5th International Symposium on Imprecise Probability: Theories and Applications, Prague, Czech Republic, 2007

Updating and Testing Beliefs: An Open Version of Bayes' Rule

Elmar Kriegler
Department of Engineering and Public Policy, Carnegie Mellon University
Potsdam Institute for Climate Impact Research
elmar@cmu.edu

Abstract

Developing models to describe real systems is a challenge because it is difficult to assess and control the residual between the two entities. Bayesian updating of a belief about model accuracy across an ensemble of available models can lead to spurious results, since the application of Bayes' rule presupposes that an accurate model is contained in the ensemble with certainty. We present a framework in which this assumption can be dropped. The basic idea is to extend Bayes' rule to the exhaustive, but unknown space of all models, and then contract it again to the known set of models by making best/worst case assumptions for the remaining space. We show that this approach leads to an ε-contamination model for the posterior belief, where the ε-contamination is updated along with the distribution of belief across available models. In essence, the ε-contamination provides an additional test on the accuracy of the overall model ensemble compared to the data, and will grow rapidly if the ensemble fails such a test. We demonstrate our concept with an example of autoregressive processes.

Keywords. Bayesian updating, prediction, model accuracy, ε-contamination model, AR process

1 Introduction

A vital part of the scientific endeavor consists in developing models for real systems. Obviously, a model can never be an identical copy of a real system, but rather a proxy to understand a limited set of system features, on the basis of which future observations of these features may be predicted. In order to construct a useful model, it is important to control the residual between model and real system in a way that allows the model to have some predictive accuracy. Therefore, it is extremely helpful if the real system can be studied in laboratory experiments where the experimenter can force her ways on it to test the model. However, controlling the residual becomes an enormous challenge if the real system is not accessible to laboratory studies. The situation is further exacerbated if available observations cover only a small part of the phase space. The climate system and computer models of it are a perfect example of this situation, and we will have this example in mind in what follows.

In such cases, model quality is usually assessed with a mixture of scientific knowledge about the system and statistical inference from system measurements. Here, we focus on model accuracy to predict certain system features. While several definitions of accuracy can be found in the literature (e.g. in terms of bias), we use our own definition tailored to an application to dynamic systems characterized by noisy time series. We say that a model S is *accurate* (to make predictions) if it can describe the observables Y of a (few) system feature(s) of interest – not of the entire system – up to an additive Gaussian iid process ϵ, i.e., $Y - S = \epsilon \sim N(0, \sigma)$. The choice of a Gaussian iid process for the residual between model and data is a common, but subjective assumption, and should be regarded as part of the model formulation. In principle, our approach could be applied with another choice of stationary stochastic process for the residual.

Assume we have an ensemble of model hypotheses $M(\theta)$, with $\theta \in \Theta$ indexing the available models, and some system data $\hat{y}$, from which we want to learn about the relationship between $M(\theta)$ and an accurate model S. If a probability $P(M(\theta) = S|\hat{y})$ is sought, we have to turn to Bayesian statistics. In our case, this requires to

1. estimate a likelihood function $\mathcal{L}(\theta; \hat{y}) \sim \rho(\hat{y}|\theta)$, i.e., the probability of observing $\hat{y}$ for a given model $M(\theta)$ under the assumption that $M(\theta)$ constitutes an accurate model (which defines the likelihood function given our knowledge about the residual $Y - S = \epsilon$), and
2. updating it with a prior probability density $\rho(\theta)$:

$\Omega \to \mathbb{R}_0^{+1}$ to derive the posterior probability density $\rho(\theta|\hat{y})$ that model $M(\theta)$ constitutes an accurate model S.

However, this approach requires us to make the assumption that S is contained in our model ensemble with certainty as evident from Bayes' rule

$$\rho(\theta|\hat{y}) = \frac{\mathcal{L}(\theta;\hat{y})\rho(\theta)}{\rho(\hat{y})} \quad (1)$$
$$\text{with} \qquad \rho(\hat{y}) := \int_\theta \mathcal{L}(\theta;\hat{y})\rho(\hat{y})d\theta\,,$$

where the denominator assures that the posterior probability is normalized. In the following, we may call the fact that $\int_\theta \rho(\theta)d\theta = \int_\theta \rho(\theta|\hat{y})d\theta = 1$ *closed world assumption*.

We believe that this assumption is at odds with the open nature of the scientific endeavor, where a set of possible models $\{M(\theta)|\theta \in \Theta\}$ imagined at some initial time is usually expanded as more data is obtained. More precisely, the model development process consists in (I) expanding the set Θ of known models, and (II) updating our belief about model accuracy across Θ. Obviously, only the later type (II) learning can be described in terms of Bayesian learning. The former type (I) may be informed by Bayesian inference, but seems to be complementary to it, since it relates to the emergence of positive belief in an area of the model space that was not supported by the prior belief.

Acknowledging this fundamental difference, we will not attempt to force type (I) learning in terms of expanding Θ into the Bayesian updating framework. Instead, we aim at the more modest goal to include an indicator for the necessity of type (I) learning into the updating process. This is important because naive application of Bayesian learning without contemplating the possibility that the entire model ensemble $\{M(\theta)|\theta \in \Theta\}$ might not contain an accurate model can lead to spurious results. As the amount of data $\hat{y}$ increases, the likelihood function tends to sharpen, and updating by means of Equation (1) will decrease the spread of the posterior belief that a given model $M(\theta)$ coincides with the accurate model S. Hence, an analyst ignoring anything else will converge in his belief on some model $M(\theta) = S$. As a consequence, his predictions of real system features, based on his converging belief, will grow more and more (over)confident – although off the mark – as the data accumulates. This paradoxical behavior is a direct consequence of the closed world assumption. It is therefore desirable to drop this assumption, and directly include an indicator for $S \notin \{M(\theta)|\theta \in \Theta\}$ in the updating process. In this paper, we present such a framework.

A similar concern about Bayesian learning on model quality and the subsequent use of posterior beliefs for prediction of future observations has been raised by Draper [2] and, more recently, Goldstein and Rougier [3, 4]. Draper criticizes the practice of neglecting structural uncertainty, and proposes to extend prior and likelihood to the space of possible model structures. His approach [2] leads to an increased spread of the posterior on the model ensemble. Goldstein and Rougier highlight the importance to assess the discrepancy between the ensemble of available models and the 'ideal' model which captures the system up to an additive noise term. They coined the term 'reified' for the 'ideal' reference model. Obviously, the idea of a 'reified model' is closely related to what we call accurate model here. In [3, 4], Goldstein and Rougier propose to address model discrepancy by including a meta-model of it in the updating process, and offer guidelines how such a meta-model might be constructed. This is a very challenging task. As indicated above, we take a different approach. We do not try to find a positive expression for model discrepancy, or the extension of prior and likelihood to the space of possible model structures, but rather seek to include an indicator for the negative result that model discrepancy impinges on the predictive accuracy of the model ensemble.

The paper is organized as follows. Section 2 presents a simple example of autoregressive (AR) processes in which the application of standard Bayesian updating is shown to fail if the model hypotheses have limited accuracy to predict the real system. Section 3 contains the core of the paper, detailing our derivation of an open version of Bayes' rule that allows to drop the closed world assumption. This rule is put into operation for our example of AR processes in Section 4. We conclude by highlighting the challenges for an application of the open Bayes' rule to real world problems in Section 5.

2 Limitations of Bayes' rule: Example of autoregressive (AR) processes

Let us assume the following dynamic 'real system' evolving over n time steps.

$$Y(n) = (\xi_1, \alpha_1^*\xi_1 + \xi_2, X_3, ..., X_n) \quad (2)$$
$$\text{with} \qquad X_t = \alpha_1^* X_{t-1} + \alpha_2^* X_{t-2} + \xi_t\,,\ t \geq 3 \quad (3)$$
$$\xi_t \sim N(0, \sigma_\xi^*) \quad \text{iid process (white noise)}\,,$$

[1] For the sake of simplicity, we assume throughout the paper that $\Omega \subseteq \mathbb{R}^n$ is a continuous space, and that a prior probability measure $P : \sigma(\Theta) \to [0, 1]$ over a σ-field of Θ is continuous, i.e., can be described by a probability density on Θ.

where we require the AR(2) process X_t to be stationary. An AR(2) process described by Equation (3) is stationary iff $\alpha_1^* + \alpha_2^* < 1$, $\alpha_2^* - \alpha_1^* < 1$, and $|\alpha_2^*| < 1$. For the sake of simplicity, we have neglected any measurement error in observing the real system, and therefore can identify it directly with the observable $Y(n)$. Let us further assume that our ensemble of model hypotheses for $Y(n)$ is restricted to a closed set of stationary AR(1)-process with noise term $\xi_t \sim N(0, \sigma_\xi^*)$:

$$\{M(\alpha_1) := (\xi_1, X_2', ..., X_n') \mid X_t' = \alpha_1 X_{t-1}' + \xi_t\,, \quad t \geq 2\,,\ -\bar{\alpha} \leq \alpha_1 \leq \bar{\alpha}\,,\ \bar{\alpha} := 0.995\}\,. \quad (4)$$

Obviously, the model ensemble contains an accurate model S if $\alpha_1^* \in [-\bar{\alpha}, \bar{\alpha}]$ and $\alpha_2^* = 0$. In this case, we find $S := M(\alpha_1^*) = Y(n)$. We will discuss below whether there can be an accurate model in the ensemble if $\alpha_2^* \neq 0$.

After having received a realization $\hat{y}(n) = (\hat{y}_1, ..., \hat{y}_n)$ of $Y(n)$, we can apply Bayesian updating to our prior belief about the accuracy of the model hypotheses $M(\alpha_1)$ as defined by a probability density $\rho(\alpha_1)$. Without loss of generality, let the prior $\rho(\alpha_1)$ be uniformly distributed on $[-\bar{\alpha}, \bar{\alpha}]$. As shown in Appendix A, the likelihood of having obtained the realization $\hat{y}(n)$ from an AR(1)-process with propagator α_1 is given by

$$\mathcal{L}(\alpha_1; \hat{y}(n)) \sim N\left(\frac{\hat{\alpha}(n)}{1-\hat{\beta}(n)}, \frac{\hat{\sigma}(n)}{\sqrt{1-\hat{\beta}(n)}}\right)\,, \quad (5)$$

where $\hat{\alpha}(n)$, $\hat{\sigma}(n)$, and $\hat{\beta}(n)$ are estimated from the observed time series $\hat{y}(n)$ as defined in Equation (26), (27), and (28), respectively. Hence, application of Bayes rule (Equation 1) with a uniform prior for α_1 yields the following posterior probability density on $[-\bar{\alpha}, \bar{\alpha}]$:

$$\rho(\alpha_1|\hat{y}(n)) = \frac{\exp\left(-\frac{1-\hat{\beta}(n)}{2\hat{\sigma}(n)^2}\left(\alpha_1 - \frac{\hat{\alpha}(n)}{1-\hat{\beta}(n)}\right)^2\right)}{\int_{-\bar{\alpha}}^{\bar{\alpha}} \exp\left(-\frac{1-\hat{\beta}(n)}{2\hat{\sigma}(n)^2}\left(\alpha_1 - \frac{\hat{\alpha}(n)}{1-\hat{\beta}(n)}\right)^2\right) d\alpha_1}\,. \quad (6)$$

Equation (6) can be used to test the effect of the closed world assumption on the Bayesian updating process. For the experiment, we generated 200 realizations of time series $\hat{y}(n)$ with length $n = 5000$ for four different AR(2)-processes with $\sigma_\xi^* = 1$, $\alpha_1^* = 0.866$ and $\alpha_2^* = \{-0.9, -0.3, 0, 0.06\}$. Note that in the asymptotic limit $n \to \infty$ any AR(k)-process is normally distributed $\sim N\left(0, \sigma/\sqrt{1 - \sum_{i=1}^k \alpha_i \rho_i}\right)$, with ρ_i the autocorrelation of lag i [7]. Therefore, removing the time index from the observations renders AR-processes of different order indistinguishable from each other. It is in this sense, that we can calculate an AR(1)-equivalent of an AR(2)-process with propagators α_1 and α_2. The AR(1)-equivalent yielding a normal distribution with identical standard deviation in the asymptotic limit has the propagator

$$\alpha_{\text{equiv}} = \sqrt{\alpha_1\rho_1 + \alpha_2\rho_2} = \sqrt{\alpha_1^2 \frac{1+\alpha_2}{1-\alpha_2} + \alpha_2^2}\,. \quad (7)$$

For the four different AR(2)-processes chosen above we find AR(1)-equivalents with propagators $\alpha_{\text{equiv}} = \{0.922, 0.703, 0.866, 0.922\}$. It can be seen that the asymptotic distribution of the two AR(2)-processes with $\alpha_2^* = -0.9$ and $\alpha_2^* = 0.06$ are indistinguishable.

We have considered AR(2)-processes with very pronounced tails compared to ξ, because we are interested in the ability of the model ensemble $M(\alpha_1)$ to predict in particular the tails of the distributions. In practice, a good prediction of the tails is often what matters most. Note that it follows from Equation (7) that there will exist an accurate model S in the ensemble of model hypotheses $M(\alpha_1)$ even if the real system is described by an AR(2)-process with $\alpha_2 \neq 0$ - if we are only interested in predicting the asymptotic distribution of future observations. It will be interesting to see whether Bayesian updating is capable to converge to the propagator of the AR(1)-equivalent model.

Figure 1 shows the result of Bayesian updating for the four different AR(2)-processes. We have updated the posterior belief about α_1 (see Equation 6) after each 20 new observations. Shown is the development of the 90% confidence limits for the mean value of the posterior distribution. The confidence limits were derived from the sample of 200 time series used in the updating process. It can be seen that the posterior mean converges to the correct value of $\alpha_1^* = 0.866$ (horizontal solid line) in the case where the real system is described by an AR(1) process ($\alpha_2^* = 0$). Convergence is still good if only a small deviation from the AR(1) assumption is considered ($\alpha_2^* = 0.06$). In this case, the posterior mean converges to the propagator $\alpha_{\text{equiv}} = 0.922$ of the AR(1)-equivalent process. However, if the deviation from the AR(1) assumption is negative and increases in magnitude ($\alpha_2^* = -0.3$), the posterior belief converges to a biased value below α_{equiv}. In the extreme case $\alpha_2^* = -0.9$, Bayesian learning leads to a spurious result. Since the posterior distribution has contracted strongly after several thousand observations (see black dots on the right axis), the updating procedure has settled on the wrong region of α_1-space with very high confidence. This is a direct consequence of the closed world assumption.

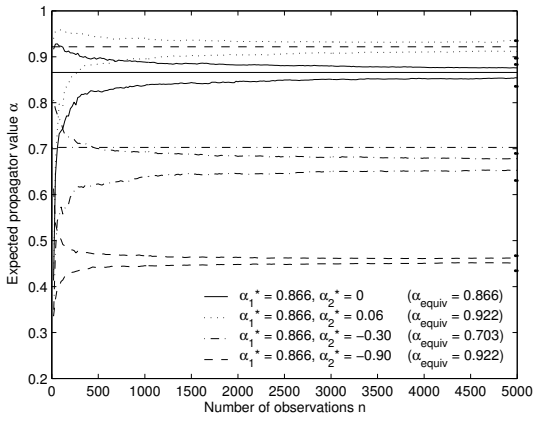

Figure 1: Updated belief about the propagator α of an hypothetical AR(1)-process after n observations. The lower and upper bound of the 90% confidence interval for the mean value of the posterior belief (derived from the sample of 200 time series) are plotted. Horizontal lines indicate the propagator value α_{equiv} of the equivalent AR(1)-process in the asymptotic limit. α_{equiv} for the AR(2)-process with $\alpha_2^* = -0.9$ is identical to the case $\alpha_2^* = 0.06$. Black dots on the right axis indicate the range between the 5% and 95% quantiles of the posterior belief after 5000 observations.

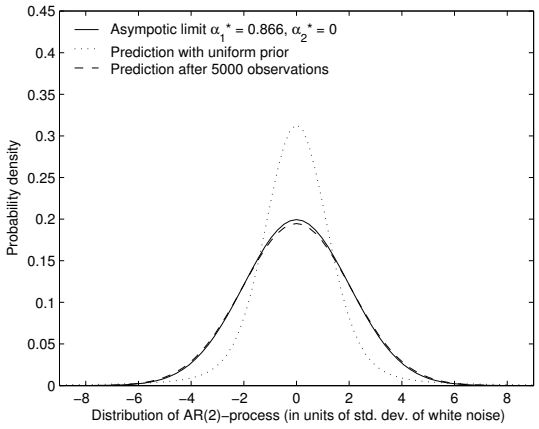

Figure 2: Predictions based on the belief about α_1 for the case $\alpha_1^* = 0.866$ and $\alpha_2^* = 0$. The solid line shows the asymptotic distribution of the 'real' AR(1) process. The dotted line shows the prediction before any learning occurred (based on a uniform prior for $\alpha \in [-\bar{\alpha}, \bar{\alpha}]$). The updated prediction after 5000 observations (dashed line) lies almost exactly on the asymptotic distribution of the 'real' system.

We briefly assess the consequences for predicting the distribution of system observations y in the asymptotic limit. As mentioned above, we know that the asymptotic distribution of an AR(1) process for given values of α_1 and σ_ξ is defined by $\rho(y|\alpha_1) \sim N(0, \sigma_\xi/\sqrt{1-\alpha_1^2})$. Hence, if our belief about α_1 is described by the posterior $\rho(\alpha_1; \hat{y}(n))$, our prediction for the distribution of system observations based on past data $\hat{y}(n)$ is given by

$$\rho(y|\hat{y}(n)) = \int_{-\bar{\alpha}}^{\bar{\alpha}} \rho(y|\alpha_1)\ \rho(\alpha_1; \hat{y}(n))\ d\alpha_1\,. \quad (8)$$

Figure 2 shows predictions for the case of learning from a realization of the AR(1)-process with $\alpha_1^* = 0.866$ and $\alpha_2^* = 0$. The dotted line depicts the prediction on the basis of the uniform prior, before any learning occurred. Interestingly, the assumption of the uniform prior strongly underestimates the probability mass in the flanks of the distribution. The example shows that in general it is not warranted to associate the uniform prior with a conservative (or non-informative) choice of belief. After the uniform prior is updated with observations $\hat{y}(n)$ the predictions converge very quickly to the asymptotic distribution of the 'real' system. Figure 2 shows that the prediction after 5000 observations is nearly identical with the 'real' distribution.

While Bayesian learning was very successful for the

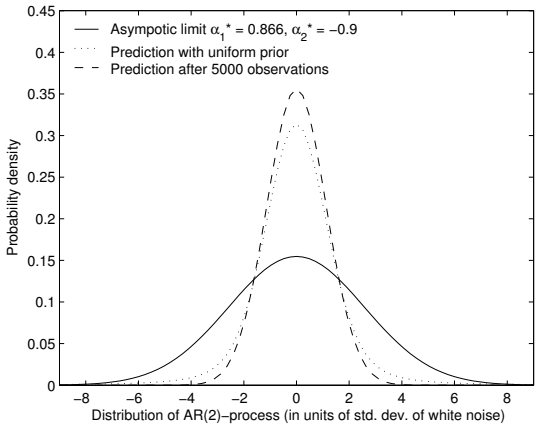

Figure 3: Predictions based on the belief about α_1 for the case $\alpha_1^* = 0.866$ and $\alpha_2^* = -0.9$. Solid, dotted and dashed lines as described in Figure 2.

case where the 'real' system is part of the ensemble of model hypotheses, the situation is markedly different for an AR(2)-process which strongly deviates from the AR(1)-assumption ($\alpha_2^* = -0.9$). As depicted in Figure 3, updating with observations $\hat{y}(n)$ leads to a further decrease in variance of the prediction as compared to the initial prediction based on the uniform prior. This is exactly the opposite of what should happen, because the asymptotic distribution of the 'real' system exhibits a much larger spread than both the initial and informed prediction. As is apparent from Figure 3, these spurious predictions strongly underes-

timate the tails of the distribution, and may therefore provide a false sense of security. What makes matters worse is that no amount of additional data will be able to rectify the situation. In contrast, the posterior belief will continue to sharpen, and the spread of the prediction will further decrease. This example shows that the closed world assumption underlying Bayes' rule can lead to spurious beliefs and predictions.

3 Extension of Bayes rule: dropping the closed world assumption

Given the spurious results that can emerge from a naive application of Bayes' rule, we are looking for an extension of Bayesian updating that includes an indicator for the overall accuracy of the model ensemble to reproduce 'real' system observations. This would allow us to drop the assumption that an accurate model S has to be included in the set of available models $M(\theta),\ \theta \in \Theta$ with certainty. A natural first step in this direction is to extend Bayes rule to a larger space $\Omega \supset \Theta$ for which the assumption $S = M(\omega)$ will be true for at least one $\omega^* \in \Omega$. Similar extensions are also the starting points for the proposals by Draper [2], and Goldstein and Rougier [4]. We assert that such a hypothetical space Ω is constituted by the space of all models, known and unknown. We think of Ω as a continuous vector space with large, but finite dimension that contains the parameter vectors ω for a large, but finite list of time-discrete[2] equations and relations. A model ensemble, i.e., a reduced list of parameterized equations, is characterized in this space by fixing the parameter values in some dimensions (collected in ψ), and allowing to vary – within bounds – the remaining parameters θ. Hence, a choice of model ensemble $M(\theta, \psi_0),\ \theta \in \Theta$, defines a Cartesian product $\Omega = \Theta \times \Psi$, where the parameters $\psi \in \Psi$ are fixed at ψ_0, and only $\theta \in \Theta$ can be varied.

So far, we have gained little because the nature of models in the residual space $\Theta \times (\Psi - \{\psi_0\})$ is completely unknown to us. Thus, our prior belief about the accuracy of unknown models in that space is vacuous. Fortunately, imprecise probability theory allows to capture a vacuous belief without having to assess the cardinality of its underlying space [9]. This is simply done by the vacuous probability model $\mathcal{V}(\Theta \times (\Psi - \{\psi_0\}))$ comprising the set of all probability distributions with support on $\Theta \times (\Psi - \{\psi_0\})$ [8, Chapter 2.9.1]. Since the complement space $\Theta \times \{\psi_0\}$ has zero measure, $\mathcal{V}(\Theta \times (\Psi - \{\psi_0\}))$ is identical to $\mathcal{V}(\Theta \times \Psi)$ almost everywhere. Therefore, we will continue to use the latter vacuous probability model in what follows.

We assume that our prior belief about the 'known' model ensemble $M(\theta, \psi_0),\ \theta \in \Theta$, i.e. more precisely, the set of models considered for our particular assessment, is described by $\nu(\theta, \psi_0)$. How should we combine this prior belief with the vacuous belief on the complementary unknown space? It seems to be a precondition of human agency that we assign non-zero probability to our conception of the 'real world' even though it exists on a space with zero measure. Thus, when it comes to considering the unknown, our prior belief on the space of all models will be degenerate,

$$\nu(\theta, \psi) \in p_0\ \nu(\theta, \psi_0)\ \delta(\psi - \psi_0) + (1 - p_0)\ \mathcal{V}(\Theta \times \Psi)\ , \quad (9)$$

where $\delta(\psi - \psi_0)$ denotes the Dirac measure which concentrates all probability mass on $\psi = \psi_0$, i.e., the set of models available to us. The probability $0 \leq p_0 \leq 1$ weighs our prior belief across the two different domains of knowledge, and may be associated with the prior level of confidence that the model ensemble $M(\theta, \psi_0),\ \theta \in \Theta$ can accurately describe the 'real' system features of interest. For $p_0 = 1$, we completely ignore the possibility that the accurate model may still be unknown. This choice reflects the closed world assumption underlying the standard application of Bayesian learning. In the other extreme, $p_0 = 0$, we are completely lost in the unknown, and cannot expect to learn anything from whatever data we receive. Here, we suggest to choose p_0 as to reflect a typical confidence level used in statistics, e.g., $p_0 = 0.95$. However, the choice of p_0 will not influence the posterior belief significantly (see Equation 15) as long as it is not set to the extreme values of 0 or 1.

Since we cannot talk in positive terms about what we do not know, we are not searching for the posterior belief $\nu(\theta, \psi|\hat{y}(n))$ on the space of all models, but rather for its marginal distribution $\rho(\theta|\hat{y}(n))$ on the subspace of known models. After receiving an observed time series $\hat{y}(n)$, Bayes' rule gives us the following expression for the marginal posterior belief:

$$\rho(\theta|\hat{y}(n)) = \frac{\int_\Psi \mathcal{L}(\theta, \psi; \hat{y}(n))\ \nu(\theta, \psi)\ d\psi}{\int_{\Psi \times \Theta} \mathcal{L}(\theta, \psi; \hat{y}(n))\ \nu(\theta, \psi)\ d\psi d\theta}\ . \quad (10)$$

We follow the usual practice to normalize the likelihood on the space of *known* models to one. Hence, we divide both the nominator and denominator by the maximum likelihood $\mathcal{L}(\theta', \psi_0; \hat{y}(n))$ that we find on the model ensemble $M(\theta, \psi_0),\ \theta \in \Theta$. Inserting the prior belief described by Equation (9) into above

[2] The assumption of time-discrete equations accounts for the numerical implementation of time-continuous differential equations. It shall also extend to other, e.g. spatial, dimensions if partial differential equations are concerned. Thus, we are thinking of computer models here.

expression, we then find

$$\rho(\theta|\hat{y}(n)) \quad \in \quad \frac{p_0\,\mu_{\mathcal{L}}(\theta) + (1-p_0)\,\mathcal{V}_{\mathcal{L}}(\Theta)}{\int_\Theta\,(p_0\,\mu_{\mathcal{L}}(\theta) + (1-p_0)\,\mathcal{V}_{\mathcal{L}}(\Theta)\,)\,d\theta}\,, \quad (11)$$

$$\mu_{\mathcal{L}}(\theta) \quad := \quad \frac{\mathcal{L}(\theta,\psi_0;\hat{y}(n))}{\mathcal{L}(\theta',\psi_0;\hat{y}(n))}\;\nu(\theta,\psi_0)\,,$$

$$\mathcal{V}_{\mathcal{L}}(\Theta) \quad := \quad \int\limits_\Psi \frac{\mathcal{L}(\theta,\psi;\hat{y}(n))}{\mathcal{L}(\theta',\psi_0;\hat{y}(n))}\;\mathcal{V}(\Theta\times\Psi)d\psi\,,$$

where $\mathcal{V}_{\mathcal{L}}(\Theta)$ is the unknown set of marginals on Θ that emerge from multiplying all prior probability distributions on $\Theta\times\Psi$ with an unknown likelihood function. Note that it is not a vacuous probability model itself, since its elements are not normalized. However, this set of marginals is contained in the set of all probability distributions on Θ multiplied by the range of values covered by the likelihood ratio, i.e.

$$\mathcal{V}_{\mathcal{L}}(\Theta) \quad \subset \quad [0,\mathcal{L}^*(n)]\cdot\mathcal{V}(\Theta)$$

$$\text{with} \quad \mathcal{L}^*(n) \quad := \quad \max_{(\theta,\psi)\in\Theta\times\Psi}\frac{\mathcal{L}(\theta,\psi;\hat{y}(n))}{\mathcal{L}(\theta',\psi_0;\hat{y}(n))}\,, \quad (12)$$

Here, the zero lower bound of the interval accounts for the fact that there will certainly be a model with zero likelihood in the space of all models. Note that the nominator of $\mathcal{L}^*(n)$ describes the likelihood function on the entire model space prior to normalization, and therefore can take any value in $\mathbb{R}_0^+$. In the following, we will replace $\mathcal{V}_{\mathcal{L}}(\Theta)$ in the extended Bayes' rule (11) by its superset $[0,\mathcal{L}^*(n)]\cdot\mathcal{V}(\Theta)$ due to greater methodological convenience. This substitution will give us outer bounds on the set of posterior probabilities, but we assert that the associated information loss is minimal. As an example, consider the asymptotic case $n\to\infty$ for which the likelihood function will concentrate around the accurate model at the point (θ^*,ψ^*) $(\mathcal{L}(\theta,\psi;\hat{y}(n))\to\delta(\theta-\theta^*)\;\delta(\psi-\psi^*))$. Then, $\mathcal{V}_{\mathcal{L}}(\Theta)$ will contain only functions proportional to $\delta(\theta-\theta^*)$, which constitutes a considerably smaller set than the functions proportional to $\mathcal{V}(\Theta)$. However, since we are completely ignorant about the location of θ^*, we need to consider $\delta(\theta-\theta^*)$ for all possible values $\theta^*\in\Theta$, which coincides with the set of extreme points of $\mathcal{V}(\Theta)$.

The set of Dirac measures $\delta(\theta-\tilde{\theta})\;\delta(\psi-\tilde{\psi})$, $(\tilde{\theta},\tilde{\psi})\in\Theta\times\Psi$ comprises the extreme points of the vacuous probability model $\mathcal{V}(\Theta\times\Psi)$ on the entire model space. They are all we need to calculate the extreme points of the imprecise posterior probability given by Equation (11) [8, Theorem 8.4.8]. They also tell us that

$$0\le\int_\Theta\;V_{\mathcal{L}}(\Theta)\;d\theta\le\mathcal{L}^*(n)\,, \quad (13)$$

where the upper bound is achieved for the Dirac prior $\delta(\theta-\theta^*)\;\delta(\psi-\psi^*)$.

We are now in the position to separate the extended Bayes rule (11) into a term concerned with updating our prior belief on the model ensemble $M(\theta,\psi_0)$, $\theta\in\Theta$ (the original Bayes' rule), and a term that summarizes the contribution from the residual space of unknown models.

$$\rho(\theta|\hat{y}(n))\in$$

$$(1-\varepsilon(\lambda,p_0))\,\frac{\mu_{\mathcal{L}}(\theta)}{\int_\Theta\,\mu_{\mathcal{L}}(\theta)\,d\theta}+\varepsilon(\lambda,p_0)\,\mathcal{V}(\Theta)\,, \quad (14)$$

$$\varepsilon(\lambda,p_0) \quad := \quad \frac{(1-p_0)\,\lambda}{p_0\,\int_\Theta\,\mu_{\mathcal{L}}(\theta)\,d\theta+(1-p_0)\,\lambda}\,, \quad (15)$$

$$\text{with} \quad \lambda\in[0,\mathcal{L}^*(n)]\,.$$

Since the contamination $\varepsilon(\lambda,p_0)$ increases with λ, the most conservative posterior belief – encompassing the set of posterior probabilities for all possible choices of λ – is obtained in the limit $\lambda\to\mathcal{L}^*(n)$. Therefore, we focus in the following on the most conservative case, for which the vacuous probability model is mixed into the posterior belief with contamination $\varepsilon(\mathcal{L}^*,p_0)$. The ε-contamination model in Equation (14) has been investigated extensively in the context of robust Bayesian and imprecise probability approaches (see, e.g., [1, 5]). It is a very tractable model, since it can be easily characterized by its set of extreme points or its coherent lower probability which constitutes a belief function. Note that we can recover the standard case of Bayesian learning under the closed world assumption from Equations (14) and (15) by choosing $p_0=1$, implying $\varepsilon(\mathcal{L}^*,p_0)=0$. For $p_0\in(0,1)$, the 'contamination' $\varepsilon(\mathcal{L}^*,p_0)$ of our posterior belief will grow with increasing $\mathcal{L}^*(n)$ (see Equation 15). What can we say about $\mathcal{L}^*(n)$, and how will it behave as a function of our observations $\hat{y}(n)$?

In general, we expect the likelihood $\mathcal{L}(\theta,\psi;\hat{y}(n))$ to be largest at an unknown point (θ^*,ψ^*) where an accurate model S is located. The probability that it will be otherwise becomes infinitesimal as the number of observations $n\to\infty$. Hence, we assert that $\mathcal{L}^*(n)$ is obtained at the point (θ^*,ψ^*). Given the definition of an accurate model in the introduction, we know that $Y_t-M(\theta^*,\psi^*)=\epsilon_t\sim N(0,\sigma)$ at this point. Thus, for a given observation $\hat{y}(n)=(\hat{y}_1,...,\hat{y}_n)$, we construct a random variable

$$L^*(n) \quad := \quad \frac{\frac{1}{\sqrt{2\pi}^n\sigma^{2n}}\;\exp\left(-\frac{1}{2\sigma^2}\sum_{t=1}^n\epsilon_t^2\right)}{\mathcal{L}(\theta',\psi_0;\hat{y}(n))}\,,$$

$$= \quad \exp\left(-\frac{1}{2}\,(\sum_{t=1}^n\frac{\epsilon_t^2}{\sigma^2}-\hat{s}(\theta'))\right) \quad (16)$$

$$\text{with} \quad \hat{s}(\theta'):=\sum_{t=1}^n\frac{(\hat{y}_t-M(\theta',\psi_0)_t)^2}{\sigma^2}\,,$$

where the denominator (respectively the second term in the exponent) includes the likelihood of the 'best' model $M(\theta', \psi_0)$ (respectively the least square sum of its residual) in our ensemble of available models (compare Equation 12). Hence, our quantity of interest, i.e., the realization $\mathcal{L}^*(n)$ of $L^*(n)$, depends on the actual realization $(\hat{\epsilon}_1, ..., \hat{\epsilon}_n)$ as well as the residual $\hat{s}(\theta')$ of the 'best' model $M(\theta', \psi_0)$. While we can calculate $\hat{s}(\theta')$ after having received the observation $\hat{y}(n)$, we cannot access the realization $\hat{\epsilon}$ of the residual between $\hat{y}(n)$ and the unknown accurate model $M(\theta^*, \psi^*)$. We only know that $\epsilon \sim N(0, \sigma)$ is an iid Gaussian process, and its variance is distributed as χ^2:

$$s(n) := \sum_{t=1}^{n} \frac{\epsilon_t^2}{\sigma^2} \sim \chi^2_{n-1} \,. \tag{17}$$

Therefore, we only can try to derive a useful estimator $E(L^*(n))$ of $\mathcal{L}^*(n)$ from the asymptotic χ^2_{n-1} distribution of $s(n)$:

$$E(L^*(n)) = \exp\left(-\frac{1}{2}\,\left(E(s(n)) - \hat{s}(\theta')\right)\right) \tag{18}$$

Such an estimator $E(L^*(n))$ will be useful for our purpose, if it discriminates between the two cases where an accurate model S is contained in the ensemble, i.e., it exists $\tilde{\theta} \in \Theta$ with $S = M(\tilde{\theta})$, and where it is not. In the former case, we can assume for large numbers of observations that S will coincide with the 'best' model $M(\theta', \psi_0)$ which exhibits the maximum likelihood on the space of available models Θ. In the latter case ($S \neq M(\theta', \psi_0)$), we assert for large numbers n of observations that the residual $\hat{s}(\theta')$ will grow faster than any estimator $E(s(n))$ constructed from a χ^2_{n-1} distribution, i.e., $E(s(n)) - \hat{s}(\theta') \to -\infty$ for $n \to \infty$, and thus $E(L^*(n)) \to \infty$ and $\varepsilon(E(L^*), p_0) \to 1$.

It remains to investigate the asymptotic behavior for the case $S = M(\theta', \psi_0)$, for which the residual between the 'best' model and the data will also be a realization of an iid Gaussian process $N(0, \sigma)$. Hence $\hat{s}(\theta')$ will constitute a draw from the same χ^2_{n-1} distribution on which $E(s(n))$ is based. Since χ^2_{n-1} becomes approx. normal for $n \to \infty$, it can be seen that $s(n) - s(\theta')$ will also be approx. normally distributed with zero mean and variance $\to \infty$. This shows that the estimator $E(s(n))$ needs to be carefully chosen in order to avoid a situation where $\varepsilon(E(L^*), p_0)$ can take any value between 0 and 1, *if* $S = M(\theta')$. Therefore, we select a q-quantile of the χ^2_{n-1}-distribution

$$\int_0^{qs} \chi^2_{n-1}(s)\; ds := q \,,$$

as estimator $E(s(n))$. The quantile qs will be larger than $\hat{s}(\theta)$ with probability q, if the accurate model S is contained in the model ensemble. We use qs to define our estimator $E(L^*(n))$ of $\mathcal{L}^*(n)$ in Equation 15), i.e.,

$$E(L^*(n)) := e^{-\frac{n-1}{2}\left(\frac{qs}{n-1} - \frac{\hat{s}(\theta')}{n-1}\right)} \,. \tag{19}$$

Equation (19) constitutes the final building block for our extension of Bayes' rule that allows us to drop the closed world assumption. This open version of Bayes' rule is summarised by Equations (14), (15) (with $\mathcal{L}^*(n)$ replaced by the estimator $E(L^*(n))$), and (19). It should be noted that the extended Bayes' rule depends on the choice of confidence level q for the upper limit of the variance of the residual $\hat{\epsilon}$. This makes it clear that in our attempt to account for the space of unspecified models, we allowed classical statistics to enter our otherwise Bayesian approach through the backdoor. For large n, the introduction of a contamination term in the posterior belief amounts to a hypothesis test on our best model $M(\theta')$. In this case, $E(L^*(n))$ will jump rapidly from zero to a very large number, when the residual of our best model $M(\theta')$ crosses the upper limit qs at the q-confidence level (see Equation 19). This will cause the contamination term $\varepsilon(E(L^*), p_0)$ to jump from 0 to 1 (see Equation 15). Therefore, our choice of $E(L^*(n))$ can lead to strong fluctuations in the contamination term if the residual of the best model $M(\theta')$ is hovering around the upper limit qs. The responsiveness of the contamination term can be reduced by replacing the linear scaling of the exponent of $E(L^*(n))$ with increasing number of observations by a sublinear function. We suggest that this is most effectively done by using the scaling of the χ^2_{n-1} distribution for increasing degrees of freedom, and offer the following heuristic expression as an alternative choice:

$$E(L^*(n)) := e^{-\frac{1}{2}(qs-n+1)\,\left(\frac{qs}{n-1} - \frac{\hat{s}(\theta')}{n-1}\right)} \,. \tag{20}$$

4 Prediction with ε-contamination: Example of AR processes continued

We now put the conceptual framework developed in the previous section into operation for our example of AR processes. The setup is identical to what was described in Section 2. For applying our open version of Bayes' rule to this updating problem, we need to calculate the development of the contamination $\varepsilon(E(L^*), p_0)$ for time series of observations $\hat{y}(n)$ with increasing length. We do this for the random sample of 200 time series from Section 2, and for both choices of $E(L^*(n))$ proposed in Equations (19) and (20). We use a prior weight $p_0 = 0.95$ on our model ensemble $M(\alpha_1)$, $\alpha \in [-\bar{\alpha}, \bar{\alpha}]$, and choose a confidence level of $q = 0.99$ to determine the upper limit qs on the residual variance of the accurate model.

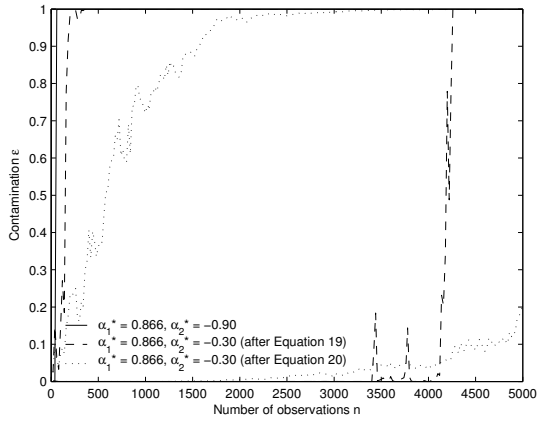

Figure 4: Behavior of lower and upper bounds of the 90% confidence interval for the ε-contamination. The ε-contamination for the models $\alpha_1^* = 0.866$ and $\alpha^2 = \{0, 0.06\}$ falls immediately to zero and stays there throughout the 5000 observations. In contrast, the ε-contamination for the model with $\alpha_2^* = -0.9$ jumps quickly to one (after 100 observations).

The lower 5% and upper 95% quantile limits (deduced from the sample of 200 time series) for the value of the contamination $\varepsilon(E(L^*), p_0)$ are shown in Figure 4. The contamination is zero for the cases in which standard Bayesian updating did well. Hence, in these cases our posterior belief about model accuracy and the associated prediction of the asymptotic distribution of system observations is identical to what we have found in Section 2. In the remaining two cases where standard Bayesian updating failed, the situation is markedly different. For $\alpha_2 = -0.9$, the contamination rapidly approaches $\varepsilon(E(L^*), p_0) = 1$, rendering our posterior belief vacuous after 100 observations at the latest. For the less extreme case $\alpha_2 = -0.3$, the increase in contamination is much slower, reflecting the results shown in Figure 1 that the posterior belief about the true propagator value remains in the vicinity of the AR(1)-equivalent propagator for several thousand observations. In that boundary case, the contamination term based on Equation (19) can fluctuate indeed strongly up to $n = 4000$ observations depending on the actual time series. The alternative contamination term based on Equation (20) offers a smoother response (see Figure 4), but on the downside responds slower to pick up the lack of accuracy in the model ensemble. We suggest that the proper choice of contamination term will depend on the application.

We now investigate the consequences of the growing contamination for the posterior belief in the case $\alpha_2^* = -0.9$. Our main question is whether the associated predictions of the asymptotic distribution of system observations can anticipate quickly the possibility of strong tails that was missed by standard Bayesian updating (see Figure 3). The analysis will also illustrate how the ε-contamination model can be used in statistical inference.

Due to the mixture with the vacuous probability model $\mathcal{V}(A_1)$, $A_1 = [-\bar{\alpha}, \bar{\alpha}]$, the posterior belief as expressed in Equation (14) is imprecise. Since it includes Dirac measures, the set of posterior probabilities can be depicted as a band of cumulative distributions (CDFs), but not as density band. The upper and lower CDFs set up by the ε-contaminated posterior belief model are given by (using the shorthand $\varepsilon^* := \varepsilon(E(L^*), p_0)$):

$$\underline{F}(\alpha_1; \hat{y}(n)) = (1-\varepsilon^*) \int_{-\bar{\alpha}}^{\alpha_1} \rho(\alpha_1'|\hat{y}(n))\, d\alpha_1' + \varepsilon^*\, H(\alpha_1 - \bar{\alpha}) \quad (21)$$

$$\overline{F}(\alpha_1; \hat{y}(n)) = (1-\varepsilon^*) \int_{-\bar{\alpha}}^{\alpha_1} \rho(\alpha_1'|\hat{y}(n))\, d\alpha_1' + \varepsilon^*\,, \quad (22)$$

where H denotes the Heavyside function which adds the missing probability mass at the upper bound of the support for α_1. It is important to note that the distribution band defined by $\underline{F}(\alpha_1; \hat{y}(n))$ and $\overline{F}(\alpha_1; \hat{y}(n))$ is not equivalent to the ε-contamination model, but a true superset of it. Every distribution contained in the ε-contamination model will be contained in the distribution band, but not vice versa

Figure 5 shows the change of posterior distribution band with increasing number of observations of an AR(2) process with $\alpha_2^* = -0.9$. It can be seen that the imprecision in the posterior belief increases quickly with observations. After $n = 80$ observations the posterior belief becomes vacuous, and the associated distribution band would cover the entire graph. At this point, any predictive power has been lost.

We take a closer look on the prediction of the asymptotic distribution of system observations $\rho(y|\hat{y}(n))$ in Figure 6. The prediction is again imprecise, and its lower and upper bound can be calculated on the basis of Equation (8) by recalling that these bounds are set up by the Dirac measures contained in the vacuous probability model $\mathcal{V}(A_1)$. Those Dirac measures allocate the probability mass carried by the contamination $\varepsilon^* := \varepsilon(E(L^*), p_0)$ at a value of α that minimizes respectively maximizes the contribution to $\rho(y|\hat{y}(n))$.

$$\underline{\rho}(y|\hat{y}(n)) = (1-\varepsilon^*) \int_{-\bar{\alpha}}^{\bar{\alpha}} \rho(y|\alpha_1)\, \rho(\alpha_1|\hat{y}(n))\, d\alpha_1 + \varepsilon^* \min_{\alpha_1 \in [-\bar{\alpha}, \bar{\alpha}]} \rho(y|\alpha_1)\,. \quad (23)$$

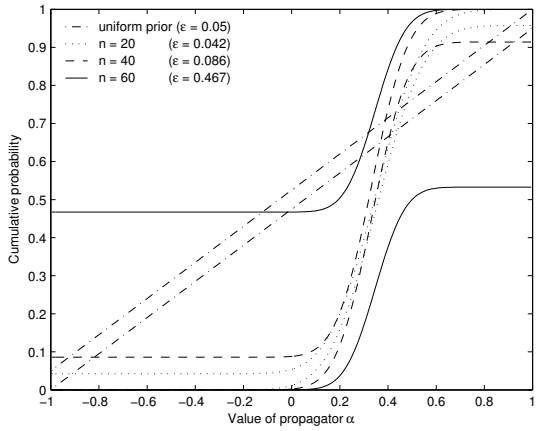

Figure 5: Cumulative posterior distribution bands for the propagator α_1 learned from a realization of the AR(2)-process with $\alpha_1^* = 0.866$ and $\alpha_2^* = -0.9$. The distribution band for $n = 80$ observations is vacuous and covers the entire graph.

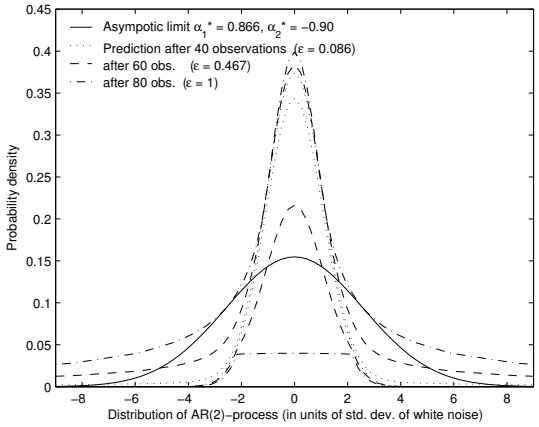

Figure 6: Predictions for the asymptotic probability distribution of system observations for the AR(2)-process with $\alpha_1^* = 0.866$ and $\alpha_2^* = -0.9$. The quickly growing ε-contamination destroys the predictive accuracy of the model ensemble after 80 observations.

$$\overline{\rho}(y|\hat{y}(n)) \;=\; (1-\varepsilon^*)\int_{-\bar{\alpha}}^{\bar{\alpha}} \rho(y|\alpha_1)\;\rho(\alpha_1|\hat{y}(n))\;d\alpha_1 + \varepsilon^* \max_{\alpha_1\in[-\bar{\alpha},\bar{\alpha}]} \rho(y|\alpha_1)\,. \qquad (24)$$

Figure 6 shows the predicted bounds on the asymptotic distribution of system observations. It can be seen that the imprecision in the prediction grows quickly, and its range covers the tails after $n = 60$ observations. The full asymptotic distribution is contained in the predicted range after $n = 80$ observations when the posterior belief has become vacuous. At this point, the analyst employing our open version of Bayes' rule will have noticed that it is time to engage in type (I) learning as defined in the introduction, and to try to extend the set of models that she considers (e.g., to the set of all AR(2)-processes).

5 Conclusions

We have presented a framework for updating belief about prediction accuracy across an ensemble of available models using observations of the system that those models are supposed to predict. While following the Bayesian approach to learning, we have dropped the assumption that an accurate model – predicting the system observations up to an iid Gaussian process – is contained in the model ensemble with certainty as would be required by Bayes' rule in its conventional form. This is an achievement because the closed world assumption can lead to spurious beliefs about model accuracy and false predictions, as was demonstrated with an example of AR processes. By drawing on elements of imprecise probability theory and the knowledge of asymptotic distributions for large samples, we established an open version of Bayes' rule that extends its consideration to the unknown space of unspecified models, and thus includes the possibility that an accurate model might not be contained in the set of available models. Under the open Bayes' rule, the posterior belief takes on the form of an ε-contamination model, where the contamination ε is updated along with the prior belief on the set of available models. A growing contamination will indicate limited accuracy of the entire model ensemble, and will eventually lead to a vacuous posterior belief. In this way, false predictions due to limitations of the models under consideration can be avoided as was demonstrated again with an example of AR processes.

Also the method presented here has proven successful – in a stylized example – to discriminate between cases where standard Bayesian updating works well, and where it fails, this paper can offer only a proof of concept. It will require further research to investigate how the open Bayes' rule works in practice. In a next step we intend to apply it to the comparison of the 20th century temperature record with a simple climate model parameterized in terms of key quantities influencing the temperature response [6]. As a matter of concern, we will have to analyze whether the open Bayes' rule in its current form is too discriminative as it may discount every model that cannot explain the data up to an additive Gaussian process. In practice, such a strong requirement is hard to fulfill, not the least because the observations might be overlaid by a systematic non-Gaussian error due to changing

measurement practices over time. This, however, is a general problem for model validation and model-based prediction, and by no means limited to the application of the open Bayes' rule. In these cases it may be unavoidable to attempt adding and updating a positive model for the discrepancies between actual measurements and 'ideal' measurements (and, if necessary, between actual model and 'ideal' model) to the analysis as has been proposed by [4]. In any case, the open Bayes' rule can be a valuable tool to assess whether such additions are bearing fruit.

A Calculation of the likelihood for the AR(1) propagator α

Let an AR(1) process be defined by $X_1 = \xi_1$, $X_t = \alpha X_{t-1} + \xi_t$, $t \geq 2$, and $\xi_t \sim N(0, \sigma_\xi)$. Estimators $\alpha(n)$ for the propagator α and $s(n)$ for the variance of the AR(1) process are defined in terms of the observation $Y(n) = (X_1, ..., X_n)$ after n time steps

$$s(n) = \frac{1}{n-1} \sum_{t=1}^{n} X_t^2 \,, \tag{25}$$

$$\alpha(n) = \frac{1}{n-1} \frac{\sum_{t=2}^{n} X_t X_{t-1}}{s(n)} \,. \tag{26}$$

Here, we deviate from the standard choice of these estimators [7] by omitting the subtraction of the sample mean ($1/n \sum_{t=1}^{n} X_t \to 0$ for $n \to \infty$) in the estimator for the variance, and by inflating the estimator for the propagator by $n/(n-1)$. The reason for this is that the distribution of those estimators for a given choice of α, σ_ξ can be calculated easily:

$$\begin{aligned} &\rho(\alpha(n), s(n) | \alpha, \sigma_\xi) \\ &\quad \sim e^{-\frac{1}{2\sigma_\xi^2}\left(\sum_{t=2}^{n} (X_t - \alpha X_{t-1})^2 + X_1^2\right)} \\ &\quad = e^{-\frac{(n-1)\ s(n)}{2\sigma_\xi^2}\left(1 + \alpha^2 - 2\alpha\alpha(n) - \frac{\alpha^2\ X_n^2}{(n-1)\ s(n)}\right)} \,. \end{aligned}$$

Once we have observed an actual realization $\hat{y}(n) = (\hat{y}_1, ..., \hat{y}_n)$, fixing the values of the estimators at $\hat{\alpha}(n)$ and $\hat{V}(n)$, we can calculate a likelihood function $\mathcal{L}(\alpha; \hat{y}(n)) \sim \rho(\hat{\alpha}(n), \hat{V}(n) | \alpha, \sigma_\xi)$ for the propagator α of the underlying AR(1) process (assuming that σ_ξ is known). With

$$\hat{\sigma}(n) := \frac{\sigma_\xi}{\sqrt{(n-1)\ \hat{s}(n)}} \,, \tag{27}$$

$$\hat{\beta}(n) := \frac{\hat{y}_n^2}{(n-1)\ \hat{s}(n)} \,, \tag{28}$$

we find

$$\begin{aligned} \mathcal{L}(\alpha; \hat{y}(n)) &\sim e^{-\frac{1}{2\hat{\sigma}(n)^2}\left((\alpha - \alpha(n))^2 - \hat{b}(n)\alpha^2\right)} \\ &\sim N\left(\frac{\hat{\alpha}(n)}{1 - \hat{\beta}(n)}, \frac{\hat{\sigma}(n)}{\sqrt{1 - \hat{\beta}(n)}}\right) \,. \end{aligned}$$

Acknowledgment: Elmar Kriegler is supported by a Marie Curie Outgoing International Fellowship funded by the European Commission under the Sixth Framework Programme (Contract #MOIF-CT-2005-008758). Infrastructure for his research was partly provided by the Climate Decision Making Center (CDMC) located in the Department of Engineering and Public Policy. This Center has been created through a cooperative agreement between the National Science Foundation (SES-0345798) and Carnegie Mellon University.

References

[1] J. O. Berger. An overview of robust Bayesian analysis. Technical Report #93-53C, Purdue University, 1993.

[2] D. Draper. Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. B*, 57:45–97, 1995.

[3] M. Goldstein and J. C. Rougier. Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM J. Sci. Comput.*, 26:467–487, 2004.

[4] M. Goldstein and J. C. Rougier. Reified Bayesian modelling and inference for physical systems. *J. Stat. Plan. and Inf.*, forthcoming as discussion paper, 2007.

[5] T. Herron, T. Seidenfeld, and L. Wasserman. Divisive conditioning: Further results on dilation. *Philosophy of Science*, 64:411–444, 1997.

[6] E. Kriegler. *Imprecise probability analysis for integrated assessment of climate change*. PhD thesis, University of Potsdam, 256pp, 2005.

[7] H. von Storch and F. W. Zwiers. *Statistical analysis in climate research*. Cambridge University Press, Cambridge, 1999.

[8] P. Walley. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London, 1991.

[9] P. Walley. Inference from multinomial data: Learning about a bag of marbles. *J. Roy. Stat. Soc. B*, 58:3–57, 1996.