Statistical Science
2016, Vol. 31, No. 4, 495–498
DOI: 10.1214/16-STS561
Main article DOI: 10.1214/16-STS592

Contextuality of Misspecification and Data-Dependent Losses

Peter Grünwald

Abstract. We elaborate on Watson and Holmes' observation that misspecification is contextual: a model that is wrong can still be adequate in one prediction context, yet grossly inadequate in another. One can incorporate such phenomena by adopting a generalized posterior, in which the likelihood is multiplied by an exponentiated loss. We argue that Watson and Holmes' characterization of such generalized posteriors does not really explain their good practical performance, and we provide an alternative explanation which suggests a further extension of the method.

It is a pleasure to comment on this stimulating paper about decision making under model misspecification. I was happy to see that it begins by pointing out that misspecification is *contextual*—a point that cannot be stressed enough, and that has also played a central part in my own work on Bayesian inconsistency under misspecification (Grünwald and Langford, 2007, Grünwald and Van Ommen, 2014). I will focus my comments on this aspect and on the developments in Section 4, which are the most closely related to my own work. While I think the paper's combined Bayes-minimax approach has substantial merit for the case of "simple" loss functions of the form $L_a(\theta)$, involving model parameters and actions (as in the synthetic example in their Section 3.5), I am more skeptical of the application to losses of the form $L_a(\theta, z)$ or $L_a(z)$ that involve data z as well, as in their Section 4.2. I do see the merit of the approaches described by the authors for such losses (indeed I have been advocating them myself), yet I do not see how their characterization can explain their practical success: the proposed formalism is rich enough to incorporate such approaches as special cases, but it does not really motivate them. Before elaborating on this in Section 3, below I first introduce data-dependent (DD from now on) losses and I then show how nicely they illustrate the contextuality of misspecification. I end by suggesting an extension to the paper's approach that may address my concerns.

Peter Grünwald is Senior Researcher, Mathematical Institute, CWI and Professor of Statistical Learning, Leiden University, Science Park 123, 1098 XG Amsterdam, The Netherlands (e-mail: pdg@cwi.nl).

1. DD LOSSES

This section recalls some standard Bayesian decision theory (Berger, 1985). It will be useful to slightly extend the authors' setup and consider models $\{f(y|x;\theta) : \theta \in \Theta\}$ of conditional densities $f(y|x;\theta)$, with a joint outcome denoted as $z = (x, y)$ and x taking values in some covariate space $\mathcal{X}$—the authors (WH from now on) only consider the unconditional case. Now, data-dependent losses come into play whenever, based on initial data $z^n := z_1, \dots, z_n$, one wants to make predictions about one or more future data items $z_{n+1}^m := z_{n+1}, \dots, z_{n+m}$ coming from the same source. In practice, one often observes x_{n+1}^m, and needs to predict y_{n+1}^m, where one predicts y_i by $a(x_i)$, and $a : \mathcal{X} \to \mathbb{R}$ is some prediction function. One measures the quality of such predictions using some DD loss function such as, for example, the squared error loss, $L_a^{(2)}((x, y)) = (y - a(x))^2$, extended to n outcomes by summing the losses, $L_a^{(2)}(z_{n+1}^m) := \sum_{i=n+1}^m (y_i - a(x_i))^2$.

If one's uncertainty about Y given X is described by density $f(Y|X)$ and one wants to predict a single outcome, one should use the *Bayes optimal act* for the squared error loss. This is the function a_f defined by, for each x,

$$a_f(x) := \arg\min_{y' \in \mathbb{R}} \mathrm{E}_f[(Y - y')^2 | X = x],$$

which turns out to be given by $a_f(x) = \mathrm{E}_f[Y|X = x]$. Similarly, if the prediction task of interest is to make good predictions with respect to the absolute loss $L_a^{(1)}((x, y)) = |y - a(x)|$, then the Bayes optimal act

a_f that minimizes, for each x, $\mathrm{E}_f[|Y - y'||X = x]$, is obtained by setting $a_f(x)$ to the median of Y under $f(Y|X = x)$.

Aside from their direct use in predictive inference, data-dependent loss functions also play a central role in Bayesian decision theory when the goal is to infer structural properties of the domain being modeled. They are then usually called (proper) *scoring rules*. For example, consider a DM (decision-maker) who represents her uncertainty about a domain by an unknown conditional density $f(Y|X)$. If one wants her to quote her true beliefs about the regression function $\mathrm{E}_f[Y|X]$, one may ask her to play a prediction game in which action a will be scored by the squared error loss, $L^{(2)}$, that is, upon observing $Z = (X, Y)$, she will be scored $L^{(2)}_a(Z) = (Y - a(X))^2$. Her optimal (Bayes) response will then be to output the function a_f given by $a_f(x) := \mathrm{E}_f[Y|X = x]$. Yet if, instead, one wants to entice a DM to quote her beliefs about the median of Y as a function of X, one should score her using the absolute loss $L^{(1)}$.

2. DD-LOSSES AND CONTEXTUALITY OF MISSPECIFICATION

The central issue here is that, if under misspecification, the Bayes posterior concentrates at all, it will tend to concentrate on distributions that assign high (log-) likelihood to the data in expectation. Let me illustrate this from a frequentist point of view, complementary to, but not in contradiction with, the explanation given by WH: assume that $Z_i = (X_i, Y_i)$ are i.i.d. under some imagined distribution with joint density f^*—note that X_i are random as well. Assume that there exists a unique $\tilde{\theta}$ such that $f(\cdot; \tilde{\theta})$ is closest, among all $\theta \in \Theta$, to f^* in KL divergence. Then the tendency of the posterior to prefer θ with high likelihood implies, under further (nontrivial) conditions, that, as more data becomes available, it concentrates on ever smaller KL neighborhoods of $\tilde{\theta}$—after all, the KL divergence $D(f^* \| f(\cdot; \theta))$ is just the minus expected log-likelihood ratio for one outcome.

Now for some combinations of models and loss functions $L_a(z)$, it holds that the smaller the KL divergence $D(f^* \| f(\cdot; \theta))$, the better the prediction performance $\mathrm{E}_{(X,Y) \sim f^*}[L_{a_\theta}(Y|X)]$, if one predicts with the action $a_\theta := a_{f(\cdot|\cdot;\theta)}$ that is Bayes-optimal for θ. We call a loss function L with this property *associated* with the model. For example, consider a standard Bayesian linear regression model $\{f_\theta | \theta \in \Theta\}$, which assumes Gaussian noise of (say) fixed variance σ^2, parameterized such that the likelihood for a sample z^n is of the form

$$f(y^n | x^n; \theta) \propto e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \theta(x_i))^2},$$

that is, θ represents the regression function corresponding to $f(\cdot|\cdot; \theta)$. Now the "best" $\tilde{\theta}$ (closest to f^* in the KL sense) will also be the best $\tilde{\theta}$ for squared error prediction purposes, minimizing the f^*-$L^{(2)}$-risk (risk = expected loss). The reason is that, for each fixed σ^2, the KL divergence as a function of θ is an affine function of the squared error risk achieved by θ:

$$\begin{aligned} D(f^* \| f(\cdot; \theta)) &= \mathrm{E}_{(X,Y) \sim f^*}\left[\log \frac{f^*(Y|X)}{f(Y|X; \theta)}\right] \\ &= \frac{1}{2\sigma^2} \cdot \mathrm{E}_{(X,Y) \sim f^*}(Y - \theta(X))^2 - C, \end{aligned}$$

where the constant C depends on σ^2 and f^*, but not on θ. So, if the goal is to make good squared-error predictions, the Bayes posterior—if it concentrates at all—will concentrate on the squared-error risk-optimal θ in the model. In the terminology of Grünwald and Van Ommen (2014), the squared error is *associated* with the Gaussian regression model; in WH's terminology, the Gaussian regression model is suitable in the *context* of squared error prediction, even under misspecification.

. . . but now consider absolute loss ($L^{(1)}$) predictions. If the standard Gaussian regression model is used and the model is correct, then $f^*(Y|X) = f(Y|X; \tilde{\theta})$, and, given enough data, the Bayes predictive distribution will converge to $f^*(Y|X)$, and hence its Bayes-optimal $L^{(1)}$-predictions will converge to the optimal predictions based on the "true" f^*, and hence become optimal, in expectation, for the $L^{(1)}$-loss, under the true distribution f^*. But under misspecification, $\tilde{\theta}$ may be quite different from the $\theta \in \Theta$ that gives the optimal $L^{(1)}$-predictions, and hence the Bayesian posterior—even if it concentrates—may not lead one to adopt the optimal $L^{(1)}$-predictions that are available within the model. To see this, note that the Bayes act under $f(Y|X; \theta)$ under $L^{(1)}$-loss is to predict Y again using $\theta(X)$ [since, according to the model, $\theta(X)$ is indeed the median of $Y|X$], but in reality, even if $\tilde{\theta}$ is the true regression function, $\tilde{\theta}(X)$ may be very far from the median if the errors are not really normally distributed. In our terminology, the $L^{(1)}$-error function is not associated with the Gaussian regression model. In WH's terminology, under misspecification, the Gaussian regression model is neither suitable in the *context* (inference goal) of predicting with the $L^{(1)}$-loss function nor in the context of inferring the true (or at least "a reasonable") median of Y as a function of X.

3. DD LOSSES AND MINIMAXITY (OR MAXIMINITY?)

Section 4 of WH's paper describes model diagnostics based on a modified posterior distribution, defined as

$$\pi^{\sup}_{a,C} = \arg\sup_{\pi\in\Gamma_C} \mathrm{E}_\pi[L_a(\theta, Z)], \tag{1}$$

where Z is observed data, and Γ_C is the KL ball of radius C around π_I. In most applications including the ones below, π_I is just the standard posterior based on prior π. WH show that $\pi^{\sup}_{a,C}$ is given by

$$\begin{aligned}\pi^{\sup}_{a,C} &\propto e^{\lambda_a(C)L_a(\theta,Z)}\pi_I(\theta)\\ &\propto e^{\lambda_a(C)L_a(\theta,Z)} f(Z;\theta)\pi(\theta),\end{aligned} \tag{2}$$

where $\lambda_a(C)$ is a nonnegative real valued monotone function of C (essentially a Lagrange multiplier).

The idea of adopting (2) is that π_I may paint an overly optimistic picture of the loss $L \equiv L_a(\theta, Z)$—therefore, one may look for a more robust assessment by specifying a neighborhood of π_I and look at the π' that gives the worst-case possible expectation of L within this neighborhood. Because of a special coherence property of Kullback–Leibler (KL) divergence, KL balls are the preferred choice for defining such neighborhoods. This all makes perfect sense—as long as L does not involve the *already observed* data. An example of such an $L \equiv L_a(\theta)$ is provided by the synthetic example of Section 3.5. But WH also consider cases in which L does depend on the data, still taking the π achieving the maximum in (1). This seems strange: if a DM wants to make a robust assessment of the loss an action makes on data, this should be new, as yet unseen data—not the data already observed, about which there is no uncertainty anyway. Both from a Bayesian and a game-theoretic (robust, minimax) point of view, adopting a distribution that is minimax for data already observed seems unnatural to me.

The issue becomes acute when a DM has prior beliefs about a set of parameters but does not know how to specify a likelihood $f(z;\theta)$. WH give the example where $z = y$ and θ represents the median of an unknown distribution, which we can extend to the conditional case with $z = (x, y)$ and $\theta(x)$ now representing the conditional median of y given covariate x. The loss function[1] whose Bayes act is the empirical median is given by the $L^{(1)}$-loss, $L^{(1)}_\theta(z^n) = \sum_i |y_i - \theta(x_i)|$, and WH suggest to adopt a posterior of the form

$$\pi^{\inf}(\theta) \propto e^{-\lambda L_\theta(z^n)} \cdot \pi(\theta), \tag{3}$$

for some $\lambda > 0$, which is compatible with previous approaches from the machine learning literature such as in prediction with expert advice (Vovk, 2001) and PAC-Bayesian style inference (Zhang, 2006a, 2006b). But (3) gives the distribution which *minimizes* expected loss among all distributions in a KL neighborhood of π, and to make it a special case of their framework, they have to change the goal, it seems, from a worst-case approach to a best-case approach [posteriors of the form (1) always induce a positive multiplier of the loss; yet here it is negative]. The paper does not really explain why it would make sense to switch goals in this case, other than noting that a DM might "wish" to do this.

I claim that the real motivation for using generalized posteriors of the form (3), with data-dependent losses, is quite different: this posterior tends to favor distributions with small empirical $L^{(1)}$-loss, and hence, will tend to assign high posterior density to those θ that will have small $L^{(1)}$-loss on future data. In fact, one can think of (3) as defining a pseudo-likelihood

$$f^\circ(y|x;\theta,\lambda) \propto e^{-\lambda|y-\theta(x)|}, \tag{4}$$

so that for each fixed λ, the corresponding KL divergence satisfies

$$\begin{aligned}D\big(f^* \| f^\circ(\cdot;\theta,\lambda)\big) &= \mathrm{E}_{(X,Y)\sim f^*}\left[\log \frac{f^*(Y|X)}{f^\circ(Y|X;\theta,\lambda)}\right]\\ &= \lambda \cdot \mathrm{E}_{(X,Y)\sim f^*}\big|Y - \theta(X)\big| - C,\end{aligned}$$

where C is a constant depending on λ and f^* but not on θ. Thus, under this model the KL divergence of θ to f^* is an affine function of the $L^{(1)}$-risk, so that good performance in terms of KL divergence implies good performance in terms of $L^{(1)}$-risk (and hence optimal estimation of the median). This makes the $L^{(1)}$-loss an associated loss (in the sense above) for the model defined by (4). Since the posterior—if it concentrates at

[1] I am deviating from WH's notation here: throughout the paper, actions for a given loss function are written as subscripts and model parameters (random variables under the posterior) are written as second argument for the loss function. Yet in the median example, WH use the notation $L(y,\theta)$ rather than $L_\theta(y)$ for $|y-\theta|$. Here, θ clearly plays the role of an action rather than a parameter though: it is equal to the action a_θ that would be Bayes-optimal under distribution $f(y;\theta)$. Note that we have $a_\theta = \theta$ because the densities specified are symmetric around the mean θ.

all—will tend to concentrate on $\tilde{\theta}$ that is closest in KL divergence to f^*, this is a desirable property if one is interested in good $L^{(1)}$-loss behavior, and so it does explain why one should have a minus in the exponent in (3).

I have been advocating the use of pseudo-likelihoods of the form $\exp(-\lambda L)$ with such a motivation ever since Grünwald (1999). WH suggest some potentially important extensions to this idea—though again, I think the answers that WH provide need additional motivation and disambiguation. Consider, for example, a case in which a probability model, and hence a likelihood are available after all. Suppose one thinks that the likelihood, while not a perfect description of the world, is sufficiently reasonable to take it into account when determining the posterior, and one is once again interested in learning the conditional median. The linear regression model above is a case in point. The paper's original minimax approach suggests to take the final posterior π' as $\pi^{\sup}_{a,C}$ in (2), which would favor θ with large $L^{(1)}$-loss; in contrast, the paper's later maximin approach suggests that one might replace the prior π in (3) by the Bayes posterior, which would amount to adopting π' again as $\pi^{\sup}_{a,C}$ in (2) but now with a negative multiplier, favoring θ with small $L^{(1)}$-loss. What to do? The motivation of generalized posteriors in terms of model-associated ("contextual") loss functions suggests a solution: one extends the original model, defining a likelihood of the form

$$f^{\circ}(y|x;\theta,\lambda) \propto e^{-\lambda|y-\theta(x)|} \cdot e^{-\frac{1}{2\sigma^2}(y-\theta(x))^2}, \tag{5}$$

and determines the η-generalized posterior for some $\eta > 0$ as

$$\pi_\eta(\theta|(x,y)) \propto \left(f^{\circ}(y|x;\theta,\lambda)\right)^{\eta} \cdot \pi(\theta,\lambda).$$

One thus has enlarged the model by an extra $L^{(1)}$-term in the exponent weighted by extra parameter λ, which determines how strong the $L^{(1)}$-loss should influence the standard likelihood. One then adds a second parameter η when determining the posterior. Grünwald and Van Ommen (2014) show that such an η is different from λ and σ^{-2} and cannot be absorbed, in general, into the likelihood itself; it needs to be added since setting $\eta = 1$ may cause the posterior never to converge at all under misspecification. While λ is then determined by standard Bayesian means (it is part of the prior and posterior), something different is needed for η: Grünwald and Van Ommen (2014) describe a data-dependent ("Safe Bayesian") method for finding it. (5) is a special case of likelihoods of the form

$$\frac{1}{Z_\lambda} \cdot e^{-\lambda L_{a_\theta}(x,y)} \cdot f(y|x;\theta), \tag{6}$$

where a_θ is the Bayes act for L under $f(\cdot;\theta)$ which, in the case of L equal to the $L^{(1)}$-loss, is the median under $f(\cdot;\theta)$. In the present case, $f(y|x;\theta)$ is a symmetric density for y with mean $\theta(x)$, hence the median is itself equal to $\theta(x)$, giving rise to (5). It would be interesting to explore whether such generalized Bayesian procedures based on extended likelihoods perform well in practice.

ACKNOWLEDGMENTS

This research was supported by the Netherlands Organization for Scientific Research (NWO) VICI Project Nr. 639.073.04.

REFERENCES

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. Springer, New York. MR0804611

GRÜNWALD, P. (1999). Viewing all models as 'probabilistic'. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory* (*Santa Cruz*, *CA*, 1999) 171–182 (electronic). ACM, New York. MR1811613

GRÜNWALD, P. and LANGFORD, J. (2007). Suboptimal behavior of Bayes and MDL in classification under misspecification. *Mach*. *Learn*. **66** 119–149. DOI 10.1007/s10994-007-0716-7.

GRÜNWALD, P. D. and VAN OMMEN, T. (2014). Inconsistency of Bayesian inference for misspecified linear models, and a proposal for repairing it. Technical report No. abs/1502.08009.

VOVK, V. (2001). Competitive on-line statistics. *Int*. *Stat*. *Rev*. **69** 213–248.

ZHANG, T. (2006a). From ϵ-entropy to KL-entropy: Analysis of minimum information complexity density estimation. *Ann*. *Statist*. **34** 2180–2210. MR2291497

ZHANG, T. (2006b). Information-theoretic upper and lower bounds for statistical estimation. *IEEE Trans*. *Inform*. *Theory* **52** 1307–1321. MR2241190