Supplement to "Yes, but Did It Work?: Evaluating Variational Inference"

Yuling Yao* Aki Vehtari † Daniel Simpson ‡ Andrew Gelman§

June 7, 2018

A Sketch of Proofs

A.1 Proof to Proposition 1: Marginal $\hat{k}$ in PSIS diagnostic

Proposition 1. *For any two distributions p and q with support Θ and the margin index i, if there is a number $\alpha > 1$ satisfying $E_q \left(p(\theta)/q(\theta)\right)^\alpha < \infty$, then $E_q \left(p(\theta_i)/q(\theta_i)\right)^\alpha < \infty$.*

Proof. Without loss of generality, we could assume $\Theta = \mathbb{R}^K$, otherwise a smooth transformation is conducted.

For any $1 \leq i \leq K$, $p(\theta_{-i}|\theta_i)$ and $q(\theta_{-i}|\theta_i)$ define the conditional distribution of $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_K) \in \mathbb{R}^{K-1}$ given θ_i under the true posterior p and the approximation q separately.

For any given index $\alpha > 1$, Jensen inequality yields

$$\int_{\mathbb{R}^{K-1}} \left(\frac{p(\theta_{-i}|\theta_i)}{q(\theta_{-i}|\theta_i)}\right)^\alpha q(\theta_{-i}|\theta_i) \geq \left(\int_{\mathbb{R}^{K-1}} \frac{p(\theta_{-i}|\theta_i)}{q(\theta_{-i}|\theta_i)} q(\theta_{-i}|\theta_i)\right)^\alpha = 1$$

Hence

$$\begin{aligned}
\int_{\mathbb{R}^K} \left(\frac{p(\theta)}{q(\theta)}\right)^\alpha q(\theta) d\theta &= \int_{\mathbb{R}^{K-1}} \int_{\mathbb{R}} \left(\frac{p(\theta_i) p(\theta_{-i}|\theta_i)}{q(\theta_i) q(\theta_{-i}|\theta_i)}\right)^\alpha q(\theta_i) q(\theta_{-i}|\theta_i) d\theta_i d\theta_{-i} \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}^{K-1}} \left(\frac{p(\theta_{-i}|\theta_i)}{q(\theta_{-i}|\theta_i)}\right)^\alpha q(\theta_{-i}|\theta_i) d\theta_{-i}\right) \left(\frac{p(\theta_i)}{q(\theta_i)}\right)^\alpha q(\theta_i) d\theta_i \\
&\geq \int_{\mathbb{R}} \left(\frac{p(\theta_i)}{q(\theta_i)}\right)^\alpha q(\theta_i) d\theta_i
\end{aligned}$$

□

A.2 Proof to Proposition 2: Symmetry in VSBC-Test

Proposition 2. *For a one-dimensional parameter θ that is of interest, Suppose in addition we have:*
(i) the VI approximation q is symmetric;
(ii) the true posterior $p(\theta|y)$ is symmetric.
If the VI estimation q is unbiased, i.e.,

$$\mathrm{E}_{\theta \sim q(\theta|y)}\, \theta = \mathrm{E}_{\theta \sim p(\theta|y)}\, \theta, \forall y$$

Then the distribution of VSBC p-value is symmetric.
If the VI estimation is positively/negatively biased, then the distribution of VSBC p-value is right/left skewed.

In the proposition we write $q(\theta|y)$ to emphasize that the VI approximation also depends on the observed data.

Proof. First, as the same logic in Cook et al. (2006), when $\theta^{(0)}$ is sampled from its prior $p(\theta)$ and simulated data y sampled from likelihood $p(y|\theta^{(0)})$, $(y, \theta^{(0)})$ represents a sample from the joint

*Department of Statistics, Columbia University, USA.
†Helsinki Institute for Information Technology, Department of Computer Science, Aalto University, Finland.
‡Department of Statistical Sciences, University of Toronto, Canada.
§Department of Statistics and Political Science, Columbia University,USA.

distribution $p(y, \theta)$ and therefore $\theta^{(0)}$ can be viewed as a draw from $p(\theta|y)$, the true posterior distribution of θ with y being observed.

We denote $q(\theta^{(0)})$ as the VSBC p-value of the sample $\theta^{(0)}$. Also denote $Q_x(f)$ as the $x-$quantile ($x \in [0, 1]$) of any distribution f. To prove the result, we need to show

$$1 - \Pr(q(\theta^{(0)}) < x) = \Pr(q(\theta^{(0)}) < 1 - x), \forall x \in [0, 1],$$

$$\begin{aligned} \text{LHS} &= \Pr\left(q(\theta^{(0)}) > x\right) \\ &= \Pr\left(\theta^{(0)} > Q_x\left(q(\theta|y)\right)\right). \end{aligned}$$

$$\begin{aligned} \text{RHS} = \Pr\left(\theta^{(0)} < Q_{1-x}\left(q(\theta|y)\right)\right) &= \Pr\left(\theta^{(0)} < 2E_{q(\theta|y)}\theta - Q_x\left(q(\theta|y)\right)\right) \\ &= \Pr\left(\theta^{(0)} < 2E_{p(\theta|y)}\theta - Q_x\left(q(\theta|y)\right)\right) \\ &= \Pr\left(\theta^{(0)} > Q_x\left(q(\theta|y)\right)\right) \\ &= \text{LHS} \end{aligned}$$

The first equation above uses the symmetry of $q(\theta|y)$, the second equation comes from the the unbiasedness condition. The third is the result of the symmetry of $p(\theta|y)$.

If the VI estimation is positively biased, $\mathrm{E}_{\theta\sim q(\theta|y)}\,\theta > \mathrm{E}_{\theta\sim p(\theta|y)}\,\theta, \forall y$, then we change the second equality sign into a less-than sign. □

B Details of Simulation Examples

In this section, we give more detailed description of the simulation examples in the manuscript. We use Stan (Stan Development Team, 2017) to implement both automatic differentiation variational inference (ADVI) and Markov chain Monte Carlo (MCMC) sampling. We implement Pareto smoothing through R package "loo" (Vehtari et al., 2018). We also provide all the source code in `https://github.com/yao-yl/Evaluating-Variational-Inference`.

B.1 Linear and Logistic Regressions

In Section 4.1, We start with a Bayesian linear regression $y \sim \mathrm{N}(X\beta, \sigma^2)$ without intercept. The prior is set as $\{\beta_i\}_{i=1}^d \sim \mathrm{N}(0, 1), \sigma \sim \text{gamma}(0.5, 0.5)$. We fix sample size $n = 10000$ and number of regressors $d = 100$. Figure I displays the Stan code.

We find ADVI can be sensitive to the stopping time. Part of the reason is the objective function itself is evaluated through Monte Carlo samples, producing large uncertainty. In the current version of Stan, ADVI computes the running average and running median of the relative ELBO norm changes. Should either number fall below a threshold `tol_rel_obj`, with the default value to be 0.01, the algorithm is considered converged.

In Figure 1 of the main paper, we run VSBC test on ADVI approximation. ADVI is deliberately tuned in a conservative way. The convergence tolerance is set as `tol_rel_obj`$=10^{-4}$ and the learning rate is $\eta = 0.05$. The predictor $X_{10^5\times 10^2}$ is fixed in all replications and is generated independently from $\mathrm{N}(0, 1)$. To avoid multiple-comparison problem, we pre-register the first and second coefficients β_1 β_2 and $\log\sigma$ before the test. The VSBC diagnostic is based on $M = 1000$ replications.

In Figure 2 we independently generate each coordinate of β from $\mathrm{N}(0, 1)$ and set a relatively large variance $\sigma = 2$. The predictor X is generated independently from $\mathrm{N}(0, 1)$ and y is sampled from the normal likelihood. We vary the threshold `tol_rel_obj` from 0.01 to 10^{-5} and show the trajectory of $\hat{k}$ diagnostics. The $\hat{k}$ estimation, IS and PSIS adjustment are all calculated from $S = 5 \times 10^4$ posterior samples. We ignore the ADVI posterior sampling time. The actual running time is based on a laptop experiment result (2.5 GHz processor, 8 cores).The exact sampling time

```
,
data {
int <lower=0> n;      //number of observations, we fix n=10000 in the simulation;
int <lower=0> d;       //number of predictor variables,  fix d=100;
matrix [n,d] x ;        // predictors;
vector [n] y;          // outcome;
}
parameters {
vector [d] b;     // linear regression coefficient;
real <lower=0> sigma;     //linear regression std;
}
model {
y ~ normal(x * b, sigma);
b ~ normal(0,1);  // prior for  regression coefficient;
sigma ~ gamma(0.5,0.5);     // prior for  regression std.
}

```

Figure I: *Stan code for linear regressions*

is based on the No-U-Turn Sampler (NUTS, Hoffman and Gelman 2014) in Stan with 4 chains and 3000 iterations in each chain. We also calculate the root mean square errors (RMSE) of all parameters $||E_p[(\beta,\sigma)] - E_q[(\beta,\sigma)]||_{L^2}$, where (β,σ) represents the combined vector of all β and σ. To account for the uncertainty, $\hat{k}$, running time, and RMSE takes the average of 50 repeated simulations.

```
,
  data {
  int <lower=0> n;      //number of observations;
  int <lower=0> d;       //number of predictor variables;
  matrix [n,d] x ;        // predictors; we vary its correlation during simulations
    .
  int<lower=0,upper=1> y[n]; // binary outcome;
  }
  parameters {
  vector[d] beta;
  }
  model {
  y ~ bernoulli_logit(x*beta);
  }

```

Figure II: *Stan code for logistic regressions*

Figure 3 and 4 in the main paper is a simulation result of a logistic regression

$$Y \sim \text{Bernoulli}\left(\text{logit}^{-1}(\beta X)\right)$$

with a flat prior on β. We vary the correlation in design matrix by generating X from $\text{N}(0, (1-\rho)I_{d\times d}+\rho 1_{d\times d})$, where $1_{d\times d}$ represents the d by d matrix with all elements to be 1. In this experiment we fix a small number $n = 100$ and $d = 2$ since the main focus is parameter correlations. We compare $\hat{k}$ with the log predictive density, which is calculated from 100 independent test data. The true posterior is from NUTS in Stan with 4 chains and 3000 iterations each chain. The $\hat{k}$ estimation, IS and PSIS adjustment are calculated from 10^5 posterior samples. To account for the uncertainty, $\hat{k}$, log predictive density, and RMSE are the average of 50 repeated experiments.

B.2 Eight-School Model

The *eight-school model* is named after Gelman et al. (2013, section 5.5). The study was performed for the Educational Testing Service to analyze the effects of a special coaching program on students'

School Index j	Estimated Treatment Effect y_i	Standard Deviation of Effect Estimate σ_j
1	28	15
2	8	10
3	-3	16
4	7	11
5	-1	9
6	1	11
7	8	10
8	12	18

Table I: *School-level observed effects of special preparation on SAT-V scores in eight randomized experiments. Estimates are based on separate analyses for the eight experiments.*

SAT-V (Scholastic Aptitude Test Verbal) scores in each of eight high schools. The outcome variable in each study was the score of a standardized multiple choice test. Each school i separately analyzed the treatment effect and reported the mean y_i and standard deviation of the treatment effect estimation σ_i, as summarized in Table I.

There was no prior reason to believe that any of the eight programs was more effective than any other or that some were more similar in effect to each other than to any other. Hence, we view them as independent experiments and apply a Bayesian hierarchical normal model:

$$y_j|\theta_j \sim \mathrm{N}(\theta_j, \sigma_j), \quad \theta_j \sim \mathrm{N}(\mu, \tau), \quad 1 \leq j \leq 8,$$
$$\mu \sim \mathrm{N}(0, 5), \qquad \tau \sim \text{half–Cauchy}(0, 5).$$

where θ_j represents the underlying treatment effect in school j, while μ and τ are the hyper-parameters that are shared across all schools.

,

```
data {
int<lower=0> J;       // number of schools
real y[J];          // estimated treatment
real<lower=0> sigma[J];      // std  of estimated effect
}

parameters {
real theta[J];    // treatment effect in school j
real mu;    // hyper-parameter of mean
real<lower=0> tau;   // hyper-parameter of sdv
}
model {
theta ~ normal(mu, tau);
y ~ normal(theta, sigma);
mu ~ normal(0, 5);    // a non-informative prior
tau ~ cauchy(0, 5);
}

```

Figure III: *Stan code for centered parametrization in the eight-school model. It leads to strong dependency between tau and theta.*

There are two parametrization forms being discussed: *centered parameterization* and *non-centered parameterization*. Listing III and IV give two Stan codes separately. The true posterior is from NUTS in Stan with 4 chains and 3000 iterations each chain. The $\hat{k}$ estimation and PSIS adjustment are calculated from $S = 10^5$ posterior samples. The marginal $\hat{k}$ is calculated by using the NUTS density, which is typically unavailable for more complicated problems in practice.

The VSBC test in Figure 6 is based on $M = 1000$ replications and we pre-register the first treatment effect θ_1 and group-level standard error $\log \tau$ before the test.

As discussed in Section 3.2, VSBC assesses the average calibration of the point estimation.

```
,
data {
int<lower=0> J;       // number of schools
real y[J];          // estimated treatment
real<lower=0> sigma[J];      // std  of estimated effect
}
parameters {
vector[J] theta_trans;    // transformation of theta
real mu;     // hyper-parameter of mean
real<lower=0> tau;  // hyper-parameter of sd
}
transformed parameters{
vector[J] theta;      // original theta
theta=theta_trans*tau+mu;
}
model {
theta_trans ~normal (0,1);
y ~ normal(theta, sigma);
mu ~ normal(0, 5);     // a non-informative prior
tau ~ cauchy(0, 5);
}

```

Figure IV: *Stan code for non-centered parametrization in the eight-school model. It extracts the dependency between tau and theta.*

Hence the result depends on the choice of prior. For example, if we instead set the prior to be

$$\mu \sim \mathrm{N}(0, 50), \qquad \tau \sim \mathrm{N}^{+}(0, 25),$$

which is essentially flat in the region of interesting part of the likelihood and more in agreement with the prior knowledge, then the result of VSBC test change to Figure V. Again, the skewness of p-values verifies VI estimation of θ_1 is in average unbiased while τ is biased in both centered and non-centered parametrization.

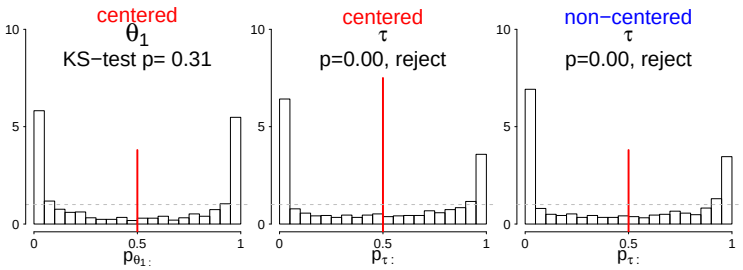

Figure V: *The VSBC diagnostic of the eight-school example under a non-informative prior* $\mu \sim N(0, 50)$, $\tau \sim \mathrm{N}^{+}(0, 25)$. *The skewness of p-values verifies VI estimation of* θ_1 *is in average unbiased while* τ *is biased in both centered and non-centered parametrization.*

B.3 Cancer Classification Using Horseshoe Priors

In Section 4.3 of the main paper we replicate the cancer classification under regularized horseshoe prior as first introduced by Piironen and Vehtari (2017).

The Leukemia microarray cancer classification dataset [1]. It contains $n = 72$ observations and $d = 7129$ features $X_{n\times d}$. X is standardized before any further process. The outcome $y_{1:n}$ is binary,

[1]The Leukemia classification dataset can be downloaded from http://featureselectiocn.asu.edu/datasets.php

so we can fit a logistic regression

$$y_i|\beta \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\sum_{j=1}^{d}\beta_j x_{ij} + \beta_0\right)\right).$$

There are far more predictors than observations, so we expect only a few of predictors to be related and therefore have a regression coefficient distinguishable from zero. Further, many predictors are correlated, making it necessary to have a regularization.

To this end, we apply the *regularized horseshoe prior*, which is a generalization of *horseshoe prior*.

$$\begin{aligned}\beta_j|\tau,\lambda,c &\sim \text{N}(0,\tau^2\tilde{\lambda}_j^2), \qquad & c^2 &\sim \text{Inv–Gamma}(2,8),\\ \lambda_j &\sim \text{Half–Cauchy}(0,1), & \tau|\tau_0 &\sim \text{Half–Cauchy}(0,\tau_0).\end{aligned}$$

The scale of the global shrinkage is set according to the recommendation $\tau_0 = 2\left(n^{1/2}(d-1)\right)^{-1}$ There is no reason to shrink intercept so we put $\beta_0 \sim \text{N}(0,10)$. The Stan code is summarized in Figure VI.

,

```
data {
int<lower=0> n;           // number of observations
int<lower=0> d;             // number of predictors
int<lower=0,upper=1> y[n];  // outputs
matrix[n,d] x;              // inputs
real<lower=0> scale_icept;  // prior std for the intercept
real<lower=0> scale_global; // scale for the half-t prior for tau
real<lower=0> slab_scale;
real<lower=0> slab_df;
}
parameters {
real beta0; // intercept
vector[d] z; // auxiliary parameter
real<lower=0> tau;      // global shrinkage parameter
vector<lower=0>[d] lambda;  // local shrinkage parameter
real<lower=0> caux; // auxiliary
}
transformed parameters {
real<lower=0> c;
vector[d] beta;       // regression coefficients
vector[n] f;        // latent values
vector<lower=0>[d] lambda_tilde;
c = slab_scale * sqrt(caux);
lambda_tilde = sqrt( c^2 * square(lambda) ./ (c^2 + tau^2* square(lambda)) );
beta = z .* lambda_tilde*tau;
f = beta0 + x*beta;
}
model {
z ~ normal(0,1);
lambda ~ cauchy(0,1);
tau ~ cauchy(0, scale_global);
caux ~ inv_gamma(0.5*slab_df, 0.5*slab_df);
beta0 ~ normal(0,scale_icept);
y ~ bernoulli_logit(f);
}

```

Figure VI: *Stan code for regularized horseshoe logistic regression.*

We first run NUTS in Stan with 4 chains and 3000 iterations each chain. We manually pick β_{1834}, the coefficient that has the largest posterior mean. The posterior distribution of it is bi-modal with one spike at 0.

ADVI is implemented using the same parametrization and we decrease the learning rate η to 0.1 and the threshold `tol_rel_obj` to 0.001

The $\hat{k}$ estimation is based on $S = 10^4$ posterior samples. Since $\hat{k}$ is extremely large, indicating VI is far away from the true posterior and no adjustment will work, we do not further conduct PSIS.

In the VSBC test, we pre-register that pre-chosen coefficient β_{1834}, $\log \lambda_{1834}$ and global shrinkage $\log \tau$ before the test. The VSBC diagnostic is based on M=1000 replications.

References

Samantha R Cook, Andrew Gelman, and Donald B Rubin. Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, 15(3): 675–692, 2006.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Matthew D Hoffman and Andrew Gelman. The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

Juho Piironen and Aki Vehtari. Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051, 2017.

Stan Development Team. *Stan modeling language users guide and reference manual*. http://mc-stan.org, 2017. Version 2.17.

Aki Vehtari, Jonah Gabry, Yuling Yao, and Andrew Gelman. loo: Efficient leave-one-out cross-validation and waic for bayesian models, 2018. URL https://CRAN.R-project.org/package=loo. R package version 2.0.0.