ICES REPORT 14-21

August 2014

Model Misspecification and Plausibility

by

Kathryn Farrell and J. Tinsley Odena

The Institute for Computational Engineering and Sciences
The University of Texas at Austin
Austin, Texas 78712

Reference: Kathryn Farrell and J. Tinsley Odena, "Model Misspecification and Plausibility," ICES REPORT 14-21, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, August 2014.

Model Misspecification and Plausibility

Kathryn Farrell[a,*], J. Tinsley Oden[a]

[a]*Institute for Computational Engineering and Sciences, The University of Texas at Austin*

Abstract

We address in this communication relationships between Bayesian posterior model plausibilities and the Kullback-Leibler divergence between the so-called "true" observational distribution and that predictable by a parametric model.

Keywords: Bayesian model calibration, Kullback-Leibler divergence, model plausibility

1. Introduction

Generally, we consider a set $\mathbb{M}(\mathcal{Y})$ of probability measures μ, defined on a (metric) space $\mathcal{Y}$ of physical observations. A target quantity of interest, $Q : \mathbb{M}(\mathcal{Y}) \to \mathbb{R}$ is selected, *e.g.* $Q(\mu) = \mu[X \geq a]$, X being a random variable and a a threshold value. We seek a particular measure, μ^*, from which the "true" value of the quantity of interest, $Q(\mu^*)$, is computed. For observational data, we draw independent and identically distributed (*i.i.d.*) samples $\mathbf{y} = \{y_1, y_2, \ldots, y_n\} \in \mathcal{Y}^n$ from μ^*. We wish to predict $Q(\mu^*)$ using a parametric model $\mathcal{P} : \Theta \to \mathbb{M}(\mathcal{Y})$, where Θ is the space of parameters. If there exists a parameter $\theta^* \in \Theta$ such that $\mathcal{P}(\theta^*) = \mu^*$, the model $\mathcal{P}$ is said to be *well-specified*. Otherwise, if $\mu^* \notin \mathcal{P}(\Theta)$, the model is *misspecified*.

Bayesian frameworks provide a general setting for the analysis of such models and their predictive capabilities in the presence of uncertainties. For any fixed *i.i.d.* observational data $\mathbf{y}$, let us consider a set $\mathcal{M}$ of m parametric models, each with its own parameter space, $\mathcal{M} = \{\mathcal{P}_1(\boldsymbol{\theta}_1),$

*Corresponding author

Email addresses: kfarrell@ices.utexas.edu (Kathryn Farrell), oden@ices.utexas.edu (J. Tinsley Oden)

$\mathcal{P}_2(\boldsymbol{\theta}_2)$, ..., $\mathcal{P}_m(\boldsymbol{\theta}_m)\}$, $\boldsymbol{\theta}_i \in \Theta_i$. Prior information regarding parameters $\boldsymbol{\theta}_i$ may be collected and characterized by the prior probability density function $\pi(\boldsymbol{\theta}_i|\mathcal{P}_i, \mathcal{M})$. This information may be updated via Bayes's Rule,

$$\pi(\boldsymbol{\theta}_i|\mathbf{y}, \mathcal{P}_i, \mathcal{M}) = \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{P}_i, \mathcal{M})\pi(\boldsymbol{\theta}_i|\mathcal{P}_i, \mathcal{M})}{\pi(\mathbf{y}|\mathcal{P}_i, \mathcal{M})}, \tag{1}$$

where the likelihood probability density, $\pi(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{P}_i, \mathcal{M})$, captures how well the parameters are able to reproduce the given data $\mathbf{y}$ and the posterior density $\pi(\boldsymbol{\theta}_i|\mathbf{y}, \mathcal{P}_i, \mathcal{M})$ contains the updated information about the parameters. The denominator in (1), called the model evidence, is the marginalization of the product of the likelihood and prior densities over the parameters,

$$\pi(\mathbf{y}|\mathcal{P}_i, \mathcal{M}) = \int_{\Theta_i} \pi(\mathbf{y}|\boldsymbol{\theta}_i, \mathcal{P}_i, \mathcal{M})\pi(\boldsymbol{\theta}_i|\mathcal{P}_i, \mathcal{M})\, d\boldsymbol{\theta}_i. \tag{2}$$

An important observation presents itself at this point: the evidence (2) can be viewed as the likelihood in a higher form of Bayes's Rule,

$$\pi(\mathcal{P}_i|\mathbf{y}, \mathcal{M}) = \frac{\pi(\mathbf{y}|\mathcal{P}_i, \mathcal{M})\pi(\mathcal{P}_i|\mathcal{M})}{\pi(\mathbf{y}|\mathcal{M})}. \tag{3}$$

The left-hand side of (3) is the *posterior model plausibility* and may be denoted $\rho_i = \pi(\mathcal{P}_i|\mathbf{y}, \mathcal{M})$ for simplicity. Taking the place of the prior density is the prior model plausibility $\pi(\mathcal{P}_i|\mathcal{M})$, and $\pi(\mathbf{y}|\mathcal{M})$ is a normalizing factor such that $\sum_{i=}^{m} \rho_i = 1$. Among the models in $\mathcal{M}$, the model with plausibility ρ_i closest to unity is deemed the most plausible model.

2. Misspecified Models

Let us now suppose that the model $\mathcal{P}$ (or $\mathcal{P}_i$) is misspecified, *i.e.* $\mu^* \notin \mathcal{P}(\Theta)$. Suppose μ^* is absolutely continuous with respect to the Lebesgue measure and that $g(\mathbf{y})$ is the probability density associated with μ^*. Then the best approximation to g in $\mathcal{P}(\Theta)$ is the model with the parameter

$$\boldsymbol{\theta}^\dagger = \operatorname*{argmin}_{\Theta} D_{KL}(g\|\pi(\cdot|\boldsymbol{\theta}, \mathcal{P}, \mathcal{M})), \tag{4}$$

where $D_{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence, or relative entropy, between two densities: if $p(y)$ and $q(y)$ are two probability densities,

$$D_{KL}(p\|q) = \int_{\mathcal{Y}} p(y) \log \frac{p(y)}{q(y)}\, dy. \tag{5}$$

The parameter $\boldsymbol{\theta}^\dagger$ yields a probability measure, $\mu^\dagger = \mathcal{P}(\boldsymbol{\theta}^\dagger)$, that is as close as possible to μ^* in the D_{KL} pseudo-measure..

It is easily shown that $\boldsymbol{\theta}^\dagger$ is the maximum likelihood estimate, *i.e.* it maximizes the expected value of the log-likelihood relative to the true density g:

$$\begin{aligned}
\boldsymbol{\theta}^\dagger &= \underset{\Theta}{\operatorname{argmin}} \left[\int_{\mathcal{Y}^n} g(\mathbf{y}) \log g(\mathbf{y})\, d\mathbf{y} - \int_{\mathcal{Y}^n} g(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta})\, d\mathbf{y} \right] \\
&= \underset{\Theta}{\operatorname{argmin}} \left[- \int_{\mathcal{Y}^n} g(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta})\, d\mathbf{y} \right] \\
&= \underset{\Theta}{\operatorname{argmax}} \int_{\mathcal{Y}^n} g(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta})\, d\mathbf{y} \\
&= \underset{\Theta}{\operatorname{argmax}}\, \mathbb{E}_g \left[\log \pi(\mathbf{y}|\boldsymbol{\theta}) \right],
\end{aligned} \tag{6}$$

where the negative self-entropy $\int g \log g\, d\mathbf{y}$ was eliminated since it does not depend on $\boldsymbol{\theta}$ and therefore does not affect the optimization.

It should be noted that the parameter $\boldsymbol{\theta}^\dagger$ is of fundamental significance in the theories of the asymptotic behavior of parameter estimates of both Bayesian and frequentist statistics. Under suitable smoothness assumptions, as more data becomes available, the posterior density characterized by $\pi(\boldsymbol{\theta}^n|\{y_1, y_2, \ldots, y_n\}, \mathcal{P}, \mathcal{M})$ through Bayes's Rule converges in $\mathcal{L}_1$ (in probability) to a normal distribution, $\mathcal{N}(\boldsymbol{\theta}^\dagger, \mathbf{V}(\boldsymbol{\theta}^\dagger))$, centered at $\boldsymbol{\theta}^\dagger$ with covariance matrix given by [2, 3]

$$\mathbf{V}(\boldsymbol{\theta}^\dagger) = -\mathbb{E}_g \left[\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \pi(\mathbf{y}|\boldsymbol{\theta}^\dagger, \mathcal{P}, \mathcal{M}) \right]. \tag{7}$$

This result is referred to as the Bernstein-von Mises Theorem for misspecified models (see *e.g.* [1, 3, 4]). Of course, if $\mathcal{P}$ is well-specified, $\boldsymbol{\theta}^\dagger = \boldsymbol{\theta}^*$ and the posterior produced by Bayes's Rule converges in probability to a normal distribution centered at $\boldsymbol{\theta}^*$ as the number of data samples increases and $V(\boldsymbol{\theta}^\dagger)$ becomes the Fisher information matrix at $\boldsymbol{\theta}^*$ [2].

3. Plausibility-D_{KL} Theory

Let us now suppose that we have two misspecified models, $\mathcal{P}_1$ and $\mathcal{P}_2$. We may compare these models in the Bayesian setting through the concept of model plausibility: if $\mathcal{P}_1$ is more plausible than $\mathcal{P}_2$, $\rho_1 > \rho_2$. In the maximum likelihood setting, the model that yields a probability measure closer to μ^* is considered the "better" model. That is, if

$$D_{KL}(g\|\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})) < D_{KL}(g\|\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})), \tag{8}$$

it can be said that model $\mathcal{P}_1$ is "better" than model $\mathcal{P}_2$. The theorems presented here define the relationship between these two notions of model comparison.

However, Bayesian and frequentist methods fundamentally differ in the way they view the model parameters. Bayesian methods consider parameters to be stochastic, characterized by probability density functions, while frequentist approaches seek a single, deterministic parameter value. To bridge this gap in methodology, we note that considering parameters as deterministic vectors, for example $\boldsymbol{\theta}_0$, is akin to assigning them delta functions as their posterior probability distributions, which result from delta prior distributions. In this case, the evidence (2) becomes

$$\pi(\mathbf{y}|\mathcal{P}_i, \mathcal{M}) = \int_\Theta \pi(\mathbf{y}|\boldsymbol{\theta}, \mathcal{P}_i, \mathcal{M})\delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\, d\boldsymbol{\theta} = \pi(\mathbf{y}|\boldsymbol{\theta}_0, \mathcal{P}_i, \mathcal{M}). \tag{9}$$

Particularly, if we consider the optimal parameter $\boldsymbol{\theta}_i^\dagger$ for model $\mathcal{P}_i$, $\pi(\mathbf{y}|\mathcal{P}_i, \mathcal{M})$ $= \pi(\mathbf{y}|\boldsymbol{\theta}_i^\dagger, \mathcal{P}_i, \mathcal{M})$. We can take the ratio of posterior model plausibilities,

$$\frac{\rho_1}{\rho_2} = \frac{\pi(\mathbf{y}|\mathcal{P}_1, \mathcal{M})\pi(\mathcal{P}_1|\mathcal{M})}{\pi(\mathbf{y}|\mathcal{P}_2, \mathcal{M})\pi(\mathcal{P}_2|\mathcal{M})} = \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})\pi(\mathcal{P}_1|\mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})\pi(\mathcal{P}_2|\mathcal{M})} = \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})} O_{12}, \tag{10}$$

where O_{12} is the ratio of prior odds and is often assumed to be one. With these tools in hand, we present the following theorems.

Theorem 1. *Let* (10) *hold. If* $\mathcal{P}_1$ *is more plausible than* $\mathcal{P}_2$ *and* $O_{12} \leq 1$, *then* (8) *holds.*

Proof. If $\mathcal{P}_1$ is more plausible than $\mathcal{P}_2$,

$$1 < \frac{\rho_1}{\rho_2} = \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})} O_{12} \leq \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})} \tag{11}$$

Equivalently, the reciprocal of the far right hand side is less than one, so

$$\log \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} < 0. \tag{12}$$

Since $g(\mathbf{y})$ is a probability measure, it is always non-negative. Thus

$$g(\mathbf{y}) \log \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} < 0 \Rightarrow \int_{\mathcal{Y}^n} g(\mathbf{y}) \log \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} \, d\mathbf{y} < 0. \tag{13}$$

This can be expanded into

$$\int_{\mathcal{Y}^n} g(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M}) \, d\mathbf{y} - \int_{\mathcal{Y}^n} g(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M}) \, d\mathbf{y} < 0, \tag{14}$$

which means

$$-\int_{\mathcal{Y}^n} g(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M}) \, d\mathbf{y} < -\int_{\mathcal{Y}^n} g(\mathbf{y}) \log \pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M}) \, d\mathbf{y}. \tag{15}$$

By adding the quantity $\int_{\mathcal{Y}^n} g \log g \; d\mathbf{y}$ to both sides, the desired result (8) immediately follows. □

This theorem demonstrates that if model $\mathcal{P}_1$ is "better" than model $\mathcal{P}_2$ in the Bayesian sense, it is also a "better" deterministic model in the sense of (8). However, the reverse implication requires much stronger conditions. The assertion (8) can be equivalently written as

$$\int_{\mathcal{Y}^n} g(\mathbf{y}) \log \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} \, d\mathbf{y} < 0. \tag{16}$$

For this inequality to hold, the relationship

$$\frac{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} < 1 \tag{17}$$

does not necessarily need to be true for *every* point $\mathbf{y} \in \mathcal{Y}^n$.

One way to proceed is to invoke the Mean Value Theorem: if $|\mathcal{Y}^n| < \infty$ and under suitable smoothness conditions, there exists some $\bar{\mathbf{y}} \in \mathcal{Y}^n$ such that

$$\int_{\mathcal{Y}^n} g(\mathbf{y}) \log \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\mathbf{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}\, d\mathbf{y} = |\mathcal{Y}^n|\, g(\bar{\mathbf{y}}) \log \frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}. \tag{18}$$

Combining (16) and (18),

$$|\mathcal{Y}^n|\, g(\bar{\mathbf{y}}) \log \frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} < 0. \tag{19}$$

Since $|\mathcal{Y}^n| > 0$ and $g(\mathbf{y}) > 0$,

$$\log \frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} < 0 \Rightarrow \frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})} < 1 \Rightarrow \frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})} > 1. \tag{20}$$

If $O_{12} \geq 1$,

$$\frac{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})}{\pi(\bar{\mathbf{y}}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M})} O_{12} > 1 \Rightarrow \frac{\rho_1}{\rho_2} > 1. \tag{21}$$

Thus $\mathcal{P}_1$ is more plausible than $\mathcal{P}_2$ for given data $\bar{\mathbf{y}}$. In summary, we have:

Theorem 2. *If* $D_{KL}(g\|\pi(\boldsymbol{y}|\boldsymbol{\theta}_1^\dagger, \mathcal{P}_1, \mathcal{M})) < D_{KL}(g\|\pi(\boldsymbol{y}|\boldsymbol{\theta}_2^\dagger, \mathcal{P}_2, \mathcal{M}))$ *and if* $|\mathcal{Y}^n| < \infty$ *and the integrand in* (18) *is continuous, then there exists a* $\bar{\boldsymbol{y}} \in \mathcal{Y}^n$ *such that* $\mathcal{P}_1$ *is more plausible than* $\mathcal{P}_2$*, given that* $O_{12} \geq 1$*.*

Acknowledgement: The support of this work by the US Department of Energy Applied Mathematics program MMICC's effort under Award Number DE-5C0009286 is gratefully acknowledged.

[1] Bernstein S. Theory of Probability. Moscow 1917.

[2] Freedman, David A. "On the so-called Huber sandwich estimator and robust standard errors." The American Statistician 60.4 (2006).

[3] Kleijn, B.J.K. and van der Vaart, A.W. The Bernstein-Von-Mises theorem under misspecification. Electronic Journal of Statistics 6 (2012), 354–381

[4] Von Mises, R. Wahrscheinlichkeitsrechnung. Springer Verlag, Berlin 1931