}_\eta \boldsymbol{H}^{-1}_{\eta\eta}\boldsymbol{f}_{\alpha\eta}, \quad \boldsymbol{f}_{\alpha\eta} = \frac{\partial}{\partial\eta^{\mathrm{T}}}\mathbb{E}_{q(\theta;\eta)}\left[\frac{\partial\rho(\theta,\alpha)}{\partial\alpha}\Big|_{\alpha_0}\right]\Big|_{\eta_0^*} \tag{12}$$

If $p(\theta|\alpha)$ is in the exponential family, i.e. log perturbation $\rho(\theta,\alpha)$ is linear, then it turns out that $\boldsymbol{f}_{\alpha\eta} = \boldsymbol{g}_\eta$. Then,

$$\boldsymbol{S}^q_{\alpha_0} = \mathrm{Cov}^{LR}_{q_0}(\theta)$$

Similarly,

$$\boldsymbol{S}_{\alpha_0} = \mathrm{Cov}_{p_0}(\theta)$$

Thus, if Condition 1 holds, $\mathrm{Cov}^{LR}_{q_0} \approx \mathrm{Cov}_{p_0}(\theta)$, and $\boldsymbol{S}^q_{\alpha_0} \approx \boldsymbol{S}_{\alpha_0}$ accordingly.

2 Result and Discussion

This paper provides extensive simulations and real data experiments to illustrate the speed and effectiveness of linear response methods. I reproduce the expository examples and they are adequate for understanding the main idea of this paper. Code can be found in the github repository xinyiyu/MATH5472-2020.

The following simulations are all based on Gaussian mixture, either univariate or bivariate, up to 3 components. Formally, the Gaussian mixture distribution takes the form:

$$p_0(\theta) = \sum_{k=1}^{K_z} \pi_k \mathcal{N}(\theta; m_k, \mathbf{\Sigma}_k)$$

We estimate the mean and variance of the first component, namely $g(\theta) = \theta_1$, using MCMC, MFVB, LRVB and Laplace approximation. The aim is to test the effectiveness of linear response method by comparing it with MFVB and Laplace approximation.

The mean-field family is defined as:

$$\mathcal{Q}_{mf} = \left\{ q(\theta) : q(\theta) = \prod_{k=1}^{K} \mathcal{N}(\theta_k; \mu_k, \sigma_k^2) \right\}$$

We minimize the KL divergence between $q_0\theta \in \mathcal{Q}$ and $p_0(\theta)$ to get the MFVB means and covariances. Note that $\mathbb{E}_{q(\theta;\eta)}[\log p(\theta)]$ is intractable, we turn to the approximated KL divergence by drawing i.i.d standard multivariate normal samples $\xi_m, m = 1, ..., M$ and using the "re-parameterization trick" $\theta_m \coloneqq \sigma \circ \xi_m + \mu$, which writes as:

$$KL_{approx}(q(\theta;\eta)||p_0(\theta)) \coloneqq -\frac{1}{M}\sum_{m=1}^{M} \log p_0(\theta_m) - \sum_{k=1}^{K} \log \sigma_k$$

The same draws ξ_m are used for both optimization and calculation of $\boldsymbol{H}_{\eta\eta}$. We use formula (5) for calculating LRVB covariances. Laplace approximation mean is given by the mode of the Gaussian mixture and its covariance is the inverse of the Hessian matrix at the mode. Here I take $M = 1000$ for fast reproducing the results.

2.1 A Univariate Skewed Distribution

First, consider a two-component skewed univariate Gaussian mixture distribution. As shown in the left penal of Figure 1, Laplace approximation only use information at the posterior mode thus failing to take the mass of the other component into account. Consequently, Laplace approximation inaccurately estimates both mean and variance. In contrast, though MFVB approximates the Gaussian mixture by normal distribution as well, it is accurate for posterior mean of θ_1.

Since local sensitivity of the expectation of θ_1 to α is the variance of θ_1, the slopes of the exact distribution, MFVB and Laplace are $\mathrm{Cov}_{p_0}(\theta_1)$, $\mathrm{Cov}_{q_0}^{LR}(\theta_1)$ and $\mathrm{Cov}_{q_{Lap}}^{Lap}(\theta_1)$, respectively. For the univariate skewed distribution, we can see from the left penal of Figure 2 that the sensitivities of exact distribution and MFVB are close while the slope of Laplace

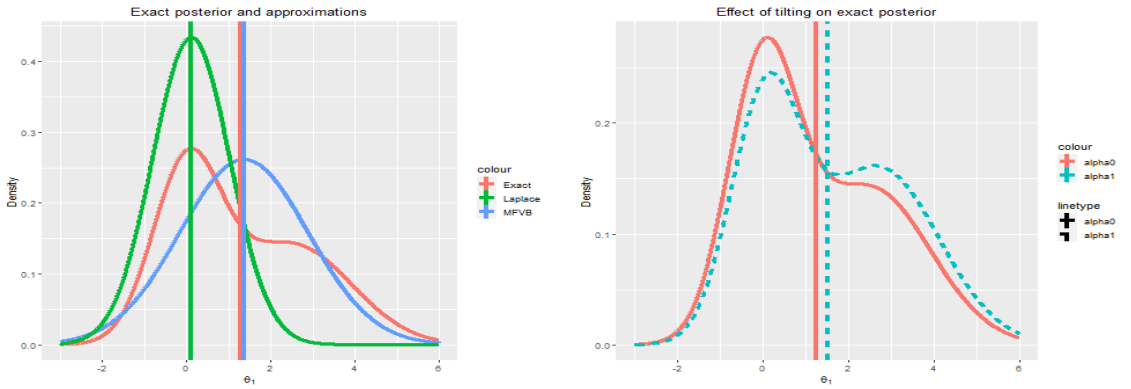
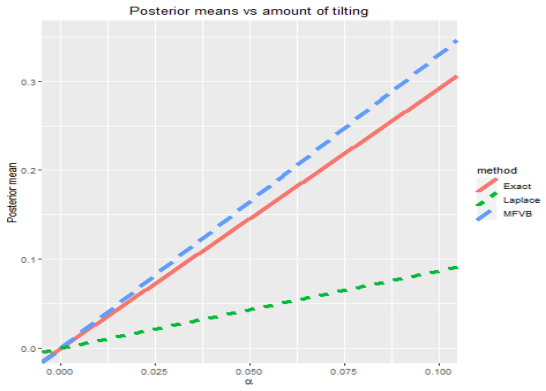


Figure 1: A univariate skewed distribution.



Metric	Exact	LRVB	MFVB	Laplace
mean	1.287		1.360	0.111
var	3.051	3.284	2.350	0.849

Figure 2: Effect of tilting on a univariate skew distribution.

line is lower. Intuitively, this phenomenon is reasonable as the mode is less sensitive to perturbation compared with mean for univariate skewed distribution.

This example mainly demonstrates the difference between MFVB and Laplace approximation while the covariances estimated by MFVB and LRVB do not differ by much and they are both close to the true covariance.

2.2 A Univariate Over-dispersed Distribution

Second, consider a three-component univariate over-dispersed Gaussian mixture distribution. Analogy to last example, MFVB is able to estimate both mean and variance well, and its ability to capture the change of mean after tilting leads to the reasonably well LRVB covariance. In other words, Condition 1 holds in this case.

By comparison, Laplace approximation only focuses on the mean of the dominant component. As a result, Laplace mean is more reluctant to change than true mean. And its lower sensitivity is in accordance with its underestimated variance.

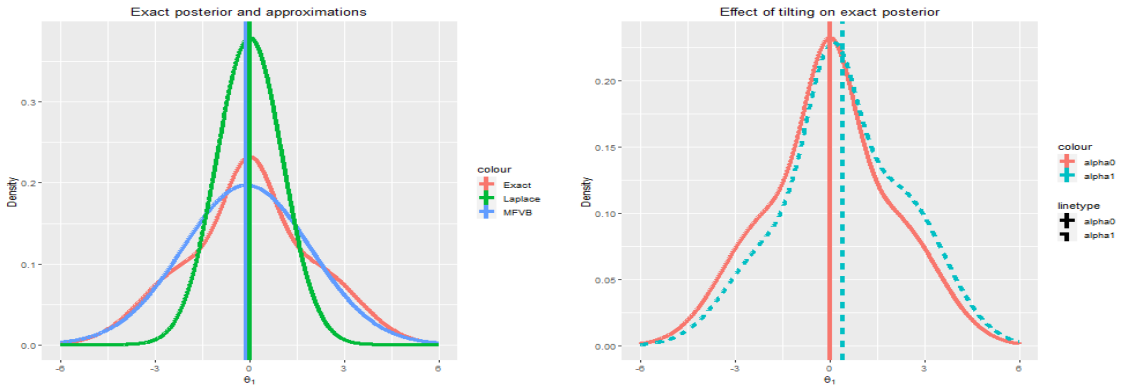
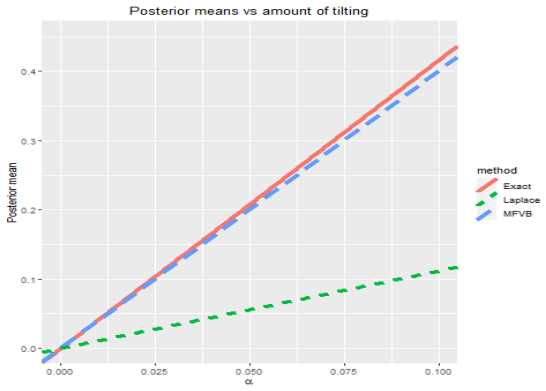


Figure 3: A univariate over-dispersed distribution.



Metric	Exact	LRVB	MFVB	Laplace
mean	0.006		-0.100	-0.000
var	4.410	4.048	4.128	1.107

Figure 4: Effect of tilting on a univariate over-dispersed distribution.

2.3 A Bivariate Over-dispersed Distribution

In the last two cases, MFVB both performs well and LRVB covariance has not show its superiority over MFVB covariance due to the univariate setting. Now consider a bivariate three-component over-dispersed Gaussian mixture distribution where the factorizable assumption of MFVB matters. As expected, MFVB gives a substantially lower covariance estimate and it is even worse than Laplace covariance which anyhow takes the correlation between the two variables θ_1 and θ_2 into account.

The remarkable discovery of this example is the significant improvement of LRVB variance as shown in Figure 6. LRVB provides a marginal variance much closer to the true margianl variance than MFVB variance and Lapalce variance by taking advantage of the fact that MFVB gives good mean estimates.

However, note that despite the improvement, LRVB marginal variance is still away from the true marginal variance. Intuitively, LRVB should give a more accurate variance since MFVB can estimate mean accurately, since the bivariate over-dispersed distribution is only an extension of the univariate over-dispersed distribution in the last subsection. I hypothesize that the reason lies in the linear log perturbation $\rho(\theta) = \alpha\theta_1$ which perturbs on the marginal distribution rather than the joint distribution. Thus, if the perturbation

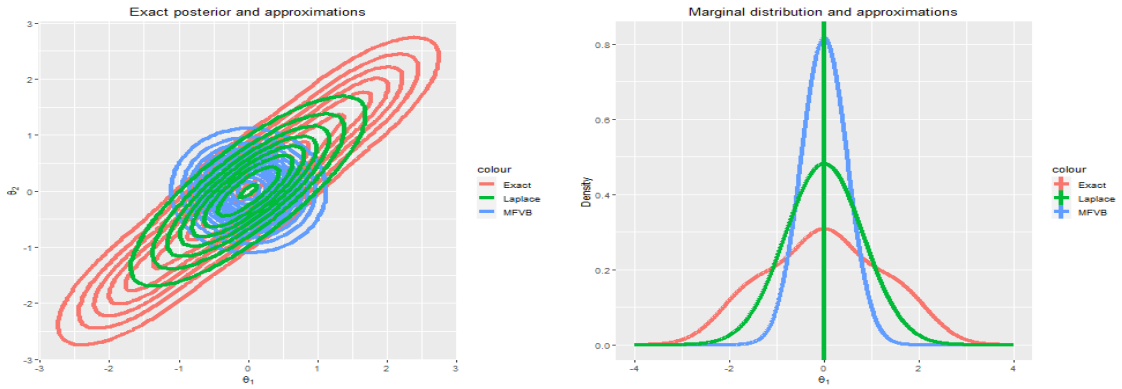
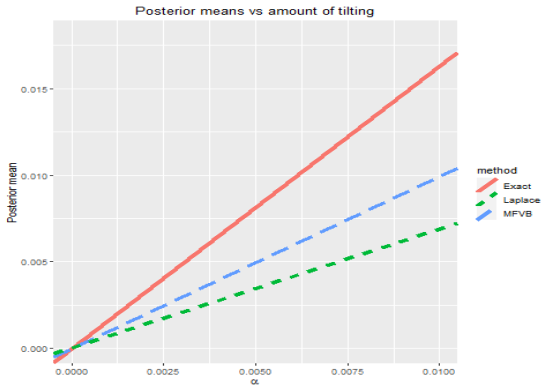


Figure 5: A bivariate over-dispersed distribution.



Metric	Exact	LRVB	MFVB	Laplace
mean	-0.024		0.001	0.000
var	1.629	0.987	0.240	0.684

Figure 6: Effect of tilting on a bivariate over-dispersed distribution.

is imposed on the joint distribution and then extract the marginal variance of θ from the LRVB covariance, we may get a better estimate of $\mathrm{Var}(\theta_1)$.

3 Conclusion

This paper provides some intuitions about the properties of mean-field variational Bayes, a different view of Laplace approximation as a special VB method and comparison between MFVB and Laplace approximation. The main message this paper wants to pass is that by linking sensitivity with covariance, one can well estimate both posterior covariances and local prior sensitivity under the condition that the posterior means are close to true means for all α near α_0. This condition generally is not hard for MFVB to achieve. In other words, we are able to improve the usually underestimated covariances of MFVB by its usually well-estimated means. Additionally, the proposed LRVB uses linear log perturbation in place of generic log perturbation. The linear log perturbation corresponds to priors in the exponential family thus actually covering a wide range of priors.