Held, Hermann; Kriegler, Elmar; Augustin, Thomas

Working Paper

Bayesian learning for a class of priors with prescribed marginals

Discussion Paper, No. 488

Provided in Cooperation with:
Collaborative Research Center (SFB) 386: Statistical Analysis of discrete structures - Applications in Biometrics and Econometrics, University of Munich (LMU)

Suggested Citation: Held, Hermann; Kriegler, Elmar; Augustin, Thomas (2006) : Bayesian learning for a class of priors with prescribed marginals, Discussion Paper, No. 488, Ludwig-Maximilians-Universität München, Sonderforschungsbereich 386 - Statistische Analyse diskreter Strukturen, München,
https://doi.org/10.5282/ubm/epub.1856

This Version is available at:
https://hdl.handle.net/10419/31157

Bayesian Learning for a Class of Priors with prescribed Marginals

Hermann Held[*], Elmar Kriegler[†], and Thomas Augustin[‡]

September 6, 2006

Abstract

We present Bayesian updating of an imprecise probability measure, represented by a class of precise multidimensional probability measures. Choice and analysis of our class are motivated by expert interviews that we conducted with modelers in the context of climatic change. From the interviews we deduce that generically, experts hold a much more informed opinion on the marginals of uncertain parameters rather than on their correlations. Accordingly, we specify the class by prescribing precise measures for the marginals while letting the correlation structure subject to complete ignorance. For sake of transparency, our discussion focuses on the tutorial example of a linear two-dimensional Gaussian model. We operationalize Bayesian learning for that class by various updating rules, starting with (a modified version of) the generalized Bayes' rule and the maximum likelihood update rule (after Gilboa and Schmeidler). Over a large range of potential observations, the generalized Bayes' rule would provide non-informative results. We restrict this counter-intuitive and unnecessary growth of uncertainty by two means, the discussion of which refers to any kind of imprecise model, not only to our class. First, we find our class of priors too inclusive and, hence, require certain additional properties of prior measures in terms of smoothness of probability density functions. Second, we argue that both updating rules are dissatisfying, the generalized Bayes' rule being too conservative, i.e., too inclusive, the maximum likelihood rule being too exclusive. Instead, we introduce two new ways of Bayesian updating of imprecise probabilities: a "weighted maximum likelihood method" and a "semi-classical method." The former bases Bayesian updating on the whole set of priors, however, with weighted influence of its members. By referring to the whole set, the weighted maximum likelihood method allows for more robust inferences than the standard maximum likelihood method and, hence, is better to justify than the latter. Furthermore, the semi-classical method is more objective than the weighted maximum likelihood method as it does not require the subjective definition of a weighting function. Both new methods reveal much more informative results than the generalized Bayes' rule, what we demonstrate for the example of a stylized insurance model.

[*]Potsdam Institute for Climate Impact Research, PO Box 60 12 03, D-14412 Potsdam, Germany, held@pik-potsdam.de

[†]Potsdam Institute for Climate Impact Research, PO Box 60 12 03, D-14412 Potsdam, Germany, and Carnegie Mellon University, Pittsburgh, USA, elmar@cmu.edu

[‡]Department of Statistics, University of Munich, Ludwigstr 33, D-80539 Munich, Germany, thomas@stat.uni-muenchen.de

Keywords Generalized Bayes rule, imprecise probabilities, known marginals, maximum likelihood update, modeling expert opinions, robust Bayesians, unknown correlation structure, updating under complex uncertainty.

1 Introduction

Since the work of Walley [21], it is increasingly recognized that, most often, subjective knowledge is better characterized by imprecise rather than precise (i.e., standard) probability measures, whereby it could be shown (see [21], Theorem 3.3.3) that any imprecise model can be interpreted as a class of precise, traditional probability measures.

Here we investigate a particular imprecise model – including its Bayesian updating – that shall represent epistemic uncertainty on a multi-dimensional parameter space for which an expert is able to specify any of the marginals in terms of probability measures, however, refuses to deliver any further information, in particular not on the correlation structure among the parameters. A couple of authors have investigated properties of such classes [3, 6, 10]. Some of the properties of those classes are also resembled by the multivariate possibility measure [7] introduced in [13]. For that possibility measure, instead of an explicit correlation structure, [13] prescribe spherical symmetry in a rather heuristic manner.

However we are not aware of any systematic investigation of *Bayesian updating* of such an imprecise model. For that reason we discuss four different types of Bayesian updating of our imprecise model: two extreme versions already to be found in the literature (a slightly modified version of Walley's Generalized Bayes' rule [21] that we denote by GBR in the following, and Gilboa's and Schmeidler's maximum likelihood update rule [12]) as well as to new hybrides that we introduce in this article and that we regard as more convincing in certain respects.

We started our investigations as we frequently observed statements by climate model developers or users that claim a lot of knowledge on individual model parameters related to specific physical processes, however, feel much less able to give any prior knowledge on the way the parameters must interact in order to obtain a reasonable model climate state. The situation is similar to constructors and users for other models used in the climatic change assessment. We put our impression on more objective grounds by setting up a questionnaire accordingly, answered by half a dozen model users (Section below). Here we would like to push the discussion "knowledge on marginals" versus "knowledge on correlations" to the extremes in order to sharpen the discussion and choose a precise probability measure for the marginals.

To our impression, such type of investigation is desperately needed as in the climate modeling community – as well as in many other research communities – the issue of prior knowledge on parameter correlations is the most neglected issue, always being represented by uncorrelated measures. On the level of individual parameters, quite the contrary, there are suggestions that come close to something like robust Bayesian analysis in a rudimentary manner (in [9] two sorts of priors are investigated) or even explicit treatments in terms of imprecise models [15].

However, the silent assumption that an expert uninformed about correla-

tions is best represented by uncorrelated parameters, seems to mimic – to our taste – the "objective Bayesians' " assumption that the situation of complete ignorance on a single parameter is best represented by a non-informative prior. Quite the contrary, we follow Walley that there is no such thing like "objective Bayesianism" and that situations of ignorance must be captured by imprecise models. Recently, the approach of Walley and others was supported by [14] who derived the neural basis of decision-making when probabilities are uncertain because of missing information (ambiguity). Therefore we proceed in setting up an imprecise model for correlations.

It is apparent that the prior correlation structure will have a strong influence on the result of Bayesian updating, in particular in high dimensions. E.g., for a non-informative likelihood and identical Gaussian marginals, the standard deviation of the posterior will scale with $\sim \sqrt{n}$ for the uncorrelated case, while with $\sim n$ for the perfectly correlated case (n denoting the number of parameters). Some first heuristic attempts to reflect such effects are made in [19] utilizing a "correlated" and an "uncorrelated" prior, however, no systematic theory of how to set up an adequate imprecise model including Bayesian learning has been developed up to now. The present article aims at closing that gap.

The article is organized as follows: in the upcoming Section, we display the outcome of our expert elicitation. In the following Section, we motivate a particular transfer function for our tutorial example that relates model parameters and a potential observation, i.e., data input. Furthermore, we select the simplest possible likelihood function then used to study Bayesian updating. An overview on the updating methods used throughout the article will be given. In Section 5, we sketch a stylized insurance situation as a potential application of the imprecise probabilities to be derived afterwards. The insurance example ought to reveal a decision problem that allows to illustrate as well as sharpen the interpretation of the various updating rules discussed below. In Section 6 we apply any of the concepts introduced before. Two new updating rules appear as much more convincing than the scheme that traditionally represents the other end (as against GBR) of the spectrum of updating methods: the *maximum likelihood update method*. Finally, in Section 7 we summarize the previous results and discuss their consequences for the future uncertainty analysis of climate models.

2 A questionnaire on the structure of prior knowledge

Above we claimed that modelers in the context of climatic change typically know more on individual parameters than on their correlations. In order to underpin that impression we developed a questionnaire on the structure of prior knowledge on model parameters. We considered users of the following models:

- The climate model of intermediate complexity CLIMBER-2 (corresponding to a system of more than 1000 ordinary differential equations) [17], [11],
- the complex ocean model MOM-3 [16],
- the dynamic vegetation model LPJ [5],

- the model of endogenous economic growth, MIND [8].

We asked them – among other items – whether for a given uncertain model parameter b, the expert would be willing to give probabilistic information in terms of a density function. *Any* of them would do so. Then we checked for the betting behavior on quantiles of b. We found certain discrepancies with the density function specified before that may suggest to utilize imprecise measures on b. However, these aspects are not central importance here and will be published elsewhere, together with the questionnaire. Here we would like to focus on the central question:

How do you judge the quality of your subjective knowledge on b compared to the quality of your knowledge on correlations of b with other unknown parameters?

Most importantly in the context of this article, we obtained the following answers:

- "For some of the parameters, I know about the sign of correlations, however my knowledge is less precise than that on individual parameters."
- "Knowledge on b is higher than knowledge on correlations with other parameters (on some specific parameters, it might be different)."
- "The parameter knowledge is relative good but the knowledge on correlation with other parameters in some cases is only an idea."
- "I have not considered the possibility of correlations."
- "Never thought about that point."
- "It would be impossible to specify anything on correlations."
- "Absolutely no comment on correlations."

We would like to stress again that any of those statements were made by an expert that at the same time was willing to specify prior knowledge in probabilistic terms on individual parameters!

Therefore we find it worthwhile to consider the somewhat extreme case of imprecise prior knowledge with prescribed marginals (i.e. knowledge on individual parameters) and fully unconstrained correlation structure.

3 Specification of the updating problem

First of all we would like to introduce the notation for traditional Bayesian learning (updating) from data y, given a single prior probability measure P:

$$P_{\text{apost}}(x) = \frac{P(x)\ L(x)}{\int dx'\ P(x')\ L(x')}, \tag{1}$$

$L(x) \equiv P(y|x)$ denoting the likelihood function for the uncertain (multivariate) parameter x. (Here we use "P" synonymously for the probability measure as well as for the accompanying density when applied to elements of the $\mathbb{R}^n$.)

3.1 Specification of a transfer function

For our purpose, the simplest non-trivial transfer functions that relates a multivariate model parameter to model output, to be compared to a real-world entity, is given by

$$F : \mathbb{R}^2 \to \mathbb{R}, \quad (x_1, x_2) \to \kappa x_1 + x_2, \quad \kappa \in \mathbb{R}. \tag{2}$$

In terms of a climate model, we may identify x_1 with a key uncertain model parameter such as climate sensitivity in reduced-form climate models and $\kappa\ x_1$ with a model output of interest, such as global mean temperature in the year 2100. Furthermore, the model constructor as well as the modeling community know that the model structure is not perfect and that there may be a systematic deviation x_2 of model output and observational data y. Hence, F, rather than $\kappa\ x_1$ ought to be compared to an observation y.

In order to keep the discussion as transparent as possible, we will focus on a single quantity of interest the posterior probability of which shall be derived in the following. As such a quantity, we choose the *probability of ruin* $P^*_{\text{apost}} = \int_{x_1^*}^{\infty} dx_1 \int_{-\infty}^{\infty} dx_2\ P_{\text{apost}}(x_1, x_2)$, i.e., the probability that x_1 (e.g. climate sensitivity) is larger than a certain threshold value x_1^*. In case a prior opinion P, κ and L were specified and y were observed, above formulas would uniquely reveal the aposteriori distribution, and thereby the desired P^*_{apost}.

However, within an imprecise setting, we do not deal with a single prior P, but with a whole class $\mathcal{P}$ of priors. We define our class of priors with prescribed marginals as

$$\mathcal{P} := \left\{ P \mid \forall_{x_1} \int dx_2\ P(x_1, x_2) = P_1(x_1), \ \ \forall_{x_2} \int dx_1\ P(x_1, x_2) = P_2(x_2) \right\}, \tag{3}$$

P_1, P_2 specified by the expert. Note that any element P of that class is – by construction – normalized to 1. We could focus on specific choices for P_1, P_2, L already here. However, some major innovative aspects of this article refer to Bayesian updating of imprecise priors. The following subsections structures various cross-relations within Bayesian updating that are independent of the choice of P_1, P_2, L and should be given in general terms in advance. We will demonstrate any of the upcoming ideas for a specific example in Section 6.

3.2 Generalizations of Bayes' formula

We now specify which types of generalizations of Bayes' formula we will employ furtheron when we need to update of $\mathcal{P}$ rather than a single P. Any sort of generalization, newly introduced in this article, will observe the following steps:

1. Select a subset $\mathcal{P}' \subset \mathcal{P}$ according to a *preselection rule*, yet to be specified.
2. Apply Bayes' formula to any of the members of $\mathcal{P}'$, thereby assembling a class of aposteriori distributions.
3. Extract the probability of interest (e.g., of crossing a certain threshold, the *probability of ruin*) for any of the posteriors.

4. Condense the so derived sets into final quantity of interest. Frequently used are the "inf" or the "sup" - operation, representing the optimist's or the pessimist's view on, e.g., the probability of ruin.

The updating rules discussed below will differ with respect to the first and the last step.

1. "Modified generalized Bayes rule (GBR)": No preselection is applied, i.e., $\mathcal{P}' := \mathcal{P}$. On the class of the posteriors, the inf- and the sup-operation are applied.

2. The maximum likelihood update rule (after [12]) preselects those priors that optimize the prior probability (density) for the measurement y, i.e. priors that have the maximum likelihood, given y. The method completely disregards expert opinions that have not foreseen the measurement y with maximum probability. In the end, the inf- and the sup-operation are applied. The maximum likelihood update method somewhat resembles Dempster's method. Although it is not unrealistic to discount expert opinions that are at odds with observations, we find it unconvincing to *completely* dismiss opinions just because they have not foreseen the measurement with maximum probability. In particular, our discomfort aims at the exclusion of those opinions that have missed the maximum by just an infinitesimal amount.

3. For that reason, we introduce a derivative of the maximum likelihood update method, the *weighted maximum likelihood update method*. We require that any of the prior opinions get a chance, i.e., $\mathcal{P}' := \mathcal{P}$. However, then we classify the members of $\mathcal{P}'$ in terms of prior probability of y and "linearly weight in influence of the according level sets" which shall be specified precisely below (see Eqs. 5 to 7).

   Let for the moment $\mathcal{P}$ be of finite power I, i.e., $\mathcal{P} = \{P_1, ..., P_I\}$ (the infinite case follows directly from that). Let $W : \mathcal{P} \to \mathbb{R}_0^+$ denote the probability of y for any prior P, i.e., for given y,

$$\forall_{P\in\mathcal{P}} \quad W(P) := \int dx\ P(x)\ L(x) = \int dx\ P(x)\ P(y|x). \tag{4}$$

   Let $\{w_1, ..., w_J\} := W(\mathcal{P})$ the set of weights generated from $\mathcal{P}$ and $P^*_{\text{apost}}(P)$ the posterior probability of the quantity of interest, given the prior P.

$$\mathcal{P}_j := \{P \in \mathcal{P} | W(P) = w_j\}, \quad j = 1, ..., J, \tag{5}$$

$$\underline{P}^*_{\text{apost.wm}} := \frac{\sum_{j=1}^{J} w_j \cdot \inf_{P\in\mathcal{P}_j} (P^*_{\text{apost}}(P))}{\sum_{j=1}^{J} w_j}, \tag{6}$$

$$\overline{P}^*_{\text{apost.wm}} := \frac{\sum_{j=1}^{J} w_j \cdot \sup_{P\in\mathcal{P}_j} (P^*_{\text{apost}}(P))}{\sum_{j=1}^{J} w_j}. \tag{7}$$

This new method would reveal results identical to those obtained from the standard maximum likelihood update if $w_1 \neq 0$, $w_2, ..., w_J = 0$.

4. The *semi-classical method* presupposes that the decision-maker is willing to pool her or his risk with equivalent imaginated potential future decision situations. Then as a preselection, we consider a certain classical volume of confidence – to be further discussed in Section 4 – within $\mathcal{P}$, depending on the measurement y. In contrast to the maximum likelihood rule, this preselection rule decides for any element of $\mathcal{P}$ solely on the basis of the element's relation to the measurement y – there is no comparative element involved (in terms of a weighting function). The remaining three learning steps (according to the list above) are then as for GBR or the maximum likelihood update method. In preparation of the upcoming paragraph, we note the following: this semi-classical rule does not imply the use of a comparative element, neither in the first, nor in the last learning step.

   In order to guarantee a coherent interpretation of this mixed classical-Bayesian procedure, that may turn out to be controversial, we will embed the method into the decision problem by a nesting-formula (see Section 4).

We will refer to these four generalizations of Bayes' formula as "learning rules" in the following.

3.3 Interpretation of overly inclusive prior classes

Finally we would like to note a further conceptual difficulty that needs to be addressed when dealing with classes of priors: our (stylized) class may be too large for a particular application, i.e. it may contain priors that correspond to incompetent expert opinions. These opinions may drastically distort the inferred upper (lower) probabilities. There are two ways how to deal with such a "contaminated class":

1. If the main sources of contamination are known, one simply would add a *filter* to the preselection step that eliminates unrealistic priors. In the following Subsection we will suggest a catalogue of such additional filters that should be observed in standard applications. E.g., we will argue that only those priors should be considered further that come with a density whose gradient does not transgress a certain norm.

2. For those updating rules that preselect element-by-element, individually-based out of $\mathcal{P}$, hence do not involve a comparative step (as maximum likelihood update does) and that involve a "sup" ("inf") - operation in the condensation step, the following obvious theorem holds:

   The upper (lower) probability derived from the overly inclusive class of priors serves as an upper (lower) boundary of the upper (lower) probability derived from the correct class.

   We note that GBR and the semi-classical rule are of that type. For both rules, the theorem conveniently implies that we are always on the safe side (i.e. we do not add spurious information) when we include also those priors we are not sure about yet.

Quite the contrary, for the (weighted) maximum likelihood update method, the probability interval derived from an overly inclusive class must not be interpreted as outer boundary of the correct probability interval. Therefore the (weighted) maximum likelihood update method can be used only *after* we have finally decided for any prior whether it should enter the class or not.

To stress this point, quite important for practical applications, inference from an overly inclusive class by the (weighted) maximum likelihood update rule is useless: the (weighted) maximum likelihood update rule tries to weigh expert opinions with respect to the prior probability with which they had anticipated the observation. If one, e.g., defined a set of priors $P_1, ..., P_M$ by "experts", who for any $i \in \{1, ..., M\}$, assign 100% chance to lottery result y_i and zero to any y_j with $j \neq i$, then accidentally opinion P_i would be highlighted if y_i was measured. So a lot of trust would be given to opinion P_i although it was just chosen for trivial reasons and not because it was characterized by higher apriori competence. This situation somewhat resembles statements notoriously outlined in popular media saying that a particular astrologist was right as he or she had correctly predicted event y_i. All the other astrologists having had predicted $y_j, j \neq i$, are not mentioned.

If we instead filter out false priors that would have two effects: GBR would become more informative and we were allowed to use the (weighted) maximum likelihood update method.

3.4 Further constraints on the class of priors

We now ask what an expert generically would be able to hold an informed opinion on, in order to narrow down the class of priors:

1. The priors should be uni-modal.

2. We assume that the typical 1D (i.e., marginal) resolution over which an expert can have an informed opinion about, reads dx_1 (here "1" for "1D") if the typical dimension of the problem is Δx (in our previous examples, $\Delta x \approx 1$). This implies that an expert can distinguish $N_1 \approx \Delta x / dx_1$ items. Our requirement is equivalent to Walley's "bounded derivative model" [23] and shall be called *gradient filter* in the following.

3. This prescription needs to be generalized to a n-dimensional parameter space.

   (a) A possible generalization that would lead to a particularly large prior class is obtained by allowing for a resolution in terms of cubes of length dx_1, i.e., $N_n \approx N_1^n$.

   (b) The other extreme may require that $N_n \approx N_1$.

   We can connect both extreme cases by $N_n := N_1^{\beta n + (1-\beta)}$, hence we construct the linear hull of the exponents of both cases, $\beta \in [0, 1]$. Such connection may turn out necessary as both extreme cases display dissatisfying features:

(a) Let $\beta = 0 \Rightarrow N_n = N_1$. As $N_n = \Delta x^n / dx_n^n$ (with dx_n denoting the length of the edge of the n-dimensional cube), we observe: $\lim_{n\to\infty} dx_n = \lim_{n\to\infty} \Delta x / N_n^{1/n} = \Delta x$. This demonstrates that the expert may not have much knowledge left on the nD parameter space, measured in terms of 1D information dx_n.

(b) On the other hand, $\beta = 1 \Rightarrow N_n = N_1^n$ would require a prior competence of the expert, exponentially growing with dimension, that seems unrealistic as well.

(Both phenomena root in the "curse of dimension.") Hence, there is urgent need for an expert elicitation, designed to obtain a meaningful intermediate value for β. For the time being we derive the consequences of various values for β.

Once β has been decided on, the current (third) prescription on prior distributions requires that the modulus of the distribution's gradient was smaller than $1/dx_n^{n+1} = \Delta x^{-(n+1)} \cdot N_1^{\beta n+1+(1-\beta)/n}$.

3.5 Strategy for the implementation of various learning rules in combination with filtering options

In this article we consider the simplest non-trivial choice of the class of priors: we require that any prior should be a bivariate Gaussian with the marginals $P_1 \equiv P_2 \equiv N(\mu, \sigma^2)$. Hereby $N(\mu, \sigma^2)(.)$ denotes a Gaussian of mean μ and standard deviation σ, i.e. variance[1] σ^2.

We sacrifice generality for an analytically elegant and transparent implementation of the otherwise intricate and potentially only numerically accessible unimodality filter. While we can rest on some tradition of focussing on parameterized classes of priors in the literature that deals with robust Bayesian analyses and imprecise models [2, 18, 22], we would like to point out that our choice is for purely pragmatic reasons and therefore does not attempt to be the most adequate model for the experts' knowledge. We expect that future sophisticated expert elicitations will find certain parameter-free models most adequate.

On that class, four learning rules on are at disposal. Any of these learning rules could be combined with choices of additional preselection filters (outlined in the preceding Subsection). This opens a diversity of combinations that we could investigate in this article.

The case appears even richer once we have noted that the semi-classical method does not only serve as an independent learning rule, but could also be interpreted as additional preselection rule for the other three learning rules. When we had previously listed the semi-classical method as a forth learning rule, it effectively served as preselection filter for the GBR.

Hence, when investigating potential classes in upcoming Sections, we could tackle any tensor element of the 3-rank "case-tensor," characterized by the indices

1. "gradient filter" on or off,

[1] For a multivariate application, the first entry would represent a vector of means, the second the symmetric covariance matrix.

2. "semi-classical filter" on or off,

3. Bayesian learning according to GBR, standard maximum likelihood update, or weighted maximum likelihood update.

In order to keep the problem practical, we focus on some key combinations, according to the following strategy:

- We investigate the most inclusive class first, i.e., both of the two filters are switched off. This appears as attractive as no additional information for the setting of the filters (such as the value for β) needs to be assumed, and from the above Theorem we know that for GBR, no spurious information is added in case our class is overly inclusive.

- The following steps are driven by the findings outlined in the following Sections. In order to give an overview on the strategy, we highlight some aspects already here: among other things, we will find that GBR with all filters switched off, does not reveal very informative results.

- In addition, we test two versions of the gradient filter, revealing three cases in total (when also considering the "switched-off case").

- For any of these three cases, we vary the learning rule (GBR vs. (weighted) maximum likelihood update) and the semi-classical filter. As both the maximum likelihood update rule and the semi-classical filter are intricate in terms of interpretation, we omit combining the two in this article (we omit the two combinations ("standard maximum likelihood update" *and* "semi-classical filter"), ("weighted maximum likelihood update" *and* "semi-classical filter"), although these would be technically possible. Hence, we stick to the original list of four learning rules in which the semi-classical filter is combined with GBR only.

4 The semi-classical method

In Subsection 3.4 we have discussed how to further constrain the class of priors in such a way that it is more adapted to what an expert *can* actually know apriori. Here we want to introduce an additional filter, based on unorthodoxly combining classical ideas on defining intervals of confidence and Bayesian learning. Our attempt to do so is motivated by the desire to find a learning rule that on the one hand is somewhat more "objective" than the weighted maximum likelihood update rule and that on the other hand is more informative than unfiltered GBR.

4.1 A volume of confidence in the set of priors

For this, we make the following strong assumption that may be controversial (noting that readers who cannot follow such an approach may skip the remainder of this Section and simply digest those results derived without utilizing the semi-classical filter):

Any prior specified by an expert can be interpreted as representing a stochastic process that describes the way in which the expert deviates from reality in the course of her or his life.

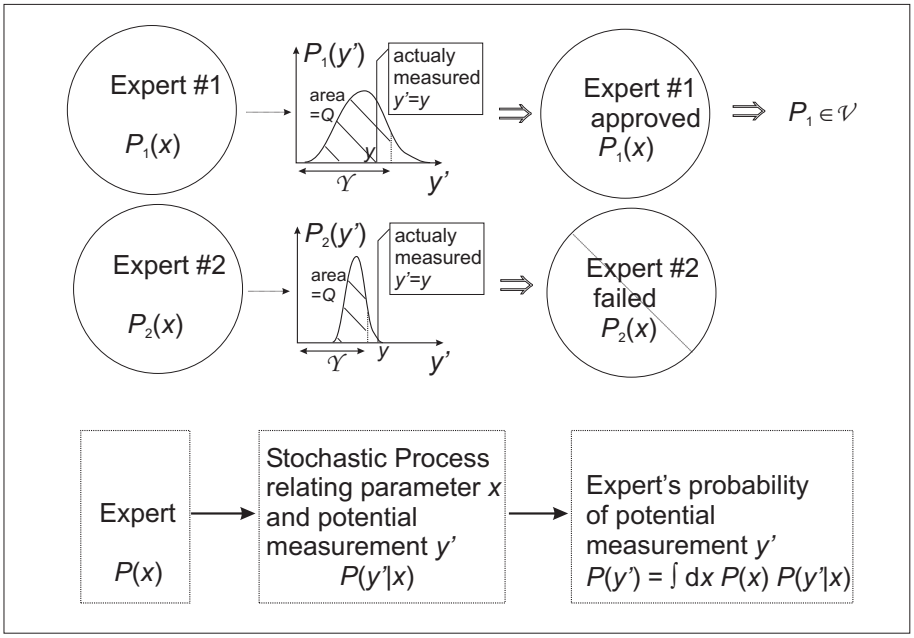

Figure 1: Scheme for the construction of a volume of confidence $\mathcal{V}$ in the class of priors. Any expert shall be characterized by one prior that induces a probability measure of the potential measurement y' (bottom). Once the measurement has been realized, i.e. $y' := y$, one can disregard priors that display y outside of a quantile, characterized by a pre-set probability Q.

We would like to elaborate on this assumption. It sees an expert as someone who performs many, N, assessments during her or his life, each assessment specified by one prior distribution. Our assumption implies that the expert's priors $P_n, n \in \{1, ..., N\}$ are consistent with her or his actual knowledge on the parameters α_n assessed. More formally: for any probability P', for any sequence of assessments P_n and of parameter subsets A_n with $\forall_n\ P_n(A_n) = P'$, $\forall_{\epsilon>0} \lim_{N\to\infty} P(|k/N - P'| < \epsilon) = 1$ if k denotes the number of "hits" in her or his life (as a "hit" we denote an assessment n for which $\alpha_n \in A_n$; hence we request the law of large numbers to hold for the expert's assessments). In a sense one may *define* an expert that way – as a person whose prior measures are sampled by the true parameter values assessed over her or his life.

That way, we choose an interpretation of subjective probability that allows us to treat it not only as epistemic uncertainty, but also as aleatoric uncertainty, i.e., as a stochastic process that governs the relation of the expert to reality during her or his life. Those users that could accept such an interpretation of experts' knowledge have the chance to interpret the combination of "choose the parameter" and "predict, given that parameter, the measurement y" as a joint stochastic process. If the former is described by $P(x)$ and the latter by $P(y|x)$, then, given the expert's P: $P(y) = \int dx\ P(x)\ P(y|x)$.

As in our interpretation, for any prior, $P(y)$ is generated by a stochastic process, it must be possible to evaluate the elements within the set of priors on the basis of the measurement utilizing classical statistics. In particular we are

interested in defining a classical volume of confidence within the set of priors as a filter, conditioned on y.

A classical interval of confidence (or, in our case, volume of confidence $\mathcal{V}$) represents a mapping from observation y onto the set of subsets of (hypotheses) $\mathcal{P}$, according to

$$\mathcal{V} : \mathbb{R} \to 2^{\mathcal{P}}, \quad \forall_{P' \in \mathcal{P}} \ \forall_{y' \in \mathbb{R}} \ P(P' \in \mathcal{V}(y')) = Q. \tag{8}$$

Here, we distinguish y' from y in order to indicate that the formula refers to *potential* measurements for a given expert opinion P'. The probability P in the formula refers to all potential parameter values x and measurements y', weighted according one fixed prior P'. The value of Q needs to be chosen *before* y is known, in an informative manner – as it is standard in classical statistics. In this article, we do not invest much in optimizing either Q or $\mathcal{V}(.)$ as we just would like to demonstrate the principle. We will show below that even with rather ad hoc settings for both we can construct a powerful semi-classical filter.

Now the volume will be constructed by the following idea: above requirement is equivalent to a prior P'-wise prescription that decides for each prior $P' \in \mathcal{P}$ which y is "compatible" with our joint stochastic process (see also Figure 1):

$$\mathcal{Y} : \mathcal{P} \to 2^{\mathbb{R}} \quad \text{with} \quad \forall_{y \in \mathbb{R}} \ \forall_{P' \in \mathcal{P}} \ P(y \in \mathcal{Y}(P')) = Q. \tag{9}$$

Obviously, the two prescriptions are equivalent:

$$\forall_{y \in \mathbb{R}} \ \forall_{P' \in \mathcal{P}} \quad \{P' \in \mathcal{V}(y) \Leftrightarrow y \in \mathcal{Y}(P')\}. \tag{10}$$

In practice, one will attempt to construct $\mathcal{Y}$ in such a way that one excludes values of y which strongly correlate with extreme values of x_1, i.e., those y, for which the probability of ruin is highest. That way, a powerful classical statistic is constructed that excludes those priors which result in non-informatively high values for $\overline{P}^*$ (an analogous argument holds for the lower limit).

Once such $\mathcal{Y}$ is constructed, it implies the mapping $\mathcal{V}$ which can be used as a filter: for any of the Bayesian updating rules, $\mathcal{P}$ may then be replaced by $\mathcal{V}(y) \subset \mathcal{P}$ for further investigations.

4.2 Proposing a nesting formula

One may now ask how a decision-maker may deal with the fact that the volume of confidence does not hold with certainty but only with probability Q. If $Q \approx 1$, in many applications of classical tests, this aspect is simply ignored and the volume of confidence is dealt with as if it were certain.

However, here we would like to suggest an exact approach that explicitly takes care of those cases for which the volume of confidence fails, appearing with probability $(1 - Q)$. We "nest" the classical uncertainty $(1 - Q)$ into the Bayesian scheme by a probability-tree argument (see Figure 2).

Let P^*_+ and P^*_- the upper and lower probabilities of ruin derived, after the semi-classical filter has been applied to GBR. In case 1, the classical volume was correct, and $\overline{P}^*_{\text{apost}} = P^*_+$, being true with probability Q. In case 2, the classical volume was wrong, and we set $\overline{P}^*_{\text{apost}} = 1$ as a conservative estimate of that quantity, with probability $(1 - Q)$. (Analogously we can proceed with the *lower* probability of ruin.)

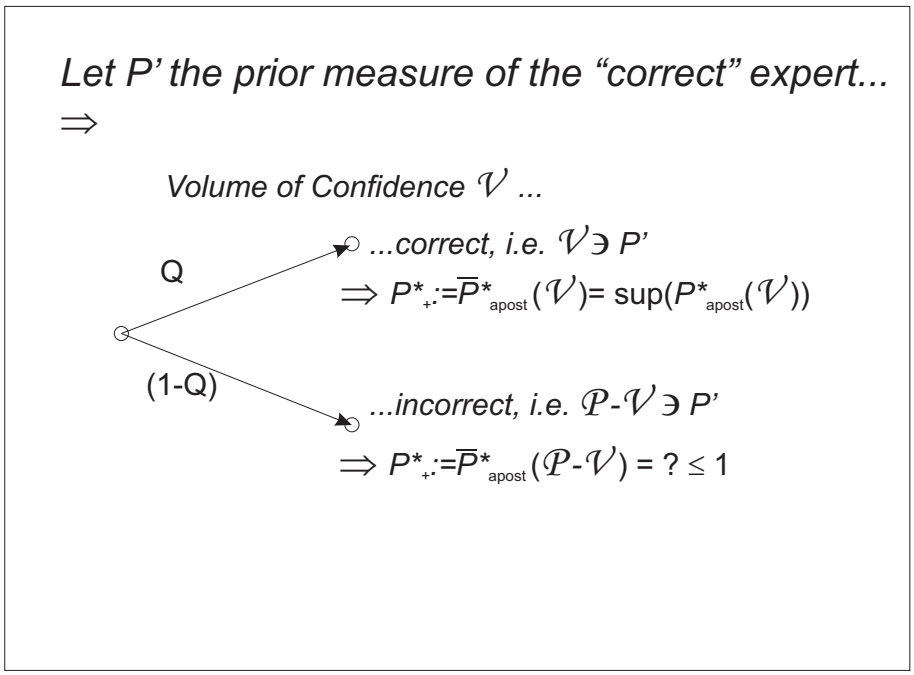

Figure 2: Nesting the classical volume of confidence $\mathcal{V}$ in a decision situation. In our frequentist's interpretation we can explicitly take care of the possibility that $\mathcal{V}$ may *not* contain the correct prior. For that we utilize a probability tree, resulting in Eqs. 11 and 12.

According to the thereby induced tree diagram,

$$\underline{P}^*_{\text{apost.nested}} = Q \cdot P^*_- + (1-Q) \cdot 0, \tag{11}$$

$$\overline{P}^*_{\text{apost.nested}} = Q \cdot P^*_+ + (1-Q) \cdot 1. \tag{12}$$

In the following, we will call the upper and lower probabilities of ruin "nested"; those based on GBR with semi-classical filter, however without nesting correction (Eqs. 11 and 12) "unnested."

4.3 Treatment of an empty $\mathcal{V}(y)$

How to proceed if y is such an "outlier" that $\mathcal{V}(y) = \emptyset$? One could proceed in saying that no expert were available, hence there were no information on P^*_{apost}. However, that lack of posterior information is counter-intuitive. If the semi-classical filter is used together with GBR, we know that adding a prior to the class does not result in spurious information. Hence if $\mathcal{V}(y) = \emptyset$ we could add a prior P_a from the original class that is most informative, e.g. the maximum likelihood prior. In the worst case, the "true" prior equals P_a and the nesting formula is too conservative. In any case, no spurious information is added by re-introducing P_a.

We would like to illustrate what updating of an imprecise prior may mean in a decisions situation. Hence, before presenting the implementation of above combinations of learning rules and filters, we now introduce a stylized potential user of our ideas.

5 A tutorial example for an insurance problem

Let us imagine an insurance company that has the choice between clients, each of which comes along with an upper and a lower probability of ruin as well as the value of a property to be insured.

To keep things simple, we assume that there is a standard loss of unit one for any of the clients. Furthermore, we assume that the insurance company as well as any of their potential clients j are pessimists, i.e., for any client j, both the company and the client focus on the *upper end* of the posterior probability of ruin $\overline{P}^*_j$ which shall be known to both parties. For a potential client, that determines her or his *willingness to pay* for an insurance premium. The company, in turn, will decide on that basis whether – and if so – for what premium it would insure the client. If the premium that the company regards as necessary, exceeds the willingness to pay, no contract will be made. Otherwise the company – whom we assume to exactly know the customers' willingness to pay in advance – will set the insurance premium $I(\overline{P}^*_j)$ equal to the client's willingness to pay. Furthermore, we assume that any of the clients share the same willingness to pay, therefore the function $I(.)$ is independent of j.

If the company attempts a positive expected gain, there must exist values for $\overline{P}^*_j$ with

$$I(\overline{P}^*_j) > \overline{P}^*_j, \tag{13}$$

otherwise no insurance contracts will be made. Let J be the number of successful contracts. Then the company faces a sure inflow of $\sum_{j=1}^{J} I(\overline{P}_j^*)$. The outflow is given by the cases of ruin that appear at the moment of contract as variables prone to epistemic uncertainty. If we assume that the clients' cases are mechanistically independent and that the company has the choice to set $\forall_j \quad \overline{P}_j^* =: p$ in order to keep their analysis simple, the upper posterior probability of ruin, i.e., for negative gain G reads (in the Gaussian approximation that reveals the exact result in the limit $J \to \infty$)

$$\overline{P}_{\text{ruin.insurance}} = \int_{-\infty}^{0} dG \; N\left(J \cdot I(p) - J \cdot p, \sqrt{J \; p \; (1-p)}\right)(G), \tag{14}$$

$N(\mu, \sigma)(.)$ denoting a Gaussian with mean μ and standard deviation σ for the remainder of this article. (Note that the company deciding on $\overline{P}_{\text{ruin.insurance}}$ rather than, e.g. $\underline{P}_{\text{ruin.insurance}}$, is consistent with focussing for any client j on $\forall_j \; \overline{P}_j^*$.)

For illustrative purposes we now specify the willingness to pay according to

$$I(p) := \left\{ \begin{array}{cccc} 2^{-1+1/\alpha} & p^{1/\alpha} & \text{for} & p \leq 1/2 \\ 1 - 2^{-1+1/\alpha} & (1-p)^{1/\alpha} & \text{for} & p > 1/2 \end{array} \right\} \quad \text{with} \quad \alpha := 3, \tag{15}$$

ensuring inequality 13 for $\forall_j \; \overline{P}_j^* \equiv p < 0.5$ (see also Figure 3, upper left graph).

Figures 3 and 4 display the maximum p that is compatible with the requirement that for the company, the probability of ruin, i.e., $P(G < 0) = 1/1000$. The so derived limits will be used in later graphs of Section 6 when once again addressing the hereby formulated stylized decision problem:

"Clients with which characterizing measurement "y" can we – as an insurance company – sign on contract if our own probability of ruin shall be below 0.1%?"

In the following, we attempt to answer that question for three major prior "correlation classes" with prescribed marginal distributions: (1) uniform marginals, (2) Gaussian marginals, (3) Gaussian marginals and the class of joint distributions being restricted to Gaussians as well.

The "uniform case" is motivated by a specific expert interview on the parameters of our in-house climate model. The expert preferred – for each parameter – an almost uniform marginal that would continuously drop to zero at the very ends of the interval. As most climate models respond in a quasi-linear manner to changes in model parameters, we regard it as a conservative approximation to choose a uniform prior over the full interval instead. That way, the extremes are pronounced. Below we will demonstrate that additional, gradient-based filters (or anything equivalent) will be necessary in order to obtain informative as well as adequate results. If we require that these rules are to be implemented in a numerically not too sophisticated a manner, then a conveniently parameterized prior class like (3) seems the class of choice. We then furthermore investigate class (2) as an intermediate step between (1) and (3).

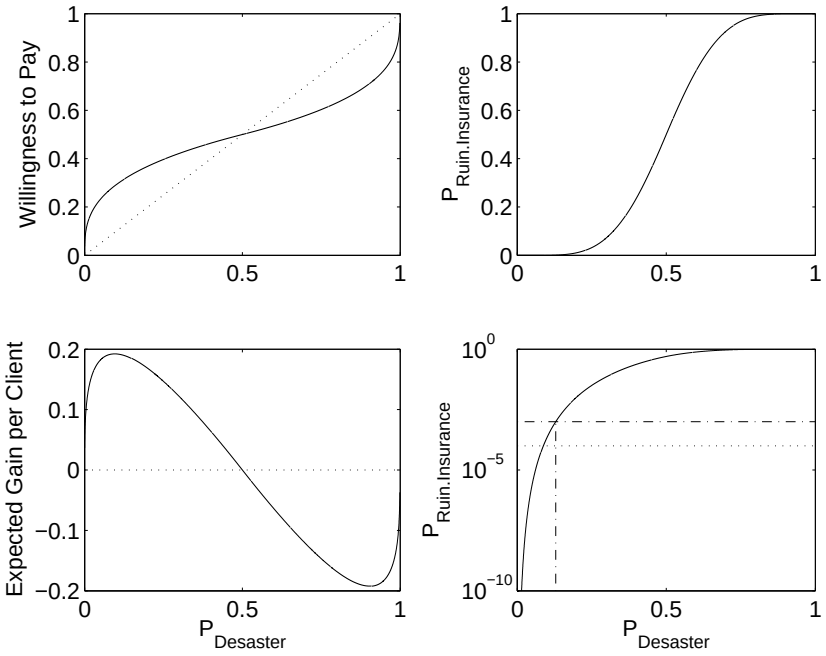

Figure 3: A stylized decision situation for an insurance company for $J = 30$ identical clients and a probability of ruin, i.e., $\overline{P}(G < 0) = 1/1000$. The graphs denote the following: upper left: $I(p)$ (with $p \equiv P_{\text{Desaster}}$), lower left: the lower expected gain per client, upper right: $P(G < 0)(p)$, lower right: the same, yet as semilog plot. The allowable values for p, given $P(G < 0) = 1/1000$, are indicated by a dashed-dotted line. The case is quite robust against changes in $P(G < 0) \equiv P_{\text{ruin.insurance}}$. The upper limit of 12.9% for $p \equiv P_{\text{Desaster}}$ that we read from the graph will be used in the upcoming tabulars as well as Figure 15.

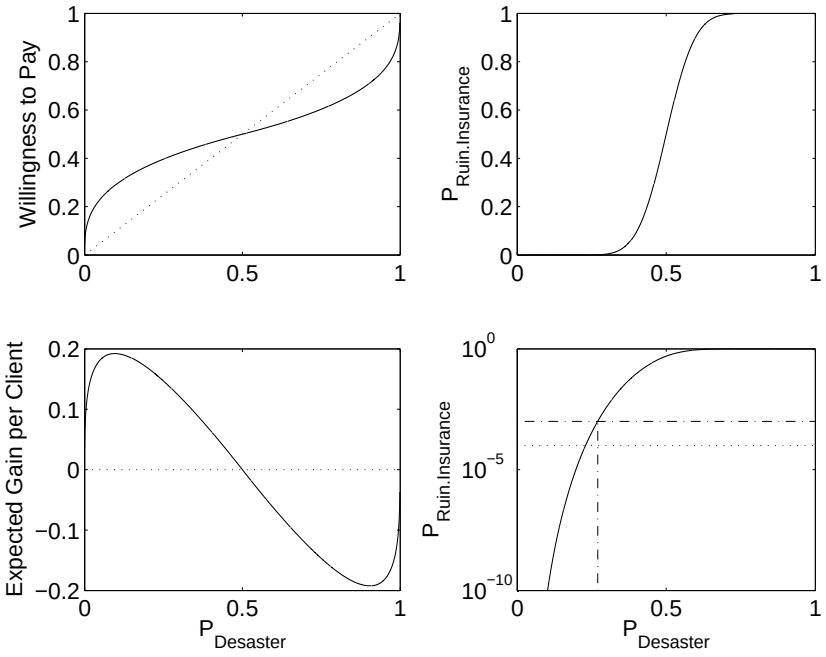

Figure 4: The same as Figure 3, yet for $J = 100$ clients, allowing for a larger upper limit of 27.0% in $p \equiv P_{\text{Desaster}}$ due to risk pooling.

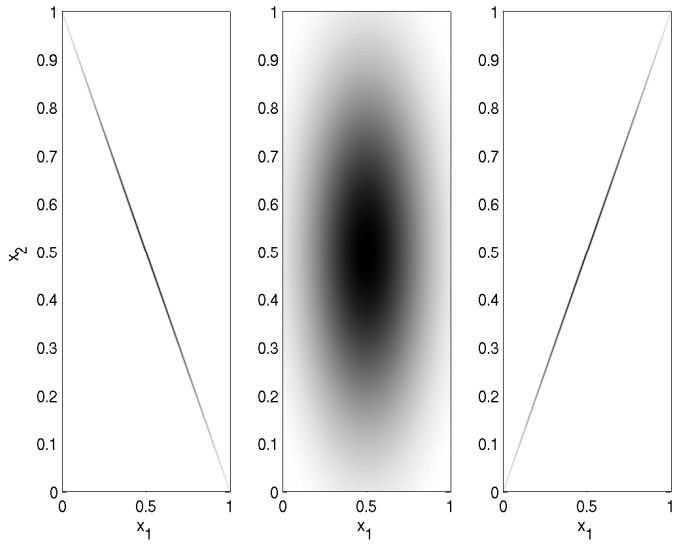

Figure 5: Three extreme representatives of the class of Gaussian priors with prescribed marginals. From left to right: maximally anticorrelated case ($f = -1$), uncorrelated case ($f = 0$), and maximally correlated case ($f = 1$) – for a definition of the parameter f see Eq. 17). The maximally (anti)correlated cases are degenerate in the sense that the supports of the distributions are one-dimensional. This will lead to paradoxical inferences during Bayesian updating.

6 Gaussian priors

We now turn to the most operational Section of this article, as we construct a class of priors which represents unimodality, and for which most of the updating methods are informative. Due to its analytical tractability we can also make transparent any of the conceptual ideas outlined before. This central Section is organized as follows: first we introduce a convenient parameterization for the class of priors. Second, we define a semi-classical filter and apply any of the four learning rules to the class of priors not yet constrained by gradient information. Third, we repeat the four learning rules for the gradient filter being switched on in two versions. Finally, we apply the nesting correction for the semi-classical method in the latter case.

6.1 Specifying the marginals and the likelihood function

We specify $P_1 \sim N(\mu, \sigma)$, $\mu = 1/2, \sigma = 1/4$, $P_2 \equiv P_1$, hence we select marginals that contain $\pm$ 2 standard deviations in [0,1].

If we consider a class of Gaussian joint distributions, any prior is unimodal. Later on we also will require bounds on the gradients, thereby avoiding degenerate, essentially lower-than-2 dimensional Gaussians (see Figure 5, left and right graphs). Before that, however, we would like to study Bayesian updating on the unrestricted class of that Gaussians.

For simplicity we assume further that the transfer function $F(x_1, x_2) = \kappa x_1 + x_2$ relates to the observation y through some additional Gaussian process

$$L(x_1, x_2) \equiv P(y|x_1, x_2) := N(\kappa x_1 + x_2, \sigma_\eta^2)(y). \tag{16}$$

If $|\kappa| \ll 1$ or if $|\kappa| \gg 1$ the transfer function were essentially one dimensional, while the non-trivial case would be obtained for $|\kappa| \approx 1$. Some quick checks reveal that $|\kappa| = 1$ reveals a degenerate exception for which reason we avoid such choice. Whenever we do not display results for κ but have to take a decision (e.g. for numerical results) we choose $\kappa := 1.05$.

σ_η represents another degree of freedom. As this article deals with the representation of imprecise prior knowledge and its updating and not so much with the objective uncertainty contained in the likelihood, we choose $\sigma_\eta \ll \sigma$, in particular, $\sigma_\eta = \sigma/10$ when we have to specify it.

Now we derive in Appendix A.1 that a 2-dimensional Gaussian prior P fulfils the constraints set by the marginals $\sim N(\mu, \sigma^2)$, iff there exists $f \in [-1, 1]$ with

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & f \\ f & 1 \end{pmatrix}, \tag{17}$$

and $P \sim N((\mu, \mu)^t, \Sigma)$, whereby Σ denotes the covariance matrix of P.

Hence, we conveniently parameterize the class of Gaussian priors by one single parameter f, and as function of $f \in [-, 1, 1]$. $f = 0$ represents the uncorrelated (standard), $f = \pm 1$ the maximally (anti)correlated case (see again Figure 5). Hence, we have parameterized $\mathcal{P}$ by f.

6.2 Posterior properties

To obtain $P_{1.\text{apost}}$, we integrate over x_2, revealing (see Appendix A.2)

$$\begin{aligned} P_{1.\text{apost}} &\sim N(\mu', \sigma'^2) \quad \text{with} \\ \mu' &= \frac{\mu(1 - (1-f)(\kappa - 1)\ \sigma^2/\sigma_\eta^2) + (f + \kappa)\ y\ \sigma^2/\sigma_\eta^2}{1 + (1 + 2f\kappa + \kappa^2)\ \sigma^2/\sigma_\eta^2}, \\ \sigma' &= \sigma\ \sqrt{\frac{1 + (1 - f^2)\ \sigma^2/\sigma_\eta^2}{1 + (1 + 2f\kappa + \kappa^2)\ \sigma^2/\sigma_\eta^2}}. \end{aligned} \tag{18}$$

We utilize this expression to calculate the posterior probability of ruin

$$P^*_{\text{apost}}(f) = \int_{x_1^*}^{\infty} N\left(\mu'(f), (\sigma'(f))^2\right)(x_1)\ dx_1. \tag{19}$$

The case of dominating likelihood uncertainty – not to be considered further – is obtained by $\sigma_\eta \to \infty$:

$$\begin{aligned} \lim_{\sigma_\eta \to \infty} \mu' &= \mu, \\ \lim_{\sigma_\eta \to \infty} \sigma' &= \sigma, \end{aligned} \tag{20}$$

i.e. if the measurement y becomes non-informative on $\kappa x_1 + x_2$, then the marginal prior and posterior on x_1 are identical.

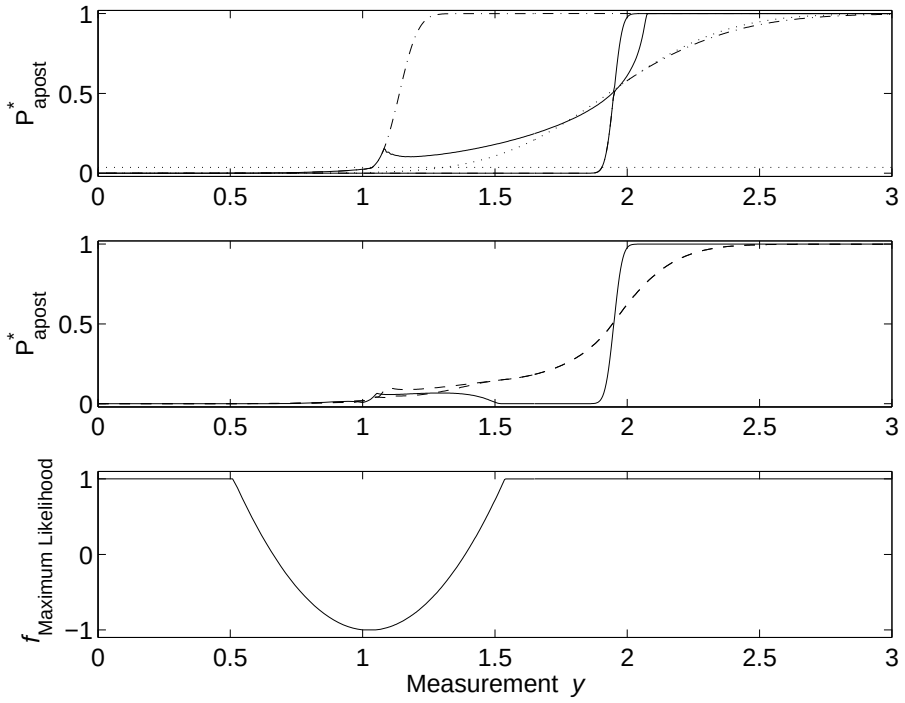

Figure 6: Probability of ruin for the correlation class parameterized by the correlation coefficient f after Eq. 17 for $\kappa = 1.05, x_1^* = 0.95$, $\sigma_\eta = \sigma/10$. Top: horizontal dotted line: apriori value, curved dotted: (standard) uncorrelated case, dashed-dotted: GBR, solid: the semi-classical method (combined with GBR), the lower curves of both coalescing. GBR displays a quasi step-function-like behavior. Furthermore, the semi- classical method reveals a lot of information beyond GBR. Center: solid line: maximum likelihood estimate, dashed lines: weighted maximum likelihood estimate. In this rather restricted class, maximum likelihood estimate lead to rather low probabilities of ruin. The fact that the solid curve show a non-monotonic relation between measurement and vastly deviates from its weighted counterparts (dashed) undermines trust in that updating method. Bottom: Correlation parameter f obtained from maximum likelihood update method, given y. For large y, the maximum likelihood update method prefers $f = 1$. Then Bayesian learning implies the intersection of two lines: the support of the likelihood, and the curve $x_1 = x_2$ (fully correlated), hence, the posterior concentrates all weight on one single point. Therefore, the probability of ruin must show a sharp transition (center graph, solid line) when that single point crosses x_1^* as a function of y.

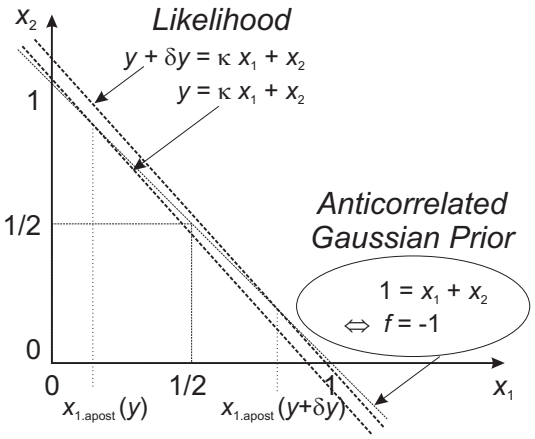

Figure 7: Discussing Bayesian learning for the double degenerate case $\sigma_\eta \to 0$, $f \to -1$ (fully anticorrelated prior). Bayesian learning simply reduces to looking up the intersection of the lines "$1 = x_1 + x_2$" and "$y = \kappa x_1 + x_2$". Note that although the aposteriori uncertainty on x_1 is zero ("a well-defined intersection of lines"), the aposteriori value for x_1 strongly varies with mild variations in y. For $f = +1$, quite the contrary is the case. Hence, in many parameter settings, these two extreme cases tend to span a large interval for the probability of ruin for GBR, leading to non-informative results.

6.3 Application of generalized Bayes' rule (GBR)

In order to derive the upper and lower probability of ruin according to GBR, we simply have to ask for the supremum and the infimum of this $P^*_{\text{apost}}(f)$ over $[-1, 1]$. We display the result as function of y in Figure 6, upper graph, dashed-dotted lines (note that the curve for the lower probability of ruin coalesces with the lower solid line (to be introduced later on), hence, is masked by it). Both curves derived from GBR display quasi step-function type behavior.

For comparison we also display the standard case of Bayesian updating of simply using the uncorrelated prior $P(x_1, x_2) = P_1(x_1) \cdot P_2(x_2)$ instead of a class of priors (Figure 6, upper graph, curved dotted line, derived by letting $f = 0$ in Eqs. 18.

Apparently, GBR reveals much less informative results than the standard method would proclaim. In particular for $y \in [1.3, 1.8]$ GBR does reveal *no information at all* on the posterior probability of ruin, i.e. we only know $P^*_{\text{apost}} \in [0, 1]$. Hence, in the GBR paradigm, utilizing the more realistic class of priors instead of the uncorrelated prior only, reveals drastically different results. The question is: how would the results change when less conservative (than GBR) methods of updating are being used? Before we discuss them we would like to highlight the underlying reason for the non-informative features of GBR.

6.4 The illustrative limit $\sigma_\eta \to 0$

In Figure 6 we display the case $\sigma_\eta = \sigma/10$, hence the uncertainty in $P(y|x_1, x_2)$ is much smaller than the prior uncertainty. For that reason we can expect to find the analytically transparent case $\sigma_\eta \to 0$ illuminating. If $\sigma_\eta \to 0$, then

the support of $L(x_1, x_2) = P(y|x_1, x_2)$ collapses to the one-dimensional linear manifold that solves the equation $y = \kappa x_1 + x_2$. Furthermore

$$\begin{aligned} \lim_{\sigma_\eta \to 0} \mu' &= \frac{\mu(1-f)(1-\kappa) + y(f+\kappa)}{1 + 2f\kappa + \kappa^2}, \\ \lim_{\sigma_\eta \to 0} \sigma' &= \sigma \sqrt{\frac{1-f^2}{1+2f\kappa+\kappa^2}}. \end{aligned} \tag{21}$$

Now consider the degenerate priors for $f = \pm 1$:

Eqs. 21 imply $\lim_{|f|\to 1} \sigma' = 0$: in that limit, the prior P and L represent two 1D lines in the 2D space spanned by x_1, x_2, intersecting only at one point, leaving no space for aposteriori uncertainty in x_1. Hence, the support of $P_{1.\mathrm{apost}}(x_1)$ collapses to μ'. Therefore it is worthwhile to explicitly note μ' for these two extreme cases:

$$\mu'(f=-1) = \frac{y - 2\mu}{\kappa - 1}, \tag{22}$$

$$\mu'(f=+1) = \frac{y}{\kappa + 1}, \tag{23}$$

which can also be interpreted as the intersection of the lines $y = \kappa x_1 + x_2$ either with $1 = x_1 + x_2$ (see Figure 7, for $f = -1$), or with $x_1 = x_2$ (for $f = 1$), i.e., as the intersection of δ-type likelihood and (anti)correlated prior, respectively. From Eqs. 22 and 23 we conclude further (compare also Figure 8)

$$P^*_{\mathrm{apost}}(f=-1) = \left\{ \begin{array}{lll} 0 & \text{for} & y < (\kappa-1)\, x_1^* \,+\, 2\mu \\ 1 & \text{for} & y \geq (\kappa-1)\, x_1^* \,+\, 2\mu \end{array} \right\}, \tag{24}$$

$$P^*_{\mathrm{apost}}(f=+1) = \left\{ \begin{array}{lll} 0 & \text{for} & y < (\kappa+1)\, x_1^* \\ 1 & \text{for} & y \geq (\kappa+1)\, x_1^* \end{array} \right\}. \tag{25}$$

These two Equations suggest the structural changes in $\overline{P}_{\mathrm{apost.GBR}}$ (at $y \approx 1$) and $\underline{P}_{\mathrm{apost.GBR}}$ (at $y \approx 2$), depicted as dashed-dotted lines in Figure 6, upper graph (the positions of the discontinuities can easily understood by noting $\kappa, x_1^* \approx 1$). Hence GBR is non-informative over a large interval of y's (in our example for $y \in [\approx 1, \approx 2]$), i.e., we "learn" from Bayesian updating over the class made-up by all $f \in [-1, 1]$ that $P^* \in [0, 1]$. This phenomenon is similar to the result found before for GBR and dissatisfying if it could be avoided by more informative learning rules.

The priors with $f \to -1$ display a further type of "instability" that are not shown by $f \to +1$: Let $\kappa =: 1 + \varepsilon$. Then $\mu'(f=-1) = (y - 2\mu)/\varepsilon$, according to Eq. 22. This implies for $\kappa \approx 1$, $\varepsilon \ll 1$. Hence Bayesian learning in the strongly anticorrelated limit is very unstable with respect to the measurement y even on the level of the *individual* prior.

For all those reasons, we feel tempted to restrict the gradient of prior densities which would exclude $|f| \to 1$. Before we do so in Subsection 6.7, we would like to implement the remaining three learning rules (besides GBR) for the non-restricted class, for the sake of illustration and completeness.

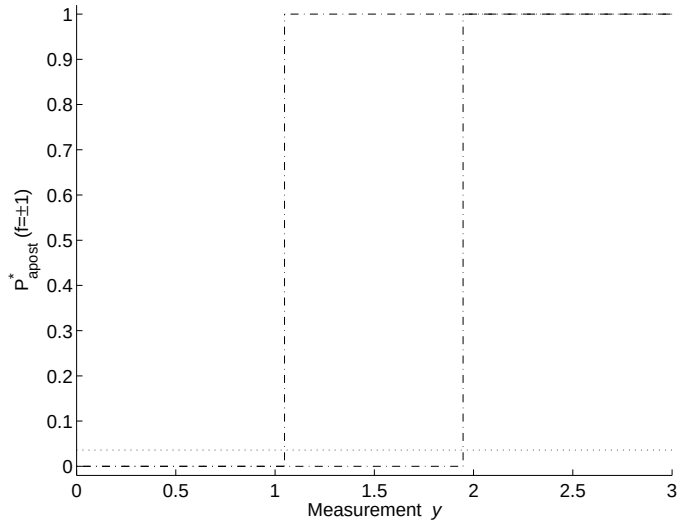

Figure 8: Illustration of Eqs. 24 and 25: Bayesian learning for the two degenerate priors $P(f = \pm 1)$ and $\sigma_\eta \to 0$ (given the standard values $\kappa = 1.05, x_1^* = 0.95$). These two priors alone are sufficient to open the rather larger non- informative window between $y \approx 1$ and $y \approx 2$ when GBR is used for updating. As before the prior probability of ruin is indicate by the dotted line.

6.5 Implementing the semi-classical learning rule

As outlined above, strongly anticorrelated priors may lead to extreme as well as unstable results. This comes along with the likelihood intersecting mostly in the tails with such prior (see Figure 5 (left graph) in combination with Figure 7), a regime of low probability. The very fact that for "most" y, the likelihood and the prior for $f = -1$ intersect in the "tails" of that prior, is the underlying reason why the classical method may successfully eliminate such type of priors from our prior class.

We define a volume of confidence as described in Subsection 4.1. First we select a fixed value for the quantile $Q := 0.98$. In this article we shall not discuss how to optimize the selection of Q. However, we have done so on a pragmatic level: we tested a couple of numerical candidates for Q on whether they would reveal informative results for the type of graphs derived below. As always in classical statistics, it is important, however, to take the decision on Q not in view of one particular measurement y but of *all potential* measurements (observations) y'.

Then by Appendix A.3 for any f we determine the lower Q-quantile $\mathcal{Y}(f)$ that would allow to decide whether y is compatible with f or not (see also Figure 1, upper graph). From this we construct the volume of confidence in the class of priors that for the present class is equivalent with an interval of confidence for f within $[-1, 1]$:

$$\mathcal{P}_{y,Q} := \mathcal{V}(y) := \{P(f) \mid f \in [-1, 1] \text{ and } y \in \mathcal{Y}(f)\}, \tag{26}$$

i.e., we assemble priors for which y does not lie outside the lower Q- quantile.

Hence we have constructed a one-sided interval of confidence such that priors with y in their tails, leading to high μ', highly sensitive to y, are omitted. The one-sidedness somewhat contradicts the classical tradition of defining symmetric intervals of confidence, unbiased tests, and so on. That was necessary as the tradition desired to decouple the statistical procedure from potential application, i.e., it aimed at *universality* of the statistical procedure. Once we drop that request and focus on a particular application, we may optimize our definition of the volume of confidence with respect to the decision situation at hand. Below we will frame such a decision as the stylized insurance problem.

In view of Subsection 4.3 we still have to specify how to proceed in cases of extreme values of y that would lead to a rejection of *any* prior under the quantile Q. As we outlined in Subsection 4.3 we do not add spurious information when we add priors to the set of priors used for updating. We decide to add $P(f = 1)$ to the set of priors if $\mathcal{V}(y) = \emptyset$. That choice appears natural as $P(f = 1)$ is the last prior to be rejected by the semi-classical filter when y were continuously to assume more and more extreme values. For $\kappa = 1.05, \sigma_\eta = \sigma/10$, that mechanism becomes activated e.g. for $y > y^{**}$, $y^{**} \in [2, 3]$.

In Figure 6, classically pre-selected estimates for GBR are added to the updating rules already introduced for the uniform marginals by solid curves in the upper graph. It becomes apparent that the interval spanned by upper and lower P^* is much narrower than for traditional GBR for most values of y. The kink at $y \approx \kappa$ (that in that Figure equals 1.05) is a "resonance-type effect" stemming from the very narrow members of the class, for the "rare" cases of y when the likelihood intersects the $(f = -1)$-prior in its center. (At this stage of analysis, the nesting correction of Subsection 4.2 has not been applied yet.)

6.6 Maximum likelihood update results

Again in Figure 6, the center graph reveals the results for the maximum likelihood update methods. In the Appendix A we outline that $\mathcal{P}_{\mathrm{m}}$ consists of exactly one element for which reason we can drop upper and lower bar for $P^*_{\mathrm{m}}(y)$. The standard maximum likelihood update method results in a non-monotonous functional relation $P^*_{\mathrm{m}}(y)$. This can be understood when relating the center to the lower graph: For extreme cases of y, the method selects $f = 1$ as the correlated prior gives the most weight to the extremes among all priors. When discussing only branches with $f = 1$, $P^*_{\mathrm{m}}(y)$ is monotonous as it must be. However, in between, around $y \approx \kappa$, the method prefers the anticorrelated prior which must lead to a sharp switch in $P^*_{\mathrm{m}}(y)$ as $y \approx \kappa$ is crossed (see Eq. 22 and remark afterwards). Hence the interplay of changing y as well as the prior does lead to a non-monotonous $P^*_{\mathrm{m}}(y)$.

As a key result, in Figure 6, center graph, it becomes apparent that standard (solid line) and weighted (dashed lines) maximum likelihood Bayesian learning qualitatively deviate for $y > 1.5$. Figure 9 illuminates the underlying reason. Large values of y force the standard method to select $f = 1$ which comes with $P^* = 0$ for $y < 1.9$ (see Eq. 25). However, the center graph of Figure 9 reveals that if one allowed for f mildly smaller than 1, $P^* = 0$ is *not structurally stable*, hence, a weighting method must result in much larger values for P^*, also found in the lower graph. The lower graph furthermore illustrates the "purifying" mechanism within any of the likelihood methods: those priors resulting in $P^* = 1$ due to dilation-type behavior come with zero weight, hence their influence on

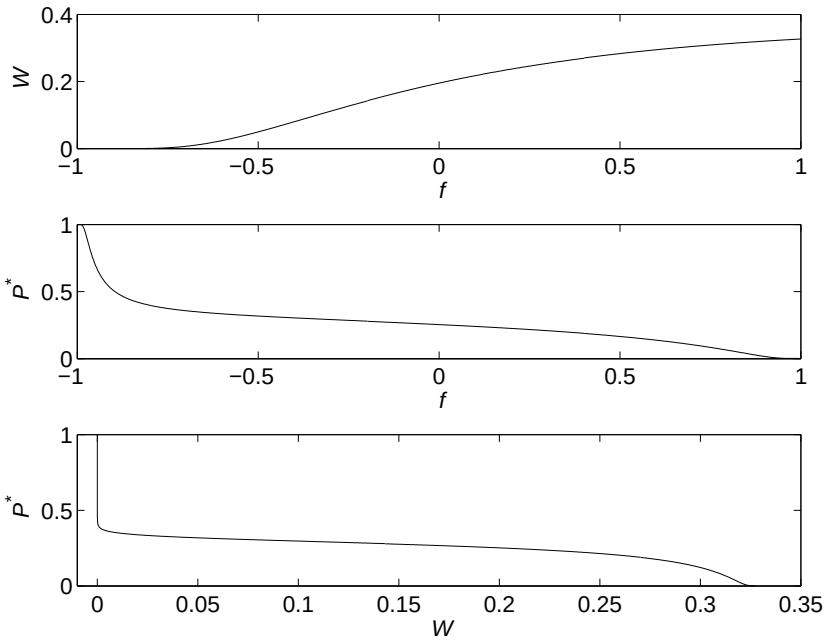

Figure 9: Relation of weight function $W(.)$, parameter f specifying the prior, and probability of ruin, for the special case $y = 1.7, \kappa = 1.05, x_1^* = 0.95, \sigma_\eta = \sigma/10$. The maximum likelihood update rule requires to select the f, i.e., the prior for which W, the prior probability of y is largest. Hence, $f_{\rm ml} = 1$ (see upper graph). This was to be expected as for the rather "large" value of y, the prior with highest correlation (i.e., $f = 1$) prefers the "extreme" y the most among all priors. However, $P^*(y = 1.7, f = 1) = 0$ (see Eq. 25 and center graph). The important point is that the case $f = 1$ is exceptional within the class of priors as for $f \in [-1, 0.5]$, $P^* > 0.1$ (center graph). The bottom graph shows that when averaging $P^*(f(W))$ over the W-scale weighted with W, an average somewhat between 0.1 and 0.3 is to be expected (weighted maximum likelihood update method), drastically differing from standard maximum likelihood update.

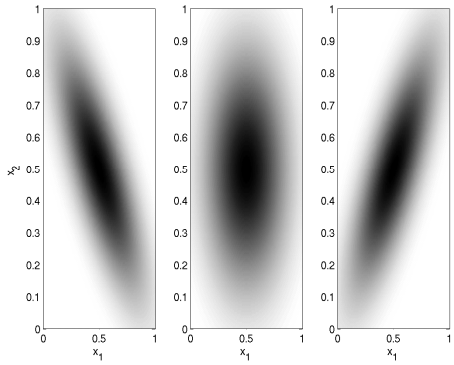

Figure 10: Extreme members of the class of priors for a bounded gradient condition, consistent with 5 blocks in the 2D parameter space ($N_2 = N_1 = 5 \Rightarrow \beta = 0$ along the notation of Subsection 3.4). Left graph for $f = -1$, centre for the (standard) uncorrelated case ($f = 0$), right graph for $f = 1$.

P^* is eliminated by both likelihood methods (and not by standard GBR).

The fact that standard maximum likelihood update drastically deviates from its weighted counterpart supports unease with standard maximum likelihood update leaving its user with the impression that it may be fundamentally "non-robust". One may argue that all these inconveniences may disappear once the class of priors is chosen more adequately – by avoiding extremely degenerate cases like $f = \pm 1$ that come along with diverging gradients. We will see, however, that this is not the case in the following Subsections; quite the contrary any effect observed so far will be found again (although in a somewhat softer version) when gradients become restricted.

6.7 Imposing constraints on gradients

Figures 10 and 11 display the most extreme members of $\mathcal{P}$ if we set $N_1 := 5$ (see Subsection 3.4) and $\beta := 0$ or $:= 1$, respectively (i.e., $N_2 := N_1$ or $N_2 := N_1^2$, respectively). In the latter case, we allow for "more" prior, mutually distinct opinions.

Figures 12 (for $\beta = 0$) and 13 (for $\beta = 1$) reveal the effects of bracketing the class of priors by those extreme elements. (We consider that type of class for our future investigations with climate models.) The $P^*(y)$-curves become smoother, in particular in the first case, and more similar. However, still drastic differences between various updating methods remain:

- For $y < 1.6$, standard Bayesian learning (i.e., the uncorrelated case, curved dotted line) results in a much more optimistic estimate of the upper probability of ruin than (classically constrained) GBR.
- For $1.4 < y < 1.8$, the upper probability of ruin according to the weighted maximum likelihood method (lower graph, dashed curve) exceeds the estimate according to the (standard) maximum likelihood update method

(lower graph, solid line), in part by an order of magnitude. That demonstrates that (standard) maximum likelihood update for the class of Gaussian priors is *not* a structurally robust pre- selection rule. This finding fuels distrust in results obtained by that method and highlights the need for alternatives to the (standard) maximum likelihood update rule.

Quite remarkably, the upper estimates according to the two new updating rules, the weighted maximum likelihood update method and the classically constrained GBR, coincide to a certain degree, much more than with the remaining methods (see Figure 14).

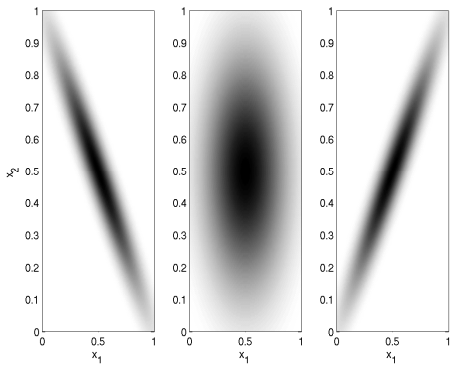

Figure 11: The same as in Figure 10, yet for dimension-adjusted resolution (i.e., $\beta = 1$): $N_2 := N_1^2 = 5^2$. Note that in higher dimensions n (here: $n = 2$) the prescriptions for N_n according to the present versus the previous graph would the more diverge, the larger n. We propose that a realistic description prescriptions for N_n would imply a compromise between these extremes of spatial resolution that are synonym with the degree of sophistication expert options may display.

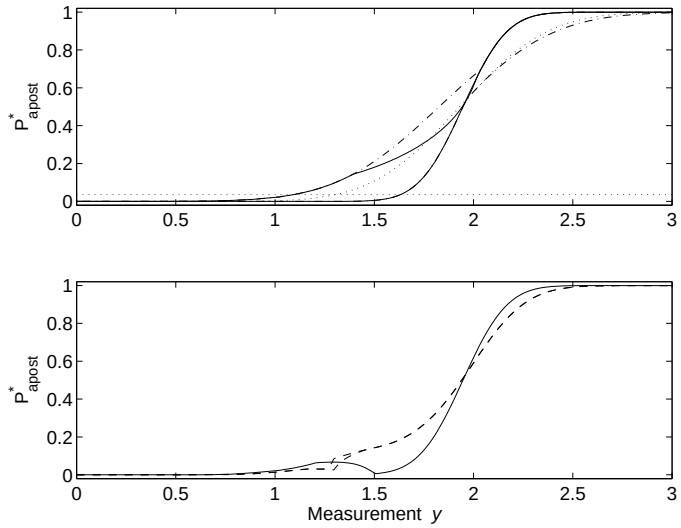

Figure 12: The same as Figure 6, however, for bounded gradients according to $N_2 := 5$. (Again, the lower curves of GBR and the semi-classical method coalesce in the upper graph.) Bounding of the gradients reveals much softer curves.

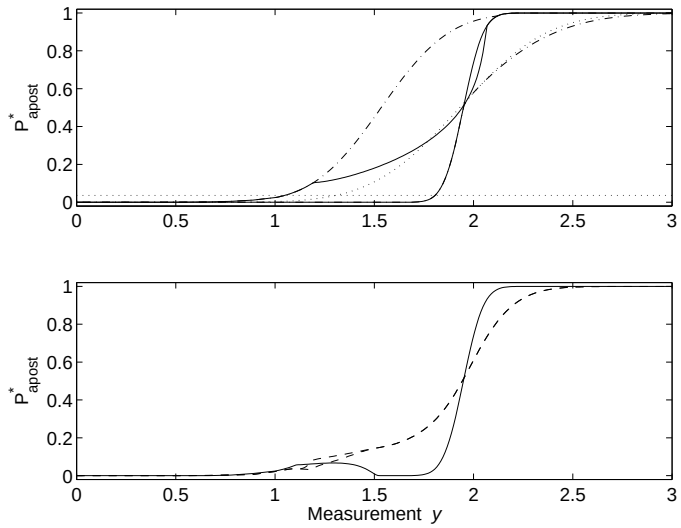

Figure 13: The same as the previous Figure, yet for more independent expert opinions, i.e. $N_2 := 5^2$ (again, the lower curves of GBR and the semi-classical method coalesce in the upper graph). The curves provide a compromise between the last two Figures of that type. Note that even for this class of priors "regularized" by the gradient filter, standard maximum likelihood update to strongly deviate from weighted maximum likelihood update. That demonstrates how questionable it may be to use standard maximum likelihood update – that is based on very few priors – *if* one desires a more balanced (through weighting) influence of all the priors.

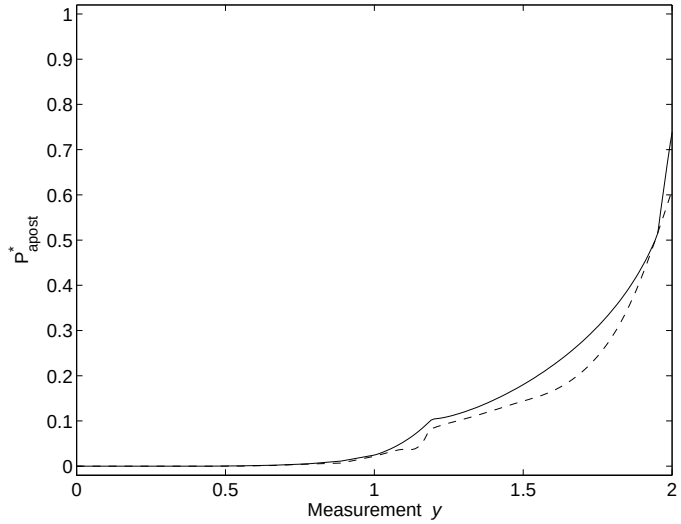

Figure 14: Extract from the previous Figure, solid line: classically constrained GBR (upper limit), dashed line: weighted maximum likelihood update method (upper limit). These two, newly introduced and favored methods, display a much larger degree of similarity among themselves than when compared to the other methods. As these two methods are the favored ones, this leaves the user with a convenient robustness of derived results.

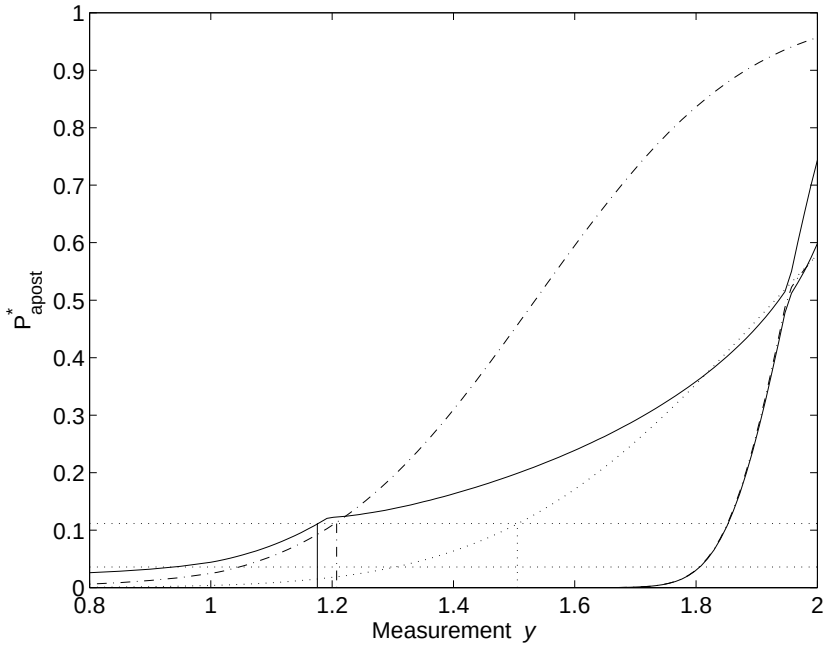

Figure 15: The same as Figure 13, upper graph, yet with the nesting-correction provided by Eqs. 27 and 28. Upper dotted line: A potential maximum upper $\overline{P}^*$, an insurance company may accept. It becomes apparent that the company would use standard GBR (upper dashed curve) rather than nested GBR (upper solid line) as a decision rule which client to insure, as the former rule would allow for a mildly larger threshold in y. For that case, the classical method has not paid off.

6.8 Utilizing the nesting formulas

We now consider the nested interpretation of the classically constrained class. Previous graphs just showed the effects of eliminating members of $\mathcal{P}$ on the basis of a classical rule, however, did not take the additional uncertainty into account that comes with injecting the classical method into the Bayesian formalism.

The ideas outlined in Subsection 4.2 read for this case as follows. Suppose that Q was chosen, *then* y measured. In the first case, the "true" prior is $\in \mathcal{P}_{y,Q}$ and it is meaningful to consider $\overline{P}^*(y,Q)$. In the second case with chance $1-Q$, P was falsely eliminated from $\mathcal{P}_{y,Q}$. In that case, a conservative estimate would be $\underline{P}^*(y) = 0$ and $\overline{P}^*(y) = 1$. By a tree diagramme on both cases, we obtain a conservative estimate

$$\underline{P}^*(y,Q,\text{nested}) = Q \cdot \underline{P}^*(y,Q) + (1-Q)\cdot 0, \tag{27}$$

$$\overline{P}^*(y,Q,\text{nested}) = Q \cdot \overline{P}^*(y,Q) + (1-Q)\cdot 1. \tag{28}$$

In Figures 15 and 16 two potential thresholds for $\overline{P}^*$ are indicated below which an insurance company may insure a client (for the more conservative case $\beta = 1$), read from Figures 3 and 4, respectively. Both Figures allow to read

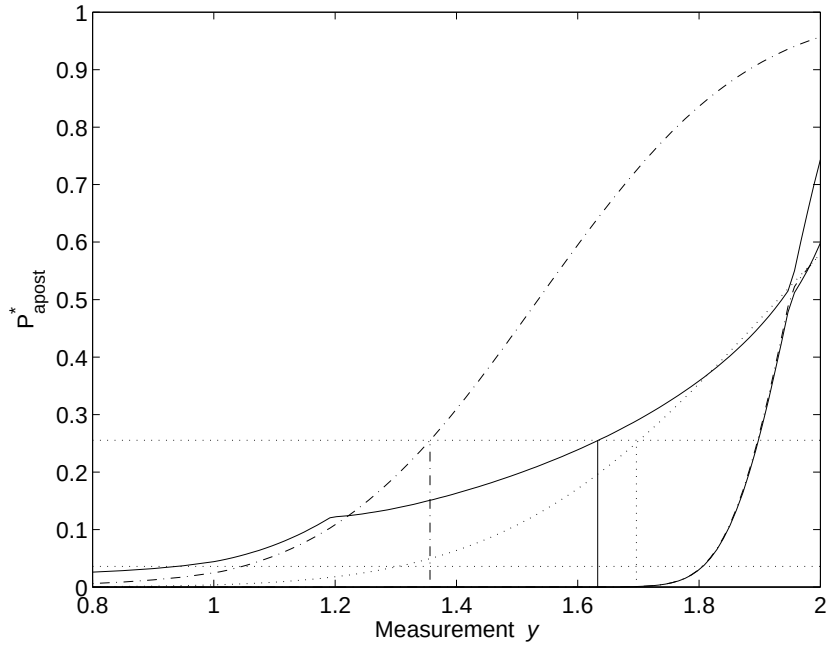

Figure 16: The same as Figure 15, yet for a higher potential threshold in $\overline{P}^*$. Here, the classical method would be advantageous over standard GBR as cases $y \in [1.34, 1.66]$ could be insured in addition.

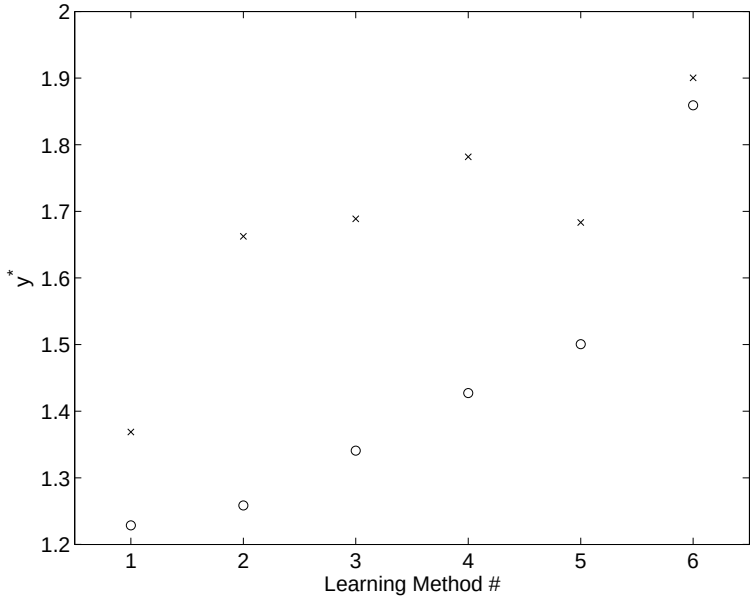

Figure 17: y^* as upper limit of y's with which clients would be insured: Circles: pooling with 30 clients; crosses: pooling with 100 clients. The abscissa indicates the six learning rules according to the tabular of this Subsection. Rules #1, 2 and 6 are "GBR", "nested semi-classical method" and "maximum likelihood update." It is remarkable that – for 100 clients – more than 50% of the gap between the objective, yet, least informative GBR and the most optimistic, yet non-robust rule (maximum likelihood update) can be regained when using the objective nested semi-classical method. (Any entry for $\kappa = 1.05, x_1^* = 0.95, \sigma_\eta = \sigma/10$.)

clients with which characterizing y may be insured and focus on the (nested) GBR as the more objective updating method (compared to (weighted) maximum likelihood update method). While for the former Figure one looses information over the standard GBR (dashed-dotted lines) due to the nesting (solid lines), in the latter Figure the insurance company and the clients face a win-win-situation by using the here introduced nested semi-classical GBR: the cases $y \in [1.37, 1.66]$ could be insured in addition! Among the learning rules depicted in the last Figure, the standard uncorrelated prior (curved dotted line) would allow for the most insurance contracts if the insurance company believed in that rather optimistic choice of a prior. However, from our expert interviews, we do not find that standard procedure very adequate.

6.9 Summary on Gaussian priors for a stylized decision-maker

We would like to summarize the results on Bayesian updating (learning) for the class of Gaussian priors that is restricted by prescribed marginals, analyzed in this Section. We do so by comparing all learning rules discussed so far. We

highlight their effects on the decision (by the insurance company introduced in Section 5) on which clients to sign on contract and which to reject. This is done by systematically performing the kind of inference indicated in the last two Figures for any of the learning rules. Among the three cases of gradient filtering analyzed so far (number of effective cells in $x_1 - x_2$-space N_2: $\infty, 5, 5^2$, compare Figures 5, 10 and 11), we select the intermediate one, i.e. $N_2 = 5^2 \Leftrightarrow \beta = 1$. The company would choose those clients that are characterized by $y \in]-\infty, y^*[$ and we display y^* in the tabular below (as well as in Figure 17), for $J = 30$ and for $J = 100$ clients:

	J updating rule	30	100
1	generalized Bayes rule (GBR)	1.23	1.37
2	semi-classical rule after nesting	1.26	1.66
3	semi-classical rule before nesting	1.34	1.69
4	weighted maximum likelihood	1.43	1.78
5	uncorrelated prior	1.50	1.68
6	standard maximum likelihood	1.86	1.90

As expected, the standard Bayesian updating (uncorrelated prior) is found more on the more optimistic (upper) end of y^*. However, standard maximum likelihood is even more optimistic. Overall, we do not find any of the rules as good approximations of one of the others. In particular can standard maximum likelihood not be related by a simple transformation as for the other two classes. Quite the contrary, it may underestimate the weighted probability of ruin by an order of magnitude. Finally, the semi-classical variant of GBR turns out as being significantly more informative than GBR for large intervals of y.

7 General discussion

We have set up a model for subjective uncertainty that aims at reflecting the opinions held by many climate modelers – or more generally, Earth system modelers, comprising climate, biosphere and economy. By means of a questionnaire we qualify that the experts assume much more confidence in marginals of priors (on model parameters) rather than in the correlation structure within those priors. We have presented a show-case accordingly, based on the simplest multi-dimensional transfer function possible, i.e. $y = \kappa x_1 + x_2$ in combination with Gaussian probability density functions. As we demonstrated, even this rather simple example unfolds a rich class of phenomena when treated under various generalizations of Bayesian updating, that refer to updating classes of priors rather than a single prior.

We have considered:

1. *The* generalized Bayes' rule (GBR) stating that each member of the prior class shall be updated, then the extremes among the posteriors shall be selected.

2. The maximum likelihood update method focuses on those priors that optimize the prior expectation for the measurement. It leads to more informative results than GBR. However, we find it hard to justify as it completely disregards even those priors who may perform only infinitesimally worse than the optimal priors.

3. For that reason we have introduced a weighted maximum likelihood update method that considers all priors, yet weights their influence on the posterior result. When applied to the class of priors that appears as most realistic in practice (i.e., the class of gradient-limited Gaussians) we find that the two likelihood methods may deviate by an order of magnitude in probability of ruin. To our taste that demonstrates how questionable (standard) maximum likelihood inference is and how desperately a generalization was needed. The weighted method carries the drawback that the weighting function is subjective. We have chosen a weighting proportional to the expectation of the measurement. Both likelihood methods are more informative than GBR, however, share the disadvantage that they may add spurious information in case the class of priors is overly inclusive (i.e., contains incompetent expert opinions), in contrast to GBR.

4. This added to our interest in an objective improvement of GBR. We opened a new dimension for designing updating rules by suggesting – for any of the above rules – to consider only a certain classical volume of confidence, depending on the measurement, within the class of priors. We applied this method to GBR and obtained a remarkable gain in information. In order to facilitate a coherent interpretation of such results, we derived an overall probability of ruin, nesting the classical and the Bayesian type of information by giving subjective knowledge a frequentist interpretation. Readers who do not follow such a potentially controversial proposal still can digest the results derived from the other three updating rules.

5. For comparison we also updated the uncorrelated prior as the standard version of Bayesian learning. This typically leads to much lower probabilities of ruin, when compared to GBR or the weighting method.

Here we propose to consider GBR as the most conservative and easiest to justify rule first. If the outcome is not informative enough, the weighted maximum likelihood update method may be used, although harder to interpret. In case one is willing to accept some classical testing, implying that one accepts to pool with similar potential future cases, one may consider a nested classical-Bayesian method, using the presented nesting formulas. The decision, whether a nested method and, if so, for which level of confidence Q it ought to be used, must be taken *before* taking notice of the actual value of the measurement. The reason for that is that classical statistics pool with potential future realizations.

Independently from these issues one has to select a level of sophistication one thinks experts may be capable of in terms of resolution in parameter space. If one allows for infinite resolution, i.e., δ-type structures, one notoriously ends up with quite non-informative results. The maximum likelihood update methods are prone to shift too much influence to experts that were right for the wrong reason. The only conservative way of dealing with too large a class of priors is to use GBR – and if not informative enough, in combination with a well-designed classical pre-selection plus nesting correction. In Figure 6 it becomes apparent that this method is much more informative over a class with unrestricted resolution than standard GBR. However, when considering the non-parameterized classes at the beginning of this Section, even the classical component could not repair for the too inclusive choice of priors.

Before more sophisticated, constraint-based classes are implemented, we suggest to pragmatically use transforms of the Gaussian class we have discussed in the end of our article. That may serve as a first iteration in order to address absent prior knowledge on multidimensional parameter spaces of complex dynamical models such as climate models.

We are pleased to note that both newly introduced updating methods, i.e. weighted maximum likelihood as well as nested classical GBR, lead to qualitatively identical results for the most realistic class of priors (of those discussed), namely the gradient-based Gaussians.

In summary, we have introduced the class of priors with prescribed marginals as a model for subjective uncertainty. We presented innovations along two dimensions: on learning rules and on further shaping the class of priors by additional filters. As we found GBR as well as maximum likelihood updating dissatisfying, we introduced two new updating rules that represent an interesting trade-off between objectivity and being informative. Finally, we introduced further restrictions on the class of priors that improve the model of subjective uncertainty and either remove spurious information or make the results more informative. For future work it will be an exciting task to see how the various newly introduced updating rules will perform under the curse of dimension, i.e. for increasing number of parameters.

8 Acknowledgement

First of all, we would like to thank half a dozen modelers from the fields of climate science, ecological modeling and economic growth theory, that volun-

teered in our survey on prior beliefs of uncertainty model parameters: E. Bauer, N. Bauer, A. Ganopolski, D. Gerten, M. Hofmann, K. Lessmann, S. Schaphoff. Furthermore we would like to thank R. Klein and J. Schellnhuber for drawing our attention to the issue of purely known correlations in complex systems. Finally, H.H. and E.K. gratefully acknowledge support by the Volkswagen Foundation under grant number II/78470, T.A. by the Deutsche Forschungsgemeinschaft within the SFB 386.

A Analytic treatment for Gaussian priors

A.1 Parameterizing the class of priors

First we recall a well-known (see e.g. [20], pages 22 and 40)

Lemma 3: Let x, γ denote n-dimensional vectors, $P(x)$ a probability density function for x. Let for all x: $y := (\gamma|x)$, $(.)$ denoting the scalar product. Then

1. mean(y)=$(\gamma|$ mean$(x))$,
2. covar(y)=$(|$ covar$(x)|\gamma)$, $(|M|v) := v^t M v$ denoting the standard symmetric quadratic form.

Let P a 2-dimensional Gaussian. Then it can be expressed as

$$P(x) = c\ \mathrm{e}^{-\frac{1}{2}\left(|\Sigma^{-1}|(x-\bar{x})\right)}, \quad c = \frac{1}{(2\pi)\sqrt{\det\Sigma}}, \tag{29}$$

$\bar{x}$ denoting the mean, Σ the covariance matrix (see, e.g., [1]).

We now relate the properties of the first marginal $\sim N(\mu, \sigma)$ to P by noting $x_1 = (\gamma|x)$ with $\gamma := (1,0)^t$. By applying item number 1 of Lemma 3, we conclude that $((1,0)|\bar{x}) = \mu$. The analogous argument holds for the second marginal, hence we establish $\bar{x} = (\mu, \mu)^t$. In analogy, by applying item number 2 of Lemma 3, we derive $\Sigma_{11} = \Sigma_{22} = \sigma^2$. Furthermore we note that Σ must be symmetric as a covariance matrix. Hence, it remains to show that $f \in [-1, 1]$.

In order to do so, we exploit the fact that a symmetric matrix Σ defines a Gaussian in Eq. 29 iff it is positive semi-definite (as a covariance matrix must be), hence its two eigenvalues $\lambda_{1,2} \geq 0$. By deriving $\lambda_{1,2} = \sigma \cdot (1 \pm |f|)$, we conclude

$$\{\Sigma \text{ positive semi-definite}\} \ \Leftrightarrow\ \{|f| \leq 1\}\,. \tag{30}$$

A.2 Derivation of the posterior

A.2.1 Derivation of the bivariate posterior

Let $h_y := 1/\sigma_\eta^2$ (i.e. the "precision" of the likelihood), $x := (x_1, x_2)^t$, $\bar{x} = (\mu, \mu)^t$, $k := (\kappa, 1)^t$. Then according to Bayes' rule, with sorting quadratic and linear terms in x

$$\begin{aligned}
P_{\text{apost}}(x_1, x_2) &= N\left((\mu,\mu), \Sigma\right)(x_1, x_2) \cdot N(\kappa x_1 - x_2, 1/h_y)(y) && (31)\\
&\propto \mathrm{e}^{-\frac{1}{2}Q'} \quad \text{with} && (32)\\
Q' &:= (|A|x) - 2(\gamma|x) \quad \text{and} && (33)\\
A &:= \Sigma^{-1} + h_y\ k \otimes k^t, && (34)\\
\gamma &:= \Sigma^{-1}\bar{x} + y\ h_y\ k. && (35)
\end{aligned}$$

As

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & f \\ f & 1 \end{pmatrix}, \tag{36}$$

$$\Sigma^{-1} = \Gamma \begin{pmatrix} 1 & -f \\ -f & 1 \end{pmatrix}, \quad \text{with} \quad \Gamma := \frac{1}{\sigma^2\ (1-f^2)}. \tag{37}$$

In order to transform Q' to standard form we make the ansatz

$$Q' = (|A|x - x_0) \ + \ C \tag{38}$$

where x_0, C do not depend on x. We determine x_0 by differentiating Eq. 38 w.r.t. x and obtain

$$\begin{aligned} x_0 &= A^{-1}\,\gamma, \quad \text{hence} && (39)\\ P_{\text{apost}}(x_1, x_2) &= N\left(A^{-1}\gamma, A\right)(x_1, x_2). && (40) \end{aligned}$$

A.2.2 Derivation of the posterior marginal in x_1

From the bivariate posterior in Eq. 40 we easily obtain the marginal in x_1 when we recall the following Lemma (see e.g. [4], page 283)

Lemma 4: The marginal density for a single element x_1 is $N(\mu_1^{'}, \Sigma_{11}^{'})$ if $N(\mu^{'}, \Sigma^{'})$ denotes the multivariate density.

Eqs. 18 are then obtained through Eq. 40 and the previous definitions by symbolic manipulation in MATHEMATICA5.2.

A.3 Derivation of the prior probability density of y

Recall that

$$P_{\text{apost}}(x_1, x_2) = N\left((\mu, \mu), \Sigma\right)(x_1, x_2). \tag{41}$$

Let again $x := (x_1, x_2)^t$, $\bar{x} = (\mu, \mu)^t$, $k := (\kappa, 1)^t$. Set $F := (k, x) = \kappa x_1 + x_2$. Utilizing Lemma 3 (see Subsection A.1) we then know

$$\begin{aligned} P_F(F) &= N(\mu_F, \sigma_F^2)(F) \quad \text{with} && (42)\\ \mu_F &= (\gamma|(\mu, \mu)^t) = \mu(1 + \kappa), && (43)\\ \sigma_F &= \sqrt{(\gamma|\Sigma|\gamma)}, && (44)\\ &= \sigma\ \sqrt{1 + 2\kappa f + \kappa^2} && (45) \end{aligned}$$

Then we recall that $P(y|x) = N(F(x), \sigma_\eta^2)(y)$, hence

$$P_y \sim N(\mu_F, \sigma_y^2) \quad \text{with } \sigma_y := \sqrt{\sigma_F^2 + \sigma_\eta^2} \tag{46}$$

So we have derived P_y and can implement Eq. 26 readily:

$$P_{\text{apriori}}(]-\infty, y[) = \int_{-\infty}^{y} dy'\ P_y(y'). \tag{47}$$

In the upcoming Subsection we will further need the following relation: by means of Eq. 45 we show readily that

$$\begin{aligned} \inf_{f\in[-1,1]}\ \sigma_y &= \sqrt{(\sigma(\kappa - 1))^2 + \sigma_\eta^2} =: \sigma_{y.\min}, && (48)\\ \sup_{f\in[-1,1]}\ \sigma_y &= \sqrt{(\sigma(\kappa + 1))^2 + \sigma_\eta^2} =: \sigma_{y.\max}. && (49) \end{aligned}$$

A.4 Derivation of the maximum likelihood priors

$P_y(y) < \infty$ sets also the weight function $W(y)$ according to which we preselect priors for the maximum likelihood updating rule: $W \equiv P_y$. Let μ_y, σ_y the mean and standard deviation of the prior distribution for y. The present class of priors is conveniently parameterized by f which influences σ_y but not μ_y. But standard curve discussion we find that for given y, $W(f)$ is maximized if

$$\sigma_{y.\mathrm{ml}} = |y - \mu_y| \tag{50}$$

in case that equation has a solution for $f \in [-1, 1]$. If we conclude from σ_y on f, we obtain

$$f_{\mathrm{ml}} = \left\{ \begin{array}{ccc} -1 & \text{for} & |y-\mu_y| < \sigma_{y.\min} \\ +1 & \text{for} & |y-\mu_y| > \sigma_{y.\max} \\ \frac{\left((y-\mu_F)^2 - \sigma_\eta^2\right)/\sigma^2 - 1 - \kappa^2}{2\kappa} & & \text{otherwise} \end{array} \right\}. \tag{51}$$

A.5 Derivation of the weighted maximum likelihood result

We note that the derivative $W'(f)$ vanishes at maximum once over $[-1, 1]$, namely for the term given in Eq. 51. Hence, the equation $W(f) = w$, w given, can have at maximum one solution left, and one right from f_{ml}. Either solution is found numerically by specifying $[-1, f_{\mathrm{ml}}]$ and $[f_{\mathrm{ml}}, 1]$ as search intervals. Then we discretize the space for w between $[0, W(f_{\mathrm{ml}})]$ and apply Eq. 7.

A.6 Derivation of the maximum-derivative condition

Let a Gaussian probability density function P be given as

$$P(x) = c\ \mathrm{e}^{-\frac{1}{2}\left(x|\Sigma^{-1}|x\right)}, \quad c = \frac{1}{(2\pi)\sqrt{\det\Sigma}}, \tag{52}$$

x being a two dimensional vector (see [1]).

A.6.1 Derivation of the maximum gradient of P

From elementary manipulations one establishes

$$G(x) := |\mathrm{grad}\ P|^2(x) = P^2(x)\ \left(x|\Sigma^{-2}|x\right) \tag{53}$$

which is the function the maximum of which will be the criterion on whether P will be considered as member of the prior class or not.

In order to determine the maximum of G, we establish the necessary condition for a local maximum:

$$\forall_{i\in\{1,2\}}\ \frac{\partial}{\partial x_i} G = 2\ P\ \frac{\partial}{\partial x_i} P\ \left(x|\Sigma^{-2}|x\right)\ + P^2\ 2\ \left(e_i|\Sigma^{-2}|x\right) = 0, \tag{54}$$

where e_i denote the unit vectors of the coordinate system. Then we note that without loss of generality we can choose the unit vectors identical with the normalized eigenvectors v_i of Σ^{-1} (to which may also belong the eigenvalues λ_i). We conclude

$$\forall_{i\in\{1,2\}} \ -\lambda_i \ (v_i|x) \ \left(x|\Sigma^{-2}|x\right) \ + \ \lambda_i \ (v_i|x) = 0 \tag{55}$$

Case 1: $\lambda_1 \neq \lambda_2$:
Without loss of generality we assume $\lambda_1 > \lambda_2$.

Case 1.1: $(v_1|x) \neq 0$:
Then we conclude

$$\left(x|\Sigma^{-2}|x\right) = \lambda_1. \tag{56}$$

Case 1.1.1: $(v_2|x) \neq 0$:
Then we conclude

$$\left(x|\Sigma^{-2}|x\right) = \lambda_2. \tag{57}$$

However, as $\lambda_1 \neq \lambda_2$, the last two equations cannot be fulfilled simultaneously, hence, Case 1.1.1 can be ruled out.

Case 1.1.2: $(v_2|x) = 0$:

In summary for Case 1 we can conclude that x must be parallel to one of the eigenvectors:

$$\exists_{i\in\{1,2\}} \quad x = \alpha_i v_i \quad \text{with} \quad \alpha_i = 1/\sqrt{\lambda_i}, \tag{58}$$

hence, the local maxima of G are along the eigenvectors at the standard deviations. It is then easily verified that the global maximum of G is along the larger eigenvector.

Case 2: $\lambda_1 = \lambda_2$:

The maximization problem can be reduced to a one-dimensional one due to rotational symmetry in x-space. By identifying the radial coordinate with α_1 of Case 1, one finds that Eq. 58 holds for Case 2 as well.

A.6.2 Operationalization of the gradient information

We now use the above information in order to preselect members of the class of priors according to their maximum norm of the gradients of their densities.

As a reference, we use the maximum gradient of the marginals which in our example are both $\sim N(\mu, \sigma)$. Their maximum gradient G^* is readily derived as

$$G^* = \frac{1}{\sqrt{2\pi \mathrm{e}}} \cdot \frac{1}{\sigma^2}. \tag{59}$$

Let p the dimension of x (in our example $p = 2$). Then for $p > 1$, the gradient of P will have a different unit than that of one of the marginals (in case x has a unit). Hence, a meaningful (in terms on units) restriction of the gradient reads as

$$|\mathrm{grad}P| \cdot (\Delta x)^{p-1} < N \cdot G^*, \tag{60}$$

where Δx denotes the typical scale per coordinate (in our example $\Delta x = 1 = 4\sigma$), N the "expert's resolution" (see Subsection 3.4). The factor $(\Delta x)^{p-1}$ can be interpreted as follows: first, it adjusts units. Second, for a P whose

density fills Δx in $p-1$ coordinates while being denser in a single coordinate, above expression in essence reveals the 1D gradient along the eigenvector v^* with the largest eigenvalue λ^* of Σ^{-1}. If P were higher concentrated in further dimensions as well, the left hand side became larger. Hence, above expression reveals the p-dimensional resolution, equivalent to N of Subsection 3.4.

In order to determine $|\mathrm{grad}P|$ at its maximum, we consider the 1D-function of α, $P((\mu,\mu)+\alpha v^*)$, if v^* is the eigenvector for the maximum eigenvalue λ^* of Σ^{-1}. Let $\sigma' := 1/\sqrt{\lambda^*}$. Then $P((\mu,\mu)+\alpha v^*) = c\sqrt{2\pi}\sigma' N(0,\sigma')(\alpha)$, hence the maximum modulus of gradient of P reads $c\sqrt{2\pi}\sigma' \cdot 1/(\sqrt{2\pi \mathrm{e}}\sigma'^2)$. Combining this information with Eq. 60 leads to a test on σ', and, in turn on Σ and on f.

References

[1] J. O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 1985.

[2] J. O. Berger, D. R. Insua, and F. Ruggeri. Bayesian robustness. In *Robust Bayesian analysis*, volume 125, pages 1–32. Lecture Notes in Statististics, Springer, New York, 2000.

[3] D. Berleant and Jianzhong Zhang. Using Pearson correlations to improve envelopes around the distribution of functions. *Reliable Computing*, 10(2):139–161, 2004.

[4] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis – Forecasting and Control, 3rd Edition*. Prentice-Hall International, Inc., New Jersey, USA, 1994.

[5] W. Cramer, A. Bondeau, F. I. Woodward, I. C. Prentice, R. A. Betts, and et al. Global response of terrestrial ecosystem structure and function to co_2 and climate change: results from six dynamic global vegetation models. *Global Change Biology*, 7:357–373, 2001.

[6] J. Dhaene, M. Denuit, M. J. Goovaerts, R. Kaas, and D. Vyncke. The concept of comonotonicity in actuarial science and finance: theory. *Insurance: Mathematics and Economics*, 31:3–33, 2002.

[7] D. Dubois and H. Prade. *Possibility Theory: An Approach to Computerized Processessing of Uncertainty*. Plenum Press, New York, 1988.

[8] O. Edenhofer, N. Bauer, and E. Kriegler. The impact of technological change on climate protection and welfare: Insights from the model MIND. *Environmental Economics*, 54:277–292, 2005.

[9] C. E. Forest, P. H. Stone, A. P. Sokolov, M. R. Allen, and M. D. Webster. Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science*, 295(5552):113–117, 2002.

[10] M. J. Frank, R. B. Nelsen, and B. Schweizer. Best-possible bounds for the distribution of a sum – a problem of Kolmogorov. *Probability Theory and Related Fields*, 74:199–211, 1987.

[11] A. Ganopolski, V. Petoukhov, S. Rahmstorf, V. Brovkin, M. Claussen, A. Eliseev, and C. Kubatzki. CLIMBER-2: a climate system model of intermediate complexity. Part II: model sensitivity. *Climate Dynamics*, 17:735–751, 2001.

[12] I. Gilboa and D. Schmeidler. Updating ambiguous beliefs. *Journal of Economic Theory*, 59:33–49, 1993.

[13] H. Held and T. Schneider von Deimling. Transformation of possibility functions in a climate model of intermediate complexity. In *Third international conference on Soft Methods in Probability and Statistics (SMPS)*, pages University of Bristol, UK, 5–7 September, 2006. to appear.

[14] M. Hsu, Meghana Bhatt, Ralph Adolphs, Daniel Tranel, and Colin F. Camerer. Neural systems responding to degrees of uncertainty in human decision-making. *Science*, 310:1680–1683, 2005.

[15] E. Kriegler and H. Held. Utilizing random sets for the estimation of future climate change. *International Journal of Approximate Reasoning*, 39:185–209, 2005.

[16] R. C. Pacanowski and S. M. Griffies. *MOM-3 manual*. NOAA / Geophysical Fluid Dynamics Laboratory, Princeton, USA, 1998.

[17] V. Petoukhov, A. Ganopolski, V. Brovkin, M. Claussen, A. Eliseev, C. Kubatzki, and S. Rahmstorf. CLIMBER-2: a climate system model of intermediate complexity. part i: model description and performance for present climate. *Climate Dynamics*, 16:1, 2000.

[18] E. Quaeghebeur and Gert de Cooman. Imprecise probability models for inference in exponential families. In *4th International Symposium in Imprecise Probabilities and Their Applications*. Pittsburgh, Pennsylvania, 2005.

[19] T. Schneider von Deimling, H. Held, A. Ganopolski, and S. Rahmstorf. Climate sensitivity estimated from ensemble simulations of glacial climates. *Climate Dynamics*, DOI 10.1007/s00382-006-0126-8:463–483, 2006.

[20] H. v. Storch and F. W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, 1999.

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

[22] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B*, 58(1):3–57, 1996.

[23] P. Walley. A bounded derivative model for prior ignorance about a real-valued parameter. *Scandinavian Journal of Statistics*, 24:463–483, 1997.