UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It

Grünwald, P.; van Ommen, T.

DOI
10.1214/17-BA1085
10.1214/17-BA1085SUPP

Publication date
2017

Document Version
Other version

Published in
Bayesian Analysis

Link to publication

Citation for published version (APA):
Grünwald, P., & van Ommen, T. (2017). Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It. *Bayesian Analysis*, *12*(4), 1069-1103. https://doi.org/10.1214/17-BA1085, https://doi.org/10.1214/17-BA1085SUPP

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Mar 2023

Supplementary material of "Inconsistency of Bayesian Inference for Misspecified Linear Models, and a Proposal for Repairing It"

Peter Grünwald* and Thijs van Ommen†

A Introduction

In the main text we demonstrated some problems for Bayesian inference under misspecification and showed that SafeBayes effectively solves them. This leaves open the questions whether there also exist problems for Bayes with misspecification that cannot be handled by generalized or SafeBayes (short answer: yes), and under what conditions generalized/SafeBayes can be useful at all (short answer: sometimes even if model is well-specified) and when one needs to resort to a Gibbs likelihood. In the first section below, we answer all these questions and more by providing an overview of all the potential problems with both standard and generalized/SafeBayes under both well- and misspecified models from a frequentist perspective. This list also provides an introduction to the additional sections in this Supplementary Material, in which the issues are sorted out in much more detail. The Supplementary Material continues with Section C which gives additional details on SafeBayes omitted in the main text; Section D which explains in more detail how hypercompression arises, and how even with well-specified models a weak form, *spurious compression*, can arise — this also motivates the differences between I- and R-log-SafeBayes. This is followed by a section that puts SafeBayes in a general context: what should be the goal of Bayesian modelling, if one acknowledges misspecification; and can one design a sort of 'Bayesian misspecification theory'? Section F discusses work related to SafeBayes and existing (in)consistency results. Finally, Section G provides more evidence that the phenomenon observed in our experiments keeps occurring for large samples, so that we have 'real' rather than just 'practical' inconsistency. Along the way, we encounter numerous *Open Problems* which in each case we highlight in italics.

B What can go wrong with standard Bayes under well- and misspecification, and what one can do about it

In the list below, W means that the issue can sometimes be a problem for Bayes even if the model is Well specified; M means that it can only be a problem under Misspecification. For simplicity, we will assume throughout the following that data are *i.i.d.* according to some distribution P^*, the non-i.i.d. case being potentially trickier since we cannot always use Barron's result (16). When we write 'generalized Bayes' we mean raising a standard likelihood to a power $\eta \neq 1$, as in (8). When we write 'Gibbs-Bayes' we mean the more radical step of using an exponentiated summed loss as a cumulative pseudo-likelihood, as in (7).

1. insufficient prior mass (W/M) *generalized Bayes won't help.* In all consistency theorems for both well- and misspecified nonparametric Bayesian methods we know of, one assumes that, for every $\epsilon > 0$, the prior mass of an ϵ-KL-neighbourhoods of either the true (W) or the KL-optimal distribution (M) $\tilde{P}$ is sufficiently large, the rate at which it decreases as $\epsilon \downarrow 0$ influencing the achievable rate of convergence. If this so-called *Kullback-Leibler property* (Walker, 2004) does not hold, then the Bayesian posterior may not concentrate on (KL or any other type of) neighbourhoods of $\tilde{P}$. This problem underlies the well-known inconsistency results of Diaconis

*CWI, Amsterdam and Leiden University, The Netherlands pdg@cwi.nl
†University of Amsterdam, The Netherlands thijsvanommen@gmail.com

and Freedman (1986); it is a problem both for standard and η-generalized Bayes, and can only be solved by using a prior satisfying the KL property — which is usually no problem, many common priors satisfying the property. Since priors with full support in parametric models automatically satisfy the KL property, the problem only plays a role with very large parametric models (overly large sample size needed) and nonparametric models (posterior may never converge). In terms of Barron's result (16), with insufficient prior mass, the term SMALL_n may actually be very large (linear in n).

2. spurious compression (W/M) *generalized Bayes can help, even if model well-specified.* The standard (well-specified) Bayesian consistency and convergence rate results for nonparametric models (by, e.g. (Barron et al., 1999, Ghosal et al., 2000, Walker, 2004, Zhang, 2006a); see also the many follow-up papers of Ghosal et al. (2000)) all require, next to the simple KL property, one or more much more complicated conditions. However, results for the well-specified case by Zhang (2006a), Grünwald and Mehta (2016) that allow η-generalized Bayes with $\eta < 1$ (taking $\eta = 1 - \delta$ for arbitrarily small $\delta > 0$ is good enough) *only* need the KL property. Indeed, Barron's result (16) implies that the complicated additional conditions invariably serve to rule out the possibility of *spurious compression*, a weak form of hypercompression that can occur even for well-specified models. We define it in Section D. In a nutshell, under spurious compression with well-specified models containing P^*, the Bayes predictive distribution converges to P^* and is therefore o.k., but the posterior does not concentrate around P^*, leading to, for example, meaningless credible sets and various other problems; see Appendix F.2 for more discussion. We thus conclude that even if the model is well-specified, generalized Bayes can help to achieve consistency and good convergence rates.

3. hypercompression (M) *generalized Bayes helps.* If the model is misspecified and nonconvex, then there is a potential for 'bad' misspecification (Section 3.1). As the examples in this paper show, in that case standard Bayes may not converge at all and a learning rate $\eta \ll 1$ may have to be chosen to get posterior concentration. Interestingly, Barron's result (16) implies that, both in the well- and the misspecified case, insufficient prior mass (lack of the KL property) and spurious/hypercompression are the *only two* problems that can lead to inconsistency in the sense that the posterior does not concentrate around the KL optimal model, both if the model is well- and if it is misspecified. Thus, assuming that SafeBayes can find the right η (see below), we *only* need the simple 'sufficient prior mass' condition for consistency. In particular, as long as (KL-) consistency and convergence rates are our sole interest, no other modifications to Bayes will ever be needed.

4. single inference task of interest is not KL-associated (M) *Generalized Gibbs-Bayes helps.* Suppose one is interested in learning to predict Y as a linear function of X under a loss function different from squared error loss; e.g., one may be interested in absolute loss

$$\ell_\beta^{\text{abs}}(X, Y) = \left| Y - \sum_{j=0}^{p} \beta_j X_j \right| \tag{B.1}$$

instead; one does not want to make strong assumptions about the noise though (here and below we consider fixed p for simplicity). One might be inclined to use generalized Bayes with the standard linear model, but this is not the right thing to do if one is interested in absolute loss, since the latter is *not* KL-associated with the standard linear model: generalized Bayes with the right η finds the KL-optimal $\tilde{\beta}_{\text{SQ}}$ within the linear model which must also be the squared error optimal predictor, but might not coincide with the optimal absolute loss predictor. To see this consider the simple case with Y independent of the X_i; then the squared error optimal predictor is the mean of Y under P^* and the absolute-loss optimal predictor is the median, and they may be quite different (Jiang and Tanner (2008) provide a beautiful example of non-KL-associated classification loss in a classification context; Hahn et al. (2013) describe a practical case where the task is conditional distribution estimation while the models contain joint distributions). In

cases like this, one can resort to using Generalized Bayes with a pseudo- ('Gibbs') likelihood as in (7), plugging in β in the role of θ and $\ell_\theta := \ell_\beta^{\text{abs}}$. With the right value of η — which can be found by SafeBayes — one can then learn the best predictor $\tilde{\beta}_{\text{ABS}}$ that minimizes, over all β considered, the expected absolute loss

$$\text{RISK}^{\text{abs}}(\beta) := \mathbf{E}_{Z\sim P^*}\left[\left|Y - \sum_{j=0}^{p} \beta_j X_j\right|\right]. \tag{B.2}$$

We stress though that this approach is not as different from a purely likelihood-based approach as one might think: just as an η-Gibbs posterior for the squared error loss is formally equivalent to an η-generalized posterior with a standard likelihood based on the linear model with fixed variance 1, a Gibbs posterior for absolute loss is *formally* equivalent to a generalized posterior with a standard likelihood based on a Laplace density (see, for example, Sriram et al. (2013)); one gets the same posterior on parameters but interprets it in a very different way. In the standard likelihood interpretation, squared error and absolute error predictions, respectively, are by construction KL-associated. We explain all this further in Section E.2 underneath (E.2).

5. multiple inference tasks of interest, not all KL-associated (M) *Generalized Bayes (both Gibbs and non-Gibbs) may help, but much more research needed* Using generalized Bayes with a Gibbs-posterior, a suitable prior and the right value of η, we can guarantee convergence to a predictor that is *optimal* for a loss function ℓ_θ of interest. By adding an extra parameter γ as described underneath (E.2), we can additionally converge to a predictor that is *reliable* in the sense of (6). Sometimes we can even learn more properties; if, for example, we are willing to assume that $\mathcal{M}$ contains the correct regression function, we know that with ℓ_θ the squared error loss and the right η will concentrate on it even if the error distribution is misspecified; similarly, for $\mathcal{M}$ containing a function giving the correct median of Y conditional on X and ℓ_θ the absolute loss. We can view all these prediction tasks as being KL-associated with the probability model implicitly used when adopting a Gibbs posterior. So, we can construct likelihoods/models so that several *related* inference tasks are at the same time KL-associated and we can learn to perform them optimally using SafeBayes. On the other hand, there seem to be severe limits to what we can do if we have several *not-so-related* inference tasks of interest. For example, it is not clear how to set up a Gibbs likelihood that guarantees that generalized Bayes concentrates on *both* the $\tilde{\beta}^{\text{ABS}}$ that is the best absolute loss predictor and the $\tilde{\beta}^{\text{SQ}}$ that is the best squared error predictor, simply because they may not be the same. In Section E we discuss several possibilities for what might be done in such a case, but this is really future work, for which new theory needs to be developed *(Open Problem 1)*. A related open problem *(Open Problem 1c)*, also further discussed underneath (E.2), occurs if one does start out with a standard probability model, and is interested not just in finding the KL-optimal $\tilde{\theta}$, but also in credible sets around it; can one, in general, give meaning to such intervals?

6. when the inference goal is "general modelling" *Generalized Bayes helps — a little* The above points are all purely technical: they tell us whether (generalized or Safe-) Bayes is able to reach a particular precise goal that we might have. But often our goal is pre-predictive and vaguer than that — we simply want to 'understand' the data, to 'model' a phenomenon. In this case, the use of SafeBayes is mostly diagnostic: if it tells us to use an η substantially different from 1, this means that the model cannot be right as it is, and may have to be modified. We return to this issue in Section E.1.

Summarizing, issues 2–4 can be overcome by generalized Bayes with the right η, which depends on the unknown ground truth and thus cannot be directly applied. Assuming that SafeBayes always leads to adoption of such an η, these issues can be successfully addressed by SafeBayes. Note, though, that this 'correctness' of SafeBayes was — up till now — only proved under somewhat restrictive conditions; hence this latter step requires a leap of faith (establishing this correctness of SafeBayes is a major

goal for future work — *Open Problem 2*); see Section C.3 for discussion. Also, other ways to set the learning rate may of course be considered and compared *(Open Problem 3)* as well — see Section F.1 for a list of existing methods.

C Additional Details about Safe Bayes

Here we provide explicit formulas (all derivable using straightforward algebra) on the implementation of SafeBayes for the linear model (Section C.1), which gives some illuminating insights as to its relation to cross-validation with the squared error loss. We provide some further background on our choice of priors for the experiments, and finally we recall existing theoretical results on consistency and optimal convergence rates of SafeBayes.

C.1 Instantiating SafeBayes to the linear model

Consider first a fixed p and σ^2, i.e. a prior that puts mass 1 on this particular p and σ^2. Then the I-log-loss is seen to become the sum, from $i = 0$ to $n-1$, of

$$\mathbf{E}_{\beta\sim\Pi|z^i,p,\eta}\left[-\log f(y_{i+1} \mid x_{i+1},\beta,\sigma^2)\right] = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(y_{i+1} - x_{i+1}\bar{\beta}_{i,\eta})^2, \tag{C.1}$$

where $\bar{\beta}_{i,\eta}$ is as in (13). Note that $\bar{\beta}_{i,\eta}$ depends on η but not on σ. Since the first term of (C.1) does not depend on the data, this version of SafeBayes thus amounts to picking the $\hat{\eta}$ minimizing just the sum of square-loss prediction errors, *which does not depend on the chosen* σ^2. It thus becomes a standard version of 'prequential model selection' as based on the square-loss, which in turn is similar to (though having different asymptotics than) leave-one-out cross validation based on the square-loss. Indeed, this version of SafeBayes can be interpreted in two ways: first, as we did in the main text, in terms of SafeBayes with ℓ_θ in (24) set to the log-loss, i.e. as a tool for dealing with misspecification; and second, with ℓ_θ in (24) set proportionally to the square-loss, as a Gibbs posterior, i.e. a generic tool to learn good square-loss predictors (not distributions) in a pseudo-Bayesian way. We will not experiment with this simplest version of SafeBayes here (we do provide a number of experiments in [GvO], summarized in Section 6) but merely include it to provide more intuition about I-log-SafeBayes.

Instead, the experiments in this paper focus on the more complex case in which σ^2 is equipped with the standard inverse gamma prior with support $\mathbb{R}^+$. Then the η-posterior on σ^2 is an inverse gamma distribution with parameters $a_{n,\eta}$ and $b_{n,\eta}$ as given at the end of Section 2.5. Then the R-log-loss is given by

$$\begin{aligned}
&\mathbf{E}_{\sigma^2,\beta\sim\Pi|z^i,p,\eta}\left[-\log f(y_{i+1} \mid x_{i+1},\beta,\sigma^2)\right] \\
&\quad= \frac{1}{2}\log 2\pi b_{i,\eta} - \frac{1}{2}\psi(a_{i,\eta}) + \frac{1}{2}\frac{(y_{i+1} - x_{i+1}\bar{\beta}_{i,\eta})^2}{b_{i,\eta}/a_{i,\eta}} + \frac{1}{2}x_{i+1}\Sigma_{i,\eta}x_{i+1}^\top \\
&\quad= \frac{1}{2}\log 2\pi\bar{\sigma}^2_{i,\eta} + \frac{1}{2}\frac{(y_{i+1} - x_{i+1}\bar{\beta}_{i,\eta})^2}{\bar{\sigma}^2_{i,\eta}} + \frac{1}{2}x_{i+1}\Sigma_{i,\eta}x_{i+1}^\top + r(i,\eta),
\end{aligned} \tag{C.2}$$

where ψ is the digamma function, $\bar{\sigma}^2_{i,\eta}$ is the η-posterior expectation of σ^2 as given by (14) and $r(i,\eta)$ is a remainder function which is $O(1/i)$ whenever $\sum_{i=1}^n (y_i - x_i\beta_{n,\eta})^2$ increases linearly in n. This final approximation follows by (14) and because we have $\psi(x) \in [\log(x-1), \log x]$. R-log-SafeBayes minimizes the sum of (C.2).

The corresponding in-model version of SafeBayes, *I-log-SafeBayes* minimizes the sum of

$$-\log f(y_{i+1} \mid x_{i+1},\bar{\beta}_{i,\eta},\bar{\sigma}^2_{i,\eta}) = \frac{1}{2}\log 2\pi\bar{\sigma}^2_{i,\eta} + \frac{1}{2}\frac{(y_{i+1} - x_{i+1}\bar{\beta}_{i,\eta})^2}{\bar{\sigma}^2_{i,\eta}}. \tag{C.3}$$

Comparing the two versions of SafeBayes, we see that R-log-SafeBayes has an additional term which decreases in η, increases in model dimensionality p (via the size of the matrix $\Sigma_{i,\eta}$), but becomes negligible for $n \gg p$.

C.2 The Within Model-Priors

The priors we used in our experiments are standard conjugate priors with specific values for hyperpriors, as described in the beginning of Section 5. We used an *informative* prior for Σ_0. This prior equals the posterior we would get by starting with an improper Jeffreys' prior on β and then observing, for each coefficient β_j, one extra point $z = (x, 0)$ with $x_j = 1$ and $x_i = 0$ for $i \neq j$. As a variation, in [GvO] we also ran experiments with a 'slightly informative' Σ_0, where we set $\Sigma_0 = 1000 \cdot \mathbf{I}_{p+1}$, comparable to observing points $z = (x, 0)$ with $x_j = 1/\sqrt{1000}$. As another variation, following the standard reference Raftery et al. (1997), we also used a prior with a level of informativeness depending on the submodel, described in more detail in [GvO]. We report on these variations in the 'Executive Summary' in the main text.

As to the prior on σ^2: Jeffreys' prior is obtained for the choice $a_0 = b_0 = 0$ in (12). We do not use this improper prior, because of the well-known issues with Bayes factors under improper priors (O'Hagan, 1995). Moreover, to calculate the posterior's reliability (defined in Section 5.1 and shown in Figure 5) and also for the I-log-loss, we need to calculate the posterior expectation of the variance σ^2 quantity as given by (14), which is only well-defined and finite for $a_n > 1$. We want to make $\pi(\sigma^2)$ as uninformative as possible while ensuring that (for any positive learning rate) this variance exists for the posterior based on at least one sample. This is accomplished by choosing $a_0 = 1$: for standard Bayes, the posterior after one observation has $a_1 = a_0 + 1/2$; for generalized Bayes, $a_1 = a_0 + \eta/2$. To set b_0, we use that b_0/a_0 represents the sample variance of a virtual initial data sequence (Gelman et al., 2013, Section 14.8). We choose $b_0 = 1/40$ so that $b_0/a_0 = 1/40$, the true variance of the noise in our data.

C.3 SafeBayes learns to predict as well as the optimal distribution

We first define the *Cesàro-averaged* posterior given data Z^n by setting, for any subset $\Theta' \subset \Theta$,

$$\Pi_{\text{CES}}(\Theta' \mid Z^n, \eta) := \frac{1}{n} \sum_{i=1}^{n} \Pi(\Theta' \mid Z^i, \eta) \tag{C.4}$$

to be the posterior probability of Θ' averaged over the n posterior distributions obtained so far. Predicting based on Cesàro-averaged posteriors was introduced independently by several authors (Barron, 1987, Helmbold and Warmuth, 1992, Yang, 2000, Catoni, 1997) and has received a lot of attention in the machine learning literature in recent years, also under the name "on-line to batch conversion of Bayes" or *progressive mixture rule* (Audibert, 2007) or *mirror averaging* (Juditsky et al., 2008, Dalalyan and Tsybakov, 2012), but is of course unnatural from a Bayesian perspective.

The main result of Grünwald (2012) essentially states the following: suppose that, under P^*, the density ratios are uniformly bounded, i.e. there is a finite v such that for all $\theta, \theta' \in \Theta$, $P^*(f_\theta(Y \mid X)/f_{\theta'}(Y \mid X) \leq v) = 1$. Suppose further that the prior Π satisfies the KL-property (Section B). Then Π_{CES} applied with the $\hat{\eta}$ learned by the SafeBayesian algorithm concentrates on the optimal $P_{\tilde{\theta}}$. That is, let Θ_δ be the subset of all $\theta \in \Theta$ with $D(P^* \| P_\theta) \geq D(P^* \| P_{\tilde{\theta}}) + \delta$. Then for all $\delta > 0$, with P^*-probability 1, as $n \to \infty$, we have that $\Pi_{\text{CES}}(\Theta_\delta \mid Z^n, \hat{\eta})$ goes to 0. Grünwald goes on to show that in several settings, one can design priors such that the rate at which the posterior concentrates is minimax optimal, i.e. no algorithm can do better in general. On the negative side, the requirement of bounded density ratio is strong, and the replacement of the standard posterior by the Cesàro one is awkward. On the positive side, the theorem has no further conditions and can be applied to parametric and nonparametric cases alike.

In recent work, Grünwald and Mehta (2016) give bounds for the convergence rates of η-generalized posteriors that also hold for unbounded density ratios, under some further, very weak conditions (in the regression setting considered here, it is sufficient to assume that the third moment of P^* exists). These results strongly suggest that the SafeBayes method for picking η also still works for unbounded density ratios. We also suspect that the need to use the Cesàro-averaged η-generalized posterior in the

results of Grünwald (2012) is an artefact of the proof technique. Establishing these results formally is a major open problem *(Open Problem 2)*. To see whether there is any practical difference, in the ridge regression experiment in the main text (Section 5.4) we included experimental results both for the Cesàro-averaged η-generalized posterior $\Pi_{\text{CES}}(\cdot \mid Z^n, \hat{\eta})$ and for the standard η-generalized posterior $\Pi(\cdot \mid Z^n, \hat{\eta})$. Briefly, the curves look as follows: Cesàroified standard Bayes performs significantly better than standard Bayes in all three quality measures in the wrong-model experiments, but is still not competitive with the two SafeBayes versions. When Cesàroified, the SafeBayes methods become a bit smoother but not necessarily better. We also considered the performance of these three Cesàro-averaged posteriors in the main experiment of Section 5.2, but for lack of space in the graphs, we did not show the corresponding curves. The behaviour we observed was qualitatively identical as in the ridge case: Cesàro makes bad methods significantly better but still not competitive, and good methods smoother, sometimes a bit worse and sometimes a bit better.

Finally, we note one major disadvantage of SafeBayes: even if the data do not have a natural ordering, the $\hat{\eta}$ selected by SafeBayes will, in general, be order-dependent. Grünwald (2011) suggested a very different (and in fact, the first) method to learn $\hat{\eta}$, that does not have this problem. However, it is only applicable to countable models, and has no obvious computationally efficient implementation, so we do not know whether it has a future. One might also try to simply use leave-one-out cross-validation with either the R-log or the I-log loss functions; this gave very similar performance to SafeBayes in additional experiments reported in [GvO]; however, the currently used theory is not at all suitable to deal with (i.e. prove results about) cross-validation like procedures. Finding a general, order-independent version of SafeBayes is thus a main avenue for future work *(Open Problem 4)*.

D Hypercompression and SafeBayes explained in more detail

In this section we explain in more detail how anomalous behaviour of the Bayesian posterior arises; for simplicity, we focus on generalized Bayes with a standard (non-Gibbs) likelihood. First we recall Figure 3 in Section 3.1, the essential picture to understand the phenomenon. If the KL property (Section B) holds, then inconsistency can only arise under a 'bad' form of misspecification, depicted by the figure, which allows for *hypercompression*, the phenomenon that the Bayes predictive distribution $\bar{P}(\cdot \mid Z^i)$ performs, at many i, as well in terms of logarithmic score as or even better than the 'best' $\tilde{P}$, even though it is a mixture of distributions most of which have much larger KL divergence from P^* than $\tilde{P}$. This can happen because the Bayes predictive distribution — a mixture of elements of $\mathcal{M}$ — may be substantially different from any of the elements of $\mathcal{M}$. Somewhat paradoxically, Bayes' overly good log-loss behaviour (hypercompression) is exactly what causes it to perform badly for the KL-associated inference tasks (squared error prediction and reliability, in our case). If one is interested in log-loss prediction, standard Bayes is just fine; the SafeBayesian algorithm should be used if one wants to optimize behaviour against the associated tasks. In contrast to the main text, in which we only spoke of 'hypercompression', we will now make a more fine-grained distinction: from now on we reserve the term *spurious* compression for the phenomenon that the Bayes predictive distribution converges to $\tilde{P}$ (in the sense that its predictions become indistinguishable from that of $\tilde{P}$ for all KL-associated prediction tasks), even though it is a mixture of distributions that are much worse than $\tilde{P}$ in terms of KL divergence to P^*; and *hypercompression* for the case that the Bayes predictive distribution even performs noticeably *better* than $\tilde{P}$. Spurious compression can even occur if the model is well-specified and can be avoided by choosing η close to, but smaller than 1, as advocated (for different reasons) by e.g. Walker and Hjort (2002), Martin et al. (2017) — see Section F.

D.1 Hypercompression

Barron (1998) showed that sequential Bayesian prediction under a logarithmic score function shows excellent behaviour in a cumulative risk sense; for a related result see (Barron et al., 1999, Lemma 4). Although Barron (1998) focuses on the well-specified case, this assumption is not required for the proof and the result still holds even if the model $\mathcal{M}$ is completely wrong. For a precise description

and proof of this result emphasizing that it holds under misspecification, see (Grünwald, 2007, Section 15.2). We now repeat the reasoning of Section 3.1 in the main text in more detail, focusing on Barron's result (15)–(16). Here, as in the remainder of this subsection, we look at the standard Bayes predictive density, i.e. $\eta = 1$. SMALL_n in (16) is the so-called *relative expected stochastic complexity* or *redundancy* (Grünwald, 2007), also known as *information complexity* (Zhang, 2006a, Grünwald and Mehta, 2016), which depends on the prior and for 'reasonable' priors is typically *small* — the more prior mass in neighbourhoods of $\tilde{P}$, i.e. the stronger the KL-property that holds, the smaller. The result thus means that, when sequentially predicting using the standard predictive distribution under a log-scoring rule, one does not lose much compared to when predicting with the log-risk optimal $\tilde{\theta}$.

When $\mathcal{M}$ is a union of a finite or countably infinite number of parametric exponential families and $\tilde{p} < \infty$ is well-defined, then, under some further regularity conditions, which hold in our regression example (Grünwald, 2007), the redundancy is, up to $O(1)$, equal to the BIC term $(\tilde{k}/2) \log n$, where $\tilde{k}$ is the dimensionality of the smallest model containing $\tilde{\theta}$. In the regression case, $\mathcal{M}_{\tilde{p}}$ has $\tilde{p}+2$ parameters $(\beta_0, \ldots, \beta_p, \sigma^2)$, so in the two experiments of Section 5, $\tilde{k} = 6$. Thus, in our regression example, when sequentially predicting with the standard Bayes predictive $\bar{f}(\cdot \mid Z^{i-1})$, the cumulative log-risk is at most $n \cdot \text{RISK}^{\log}(\tilde{\theta})$ which is linear in n, plus a logarithmic term that becomes comparatively negligible as n increases. This is confirmed by Figure 7 in Section 5.3. Now, for each individual $\theta = (p, \beta, \sigma^2)$ we know from Section 2.3 that, if its log-risk is close to that of $\tilde{\theta}$, then its square-risk must also be close to that of $\tilde{\theta}$; and $\tilde{\theta}$ itself has the smallest square-risk among all $\theta \in \Theta$. Hence, one would expect the reasoning for log-risk to transfer to square-risk: it seems that when sequentially predicting with the standard Bayes predictive $\bar{f}(\cdot \mid Z^{i-1})$, the cumulative square-risk should at most be n times the instantaneous square-risk of $\tilde{\theta}$ plus a term that hardly grows with n; in other words, the cumulative square-risk from time 1 to n, averaged over time by dividing by n, should rapidly converge to the constant instantaneous risk of $\tilde{\theta}$. Yet the experiments of Section 5 clearly show that this is *not* the case: Figure 5 shows that, until $n = 100$, it is about 3 times as large. As outlined in the main text, the reason is that the Bayesian predictive density $\bar{f}(\cdot \mid Z^{i-1})$ is a *mixture* of various f_θ, and not necessarily similar to f_θ for any individual θ — the link between log-risk and square-risk (5) only holds for individual $\theta = (p, \beta, \sigma^2)$, not for mixtures of them. Indeed, if at each point in time i, $\bar{f}(\cdot \mid Z^i)$ would be very similar (in terms of e.g. Hellinger distance) to some particular f_{θ_i} with $\theta_i \in \Theta$, then there would really be a contradiction. Thus, the discrepancy between the good log-risk and bad square-risk results in fact *implies* that at a substantial fraction of sample sizes i, $\bar{f}(\cdot \mid Z^i)$ must be substantially different from *every* $\theta \in \Theta$. In other words, *the posterior is not concentrated at such i*. A cartoon picture of this situation was given in Figure 3 in Section 3.1: the Bayes predictive achieves small log-risk because it mixes together several distributions into a single predictive distribution which is very different from any particular single $P_\theta \in \mathcal{M}$. As shown in Section 5.3, this indeed happens in our example, to such an extent that we actually get hypercompression until about $n = 100$.

D.2 I-log-SafeBayes avoids hypercompression; R-log-SafeBayes avoids spurious compression

Taking η smaller than 1 has the effect of additional regularization. As indicated by earlier works such as Audibert (2004), Barron and Cover (1991), Zhang (2006b), and as shown in great generality by Grünwald and Mehta (2016), there usually exists a critical $\bar{\eta}$ such that, for any $\eta < \bar{\eta}$, the η-generalized posterior concentrates on $\tilde{P}$ as sample size increases, often even at the best possible (minimax) rate. This $\bar{\eta}$ coincides with the $\bar{\eta}$ at which the generalized posterior has a second interpretation as a standard posterior for a reweighted model with densities (17), as explained in the main text. However, none of the just mentioned papers says anything about how to find such an η. Dawid's prequential point of view suggests to take the η minimizing

$$\sum_{i=1}^{n} \text{LOSS}(x_i, y_i; \Pi \mid z^{i-1}, \eta), \tag{D.1}$$

for some loss function which measures, at each i, how well the posterior based on data z^{i-1} predicts y_i given x_i. The idea is that the best η on the data so far will also lead to good predictions in the future.

The most obvious choice to fill in for LOSS would be $-\log \mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}[f(y_i \mid x_i, \theta)]$ as in (22); as explained above (22), that would make inference of η be very similar to empirical Bayes. If we want to use our posterior, in the end, to make predictions scored by log-loss (as we would if we were to use it for data compression or sequential gambling), this would be perfectly fine. However, as already indicated in the main text, as log-loss is not the primary interest, this will not work due to the hypercompression phenomenon, which, as we confirmed in separate experiments [GvO, Figure 13] is at its strongest at $\eta = 1$ in our model-wrong setting: the predictive $\mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}[f(Y_i \mid x_i, \theta)]$ will have smaller log-loss, but larger squared loss as η increases to 1, and if we use (22), we will tend to select $\eta = 1$ (as an aside, this also means that if we are only interested in sequential log-loss prediction, there is nothing wrong with using $\eta = 1$, i.e. standard Bayes works fine for log-loss prediction under misspecification).

Least ambitious goal: Good squared error performance Now consider the case that we are solely interested in squared loss predictions. Then a straightforward alternative would be to take LOSS in (D.1) as $(y_i - x_i\bar{\beta}_{i-1,\eta})^2$ with $\bar{\beta}_{i-1,\eta}$ as in (13). In fact, this is exactly what the I-version of SafeBayes does if we put a point prior on an arbitrary but fixed σ^2 (Eq. (C.1)), and as shown in [GvO], we do get a competitive method for learning the regression function; however, the method can obviously not be used for reliable estimation of its own prediction error, since the 'true' variance may be very different from σ^2.

Intermediate ambition: Good performance for all KL-associated prediction tasks by avoiding hypercompression If we are interested in good prediction for *all* associated prediction tasks as listed in Section 2.3, then it makes sense to set LOSS in (D.1) as $-\log f(y_i \mid x_i, \mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}[\theta])$, i.e. to minimize (25), which is just I-log-SafeBayes. With this choice, we prefer η's such that the η-posterior mean rather than the predictive has small log-loss. Small log-loss of the posterior mean implies that it is likely to have small log risk. Because, unlike the posterior predictive, the posterior mean is a member of the model $\mathcal{M}$ (here we assume the model is parameterized in such a way that the parameter set is convex, as is the case for our regression problem), small log risk implies good performance in terms of all associated prediction tasks, i.e. small squared error risk and good reliability. This motivates the use of I-log-SafeBayes.

Strongest ambition: Posterior concentration on distributions close to $\tilde{\theta}$ by avoiding spurious compression Focusing on the posterior mean as in I-log-SafeBayes still allows for the milder form of hypercompression which we have called *spurious compression* above, and to which we get back in Section F.2: it may be that the posterior *mean* is almost optimal because the posterior predictive is almost equal to $P_{\tilde{\theta}}$, yet this posterior predictive is still constituted as a weighted average of distributions all of which are very far from $\tilde{\theta}$. If we want a posterior with, e.g., meaningful credible sets, then this is highly undesirable. We may thus try to be even more ambitious and require finding an η such that the posterior is not just close to $\tilde{\theta}$ 'on average' but rather puts most if its mass on distributions which are close to the KL-optimal $\tilde{\theta}$. We can achieve this by using the prequential approach with a loss function that only gets small if we achieve small log-loss by predicting using a random draw from our posterior, rather than plugging in the posterior mean. And this is exactly what is achieved by setting LOSS in (D.1) to be the Gibbs loss $\mathbf{E}_{\theta\sim\Pi|z^{i-1},\eta}[-\log f(y_i \mid x_i, \theta)]$, i.e. minimizing η as in (23) which is just R-log-SafeBayes. This motivates R-log-SafeBayes as the version with the most ambitious goal... at least in this paper. In the next section we describe an initial idea, based on a modification of the likelihood as in (E.2), which might be used if we, even more ambitiously, aim to get good predictions for tasks that are not KL-associated.

Summarizing, the goals of R-log-SafeBayes are more ambitious, but for some applications, they may be overkill — we notice in our experiments that we need more data before getting good results than

with I-log-SafeBayes. I-log-SafeBayes needs less data, but, in addition to being less ambitious, has some other weaknesses: in contrast to R-log-SafeBayes, it is dependent on the chosen parameterization, and cannot even be applied whenever parameterizations are not convex. R-log-SafeBayes can be used whatever the model, and is fully parameterization independent.

Remark on hypercompression and concentration The above suggests that the R-versions of SafeBayes select an η at which the posterior will be more concentrated than at $\eta = 1$. Upon closer inspection, things are more subtle though — for simplicity we illustrate this only for the posterior in ridge regression (single high-dimensional model) with a degenerate prior on a fixed σ^2. As can be seen from the definition of the covariance matrix $\sigma^2\Sigma_{n,\eta}$ underneath (13), as one decreases η, the determinant of the posterior covariance matrix based on data Z^n will grow, indicating *less* concentration. However, the posterior mean $\bar{\beta}_{n,\eta}$ will change as well. When η goes below some critical value, then $\bar{\beta}_{n,\eta}$ becomes approximately $\tilde{\beta}$. At that point the posterior, although not very concentrated, will lead to good predictions. At $\eta = 1$, at sample sizes smaller than $n = 300$, $\bar{\beta}_{n,\eta}$ has a ridiculous value leading to bad predictions, yet the posterior is more concentrated than at small η. Still, it is *not concentrated enough* to avoid hypercompression — the Bayes predictive can have substantially smaller log-loss than the posterior mean $\bar{\theta}$. If we allowed $\eta > 1$, we would actually even get more concentration, with $\eta = \infty$ corresponding to a posterior concentrated in one point — but note that, as soon as $\eta > 1$, Barron's bound is not valid any more, so this would not lead to any contradiction.

So, roughly speaking, one might say that standard Bayes can perform badly if the standard posterior is centred at, but not too much concentrated at, some 'bad' $\bar{\theta}_{n,\eta}$; whereas with the $\eta \ll 1$ selected by R-log-SafeBayes, generalized Bayes will perform better because the η-posterior is centred at, yet even less concentrated at, a much better $\bar{\theta}_{n,\eta}$. Because $\bar{\theta}_{n,\eta}$ has such small KL-risk, the set of densities around it that now receive substantial posterior mass all perform quite well in terms of KL-risk, so even though we have less concentration (larger posterior variance) than at $\eta = 1$, we are still able to say that 'the posterior is now concentrated on good densities'.

E Discussion

In this section, we discuss the place of SafeBayes in Bayesian methodology, and we speculate what, on top of SafeBayes, may be needed for a general Bayesian 'theory of misspecification'.

E.1 SafeBayes and the Data-Analysis Cycle

> "If a subjective distribution Π attaches probability zero to a non-ignorable event, and if this event happens, then Π must be treated with suspicion, and *modified* or replaced" (emphasis added)
> — A.P. Dawid (1982).

Following the well-known Bayesian statisticians Box (1980), Good (1983), Dawid (1982, 2004) and Gelman (2004) (see also Gelman and Shalizi, 2012), we take the stance that model checking is a crucial part of successful Bayesian practice. When there is a large discrepancy between a model's predictions and actual observations, it is not merely sufficient to keep gathering data and update one's posterior: something more radical is needed. In many such cases, the right thing to do is to either expand or adjust the model (in directions inspired by the observed discrepancies), or sometimes even to go back to the drawing board and try to devise a more realistic model from scratch. This is definitely the situation we are in if we are mainly interested in *understanding* the phenomena underlying the data.

However, we think this story is incomplete: in machine learning and pattern recognition, one often encounters situations in which the model employed is *obviously* wrong in some respects, yet there is a model instantiation (parameter vector) that is *pretty adequate* for the specific prediction task one is interested in. Examples of such obviously-wrong-yet-pretty-adequate models are, like in this paper, assuming homoskedasticity (and unimodality) in linear regression when the goal is to approximate

the true regression function and the true noise is heteroskedastic (and/or perhaps multimodal), but also the use of N-grams in language modelling (is the probability of a word given the previous three words really independent of everything that was said earlier?), and independence in text modelling as used in Bayesian *topic models* (Blei, 2012), logistic regression in e.g. spam filtering, and every single successful data compression method that we know of (see *Bayes and Gzip* (Grünwald, 2007, Chapter 17, page 537)). The difference with the more standard statistical (be it Bayesian or frequentist) mode of reasoning is eloquently described in Breiman's (2001) *the two cultures*.[1] Bayesian inference is among the most successful methods currently used in the obviously-wrong-yet-pretty-adequate-situation (to witness, state-of-the-art data compression methods such as Context-Tree-Weighting (Willems et al., 1995) have a Bayesian interpretation). Yet the present paper shows that there is a danger: even *if* the employed model is pretty adequate (in the sense of containing a pretty good predictor), the Bayesian machinery might not be able to find it. The SafeBayesian algorithm can thus be viewed as an attempt to provide an alternative for the *data-analysis cycle* (Gelman and Shalizi, 2012) to this, in some sense, less ambitious setting: just like in the standard cycle, we do a model check, albeit a very specific one: we check whether one can outperform Bayesian predictions by a different, pseudo-Bayesian method (i.e. η-generalized Bayes); this should be impossible if model and prior are adequate for the data at hand. If one can, then we know that we may not be learning to predict as well as the best predictor in our model, so we *modify* our posterior. Not in the strong sense of 'expanding the model' or even 'going back to the drawing board', but in the much weaker sense of making the learning rate smaller — we cannot hope that our model of reality has improved, because we still employ the same model — but we can now guarantee that we are doing the best we can with our given model, something which may be enough for the task at hand and which, as our experiments show, cannot always be achieved with standard Bayes.

Summarizing, if the main goal of inference is to provide specific, KL-associated predictions, then SafeBayes can be used to select η. If the main goal is simply to understand the data, then SafeBayes can still be employed, not to use a different η but mostly as a (prior) predictive check, itself suggesting other (prior or posterior) predictive checks: if $\eta \ll 1$ leads to better predictions, this means that something is inherently wrong and an inspection of the full posterior or the posterior predictive can suggest useful model expansions and modifications.[2] A case in point of a posterior predictive check/diagnostic tool is Figure 4, which shows that the predictive distribution is very different from any of the distributions in the model, clearly indicating a need for model revision or expansion.

But what if one has an imperfect model at hand and the inference task of interest is not KL associated? Or if the goal of inference is 'somewhere inbetween' doing specific predictions/inferences and 'simply understanding the data'? These are the points we now turn to.

E.2 Towards a general theory of statistical inference under misspecification

The standard Bayesian approach is very ambitious: it can be used to solve every conceivable type of prediction or inference task. Every such task can be encoded as a loss or utility function, and, given the data and the prior, one merely has to calculate the posterior, and then make an optimal decision by taking the act that minimizes expected loss or maximizes expected utility according to the posterior. Crucially, one uses the same posterior, independently of the utility function at hand, implying that one believes that one's own beliefs are correct *in every possible respect*. We envision a more modest approach, in which one acknowledges that one's beliefs are only adequate in some respects, not in others; how one proceeds then depends on how one's model and inference task(s) of interest interact.

[1] The 'two cultures' does *not* refer to the Bayesian-frequentist divide, but to the modelling vs. prediction-divide. We certainly do not take the extreme view that statisticians should *only* be interested in prediction tasks such as classification and square-error prediction rather than density estimation and testing; our point is merely that in some cases, the goal of inference is clearly defined (it could be classification, but it could also be determination whether some random variables are (conditionally) (in)dependent), whereas part of our model is unavoidably misspecified; and in such cases, one may want to use a generalized form of Bayesian inference.

[2] We are using 'prior/posterior predictive checks' in a loose, nonformal sense here. We are thankful for one of the referees for bringing this point to our attention.

In this paper, we mostly looked at cases in which the main inference tasks were simply to predict well for a given loss function such as squared error, and to get a reliable estimate of one's own prediction error. Often times, though, one may be interested in more general tasks such as 'assessment of every random variable in some large class of random variables $\mathcal{U}$ should be *reliable*,' in the sense that the posterior concentrates on a distribution $\tilde{P}$ satisfying, for every random variable $U : \mathcal{Z} \to \mathbb{R}$ with $U \in \mathcal{U}$,

$$\mathbf{E}_{Z\sim P^*}[U] = \mathbf{E}_{Z\sim \tilde{P}}[U], \tag{E.1}$$

i.e. with respect to U, the world behaves as if $\tilde{P}$ were correct even though in reality $\tilde{P}$ is not correct. Well-known probabilistic notions such as distributions which may be incorrect yet still *calibrated* can be understood as specifying a set $\mathcal{U}$ for which (E.1) holds. A first step towards such a *general theory of misspecification* is given by the paper *Safe Probability* (Grünwald, 2017). The upshot is that, (a), if one thinks one's model may be misspecified, one always explicitly states for which random variables/inference tasks it is still *safe*; and that (b), one can, for some types of models, *guarantee* safety relative to some tasks, since (essentially) these are KL associated.

(Grünwald, 2017) focuses on reliability for general random variables, and does not consider the particular case of data-dependent prediction for given loss functions, which is the focus of this paper and which we consider to be a crucial part of a future theory of misspecification. Here, a main open question is what to do if one has a probability model at hand which one does believe captures some aspects of the domain reasonably well, yet one does not necessarily think that the predictions, based on the KL optimal distribution in the model, are necessarily also the best in terms of one or more other loss functions of interest that are not KL-associated. In that case, one can either *modify* the model to make the prediction task of interest KL associated, or otherwise one can *expand* the model to make it behave better in terms of the desired prediction task. The latter then has a dual interpretation, as a pragmatic improvement for just a task of interest but also as an improvement of the model in terms of the data-analysis cycle referred to above.

Modifying the Model Suppose the inference task at hand is simply prediction under some loss function $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbf{R}$. In this case, if the ℓ-risk is not KL-associated this simply means that the log risk is not a monotonic function of the risk in terms of the loss ℓ. To get the desired association, we may associate each conditional distribution $P_\theta(Y \mid X)$ in the model with its associated Bayes act δ_θ: $\delta_\theta(x)$ is defined as the act $\hat{y} \in \hat{\mathcal{Y}}$ which minimizes $P_\theta \mid X = x$-expected loss $\mathbf{E}_{Y\sim P_\theta|X=x}[\ell(y, \hat{y})]$. We can then define a new set of densities

$$f^{\text{NEW}}_{\theta,\gamma}(y \mid x) = \frac{1}{Z(\gamma)} e^{-\gamma \ell(y, \delta_\theta(x))}, \tag{E.2}$$

and perform (generalized) Bayesian inference based on these. Note that this effectively replaces, for each θ, the full likelihood by a 'likelihood' in which some information has been lost, and is thus reminiscent of what is done in *pseudo-likelihood* (Besag, 1975), *substitution likelihood* (Jeffreys, 1961, Dunson and Taylor, 2005), or *rank-based likelihood* (Gu and Ghosal, 2009) approaches (as a Bayesian, one may not want to lose information, but whether this still applies in nonparametric problems (Robins and Wasserman, 2000) let alone under misspecification (Grünwald and Halpern, 2004) is up to debate).

(E.2) can be made precise in two ways: either one just sets γ and $Z(\gamma)$ to 1, and allows the f^{NEW}_θ to be pseudo-densities, not necessarily integrating to 1 for each x. This is now effectively equivalent to adopting Gibbs-type pseudo-likelihood approach, the posterior based on (E.2) coinciding with the Gibbs likelihood (7); setting up a likelihood in this way is a standard approach in learning theory (Zhang, 2006b, Catoni, 2007). One could then learn η by, e.g., the basic SafeBayes algorithm with $\ell_\theta(x, y) := \ell(y, \delta_\theta(x))$ instead of log-loss. For example, if, as in the main text, we start with a linear model but we are interested in predicting not with squared error but rather with absolute loss as in (B.1), then following the procedure above, we will associate each linear predictor $\theta = (p, \beta, \sigma)$ with $\delta_{(p,\beta,\sigma)} := \beta$, where predictor β incurs loss $|y_i - \sum \beta_j x_{ij}|$ at time i. The reason why β is not modified is that, according to the linear model, the noise is Gaussian, which is symmetric, so that for every symmetric loss function, including the absolute loss, the optimal predictions according to any $\theta = (p, \beta, \sigma)$ are given by predicting Y using the conditional mean $\mathbf{E}_{P_\theta}[Y|X]$.

Modifying and/or Expanding the Model There is, however, an alternative way to deal with and interpret (E.2): one could define $Z(\gamma)$ so that the densities normalize (how to achieve this if $\int_y e^{-\gamma \ell(y,\delta_\theta(x))} dy$ depends on x is nontrivial but possible, as explained by Grünwald (2008)) and putting a prior on γ as well (for the squared error loss, this is akin to putting a prior on the variance). This will make the loss ℓ KL-associated and the $\tilde{\theta}$-component of the KL-optimal $(\tilde{\theta}, \tilde{\gamma})$ optimal in the sense of minimizing ℓ_θ-risk. If one uses a degenerate prior on a single value of γ, then the resulting procedure will be formally identical to the Gibbs posterior, showing that the Gibbs likelihood can also be interpreted as a standard likelihood with a special probability model for which the loss function of interest is KL associated.

If we put a nondegenerate prior on γ, we achieve more: we will also have the reliability property, $\tilde{\gamma}$ playing a role analogous to $\tilde{\sigma}$ in the squared error case; see again (Grünwald, 2008) for details. In this case we will get, with $z_i = (x_i, y_i)$, $\ell_\theta(z_i) := \ell(y_i, \delta_\theta(x_i))$, and using a prior on Θ and the scaling parameter γ, that the η-generalized posterior becomes

$$\pi(\theta, \gamma \mid z^n, \eta) \propto \frac{1}{Z(\gamma)^{\eta n}} e^{-\eta\gamma \sum_{i=1}^n \ell_\theta(z_i)} \cdot \pi(\theta, \gamma). \tag{E.3}$$

The idea of adding γ was, in essence, already suggested by (Grünwald, 1998, Example 5.4) (see also Grünwald (1999)) under the name of *entropification* (however, Grünwald's papers wrongly suggest that, by introducing the scale parameter γ, it would be sufficient to only consider $\eta = 1$); for related ideas see also (Lacoste-Julien et al., 2011) and (Friel and Stoehr, 2015).

But now we started adding parameters to the model, we could even go one step further. If we start out with the probability model $\{f_\theta : \theta \in \Theta\}$ but are interested in non-KL-associated loss ℓ_θ, then another proposal, briefly alluded to by Grünwald (2016) (in response to Watson and Holmes (2016)), is to use, instead of (E.2),

$$f^{\text{NEW}}_{\theta,\gamma_1,\gamma_2}(y \mid x) = \frac{1}{Z(\gamma_1, \gamma_2)} e^{-\gamma_1 \ell(y,\delta_\theta(x))} \cdot f_\theta(y|x)^{\gamma_2}, \tag{E.4}$$

where f_θ is the density of P_θ and $Z(\gamma_1, \gamma_2)$ now normalizes over both factors; then the KL-optimal $(\tilde{\theta}, \tilde{\gamma}_1, \tilde{\gamma}_2)$ will still be reliable, and, no information about the data is lost in this new likelihood; but it is not clear whether $\tilde{\theta}$ will still be optimal for the ℓ-risk — whether in practice to use (E.3) or (E.4) or something else is just one of many open questions *(Open Problem 1a)* in what we would like to call 'misspecification theory': a to-be-developed theory about how to adjust models such that guarantees for desired inference tasks can be achieved even if the model is (and stays) wrong *((general) Open Problem 1)*.

As suggested by a referee, another major open question here is whether it is possible to get inference in which one of the tasks of interest is to derive valid *credible sets* around the KL-optimal parameter. While we do not know whether this is possible in full generality, it may definitely be possible to get meaningful substitutes for credible sets. One can, in fact, think of the reliability property in these terms: instead of reporting a credible set such as 'with posterior probability 0.95 the squared-error optimal $\tilde{\beta}$ lies in region $\mathcal{E}_{\bar{\beta}}$ around the posterior mean $\bar{\beta}$', we report 'in posterior expectation, the prediction error we make by using $\bar{\beta}$ is so-and-so-large'. We can then prove that, at least for large samples, this statement will be correct also in a frequentist sense; just as in well-specified Bayes, under regularity conditions we have coverage: credible sets are asymptotically also confidence sets. Using PAC-Bayesian methods (McAllester, 2003), one might even hope to prove *finite sample bounds* that say that, for given n, with high probability, the estimated prediction error is not farther off from the real future prediction error than some explicitly given ϵ_n. Like a credible-set statement, the reliability-statement is a meta-statement about the inferred $\bar{\beta}$, telling us how much faith we can put in it; yet the detailed meaning is, of course, very different *(Open Problem 1b)* – see also the discussion of Syring and Martin (2017) in Section F.1 below.

More generally, what would one like to ideally develop here is a general theory of substitution likelihoods so that likelihoods can be adjusted based on the inference task of interest for *arbitrary* sets

of prediction tasks. Existing approaches such as pseudo-likelihood and rank-based likelihood would also become a special case of this method *(Open Problem 1d)*. If this can be done, we would have a truly generalized Bayesian method.

Potential for Criticism Now both 'pure' subjective Bayesians and 'pure' frequentists might dismiss this programme as severe ad-hockery: the strict Bayesian would claim that nothing is needed on top of the Bayesian machinery; the strict frequentist would argue that Bayesian inference was never designed to 'work' under misspecification, so in misspecified situations it might be better to avoid Bayesian methods altogether rather than trying to 'repair' them. We strongly disagree with both types of purism, the reason being the ever-increasing number of successful applications of Bayesian methods in machine learning in situations in which models are obviously wrong. We would like to challenge the pure subjective Bayesian to explain this success, given that the statistician is using a priori distributions that reflect beliefs which she knows to be false, and are thus not really her beliefs. We would like to challenge the pure frequentist to come up with better, non-Bayesian methods instead. In summary, we would urge both purists not to throw away the Bayesian baby with the misspecified bath water!

One variation of such criticism deserves further attention: if there is just a single loss function of interest, such as e.g. squared loss, then why don't we just adopt as our 'model' a set of predictors for that loss function, and infer a good predictor by penalized error minimization, such as, e.g., — for squared error loss — the Lasso? The answer is threefold: first, if a single loss function is really our *only* task of interest, then indeed standard penalization approaches are totally fine, but the I-log version of SafeBayes will actually give us something very similar to such standard methods — for the squared error loss this was explained in Section C.1. As shown by De Heide (2016), the I-log-version of SafeBayes with a Gibbs likelihood for squared error tends to select an η such that $1/\eta$ is close to the parameter selected by the standard Lasso with leave-one-out cross validation, and gives comparable performance on real-world data sets, in some cases significantly outperforming the Bayesian Lasso.

The second part of the answer is that we *are* often interested in more than one inference task. For example, even if the main goal is squared error prediction, one may want the additional meta-property of *reliability*, and it is not clear how to do this using simple penalized likelihood — whereas with SafeBayes one gets it for free by putting a prior on γ (i.e. σ^2, in the special case of squared error). The alternative would be to equip the Lasso with frequentist confidence guarantees, which in practice are often too conservative to be useful.

The third part of the answer is that in practice it is often the most intuitive by far to work with 'real' likelihoods, probability models that may be known to be misspecified and useful at the same time. Examples are the aforementioned topic models, but also e.g. phylogenetic tree models which are known to be affected by misspecification issues (Yang, 2007). Here there seems no easy way to meaningfully represent the inference problem at hand (establishing a distribution over topics given a document or a 'likely' phylogenetic tree given DNA data) as a penalized loss minimization problem, and it seems best to go with Bayesian inference with a modification or expansion of the model to make inference reliable relative to the inference problem at hand, as well as using SafeBayes to make sure that the posterior concentrates.

A final point here is that from a learning theory (citations see below) and Minimum Description Length (MDL (Barron et al., 1998)) perspective, the extension from Bayes to SafeBayes is, in fact, perfectly natural, as we now explore.

F Related work

F.1 Related Work I: Learning theory, MDL, and Finding η by Other Methods

Learning Theory From the learning theory perspective, generalized Bayesian updating as in (E.3) with $Z(\gamma)$ set to 1 can be seen as the result of a simple regularized loss minimization procedure (this

was probably first noted by Williams (1980); see in particular (Zhang, 2006b)), which means that it continues to make sense if $\exp(-\gamma\ell_\theta)$ as in (E.2) does not have a direct probabilistic interpretation. Variations of such generalized Bayesian updating are known as "aggregating algorithm", "Hedge" or "exponential weights", and often are provably almost-worst-case optimal in nonstochastic settings (Vovk, 1990, Cesa-Bianchi and Lugosi, 2006) — but to get these the learning rate must often be set as small as $O(1/\sqrt{n})$. Similarly, PAC-Bayesian inference (Audibert, 2004, Zhang, 2006b, Catoni, 2007) (for a variation, see (Freund et al., 2004)) is also based on a posterior of form (E.2) and can achieve minimax optimal rates in e.g. classification problems by choosing an appropriate η, usually also very small. From this perspective, SafeBayes can be understood as trying to find a *larger* η than the worst-case optimal one, if the data indicate that the situation is not worst-case and faster learning is possible. The question how to choose η is often left open; several authors (e.g. Jiang and Tanner (2008)) suggest cross-validation. Similarly, essentially the same algorithm as SafeBayes was proposed independently, without convergence proofs, as a pragmatic way for determining learning rates when combining expert predictions, by V. Mallet (Gerchinovitz et al., 2008), and was shown to work excellently in practice by Devaine et al. (2013). Besides SafeBayes, we know of four other methods for determining η for which convergence results have been established: Audibert's (2004) method, which is restricted to classification applications; the method of Grünwald (2011), which however only works for countable models, and the LLR (*learning the learning rate*) (Koolen et al., 2014, De Rooij et al., 2013) and SQUINT (Koolen and Van Erven, 2015) methods which were both designed for a sequential prediction context. The latter is especially interesting since, pleasingly from a Bayesian perspective, it *integrates out* η. This would not work if one used the standard likelihood (our experiments suggest it would perform similarly to determining η by empirical Bayes, which works very badly — see Section 5.4). Yet, by a subtle 'second order' modification of the likelihood, it is possible after all.

Other Methods for Setting η The above approaches all give provable frequentist performance guarantees for fixed loss functions; there also exist other approaches, mostly developed within the Bayesian statistics community (though not necessarily 'Bayesian') for setting η with different desiderata. We mention Bissiri et al. (2016) who present several methods for determining η based on coherence-type arguments; one of these ('unit information loss approach', Section 3.2) is, when ℓ_θ is the squared error, very similar to setting η equal to $1/(2\hat{\sigma}^2)$, where $\hat{\sigma}^2$ is the empirical Bayes estimate of the variance. As we show in [GvO], this is not competitive in our setting, having the same problems as the problems with setting η by empirical Bayes explained in Section 5.4. Other interesting techniques and motivations are given by Miller and Dunson (2015), Holmes and Walker (2017), Syring and Martin (2017). Especially the latter is interesting since it sets η so as to get frequentist coverage of Bayesian credible intervals for misspecified models, which, as we hinted above, is an important desideratum. Their encouraging empirical results show that this may often be possible; the price is to pay is that they need to assume consistency for all fixed η, something that, as our work shows, may fail in general. Another major goal (not necessarily done by us!) for future work is *Open Problem Nr. 3*: to provide a thorough comparison of all these methods for setting η.

MDL Of particular interest is the interpretation of the SafeBayesian method in terms of the MDL principle for model selection, which views learning as data compression. When several models for the same data are available, MDL picks the model that extracts the most 'regularity' from the data, as measured by the minimum number of bits needed to code the data *with the help of the model*. This is an interpretation that remains valid even if a model is completely misspecified (Grünwald, 2007). The resulting procedure (based on so-called *normalized maximum likelihood* codelengths) is operationally almost identical to Bayes factor model selection. Thus, it provides a potential answer to the question 'what does a high posterior belief in a model really mean, since one knows all models under consideration to be incorrect in any case?' (asked by, e.g., Gelman and Shalizi (2012)): even if all models are wrong, the information-theoretic MDL interpretation stands. However, our work implies that there is a serious issue with these NML codes: note that any distribution P in a model $\mathcal{M}$ can be mapped to a code (the *Shannon-Fano code*) that would be optimal in expectation if data

were sampled from P. Now, our work shows that if the data are sampled from some $P^* \notin \mathcal{M}$, then the codes based on Bayesian predictive distributions can sometimes compress substantially *better* in expectation than can be done based on any $P \in \mathcal{M}$ — this is the hypercompression phenomenon of Section 3.2. The same thing then holds for the NML codes, which tend to assign almost the same codelengths as the Bayesian ones. Our work thus invalidates the interpretation of NML codelengths as 'compression with the help of (and only of!) the model', and suggests that, similarly to in-model SafeBayes one should design and use 'in-model' versions of the NML codes instead — codes that are guaranteed not to outperform, at least in expectation, the code based on the best distribution in the model.

F.2 Related work II: Frequentist analysis of Bayesian behaviour with and without misspecification

Consistency theorems The study of consistency and rate of convergence under misspecification for likelihood-based and specifically Bayesian methods goes back at least to Berk (1966). For recent state-of-the-art work on likelihood-based, non-Bayesian methods see e.g. Dümbgen et al. (2011) and the very general Spokoiny (2012). Recent work on Bayesian methods includes Kleijn and Van der Vaart (2006), De Blasi and Walker (2013) and Ramamoorthi et al. (2015) who obtained results in quite general, i.i.d. nonparametric settings and Grünwald and Mehta (2016) who substantially generalized Zhang's results for the misspecification setting. For results in non-i.i.d. settings, see (Shalizi, 2009), and for results in more specific settings, see (Sriram et al., 2013).

Yet, as explicitly remarked by De Blasi and Walker (2013), the conditions on model and prior needed for consistency under misspecification are generally stronger than those needed when the model is correct. Essentially, if the data are i.i.d. both according to the model and the sampling distribution P^*, then Theorem 1 (in particular its Corollary 1) of De Blasi and Walker (2013) implies the following: if, for all $\epsilon > 0$, the model can be covered by a finite number of ϵ-Hellinger balls, then the Bayesian posterior eventually concentrates: for all $\delta, \gamma > 0$, the posterior mass on distributions within Hellinger distance δ of the $P_{\tilde{\theta}}$ that is closest to P^* in KL divergence will become larger than $1 - \gamma$ for all n larger than some n_γ. This implies that both in the ridge regression (finite p) and in the model averaging experiments (finite $p_{\max}$), Bayes eventually 'recovers' — as we indeed see in our experimental results. However, in Appendix G.2 we formally show (which requires substantial work!) that if $p_{\max} = \infty$, then the model has no finite Hellinger cover any more for small enough ϵ, so that indeed the conditions for Theorem 1 of De Blasi and Walker (2013) do not apply any more. Our results suggest that in such a case we can indeed have inconsistency if the model is incorrect. On the other hand, even if $p_{\max} = \infty$, we do have consistency and a parametric $O(1/\sqrt{n})$ Hellinger convergence rate in the setup of our correct-model experiment for the 'almost' standard Bayesian posterior, namely for any $\eta < 1$, as follows from the results by Grünwald and Mehta (2016), Zhang (2006a) — see below.

Spurious compression, Rates of Convergence and the limiting $\eta = 1$ Like several earlier results (Barron and Cover, 1991, Walker and Hjort, 2002), Zhang's consistency results for correct models hold under very weak conditions for generalized Bayes with any $\eta < 1$, and only under much stronger conditions on the prior for $\eta = 1$. Zhang provides an example of inconsistency-like behaviour in the well-specified case with $\eta = 1$ that automatically disappears as soon as one picks $\eta < 1$, leading Zhang (2006a) to claim that in general, generalized Bayesian methods ($\eta < 1$) are more stable than standard Bayesian ones; the use of η close to but not equal to 1 is also advocated in another well-specified context by Martin et al. (2017). Zhang's example, and the example of Bayesian model selection inconsistency in a well-specified model by Csiszár and Shields (2000), are closely related to ours: upon close inspection they are both examples of *spurious compression*, a milder form of hypercompression that can still occur for correct models, as described in Section D.2 and Figure 3. What happens is that the Bayes predictive $\bar{P}$ distribution for $\eta = 1$ does converge to the true distribution, but the posterior does not concentrate on it: it spreads its mass over distributions that are all very different from the true P^*, yet their posterior-weighted average is close to P^* after all. In their well-specified examples, the problem

is resolved by taking any $\eta < 1$; in our misspecification case, to avoid spurious compression, η should even be taken much smaller.

In fact, Zhang (2006a) shows that, in the well-specified case, under a very weak KL-property condition stating merely that there is sufficient prior mass in KL-neighbourhoods of the true distribution, for any $\eta < 1$, the η-posterior asymptotically concentrates in Hellinger distance around the true distribution, often at the minimax optimal frequentist rate. The result implies that spurious compression and its associated problems cannot happen for any $\eta < 1$ if the model is correct. In other well-known works on posterior convergence for well-specified models, one always imposes an additional, more complicated condition (Assumption 2 in (Barron et al., 1999), and the 'entropy number' or 'testing condition' (2.2) in (Ghosal et al., 2000)), which is needed to get posterior consistency for standard ($\eta = 1$) Bayes — Zhang's work shows that they are not needed as soon as $\eta < 1$. Grünwald and Mehta (2016) provides a version of Zhang's result which holds under misspecification, saying again that, under a weak prior mass condition, for any $\eta < \eta^*$, the η-posterior concentrates around the KL optimal $\tilde{\theta}$. Here η^* is some critical value which is 1 if the model is correct, convex or the misspecification is 'benign'; yet, under 'bad' misspecification, η^* can be $\ll 1$. In light of these results, one can think of R-log-SafeBayes as finding the largest η such that spurious compression cannot occur — which may be a useful goal even if the model is correct, for one can then get posterior concentration under Zhang's minimal condition.

Anomalous behaviour and modifications of Bayesian posterior under misspecification Anomalous behaviour of Bayesian inference under misspecification was, of course, observed before, e.g. (less dramatically than here) by Yang (2007), Müller (2013) and (as dramatically, but involving a very artificial model) Grünwald and Langford (2007). Presumably also related is the 'brittleness' of Bayesian inference that has been observed by Owhadi and Scovel (2013). Not surprisingly then, we are not the first to suggest modification of likelihood-based estimators (see e.g. White, 1982, Royall and Tsou, 2003, Kotłowski et al., 2010) and posteriors (Royall and Tsou, 2003, Hoff and Wakefield, 2012, Doucet and Shephard, 2012, Müller, 2013) under misspecification. The latter three approaches (that extend the first) employ the so-called *sandwich posterior*, in which the covariance matrix of the posterior is changed based on a 'sandwich formula' involving the empirical variance; Müller (2013) provides extensive explanation and experimentation. Compared to the sandwich approach, our proposal, besides being applicable in fully nonparametric contexts, is substantially more radical. This can be seen from the regression applications in Müller (2013), which involve a noninformative Jeffreys' prior on the regression coefficient vector β. With such a prior (as well as any normal prior scaled by variance σ^2), the posterior *mean* of β, and thus also the frequentist square-risk (which only depends on the posterior mean) remains unaffected by the sandwich modification, so for square-risk the method would perform like standard Bayes in our model-wrong experiments, and would therefore not 'save' Bayes. (Müller (2013, Section 2.4) demonstrates its usefulness on loss functions other than square risk.) Nevertheless, both the sandwich and the SafeBayesian methods can be thought of as methods for measuring the spread of a posterior, and it would be useful to compare the two in detail, both in theory and practice. Specifically, it seems that Müller's modification of the posterior can be used in parametric models to adjust *credible sets* around a given posterior mean, thus tying in to our aforementioned *Open Problem 1b*, whereas our method is suited to arrive, in large, nonparametric models, at entirely different posterior means.

G More Evidence for 'Real' Inconsistency

Here we first provide a repetition of the main experiment from Section 5.2 in the main text with a higher dimensional model, $p_{\max} = 100$ instead of $p_{\max} = 50$. This additional experiment has results that are qualitatively very similar to that with $p_{\max} = 50$, but it is crucial to substantiate our claim that we are really observing a form of inconsistency, because it shows that standard Bayes needs a much larger sample size to 'recover' at $p_{\max} = 100$ when the model is wrong — suggesting that if we used $p_{\max} = \infty$ and let $n \to \infty$, Bayes would never recover. This is further suggested (but not proved

of course) by the fact that, with $p_{\max} = \infty$, the condition for a recent, advanced consistency-under-misspecification result does not hold, as we show in Section G.2.

G.1 More graphs from main experiment

Figures G.1 and G.2 report the results for wrong-model and correct-model experiments analogous to those seen in Figures 5 and 6 in the main paper, but with $p = 100$ instead of $p = 50$. As mentioned in the main paper, we see that it takes Bayes much longer to recover in the wrong-model experiment than for $p = 50$, while both versions of SafeBayes are hardly affected.

G.2 $\mathcal{M}$ with $p_{\max} = \infty$ violates the Condition of De Blasi and Walker (2013)

While we cannot experiment with $p_{\max} = \infty$ and therefore cannot formally establish mathematical inconsistency, one thing we can do is show that the regularity conditions for an existing general theorem that establishes consistency of standard Bayes under misspecification do not hold for our model and prior. Of course, this still does not prove inconsistency but at least it shows that our claim does not contradict a fact. We chose the theorem of De Blasi and Walker (2013) because it only establishes consistency, and no convergence rates, allowing the authors to impose a single, general, elegant regularity condition that, jointly with the KL property, is sufficient for consistency under misspecification — as phrased in the Corollary 1 to their Theorem 1. As is immediate from their result, their condition implies the following substantially weaker condition: for every $\alpha \in (0, 1/2)$, there exists a covering of the model $\mathcal{M}$ by ϵ-Hellinger balls $\{\bar{B}_{j,\epsilon} : j = 1, 2 \ldots\}$ where $\epsilon \leq \alpha$ such that, with $\bar{\Pi}$ denoting the prior measure defined on distributions in $\mathcal{M}$,

$$\sum_j \bar{\Pi}(\bar{B}_{j,\epsilon})^\alpha < \infty. \tag{G.1}$$

We will now show that in our setting, with $p_{\max} = \infty$, no such coverings exist for small enough $\alpha > 0$, as long as the prior on the model dimension p has polynomial tails, thus implying that their condition does not hold. To match our setting to theirs, we extend each conditional distribution for $Y \mid X$ in each of our models $\mathcal{M}_p, p = 0, 1, 2, \ldots$ to a joint distribution for (X, Y) by extending it with the same (well-specified) marginal distribution P^*_X on X as the one used to generate data in our experiments, i.e. (beginning of Section 5) the (infinitely many) components of X are independent Gaussians with variance 1. In this way all $\mathcal{M}_p$ and their union $\mathcal{M}$ are now sets of joint distributions on (X, Y), extended to n outcomes by independence as before.

Consider $\mathcal{M}$ and a prior as defined in Section 2.5. We slightly simplify the setup and only consider the case with a degenerate prior that puts all its mass on $\sigma^2 := 1$ (the experiments in [GvO] show that with this prior, too, we get bad empirical results for standard Bayes). As in our experiments, we take the covariance matrix Σ of the prior on β to be the identity matrix, making the prior a unit variance spherical Gaussian. We use the notation Π as in the main text for the prior on parameters $\theta = (p, \beta)$ (the variance being equal to 1 is not included as a free parameter any more). We use the notation $\bar{\Pi}$ as the corresponding distribution on probability measures — note that Π uniquely determines $\bar{\Pi}$.

Let $\{\bar{B}_{j,\epsilon} : j = 1, 2 \ldots\}$ be an arbitrary cover of $\mathcal{M}$ with ϵ-Hellinger balls, and, for $p \geq 0$, let $\bar{B}_{p,j,\epsilon} := \mathcal{M}_p \cap \bar{B}_{j,\epsilon}$. We must have that

$$\sum_{j \geq 1} \bar{\Pi}(\bar{B}_{j,\epsilon})^\alpha \geq \sup_{p=0,1,2,\ldots} \sum_{j \geq 1} \bar{\Pi}(\bar{B}_{p,j,\epsilon})^\alpha$$

so for us it is sufficient to show that, for arbitrarily chosen cover, the right hand side must be infinite for all $\epsilon > 0$ smaller than some fixed constant independent of α and p, for this implies that the condition (G.1) cannot hold. We now consider $\bar{\Pi}(\bar{B}_{p,j,\epsilon})^\alpha$ for fixed p and set d (dimension) as $d := p + 1$. Let G be a unit variance spherical Gaussian distribution of dimension d. According to the *Gaussian Annulus*

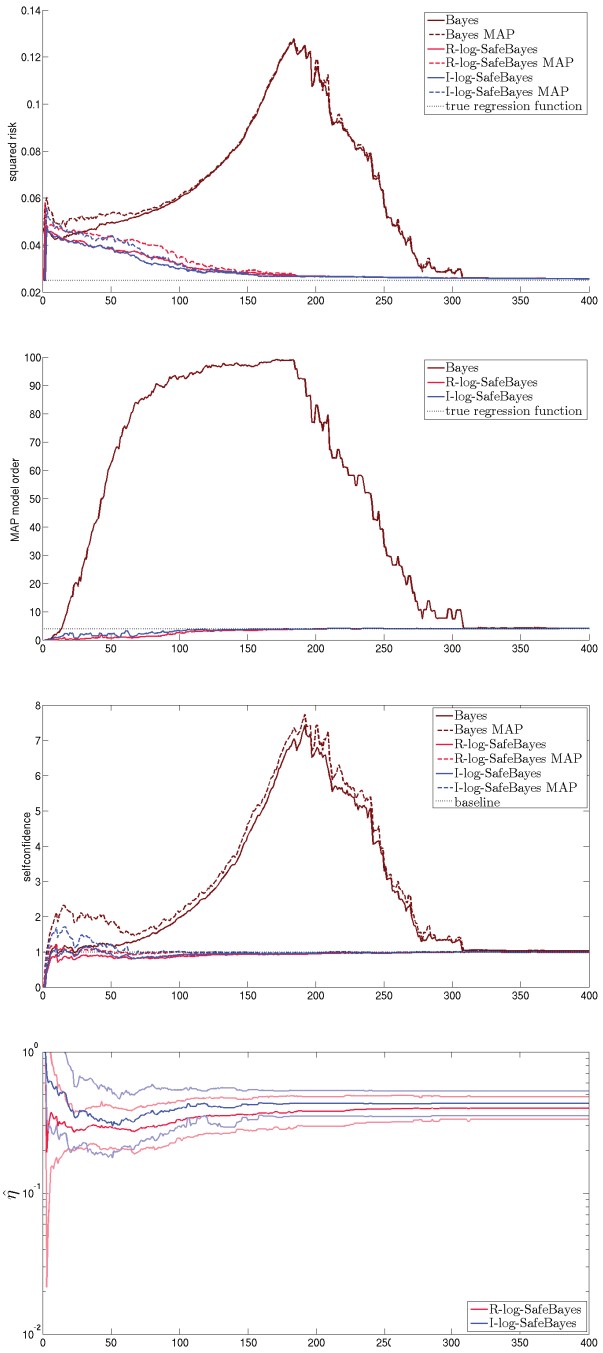

Figure G.1: Same four graphs as in Figure 5, for the wrong-model experiment with $p_{\max} = 100$

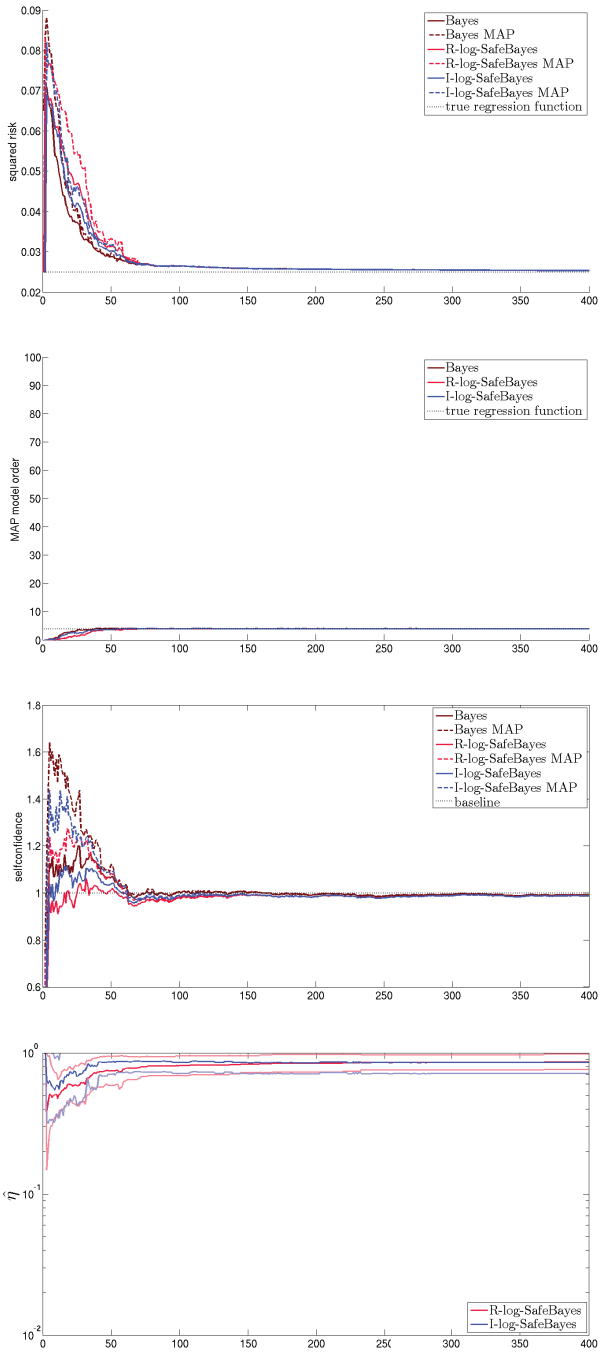

Figure G.2: Same graphs as in Figure 5 for the correct-model experiment with $p_{\max} = 100$

Theorem (see for example Blum et al. (2017)), there is a fixed constant $c > 0$ such that for all d, the G-probability mass of the Euclidean ball in $\mathbb{R}^d$ with radius $2\sqrt{d}$ is at least $1 - 3e^{-cd}$. In particular, for all large d, the mass will be at least $1/2$.

Below we use that, for all $\beta \in \mathbb{R}^{p+1}$, $p' \neq p$, we have $\pi((p,\beta) \mid p') = 0$, so that $\bar{\Pi}(\bar{B}_{p,j,\epsilon}) \mid p') = 0$. We must thus have $\sum_j \bar{\Pi}(B_{p,j,\epsilon}) \geq \pi(p)$. Letting $\|\cdot\|$ denote Euclidean distance, we let $\mathcal{J}_p \subset \mathbb{N}$ be the set of j with for some $\beta \in B_{p,j,\epsilon}, \|\beta\| \leq 2\sqrt{d}$; here $B_{p,j,\epsilon}$ represents the set of $\beta \in \mathbb{R}^{p+1}$ corresponding to the set of measures $\bar{B}_{p,j,\epsilon}$. We must have for $j \notin \mathcal{J}_p$, for all $\beta \in B_{p,j\epsilon}$, $\|\beta\| > 2\sqrt{d}$. Therefore, the Gaussian annulus theorem above gives that for large p and hence large d, the prior mass $\bar{\Pi}(\mathcal{J}_p \mid \mathcal{M}_p)$ of the set $\mathcal{J}_p$ given $\mathcal{M}_p$ is at least $1/2$. This gives for all large p,

$$\begin{aligned}\sum_j \Pi(B_{p,j,\epsilon})^\alpha &= \sum_j \Pi(B_{p,j,\epsilon}) \cdot \left(\frac{1}{\Pi(B_{p,j,\epsilon})}\right)^{1-\alpha} \geq \sum_{j \in \mathcal{J}_p} \Pi(B_{p,j,\epsilon}) \cdot \inf_{j' \in \mathcal{J}_p} \left(\frac{1}{\Pi(B_{p,j',\epsilon})}\right)^{1-\alpha} \\ &\geq \frac{1}{2} \cdot \pi(p) \cdot \left(\frac{1}{\sup_{j' \in \mathcal{J}_p} \Pi(B_{p,j',\epsilon})}\right)^{1-\alpha}. \end{aligned} \tag{G.2}$$

Now let $A \subset \mathbb{R}^{p+1}$ and let $r = \sup_{\beta \in A} \|\beta\|_\infty$. We have, for arbitrary such A and every $\beta, \beta' \in A$, with $h(\beta, \beta')$ denoting Hellinger distance between $P_{(p,\beta)}$ and $P_{(p,\beta')}$, both in $\mathcal{M}_p$, and $\|\cdot\|$ denoting the Euclidean norm,

$$\|\beta - \beta'\| = \left(\mathbf{E}_{X \sim P_X^*}[(\beta X - \beta' X)^2]\right)^{1/2} < 2.03 \cdot \sup\{e, r\} \cdot h(\beta, \beta') \tag{G.3}$$

where the equality is straightforward by independence of the X-components and the inequality is implied by Equation 2.6 in Birgé (2004). Fix some $\epsilon > 0$. We now first apply the above inequality with the set $A = \{\beta, \beta'\}$ where $\beta, \beta' \in \mathbb{R}^{p+1}$ satisfy $h(\beta, \beta') \leq \epsilon$ but otherwise are chosen arbitrarily. A straightforward case-by-case analysis of the three possibilities for the supremum $(e, \|\beta\|_\infty, \|\beta'\|_\infty)$ gives, for $\epsilon < 1/2.03$,

$$\|\beta'\| \leq \|\beta\| + \sup\left\{2.03e\epsilon, \|\beta\|2.03\epsilon, \|\beta\|\frac{2.03\epsilon}{1 - 2.03\epsilon}\right\}. \tag{G.4}$$

We will (for reasons to become clear below) take $0 < \epsilon < 1/(2.5 \cdot \sqrt{2\pi e}) \approx 0.097$. Here $\pi = 3.14..$; to avoid clutter with our notation for prior distribution, we will henceforth use so-called τ-*notation* with $\tau = 2\pi = 6.28..$ and reserve the symbol π for densities for the prior, so that $0 < \epsilon < 1/(2.5)\sqrt{\tau \cdot e}$. For every j, we know that there exists a $\beta \in B_{p,j,\epsilon}$ satisfying $\|\beta\| \leq 2\sqrt{d}$. Thus for this β, for all $\beta' \in B_{p,j,\epsilon}$, and for ϵ as just defined, the above further implies, using (G.4) and $\|\beta\| \leq 2\sqrt{d}$, that $\|\beta'\| < 2.5\sqrt{d}$. This in turn implies that $\sup\{\|\beta\|_\infty : \beta \in B_{p,j,\epsilon}\} < 2.5\sqrt{d}$. We can thus use (G.3) again, with $A = B_{p,j,\epsilon}$, to give, for every $\beta', \beta'' \in A$,

$$\|\beta' - \beta''\| \leq 2.5\sqrt{d} \cdot h(\beta', \beta'') \leq 2.5\sqrt{d}\epsilon$$

Hence, for each $j' \in \mathcal{J}_p$, $B_{p,j',\epsilon}$ is contained in a Euclidean ball of radius $2.5\sqrt{d}\epsilon$. The density of the d-dimensional spherical unit variance 0-mean Gaussian being maximized at the origin and there equal to $(\tau \cdot d)^{-1/2}$, the prior mass of such a Euclidean ball under $\Pi(\cdot \mid p)$ is at most $(\tau \cdot d)^{-1/2} V(d, 2.5\sqrt{d}\epsilon)$, where $V(d, r)$ denotes the volume of a d-dimensional ball with radius r. (G.2) thus becomes at least (where we used a Stirling approximation for $V(d, r)$ that holds for large d):

$$\begin{aligned}\sum_j \Pi(\bar{B}_{p,j,\epsilon})^\alpha &\geq \frac{\pi(p)}{2} \cdot \left(\frac{\sqrt{\tau \cdot d}}{V(d, 3.2\sqrt{d}\epsilon)}\right)^{1-\alpha} \\ &\geq \frac{\pi(p)}{2} \cdot \left(\frac{\tau \cdot d}{\sqrt{2}(\tau \cdot e(2.5\epsilon)^2)^{d/2}}\right)^{1-\alpha} \geq \frac{\pi(p)}{2} \cdot \left(\frac{\tau \cdot (p+1)}{\sqrt{2}(\epsilon')^{(p+1)/2}}\right)^{1-\alpha}.\end{aligned}$$

where ϵ' is, by our definition of ϵ, a number strictly smaller than 1. Since we assume that $\pi(p)$ decreases polynomially, i.e. there is $k > 1$ such that $\limsup_{p\to\infty} \pi(p)p^k > 0$, the supremum over p of the final expression above is thus infinite, which is what we had to prove.

We note that the same proof also establishes that the regularity condition needed for consistency in the well-specified case in the consistency result of Walker (2004) also does not hold for our model. On the other hand, if the model is correct, then as soon as one takes any $\eta < 1$, the results of Zhang (2006a), Grünwald and Mehta (2016) can be used to establish consistency after all in the well-specified case; and, if the sampling distribution P^* is as in the main text, the same result can be used to establish consistency of generalized Bayes for sufficiently small η in the misspecified case as well.

References

Audibert, J. Y. (2004). "PAC-Bayesian statistical learning theory." Ph.D. thesis, Université Paris VI. 7, 14

— (2007). "Progressive mixture rules are deviation suboptimal." In *NIPS*. 5

Barron, A. R. (1987). "Are Bayes rules consistent in information?" In *Open Problems in Communication and Computation*, 85–91. Springer. 5

— (1998). "Information-Theoretic Characterization of Bayes Performance and the Choice of Priors in Parametric and Nonparametric Problems." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics*, volume 6, 27–52. Oxford: Oxford University Press. 6

Barron, A. R. and Cover, T. M. (1991). "Minimum Complexity Density Estimation." *IEEE Transactions on Information Theory*, 37(4): 1034–1054. 7, 15

Barron, A. R., Rissanen, J., and Yu, B. (1998). "The Minimum Description Length principle in coding and modeling." *IEEE Transactions on Information Theory*, 44(6): 2743–2760. Special Commemorative Issue: Information Theory: 1948-1998. 13

Barron, A. R., Schervish, M. J., and Wasserman, L. (1999). "The consistency of posterior distributions in nonparametric problems." *The Annals of Statistics*, 27(2): 536–561. 2, 6, 16

Berk, R. (1966). "Limiting behavior of posterior distributions when the model is incorrect." *Annals of Mathematical Statistics*, 37: 51–58. 15

Besag, J. (1975). "Statistical analysis of non-lattice data." *The statistician*, 179–195. 11

Birgé, L. (2004). "Model selection for Gaussian regression with random design." *Bernoulli*, 10(6): 1039–1051. 20

Bissiri, P. G., Holmes, C., and Walker, S. G. (2016). "A general framework for updating belief distributions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5): 1103–1130. 14

Blei, D. (2012). "Probabilistic topic models." *Communications of the ACM*, 55(4): 77—84. 10

Blum, A., Hopcroft, J., and Kannan, R. (2017). "Foundations of Data Science." Preliminary version of a textbook, available via `http://www.cs.cornell.edu/jeh/book May 22 2017.pdf`. 20

Box, G. E. P. (1980). "Sampling and Bayes' inference in scientific modelling and robustness." *Journal of the Royal Statistical Society. Series A (General)*, 383–430. 9

Breiman, L. (2001). "Statistical Modeling: The Two Cultures (with discussion)." *Statistical Science*, 16(3): 199 –215. 10

Catoni, O. (1997). "A mixture approach to universal model selection." Preprint LMENS-97-30. Available from `http://www.math.ens.fr/edition/publis/Index.97.html`. 5

— (2007). *PAC-Bayesian Supervised Classification*. Lecture Notes-Monograph Series. IMS. 11, 14

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge, UK: Cambridge University Press. 14

Csiszár, I. and Shields, P. (2000). "The consistency of the BIC Markov order estimator." *Annals of Statistics*, 28: 1601–1619. 15

Dalalyan, A. S. and Tsybakov, A. B. (2012). "Mirror averaging with sparsity priors." *Bernoulli*, 18(3): 914–944. 5

Dawid, A. P. (1982). "The well-calibrated Bayesian." *Journal of the American Statistical Association*, 77: 605–611. Discussion: pages 611–613. 9

— (2004). "Probability, causality and the empirical world: A Bayes–de Finetti–Popper–Borel synthesis." *Statistical Science*, 19: 44–57. 9

De Blasi, P. and Walker, S. G. (2013). "Bayesian asymptotics with misspecified models." *Statistica Sinica*, 23: 169–187. 15, 17

Devaine, M., Gaillard, P., Goude, Y., and Stoltz, G. (2013). "Forecasting electricity consumption by aggregating specialized experts; a review of the sequential aggregation of specialized experts, with an application to Slovakian and French country-wide one-day-ahead (half-)hourly predictions." *Machine Learning*, 90(2): 231–260. 14

Diaconis, P. and Freedman, D. (1986). "On the Consistency of Bayes Estimates." *The Annals of Statistics*, 14(1): 1–26. 1

Doucet, A. and Shephard, N. (2012). "Robust inference on parameters via particle filters and sandwich covariance matrices." Technical Report 606, University of Oxford, Department of Economics. 16

Dümbgen, L., Samworth, R., and Schuhmacher, D. (2011). "Approximation by log-concave distributions, with applications to regression." *The Annals of Statistics*, 39(2): 702–730. 15

Dunson, D. B. and Taylor, J. A. (2005). "Approximate Bayesian inference for quantiles." *Nonparametric Statistics*, 17(3): 385–400. 11

Freund, Y., Mansour, Y., and Schapire., R. E. (2004). "Generalization bounds for averaged classifiers (how to be a Bayesian without believing)." *Annals of Statistics*, 32(4): 1698–1722. 14

Friel, N. and Stoehr, J. (2015). "Calibration of conditional composite likelihood for Bayesian inference on Gibbs random fields." In *18th International Conference on Artificial Intelligence and Statistics (AISTATS), San Diego, California, USA, 9-12 May 2015*, volume 38, 921–929. Microtome Publishing. 12

Gelman, A. (2004). "Bayes and Popper." Entry in A. Gelman's blog on Statistical Modeling, Causal Inference, and Social Science. 9

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. Boca Raton, FL: CRC Press, third edition. 5

Gelman, A. and Shalizi, C. (2012). "Philosophy and the practice of Bayesian statistics." *British Journal of Mathematical and Statistical Psychology*. 9, 10, 14

Gerchinovitz, S., Mallet, V., and Stoltz, G. (2008). "A further look at sequential aggregation rules for ozone ensemble forecasting." Technical report, INRIA and École normale supérieure. 14

Ghosal, S., Ghosh, J., and Van der Vaart, A. (2000). "Convergence rates of posterior distributions." *Annals of Statistics*, 28(2): 500–531. 2, 16

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. University of Minnesota Press. 9

Grünwald, P. D. (1998). "The Minimum Description Length Principle and Reasoning under Uncer-

tainty." Ph.D. thesis, University of Amsterdam, The Netherlands. Available as ILLC Dissertation Series 1998-03; see www.grunwald.nl. 12

— (1999). "Viewing all Models as "Probabilistic"." In *Proceedings of the Twelfth ACM Conference on Computational Learning Theory (COLT' 99)*, 171–182. 12

— (2007). *The Minimum Description Length Principle*. Cambridge, MA: MIT Press. 7, 10, 14

— (2008). "That Simple Device Already Used By Gauss." In Grünwald, P. D., Myllymäki, P., Tabus, I., Weinberger, M., and Yu, B. (eds.), *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, 293–304. Tampere, Finland: Tampere University Press. 12

— (2011). "Safe Learning: Bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity." In *Proceedings of the Twenty-Fourth Conference on Learning Theory (COLT' 11)*. 6, 14

— (2012). "The Safe Bayesian: Learning the learning rate via the mixability gap." In *Proceedings 23rd International Conference on Algorithmic Learning Theory (ALT '12)*. Springer. 5, 6

— (2016). "Contextuality of Misspecification and Data-Dependent Losses." *Statistical Science*, 31(4): 495–498. Comment on Watson and Holmes (2016). 12

— (2017). "Safe probability." *Journal of Statistical Planning and Inference*. To appear. 11

Grünwald, P. D. and Halpern, J. Y. (2004). "When ignorance is bliss." In *Proceedings of the Twentieth Annual Conference on Uncertainty in Artificial Intelligence (UAI 2004)*. Banff, Canada. 11

Grünwald, P. D. and Langford, J. (2007). "Suboptimal behavior of Bayes and MDL in classification under misspecification." *Machine Learning*, 66(2-3): 119–149. DOI 10.1007/s10994-007-0716-7. 16

Grünwald, P. D. and Mehta, N. (2016). "Fast Rates with Unbounded Losses." *arXiv preprint* 1605.00252. 2, 5, 7, 15, 16, 21

Gu, J. and Ghosal, S. (2009). "Bayesian ROC curve estimation under binormality using a rank likelihood." *Journal of Statistical Planning and Inference*, 139(6): 2076–2083. 11

Hahn, P. R., Carvalho, C. M., and Mukherjee, S. (2013). "Partial factor modeling: predictor-dependent shrinkage for linear regression." *Journal of the American Statistical Association*, 108(503): 999–1008. 2

De Heide, R. (2016). "The Safe–Bayesian Lasso." Master's thesis, Leiden University. 13

Helmbold, D. P. and Warmuth, M. K. (1992). "Some weak learning results." In *Proceedings of the fifth annual workshop on Computational learning theory*, 399–412. ACM. 5

Hoff, P. and Wakefield, J. (2012). "Bayesian sandwich posteriors for pseudo-true parameters." *arXiv preprint arXiv:1211.0087*. 16

Holmes, C. and Walker, S. (2017). "Assigning a value to a power likelihood in a general Bayesian model." *Biometrika*, 104(2): 497–503. 14

Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press, 3rd edition. 11

Jiang, W. and Tanner, M. (2008). "Gibbs posterior for variable selection in high-dimensional classification and data mining." *Annals of Statistics*, 36(5): 2207–2231. 2, 14

Juditsky, A., Rigollet, P., and Tsybakov, A. B. (2008). "Learning by mirror averaging." *The Annals of Statistics*, 36(5): 2183–2206. 5

Kleijn, B. and Van der Vaart, A. (2006). "Misspecification in infinite-dimensional Bayesian statistics." *Annals of Statistics*, 34(2). 15

Koolen, W. M. and Van Erven, T. (2015). "Second-order Quantile Methods for Experts and Combin-

atorial Games." In *Proceedings of the Twenty-Eighth Conference on Learning Theory (COLT' 15)*. URL http://arxiv.org/abs/1502.08009 14

Koolen, W. M., Van Erven, T., and Grünwald, P. D. (2014). "Learning the learning rate for prediction with expert advice." In *Advances in Neural Information Processing Systems*, 2294–2302. 14

Kotłowski, W., Grünwald, P. D., and De Rooij, S. (2010). "Following the flattened leader." In *Conference on Learning Theory (COLT)*, 106–118. 16

Lacoste-Julien, S., Huszár, F., and Ghahramani, Z. (2011). "Approximate inference for the loss-calibrated Bayesian." *AISTATS 2011 - Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 15: 416–424. 12

Martin, R., Mess, R., and Walker, S. G. (2017). "Empirical Bayes posterior concentration in sparse high-dimensional linear models." *Bernoulli*, 23(3): 1822–1847. 6, 15

McAllester, D. (2003). "PAC-Bayesian Stochastic Model Selection." *Machine Learning*, 51(1): 5–21. 12

Miller, J. and Dunson, D. (2015). "Robust Bayesian Inference via Coarsening." Technical report, arXiv. Available at arXiv:1506.06101. 14

Müller, U. K. (2013). "Risk of Bayesian inference in misspecified models, and the sandwich covariance matrix." *Econometrica*, 81(5): 1805–1849. 16

O'Hagan, A. (1995). "Fractional Bayes Factors for Model Comparison." *Journal of the Royal Statistical Society, Series B*, 57(1): 99–138. With discussion. 5

Owhadi, H. and Scovel, C. (2013). "Brittleness of Bayesian inference and new Selberg formulas." *arXiv preprint arXiv:1304.7046*. 16

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association*, 92(437): 179–191. 5

Ramamoorthi, R. V., Sriram, K., and Martin, R. (2015). "On posterior concentration in misspecified models." *Bayesian Analysis*, 10(4): 759–789. 15

Robins, J. and Wasserman, L. (2000). "The Foundations of Statistics: A Vignette." *Journal of the American Statistical Association*. 11

De Rooij, S., Van Erven, T., Grünwald, P. D., and Koolen, W. M. (2013). "Follow the leader if you can, Hedge if you must." *arXiv preprint arXiv:1301.0534*. 14

Royall, R. and Tsou, T.-S. (2003). "Interpreting statistical evidence by using imperfect models: Robust adjusted likelihood functions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2): 391–404. 16

Shalizi, C. (2009). "Dynamics of Bayesian updating with dependent data and misspecified models." *Electronic Journal of Statistics*, 3: 1039–1074. 15

Spokoiny, V. (2012). "Parametric estimation. Finite sample theory." *The Annals of Statistics*, 40(6): 2877–2909. 15

Sriram, K., Ramamoorthi, R. V., and Ghosh, P. (2013). "Posterior consistency of Bayesian quantile regression based on the misspecified asymmetric Laplace density." *Bayesian Analysis*, 8(2): 479–504. 3, 15

Syring, N. and Martin, R. (2017). "Calibrating general posterior credible regions." *arXiv preprint arXiv:1509.00922*. 12, 14

Vovk, V. G. (1990). "Aggregating strategies." In *Proc. COLT' 90*, 371–383. 14

Walker, S. (2004). "New approaches to Bayesian consistency." *Annals of Statistics*, 2028–2043. 1, 2, 21

Walker, S. and Hjort, N. L. (2002). "On Bayesian consistency." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(4): 811–821. 6, 15

Watson, J. and Holmes, C. (2016). "Approximate models and robust decisions." *Statistical Science*, 31(4): 465–489. 12

White, H. (1982). "Maximum likelihood estimation of misspecified models." *Econometrica: Journal of the Econometric Society*, 1–25. 16

Willems, F., Shtarkov, Y., and Tjalkens, T. (1995). "The context-tree weighting method: basic properties." *IEEE Transactions on Information Theory*, 41: 653–664. 10

Williams, P. M. (1980). "Bayesian Conditionalisation and the Principle of Minimum Information." *British Journal for the Philosophy of Science*, 31(2): 131–144. 14

Yang, Y. (2000). "Combining different procedures for adaptive regression." *Journal of multivariate analysis*, 74: 135–161. 5

Yang, Z. (2007). "Fair-Balance Paradox, Star-Tree Paradox, and Bayesian Phylogenetics." *Journal of Molecular Biology and Evolution*, 24(8): 1639–1655. 13, 16

Zhang, T. (2006a). "From ϵ-entropy to KL entropy: Analysis of minimum information complexity density estimation." *Annals of Statistics*, 34(5): 2180–2210. 2, 7, 15, 16, 21

— (2006b). "Information Theoretical Upper and Lower Bounds for Statistical Estimation." *IEEE Transactions on Information Theory*, 52(4): 1307–1321. 7, 11, 14