This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2945077, IEEE Journal of Biomedical and Health Informatics

Practical priors for Bayesian inference of latent biomarkers

Hennadii Madan, Rok Berlot, Nicola J. Ray, Franjo Pernuš and Žiga Špiclin*

***Abstract*—Latent biomarkers are quantities that strongly relate to patient's disease diagnosis and prognosis, but are difficult to measure or even not directly observable. The objective of this study was to develop, analyze and validate new priors for Bayesian inference of such biomarkers. Theoretical analysis revealed a relationship between the estimates inferred from the model and the true values of measured quantities, and the impact of the priors. This led to a new prior encoding scheme that incorporates objectively measurable domain knowledge, i.e. by performing two measurements with a reference method, which imply scale of the prior distribution. Second, priors on parameters of systematic error are non-informative, which enables biomarker estimation from a set of different quantities. Analysis showed that the volume of nucleus basalis of Meynert, which is reduced in early stages of Alzheimer's dementia and Parkinson's disease, is inter-related and could be inferred from compartmental brain volume measurements performed on routine clinical MR scans. Another experiment showed that total lesion load, associated to future disability progression in multiple sclerosis patients, could be inferred from lesion volume measurements based on multiple automated MR scan segmentations. Besides, figures of merit derived from the estimates could, without comparing against reference gold standard segmentations, identify the best performing lesion segmentation method. The proposed new priors substantially simplify the application of Bayesian inference for latent biomarkers and thus open an avenue for clinical implementation of new biomarkers, which may ultimately advance the evidence-based medicine.**

***Index Terms*—Quantitative Imaging Biomarkers, Latent Variable Prediction, Bayesian Inference, Markov Chain Monte Carlo, Brain Segmentation, Magnetic Resonance Imaging, Validation**

I. Introduction

Advances in evidence-based medicine are based on finding highly sensitive and specific quantities of interest or biomarkers, which strongly relate to patient's disease diagnosis and prognosis. Unfortunately, many such quantities cannot be measured routinely, for reasons such as invasiveness, high costs, the need for special equipment, etc. For instance, brain lesion volume at onset of multiple sclerosis is prognostic of the future disability progression [1]. However, as routine magnetic resonance (MR) scans are highly biased due to poor lesion contrast, partial volume effects, noise, and other artifacts the automated and even manual lesion volume measurements are substantially biased. Hence, such a quantity is difficult to measure accurately or even not directly observable. We refer to such quantities as *latent biomarkers*.

The aim of this work is to estimate the latent biomarker values in at least two common scenarios: (A) in case of excessive measurement variability and (B) in case the biomarker is not observable. The corresponding strategies adopted in these two scenarios are (A') to apply multiple methods to measure the same quantity and use the measurements to recover its unbiased values and (B') to measure dependent or proxy surrogate quantities and use the measurements to estimate the latent quantity.

The main contribution is the development and validation of practical priors to be used in the context of a Bayesian inference framework, which can estimate the biomarker values in the context of the two aforementioned scenarios.

A. Background

Regression without truth (RWT) [2] is a family of techniques for comparison and validation of multiple measurement methods (MMs) of a certain quantity. These techniques estimate systematic and random errors of the MMs applied to a common set of subjects, while the reference measurements are not required. In this way the high cost, cumbersome and labour intensive creation of reference measurements is eliminated.

Applying the RWT is possible under certain assumptions about the prior distribution of the latent true values and assumption of statistical independence of random errors of different MMs or measurements. With minor variations the original RWT was applied to various biomarkers [3]–[5]. Unfortunately, the obtained estimates were never validated, i.e. were never compared to those based on regression with respect to reference measurements.

The main practical limitation of RWT is that the estimation procedure relies on substantial prior knowledge about the measurand distribution, which should either be known or at least constrained to a certain parametric family. This knowledge may require one to establish reference measurements on a sample from a related population, which is difficult or even impossible in certain applications. Besides, the RWT methodology is based on an *ad hoc* computational procedure that is not derived from the first principles.

In [6] the authors reformulated and expanded the RWT into a full Bayesian inference framework, referred to as Bayesian reference-free error estimation (BRFEE). This framework was successfully validated against a known gold standard (GS)

This research was supported by the Slovenian Research Agency (Core Research Grant No. P2-0232 and Research Grants Nos. J2-5473, J7-6781, J2-7211, J2-8173 and L3-4255). *Asterisk indicates corresponding author.*

Ž. Špiclin is with the University of Ljubljana, Faculty of Electrical Engineering, SI-1000 Ljubljana, Slovenia (e-mail: ziga.spiclin@fe.uni-lj.si).

H. Madan, F. Pernuš, are with Faculty of Electrical Engineering, University of Ljubljana, Slovenia.

R. Berlot is with Department of Neurology, University Medical Centre Ljubljana, Slovenia.

N. J. Ray is with Department of Psychology, Manchester Metropolitan University, United Kingdom.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2945077, IEEE Journal of Biomedical and Health Informatics

on a clinical *in vivo* dataset. The framework drops RWT's assumptions about the shape of the measurand distribution in the sample, and the assumption of statistical independence of random errors of measurement methods. Instead, it introduces assumptions that (i) systematic error of a genuine MM has small magnitude, and (ii) it is encoded in peaked informative priors on the parameters of model. There is no rigorous justification for this particular assumptions and setting the widths of these priors is left upon the operator's subjective experience and intuition.

B. Contributions

In this paper, based on previous conference paper [7], we perform a theoretical analysis of the error model used in BRFEE and its interplay with the priors. The analysis leads us to propose a new scheme for prior specification, which encodes a certain minimal knowledge about location and scale of the distribution of true values of the measurand. The significant contributions are (1) that the theoretical analysis establishes a relationship between BRFEE estimates and the true values of measured quantities and (2) that the new scheme lifts assumptions on the magnitude of systematic errors. Furthermore, the knowledge required to run the method can now be acquired in an objective manner with only two measurements performed with a reference method.

The proposed approach is validated on synthetic and clinical data measured *in vivo* in the context of imaging biomarker extraction. Both aforementioned strategies A and B, one involving multiple measurements of a single quantity and the other multiple measurements of dependent or proxy quantities, were successfully demonstrated.

II. Bayesian inference framework

Consider a set of real scalar functions $\{g_m, m = 1..M\}$ of a single variable q analytic on $(\underline{q}, \overline{q})$. Let observables y_{pm} be defined by values of these functions at a finite set of points $q_p, p = 1..N$, with additive "random noise" ϵ:

$$y_{pm} = g_m(q_p) + \epsilon_{pm} \tag{1}$$

For example, in the context of comparison of M measurement methods for a biomarker, y_{pm} should be interpreted as individual measurements in p-th patient with m-th method, q_p as the true value of the quantity being measured (the measurand), $(g_m(q_p) - q_p)$ is then the systematic error ("bias") of the method m and ϵ_{pm} is the random error ("noise").

Assuming that we are dealing with values of q_p from a finite interval $[\underline{q_p}, \overline{q_p}]$, we may approximate g_m with a $K-th$ degree polynomial representing truncated Taylor series about a point $q_o \in [\underline{q_p}, \overline{q_p}]$ as

$$y_{pm} = \sum_{k=0}^{K} b_{km}(q_p - q_0)^k + \epsilon_{pm} \tag{2}$$

The point q_0 is the value where minimal error of approximation is desired and, for instance, may be fixed in the middle of the interval of interest as $(\underline{q_p} + \overline{q_p})/2$. For the remainder of the paper we introduce the following notation

$$x_p \triangleq q_p - q_0 \tag{3}$$

in order to make expressions clearer and shorter, so that (2) now becomes

$$y_{pm} = \sum_{k=0}^{K} b_{km} x_p^k + \epsilon_{pm} \tag{4}$$

We further assume that the random errors of MMs are jointly distributed as multivariate Gaussian (MVG)

$$\boldsymbol{\epsilon_p} \sim \mathcal{N}(0, \Sigma) \tag{5}$$

where $\boldsymbol{\epsilon_p} = (\epsilon_{p1}, ..., \epsilon_{pM})^\top$, Σ a covariance matrix. It must be stressed here that this assumption is the easiest to violate and must be justified from physical considerations. Indeed the assumption of normality is violated when "outliers" are expected or when the physical mechanism underlying the measurement leads to, e.g., binomial distributions with low number of samples and class imbalance.

From equations (4) and (5) the likelihood of observing $\boldsymbol{y_p} \triangleq [y_{p1}, ..., y_{pM}]$ is

$$l_p \triangleq f(\boldsymbol{y_p} \mid B, \Sigma, x_p) = \mathcal{N}(\boldsymbol{\epsilon_p}, \Sigma) \tag{6}$$

where f denotes probability density, $B \triangleq [b_{km}] \in \mathbb{R}^{K \times M}$. The likelihood for the entire set of observations is then

$$l \triangleq f(Y \mid \theta) = \prod_{p=1}^{N} l_p \tag{7}$$

where $Y = [y_{pm}] \in \mathbb{R}^{N \times M}$, $\theta = \{B, \Sigma, \boldsymbol{x}\}$, $\boldsymbol{x} = [x_1, ..., x_N]$.

By Bayes' Theorem the posterior probability density of θ given the observed values Y is

$$f(\theta \mid Y) \propto l \cdot f(\theta) \tag{8}$$

where $f(\theta)$ is the prior probability density of model parameters. The joint dependence between components of θ is captured in the likelihood and individual components of $f(\theta)$ may be defined separately:

$$f(\theta) = f(B) \cdot f(\Sigma) \cdot f(\boldsymbol{x}). \tag{9}$$

The prior $f(\theta)$ encodes our knowledge of the system before observing Y. When both l and $f(\theta)$ are specified, we can draw samples from $f(\theta \mid Y)$ using Markov chain Monte-Carlo (MCMC) and calculate estimates of quantities of interest and their uncertainties using those samples.

III. Theoretical analysis

The estimates $\tilde{\theta}$ of model parameters θ are determined by the interplay between the likelihood l and the priors $f(\theta)$. The former are defined by the data and the model, while the latter are defined by our state of knowledge about model parameters before we even observe the data. This knowledge is usually scant and amounts to vague estimates of b_{km} and of the range of x_p. Nevertheless, even this much plays a crucial role in this particular mathematical system due to certain properties of the likelihood function. Here we provide two theorems describing the peculiar behavior of the likelihood and use them to analyze the role of various priors on b_{km} and x_p, namely: (1) uniform $f(x_p)$ with flat $f(b_{km})$; and (2) same uniform $f(x_p)$ in conjunction with informative $f(b_{km})$, as used in our earlier works. Finally, we introduce (3) the new prior on x_p that is designed to work with flat $f(b_{km})$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2945077, IEEE Journal of Biomedical and Health Informatics

A. Properties of the likelihood

The likelihood (7) does not have a unique maximum point, instead maximum is defined up to a linear transformation of x_p as shown by the following theorem.

Theorem 1. *Let*

$$x'_p = \alpha_1 x_p + \alpha_0, \quad \alpha_0, \alpha_1 = const,\ \alpha_1 \neq 0 \quad \forall p \tag{10}$$

then there exist b'_{km} such that

$$\sum_{k=0}^{K} b_{km} x_p^k = \sum_{k=0}^{K} b'_{km} x_p'^k \quad \forall m \tag{11}$$

and thus

$$f(Y \mid B, \Sigma, \boldsymbol{x}) = f(Y \mid B', \Sigma, \boldsymbol{x}') \tag{12}$$

Thus we say that the likelihood is degenerate. Problems with degenerate likelihood are not amenable to orthodox statistical analysis, however, they are treatable with Bayesian approach with sufficiently informative priors. We will discuss various informative priors further below. Here we first prove that, in the general case, the likelihood degeneracy is such that the relationship between x'_p and x_p cannot be non-linear.

Theorem 2. *Let h be analytic on $(\underline{x}, \overline{x})$ and*

$$x'_p = h(x_p) \quad \forall p \tag{13}$$

Then (11) has a solution in b'_{km} for general b_{km} only when

$$h(x_p) = \alpha_1 x_p + \alpha_0, \quad \alpha_0, \alpha_1 = const,\ \alpha_1 \neq 0 \tag{14}$$

A consequence of the above theorems is that the likelihood alone, without priors, "recovers" x_p values only up to a linear transformation. The priors, hence, play a crucial role in that they must supply additional information to disambiguate the *scale* (α_1) and *location* (α_0) of the set of x_p.

We must stress here that there are important special configurations of b_{km} when Theorem 2 does not work b_{km} that may be encountered in practice. In such cases the estimates of x_p can be biased non-linearly and the framework will break in general. Two of these important configurations are:

1) when b_{km} ratio for different k is the same across m. This could happen when MM are all unbiased or biased only linearly with zero offset,
2) when $b_{km} = 0\ \forall(k, m)$. This could happen if the quantities y_{pm} that we try to use as proxy measurements of the latent biomarker values q_p are unrelated to it.

B. Uniform priors for x_p

Let x_p^* denote the true values of x_p. According to Theorems 1 and 2 a set $x'_p = \alpha_1 x_p^* + \alpha_0$ produces the same value of the likelihood as x_p^*. Imposing a uniform prior $f(x_p) \sim \mathcal{U}(\underline{x_p}, \overline{x_p})$ limits the values that α_1 and α_0 can take:

$$\underline{x_p} \leqslant \alpha_1 x_p^* + \alpha_0 \leqslant \overline{x_p} \quad \forall p \tag{15}$$

This is equivalent to

$$\begin{aligned} \underline{x_p} &\leqslant \begin{cases} \alpha_1 \underline{x_p^*} + \alpha_0, & \alpha_1 > 0 \\ \alpha_1 \overline{x_p^*} + \alpha_0, & \alpha_1 < 0 \end{cases} \\ \overline{x_p} &\leqslant \begin{cases} \alpha_1 \overline{x_p^*} + \alpha_0, & \alpha_1 > 0 \\ \alpha_1 \underline{x_p^*} + \alpha_0, & \alpha_1 < 0 \end{cases} \end{aligned} \tag{16}$$

where $\overline{x_p^*} = \max_p x_p^*$ and $\underline{x_p^*} = \min_p x_p^*$.

Solving for α_0 and α_1:

$$\begin{aligned} \alpha_0 &\leqslant \begin{cases} \overline{x_p} - \alpha_1 \overline{x_p^*}, & \alpha_1 > 0 \\ \overline{x_p} - \alpha_1 \underline{x_p^*}, & \alpha_1 < 0 \end{cases} \\ \alpha_0 &\geqslant \begin{cases} \underline{x_p} - \alpha_1 \underline{x_p^*}, & \alpha_1 > 0 \\ \underline{x_p} - \alpha_1 \overline{x_p^*}, & \alpha_1 < 0 \end{cases} \end{aligned} \tag{17}$$

This represents a parallelogram in α_0, α_1 space (Figure 1). The point $(\alpha_0, \alpha_1) = (0, 1)$ represents $x'_p = x_p^*$. This point lies near one of the corners inside the parallelogram. The posterior sample then represents the correct solution "smudged" over the parallelogram. This renders the estimates very uncertain (wide) and, more importantly, biased: the expected value of (α_0, α_1) is at the center of the parallelogram at $(\frac{\underline{x_p} + \overline{x_p}}{2}, 0)$ corresponding to $x'_p = \frac{\underline{x_p} + \overline{x_p}}{2} = const\ \forall p$ — an utterly useless result that does not even depend on x_p^*.

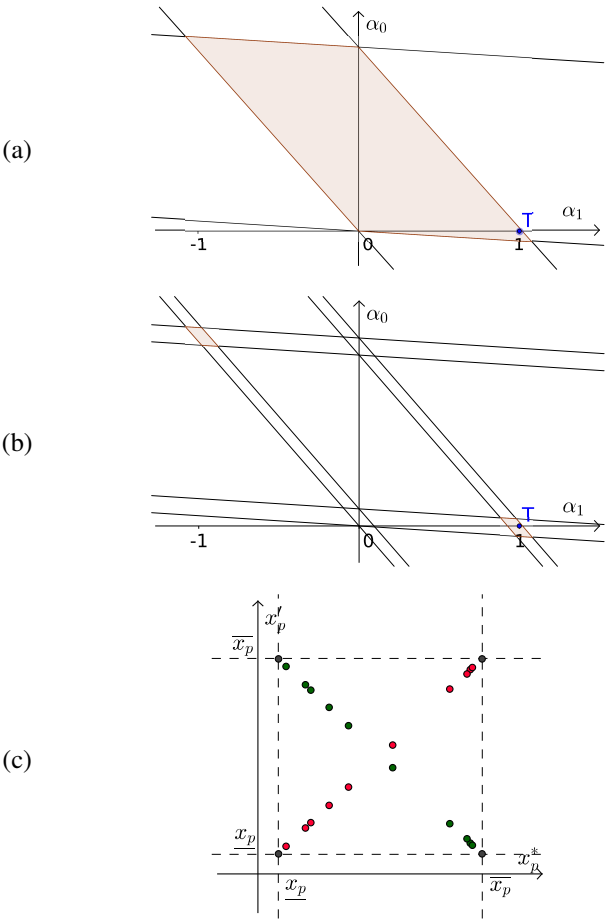

Fig. 1. (a) Regions of equal posterior probability density in (α_0, α_1) space, when using uniform priors on x_p and flat priors on b_{km}. T marks the truth point $(0, 1)$ (b) Similar to (a), but with specifying the uncertainty of the boundaries $\underline{x_p}$, $\overline{x_p}$ of the uniform prior on on x_p. (c) Relationship between estimates of $\boldsymbol{x}$ from the posterior modes in (b): mode containing the truth point $T = (0, 1)$ (*red dots*) and the one that does not (*green dots*).

C. Informative priors for b_{km} (old priors)

The previous subsection demonstrates that, taken alone, uniform priors on x_p are not sufficiently informative to provide

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2945077, IEEE Journal of Biomedical and Health Informatics

useful inferences. In our previous work [6] we supplemented these with informative priors on b_{km}, which effectively limit the high likelihood region to the vicinity of the true point. We describe the full prior specification below for completeness:

$$x_p \sim \mathcal{U}(\underline{x_p}, \overline{x_p}), \quad \forall p, \tag{18}$$

$$b_{1m} \sim \mathcal{N}(1, \sigma_{b1}), \quad \forall m, \tag{19}$$

$$b_{km} \sim \mathcal{N}(0, \sigma_{bk}), \quad \forall m, k \neq 1, \tag{20}$$

where $\underline{x_p}$ and $\overline{x_p}$ are either physical or physiological bounds on x_p and $\sigma_{bk} = c_k(\overline{x_p} - \underline{x_p})^{-k}$, $c_k \approx 1$. This choice was justified from the assumption that genuine MM's response should not deviate too much from the identity function, at least in the vicinity of q_0. Separation strategy [8] was chosen for parametrization of the covariance matrix. First, a decomposition

$$\Sigma = SRS \tag{21}$$

was applied, where $S = diag([\sigma_1, ..., \sigma_M])$ is a diagonal matrix of random error standard deviations (STDs) and $R = [r_{ij}]$ is a symmetric correlation matrix. Then STDs were assigned Jeffreys priors truncated to ensure that the posterior is proper

$$f(\sigma_m) = \begin{cases} \frac{1}{\sigma_m}, & \sigma_{min(m)} < \sigma_m < \sigma_{max(m)} \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

where $\sigma_{min(m)}$ was set to MM's resolution and $\sigma_{max(m)}$ was limited by the span of the measurements. The correlation coefficients were Lewandowski Kurowicka Joe (LKJ) priors with $\eta = 1$ [9].

This approach was successfully validated for MMs of total lesion load (TLL) on a clinical in vivo dataset and several synthetic sets in [6]. Despite the apparent usefulness of informative priors on b_{km} in the context of MM comparison, these priors are somewhat arbitrary. Besides, they limit the class of $g(x)$ to functions close to identity.

D. Informative priors for $\boldsymbol{x}$ jointly (new priors)

To avoid the shortcomings of the uniform and old priors and based on the analysis in sections III-B and III-C we propose a new prior specification scheme. The aim is to limit the region of equal likelihood in (α_0, α_1) space to a small neighborhood of $(0, 1)$ without constraining the polynomial coefficients b_{km}. This is achieved using the following two-step procedure.

We start by recalling that in Bayesian interpretation the probabilities encode our (human) knowledge of the world. When a person specifies a $\mathcal{U}(\underline{x_p}, \overline{x_p})$ prior on the components of $\boldsymbol{x}$, what is usually meant is not only that x_p lie in $[\underline{x_p}, \overline{x_p}]$, $\forall p$, but also that they actually *span* this interval. I.e. that $\min_p x_p$ is close to $\underline{x_p}$ and $\max_p x_p$ is close to $\overline{x_p}$ — a certain *scale* is implied. This implied scale is, however, *not captured* by uniform distribution, but only limited from above: a sample where all points cluster in a small region around, e.g., the center of $[\underline{x_p}, \overline{x_p}]$ is just as likely as a more spread-out sample or, indeed, any other sample falling completely within $[\underline{x_p}, \overline{x_p}]$. The implied scale information can be encoded as the following conditions:

$$\begin{aligned} \max_p x'_p &\triangleq \max_p(\alpha_1 x^*_p + \alpha_0) \geqslant \overline{x_p} - \overline{\epsilon} \\ \min_p x'_p &\triangleq \min_p(\alpha_1 x^*_p + \alpha_0) \leqslant \underline{x_p} + \underline{\epsilon} \end{aligned} \tag{23}$$

where $\overline{\epsilon}, \underline{\epsilon} > 0$ are a priori limits on how far the smallest and the largest values of $\boldsymbol{x}$ might be from the boundaries of the specified uniform prior on x_p. These equations effectively define a new prior on $\boldsymbol{x}$. When taken into account, they reduce the feasible region in (α_0, α_1) space to two small parallelograms (Figure 1) one of which contains the $(0, 1)$ point. The expected value of the feasible region is still at $(\alpha_0, \alpha_1) = (\frac{\underline{x_p} + \overline{x_p}}{2}, 0)$, but the posterior will now contain (at least) two well-separated modes.

The second step is to select the correct mode — the one containing $(\alpha_0, \alpha_1) = (0, 1)$. In principle this can be done after sampling from the posterior. For instance manually — the incorrect mode estimates will most likely be nonsensical; but of course, there are problems with this. First, the choice of the mode to reject will be based on subject matter knowledge and/or intuition; this is hard to justify formally. Second, MCMC algorithms encounter various problems (poor convergence, mixing and sensitivity to initialization) when sampling from such multimodal posteriors with well-separated modes.

To specify this information a priori and reduce the effect of the above problems we note that selecting the correct region in (α_0, α_1) space is a binary choice and as such requires only one additional bit of information. In order to understand how to acquire this information we must understand what is the fundamental difference between the two components in (α_0, α_1) space. The incorrect component contains the point $(\alpha_0, \alpha_1) = (\underline{x_p}, -1)$, for which

$$x'_p = \overline{x_p} - x^*_p \tag{24}$$

This means that x'_p values occupy the same span as x^*_p, but in reverse order (Figure 1). To resolve between the two components it is sufficient to specify the correct order of x_p. A simple way to accomplish this is to find certain $\underline{p}$ and $\overline{p}$, for which $x^*_{\underline{p}} < x^*_{\overline{p}}$, and demand that

$$x'_{\underline{p}} < x'_{\overline{p}} \tag{25}$$

Conditions in equations (23) and (25) are straightforward to implement and we do not pursue an explicit expression for the prior density thus defined. The expected value of the feasible region in (α_0, α_1) space is now at

$$\begin{aligned} \langle \alpha_0 \rangle &= \frac{\overline{x^*_p}(\underline{x_p} + \underline{\epsilon}/2) - \underline{x^*_p}(\overline{x_p} - \overline{\epsilon}/2)}{\overline{x^*_p} - \underline{x^*_p}} \\ \langle \alpha_1 \rangle &= \frac{\overline{x_p} - \underline{x_p}}{\overline{x^*_p} - \underline{x^*_p}} - \frac{1}{2}\frac{\overline{\epsilon} + \underline{\epsilon}}{\overline{x^*_p} - \underline{x^*_p}}. \end{aligned} \tag{26}$$

This is equal to exactly 0 and 1 if $\overline{x_p} - \overline{\epsilon}/2 = \overline{x^*_p}$ and $\underline{x_p} + \underline{\epsilon}/2 = \underline{x^*_p}$. Even in this case, however, there will still remain a certain residual error in the estimates since generally x^*_p would not coincide with their posterior expectations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2945077, IEEE Journal of Biomedical and Health Informatics

IV. Experimental validation

Validation of the BRFEE framework with the proposed new priors was based on two sets of experiments. The first set is in the context of genuine MM comparison (scenario A) where the old priors were established [6]. For validation we apply both the old and the new priors to a synthetic and clinical dataset and compare the results.

In the second set of experiments we consider applications, in which one can only assume a polynomial regression model with MVG random error, but has no prior knowledge of the coefficients of the polynomials. For example, if several quantities that are easy to measure in a living organism and are at the same time physically, chemically or biologically related to an unobservable or otherwise difficult to measure quantity of interest in such a way that the connection between them can be approximated with (2) and (5), then we may treat them as surrogate measurement methods in the context of our framework (scenario B) and attempt the following use cases:

- B1: determine which of the surrogate measurement methods are the best predictors of the quantity of interest;
- B2: estimate the values of quantity of interest in the sample from the values of surrogate measurements.

We perform an experiment with a synthetic dataset to demonstrate how old priors fail in such application and follow with an experiment with new priors on a clinical dataset of in vivo volumetric brain measurements.

All experiments were performed with NUTS sampler from `pymc3` Python package using a model with explicit parametrization of (α_0, α_1) and a custom post-processing of the samples. Details may be found in the documentation accompanying our code available on GitHub: http://github.com/madanh/practical_priors.

A. Performance metrics

A good MM allows to find q_p from y_{pm} with low error. The magnitude of the systematic error by itself is of little concern, since once it is known it can be compensated for. What is important is how sensitive the measurements are to the changes of the measurand compared to the random error. A method with $g_m(q) = const$ would be utterly useless for recovering q_p, conversely, the higher g'_m is relative to ϵ_{pm} the less error we would make (Figure 2). The quantity

$$Q_m(q) = \frac{\mid g'_m(q) \mid}{\sigma_m} \tag{27}$$

can be used as a MM's quality measure at point q in the context of model (1). MMs with non-linear g_m would generally be better in some regions of their range than others. As a consequence one might want to use different methods for different ranges of the measurand (Figure 2, right).

We focus on the ability of the framework to estimate the following figure of merit of an MM as a predictor of q

$$F_m = \max_q Q_m(q) \triangleq \max_{q \in [\underline{q_p}, \overline{q_p}]} \frac{\mid g'_m(q) \mid}{\sigma_m} \tag{28}$$

Under BRFEE model's assumption, this figure of merit is equal to the root mean square error (RMSE) one would obtain

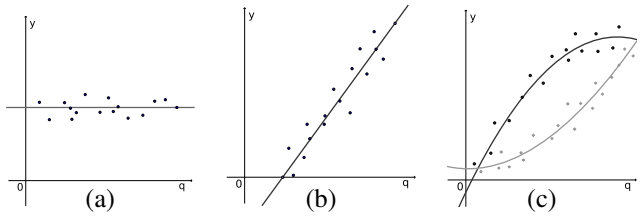

Fig. 2. Utility of hypothetical quantity y as a measurement method for quantity q: (a) y is useless as a measurement method for q; (b) y is useful as measurement method for q; (c) two quantities that are useful as measurement methods for q in different ranges.

if he or she used m with known bias coefficients within their optimal range in order to estimate q. For F_m we provide Mean absolute error (MAE) and correlation coefficient w.r.t the reference values as simple summary measures of quality of estimation defined as follows

$$MAE \triangleq \frac{1}{M} \sum_{m=1}^{M} \mid \tilde{F_m} - F_m^* \mid, \tag{29}$$

$$Corr \triangleq \frac{\sum_{m=1}^{M} (\tilde{F_m} - \langle \tilde{F_m} \rangle)(F_m^* - \langle F_m^* \rangle)}{\sqrt{\sum_{m=1}^{M} (\tilde{F_m} - \langle \tilde{F_m} \rangle)^2} \sqrt{\sum_{m=1}^{M} (F_m^* - \langle F_m^* \rangle)^2}}, \tag{30}$$

where $\tilde{F_m}$ is the estimate and F_m^* is the reference value that is either known by construction in synthetic data or estimated using least squares regression and reference q measurements that were available for BRFEE validation in clinical datasets.

We also introduce additional performance metrics for the clinical application that uses surrogate measurements in the second set of experiments. To evaluate BRFEE performance in use case B1 we use $Corr$ in (30). To gauge the performance of our approach in the use case B2 we compare RMSE of BRFEE estimates of the unobservable quantity $(\tilde{q_p})$

$$A \triangleq \sqrt{\sum_{p=1}^{n} (\tilde{q_p} - q_p^*)^2 / N} \tag{31}$$

to the smallest RMSE obtainable with only the best predictor with known g_m

$$A_1 \triangleq \min_m 1/F_m^*. \tag{32}$$

B. Genuine measurement methods

The aim of the following two experiments is to compare the previous old priors and the proposed new priors in a situation where they are expected to have similar performance, i.e., in the context of genuine measurement method comparison.

1) Synthetic data: In order to be able to disambiguate the correct and the reverse-order mode with the new priors (section III-D) we need to implement condition (25), i.e. to provide a pair of points $\underline{p}$ and $\overline{p}$ such that $x_{\underline{p}}^* < x_{\overline{p}}^*$ and therefore $q_{\underline{p}}^* < q_{\overline{p}}^*$. For all synthetic data we allowed ourselves the following convenience which does not result in loss of generality. We fixed $\underline{p}$ to be equal to one and $\overline{p}$ to be equal to N. We generated a set $\boldsymbol{q} \triangleq \{q_1, \ldots, q_N\}$ by sampling N

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2945077, IEEE Journal of Biomedical and Health Informatics

points from $\mathcal{U}(\underline{q_p}, \overline{q_p})$. Then, if q_1 happened to be greater than q_N, we simply swapped the values at the indices 1 and N.

For this experiment we had $N = 30$, $\underline{q_p} = 0$ and $\overline{q_p} = 55$. Four quadratic polynomials ($M = 4, K = 2$, coefficients given in Table I) were evaluated at those points ($x_p = q_p$). MVG noise with standard deviations given in Table I and correlation matrix

$$R = \begin{pmatrix} 1 & 0.9 & 0 & 0 \\ 0.9 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{33}$$

was added to produce "measurements" y_{pm}.

TABLE I
PARAMETERS THAT WERE USED TO GENERATE THE SYNTHETIC DATA THAT ARE NUMERICALLY EQUAL TO b_{km} AND σ_m ESTIMATED WITH LEAST SQUARES (LS) REGRESSION AGAINST CONSENSUS SEGMENTATION FOR THE TLL EXPERIMENT.

m	b_{0m}	b_{1m}	$b_{2m} \cdot 10^3$	σ_m
1	3.7	0.99	−19.2	7.0
2	2.7	1.15	−20.1	6.2
3	1.2	0.55	−3.6	4.3
4	11.6	0.35	4.0	2.2

We obtained a posterior sample with the old priors with $\underline{x_p} = 0$, $\overline{x_p} = 55$, $c_k = 1$, $\sigma_m \sim 1/\sigma_m, 0.001 < \sigma_m < 55$, and with the new priors with $\underline{x_p} = 0$, $\overline{x_p} = 55$, $\underline{\epsilon} = \overline{\epsilon} = 5$ and the first and the last sample points as order-disambiguating pair. Both setups returned similar estimates with MAE of the order of reported credible region (CR) width and $Corr$ close to one. CR of F_m are slightly wider for the new priors meaning that they are less informative than the old ones in this experiment.

TABLE II
FIGURE OF MERIT ESTIMATES IN THE EXPERIMENT WITH SYNTHETIC DATA MODELING MEASUREMENT METHODS.

m	Old priors	New Priors	Truth
1	0.12 ± 0.08	0.17 ± 0.12	0.160
2	0.13 ± 0.08	0.20 ± 0.12	0.185
3	0.24 ± 0.10	0.22 ± 0.12	0.128
4	0.54 ± 0.19	0.59 ± 0.20	0.359
MAE	0.09	0.09	
Corr	0.89	0.96	

2) Total lesion load data: experiment on the clinical dataset used in our previous works [6] and with the old priors is repeated with the new priors.

Dataset consisted of total lesion load (TLL) measurements of 22 multiple sclerosis patients obtained from corresponding magnetic resonance (MR) images based on manual consensus-based lesion segmentations [10] and four automated segmentation methods [11]–[14]. For each the TLL was measured as the lesion voxel count times voxel volume. The consensus-based segmentation [10] and were used as gold standard reference in LS regression to obtain reference values for B and Σ (Table I). The priors were setup as in the previous experiment (section IV-B1).

From Table III we see that the estimates are very similar, with the ones provided by the new priors being slightly less accurate and certain. Figures of merit shown in Figure 3 are consistent, from which we conclude that the estimates obtained with the new priors are *similar* in the domain, where the assumptions underlying the old priors are justified. The observed small decrease in nominal performance is because for the particular scenario the new priors are less informative, albeit more objective.

Figure 4a shows that the obtained model parameter estimates agree with the LS estimates. Furthermore, the estimates of TLL values are closer to the gold standard reference values as compared to any of the individual methods (Figure 4b).

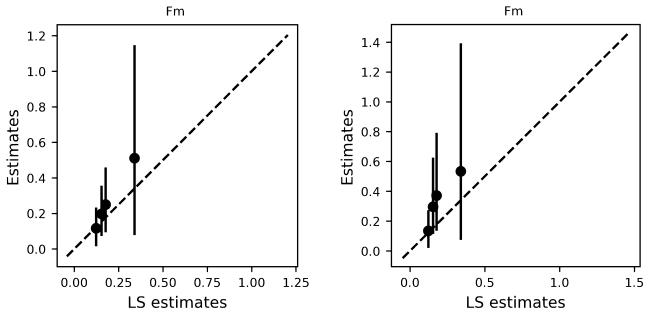

Fig. 3. Figure of merit estimates in the experiment with total lesion load data when using (a) old priors and (b) new priors.

TABLE III
FIGURE OF MERIT ESTIMATES IN THE EXPERIMENT WITH TOTAL LESION LOAD DATA.

m	Old priors	New priors	LS regression
1	0.18 ± 0.10	0.26 ± 0.12	0.153
2	0.22 ± 0.10	0.31 ± 0.14	0.177
3	0.35 ± 0.19	0.35 ± 0.21	0.34
4	0.11 ± 0.11	0.13 ± 0.13	0.123
MAE	0.03	0.07	
Corr	0.98	0.75	

Again the new priors are somewhat less informative than the old ones, which is reflected in larger CR widths and higher values of MAE and and lower $Corr$.

C. Surrogate measurement methods

As mentioned before, we envision a new class of practical applications of BRFEE framework with the new priors. With the restrictive assumptions on the shape of g_m dropped we may now attempt to use observable quantities as potentially highly biased surrogate measurement methods for an unobservable (or simply hard to measure) quantity of interest. For instance, we might treat intracellular concentration of some molecule as an unobservable quantity q_p and blood concentrations of this molecule and its metabolites as y_{pm}. Or treat disease severity as q_p and values of biomarkers for this disease as y_{pm}. Or, as we showcase below, treat volumes of large easy-to-segment brain structures as surrogate measurements y_{pm} of the volume q_p of a small structure of interest. We might be interested in BRFEE estimates of q_p (use case B1) or in comparing predictive power of surrogate measurements w.r.t. q_p in order to use only the best ones in future applications (use case B2).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2945077, IEEE Journal of Biomedical and Health Informatics

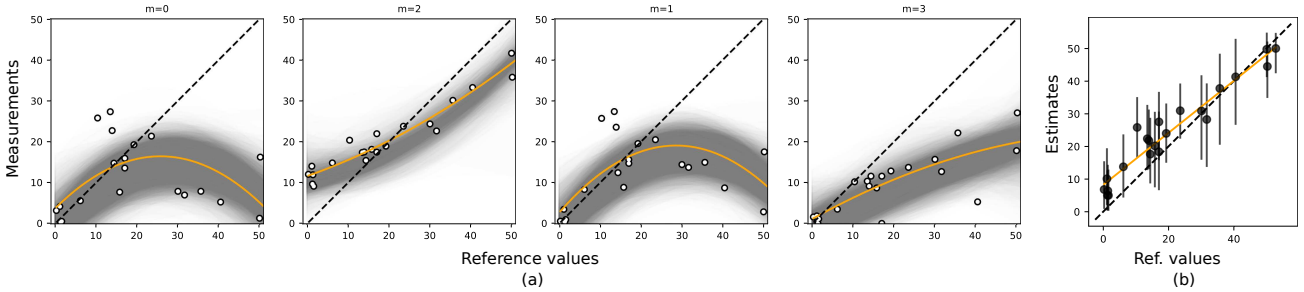

Fig. 4. (a) Measurements of TLL based on four automated segmentation methods ($m = 1, \ldots, 4$) versus the reference values based on gold standard reference segmentation (*dots*). The *orange lines* represent models obtained by LS fit, *gray densities* the models obtained from the posterior and *dashed lines* the identity. (b) Estimated versus the reference TLL. The *bars* show the associated uncertainty, while the *orange line* shows the LS fit.

1) Synthetic data: As in the previous synthetic experiment (section IV-B1) we generated $N = 30$ points from $\mathcal{U}(0, 55)$ and evaluated polynomials with coefficients given in Table IV at those points. The values of the polynomials were then perturbed with MVG noise with standard deviations from Table IV and correlations as per (33). Note that polynomials are now much further from the identity function than previously.

TABLE IV
PARAMETERS USED TO GENERATE SYNTHETIC DATA FOR THE UNOBSERVED QUANTITY ESTIMATION EXPERIMENT.

m	b_{0m}	b_{1m}	b_{2m}	σ_m
1	80	-4.0	0.01	7.0
2	-80	4.0	-0.02	6.2
3	-80	12.0	-0.20	4.3
4	40	-3.5	0.04	2.2

Priors were setup as in section IV-B1. Results are presented in Table V and show that estimates with the old priors were unrepresentative of the true values, while the new priors yielded the estimates close to the truth. Figure of merit estimates with the old priors have negative correlation, which means they are not suitable for validation nor for comparison of MMs.

The new priors, on the other hand, maintain good estimation performance. Figure of merit estimates are highly correlated with the true values. A graphical representation of the results obtained with the new priors can be seen in Figure 5. Apparently the estimates of q are very close to the true values.

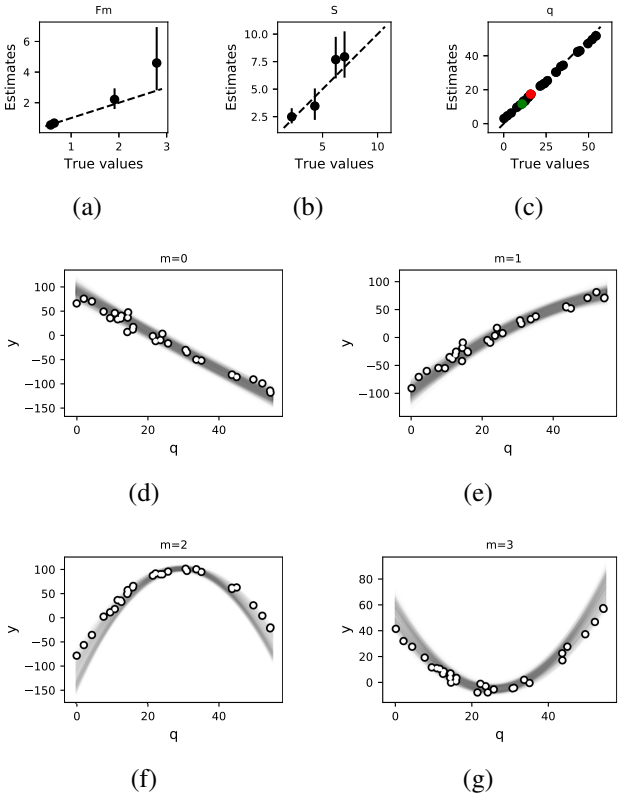

Fig. 5. Synthetic data experiment emulating surrogate measurement methods for an unobserved quantity. Comparison of (a) the figure of merit, (b) standard deviations of the random error and (c) unobserved quantity estimates with their true values. (d, e, f, g) posterior predictive curves and the actual "measurement" values for four methods. *Red and green points* mark the order-disambiguating pair.

TABLE V
FIGURE OF MERIT ESTIMATES IN THE EXPERIMENT WITH SYNTHETIC DATA MODELING OBSERVED QUANTITIES DEPENDENT ON THE UNOBSERVED QUANTITY AS POLYNOMIAL WITH ARBITRARY MAGNITUDE OF THE COEFFICIENTS.

m	Old priors	New Priors	Truth
1	0.33 ± 0.67	0.53 ± 0.11	0.571
2	0.23 ± 0.64	0.65 ± 0.11	0.645
3	0.052 ± 0.038	4.30 ± 0.59	2.79
4	0.20 ± 0.57	2.15 ± 0.29	1.91
MAE	1.27	0.45	
Corr	-0.92	0.98	

2) Nucleus basalis of Meynert volume: The nucleus basalis of Meynert (NBM) acts as the principal source of the neurotransmitter acetylcholine for the cerebral cortex. Degeneration of the NBM might herald the onset of a progressing dementia syndrome, such as Alzheimer's disease or Parkinson's disease with dementia [15], [16]. The region is small and cannot be segmented from routine MR images using conventional procedures.

On the other hand, the volume of NBM or any other structure in an intact brain can be expected to depend on the volume of structures adjacent to it, and, to a lesser degree, on the volumes of more distant structures. Another reason for dependence between volume of some structures is that they might be affected by a common pathophysiological process. In this experiment we attempt to use whole brain volumetric

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2945077, IEEE Journal of Biomedical and Health Informatics

measurements and measurements of hippocampal subfields obtained with automatic segmentation procedures to estimate these dependencies (use case B1) as well as to estimate the NBM volumes in the sample (use case B2).

We analyzed structural T1-weighted and T2-weighted MRI scans of 40 participants, including 20 healthy elderly and 20 patients with mild cognitive impairment, the prodromal stage of Alzheimer's disease. Automated segmentation of structural images and volumetric measurements of the whole brain, the hippocampus and its subfields were performed using FreeSurfer. All measurements were normalized to the total intracranial volume to account for the differences in head size between subjects. The FreeSurfer volumes were treated as surrogate measurements for NBM volume in BRFEE with the new priors. For validation purposes reference NBM volumes were obtained using morphometric analysis of T1-weighted images and a detailed stereotactic atlas that has been validated against postmortem anatomy (see [16] for the detailed procedure).

Since the old priors are inapplicable in this case we only apply the new priors. The minimum and the maximum points were determined from normalized reference NBM volumes and provided indices $\underline{p}$ and $\overline{p}$, based on which the remaining parameters of the prior were setup: $\underline{\epsilon}$ and $\overline{\epsilon}$ were set to 0.02, while $\underline{q_p}$ and $\overline{q_p}$ were set so that the respective minimum and maximum values were approximately at $\underline{q_p}+\underline{\epsilon}/2$ and $\overline{q_p}+\overline{\epsilon}/2$. K was set to 1, q_0 was set to $(\underline{q_p}+\overline{q_p})/2$.

The resulting estimates in Figure 6 show good agreement with the reference. Generally, estimates of F_m are in good agreement with those obtained using least squares on reference values, taking into account the associated uncertainty, and therefore enable use case B1. The BRFEE RMSE for NBM volume is slightly lower as compared to the RMSE obtainable from a single predictor (A=0.038, A_1=0.049 for left hemisphere, A=0.027, A_1 = 0.034 for right hemisphere). This means that BRFEE performance in use case B2 is comparable or better than the best surrogate measurement method, but does not require knowing the coefficients of systematic response in advance.

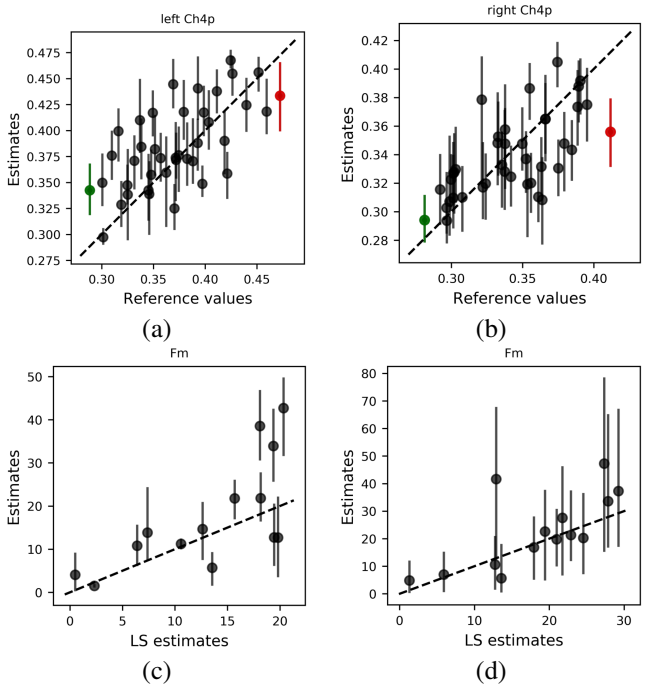

Fig. 6. Experiment with clinical data where the volumes of segmented brain structures serve as surrogate measurements for NBM volume. Estimates of normalized NBM volume plotted against reference values for (a) left and (b) right brain hemisphere. *Green and red points* denote the order-disambiguating pair $(\underline{p}, \overline{p})$. Figure of merit of the measured volumes as predictors for (a) left and (b) right NBM volume.

V. Discussion

A major contribution of this work are the theoretical results that clarify the relationship between BRFEE estimates and the true values of measured quantities and explain how to acquire experimental data such that the uncertainty of error estimation is minimized. Moreover, the proposed prior specification scheme extends the class of problems to which the framework can be applied. For instance, one can estimate a certain dependent quantity that is either unobservable in principle or inherently difficult to measure from a given a set of observable quantities. A pertinent example presented is estimation of the volume of small brain nucleus that is an important biomarker of dementia onset based on volumes of larger brain structures routinely segmented from MR images.

The experiments show that the new priors for BRFEE can be used as a drop-in replacement for the old ones for the purpose of MMs comparison. Besides, they enable a new kind of application where several easy to measure quantities, potentially dependent on a difficult to measure or unobservable quantity, are treated as surrogate measurements in order to estimate its value in the sample and estimate the said dependencies.

For the purposes of comparing MMs the new priors can be thought of as being more practical and objective — they don't require guesswork regarding coefficients of the polynomials, only sufficient knowledge of the measurand range and ordering, in the form of additional tolerance parameters $\underline{\epsilon}$ and $\overline{\epsilon}$ and the order-disambiguating pair of indices $(\underline{p}, \overline{p})$, respectively. In some situations additional parameters can be measured or inferred with certainty. For example, in case of TLL measurements, by including a healthy control subject in the dataset $\underline{\epsilon}$ is reduced to exactly zero. Furthermore, if the patient with the highest TLL can be identified, then by a single application of a gold standard MM, $\overline{\epsilon}$ can be reduced to a small multiple of this method's (nominal) accuracy. As for the order-disambiguation pair, we can think of three practical ways to select $\underline{p}$ and $\overline{p}$. The first way is to use controls: for a large class of biomarkers it is known that a healthy control subject has the true value of the biomarker exactly equal to zero and is guaranteed to be less than the corresponding value for a patient who has the relevant medical condition. The second way is to apply a GS method to obtain two reference measurements and thus infer the order of the estimates. The third way is an educated guess: if m are genuine measurement methods there might be a pair of patients, for which the measurements are such as to virtually guarantee a certain ordering of the underlying measurand values.

Since the new priors lift all assumptions on the coefficients

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2019.2945077, IEEE Journal of Biomedical and Health Informatics

of polynomials, which model the systematic error, the class of problems that can be addressed by the reference-free approach is significantly extended. The framework becomes applicable to the problems of exploratory factor analysis and unobserved quantity estimation. This is exemplified to some extent by our experiments in section IV-C. The synthetic experiment showed that despite highly non-linear g_m it was possible to recover q_p with high accuracy as long as σ_m remained low relative to the derivative of at least one of g_m in all subregions of $[\underline{q}_p, \overline{q}_p]$. For the experiment with clinical data, the estimates were less accurate, but still useful as a synthetic biomarker for diagnostic purposes with costs much lower than those of reference measurements.

In both demonstrated scenarios the BRFEE provides significant savings of time and costs. For instance, examination of F_m or Q_m allows to determine the surrogate measurement methods whose contribution to the estimates is negligible and thus discontinue their assessment, thereby eliminating future costs associated with them. Furthermore, savings and costs associated with the creation of reference dataset may be reduced or eliminated, including the costs of human time, but also the costs of non-standard acquisition protocols, high-end acquisition equipment, material costs (e.g. contrast agents, phantom), instrumentation costs (frames, fiducial markers), administrative overhead and possible side effects for the patients.

The theoretical analysis provided in the present work uncovers the essence of the reference-free approach: the measurement data alone *does contain* significant information to recover the underlying measurand, but only up to a linear transformation. Information about the scale and location of the measurand distribution must be supplied separately in the priors in order to enable effective inference.

With the proposed new priors we have shown one of the ways in which scale and location information can be encoded. Other ways can be imagined and might even be preferable in certain applications. It is important to note that, regardless of the exact encoding, the quality (accuracy and precision) of scale and location information will directly influence the quality of inference. This is exemplified in our synthetic experiments, where dependence of estimation accuracy on the number of measurand points N was studied. Constant and relatively low precision of the specified scale and location as encoded by $\underline{\epsilon}$ and $\overline{\epsilon}$ prevented an improvement of the accuracy of systematic error estimation with increasing dataset size.

In order to apply the proposed framework to a given dataset the *a priori* relationships between the observable quantities (i.e. measurements) and the unobservable quantity (i.e. the one being estimated) should adhere to Equation (2). For a meaningful inference it is also desirable that we have *a priori* grounds to believe that Equation (5) is a good approximation of the actual error distribution. Whether these assumptions hold for a given real-word problem cannot be known from the data alone, but should be established from domain knowledge.

In this work we used a generic polynomial approximation of MMs systematic error. When physical considerations suggest a different functional form it should be used instead. The proofs that we have provided are not applicable in such cases, however, we conjecture that the general property that the measurand is encoded in the likelihood up to some transformation would still hold. It might even be that, if for different m the g_m belong to different parametric families, then the measurand would be identified by the likelihood uniquely.

The results in this work showcase the power of Bayesian method and demonstrate what can be achieved by means of even a very primitive theoretical analysis. The area of possible application of BRFEE is truly wide and encompasses some important use cases. Apart from obvious use in medical image analysis and biomarker research it has implications for metrology in general. It shows how MMs can be treated symmetrically without arbitrarily declaring any method to represent the reference or "gold standard".

A software implementation of the framework, software used to generate the synthetic data and analyze the MCMC samples along with the clinical datasets used in this work are available on GitHub: http://github.com/madanh/practical_priors.

CONFLICT OF INTEREST

The Author(s) declare(s) that there is no conflict of interest.

REFERENCES

[1] T. Uher, M. Vaneckova, L. Sobisek, M. Tyblova, Z. Seidl, J. Krasensky, D. Ramasamy, R. Zivadinov, E. Havrdova, T. Kalincik, and D. Horakova, "Combining clinical and magnetic resonance imaging markers enhances prediction of 12-year disability in multiple sclerosis," *Multiple Sclerosis*, vol. 23, no. 1, pp. 51–61, 2017.

[2] M. A. Kupinski, J. W. Hoppin, E. Clarkson, H. H. Barrett, and G. A. Kastis, "Estimation in medical imaging without a gold standard," *Acad. Radiol.*, vol. 9, no. 3, pp. 290–297, Mar. 2002.

[3] A. K. Jha, M. A. Kupinski, J. J. Rodríguez, R. M. Stephen, and A. T. Stopeck, "Task-based evaluation of segmentation algorithms for diffusion-weighted MRI without using a gold standard," *Phys. Med. Biol.*, vol. 57, no. 13, pp. 4425–4446, 2012.

[4] A. K. Jha, B. Caffo, and E. C. Frey, "A no-gold-standard technique for objective assessment of quantitative nuclear-medicine imaging methods," *Physics in Medicine and Biology*, vol. 61, no. 7, pp. 2780–2800, Apr. 2016.

[5] J. Lebenberg, A. Lalande, P. Clarysse, I. Buvat, C. Casta, A. Cochet, C. Constantinidès, J. Cousty, A. d. Cesare, S. Jehan-Besson, M. Lefort, L. Najman, E. Roullot, L. Sarry, C. Tilmant, F. Frouin, and M. Garreau, "Improved Estimation of Cardiac Function Parameters Using a Combination of Independent Automated Segmentation Results in Cardiovascular Magnetic Resonance Imaging," *PLOS ONE*, vol. 10, no. 8, p. e0135715, Aug. 2015.

[6] H. Madan, F. Pernuš, and Žiga Špiclin, "Reference-free error estimation for multiple measurement methods," *Statistical Methods in Medical Research*, vol. 0, no. 0, p. 0962280217754231, 2018, pMID: 29384043. [Online]. Available: https://doi.org/10.1177/0962280217754231

[7] H. Madan, R. Berlot, N. J. Ray, F. Pernuš, and Z. Špiclin, "Predicting Nucleus Basalis of Meynert Volume from Compartmental Brain Segmentations," in *PRedictive Intelligence in MEdicine*, ser. Lecture Notes in Computer Science, I. Rekik, G. Unal, E. Adeli, and S. H. Park, Eds. Springer International Publishing, 2018, pp. 68–75.

[8] J. Barnard, R. McCulloch, and X.-L. Meng, "Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage," *Statistica Sinica*, pp. 1281–1311, 2000.

[9] D. Lewandowski, D. Kurowicka, and H. Joe, "Generating random correlation matrices based on vines and extended onion method," *Journal of multivariate analysis*, vol. 100, no. 9, pp. 1989–2001, 2009.

[10] Ž. Lesjak, A. Galimzianova, A. Koren, M. Lukin, F. Pernuš, B. Likar, and Ž. Špiclin, "A novel public mr image dataset of multiple sclerosis patients with lesion segmentations based on multi-rater consensus," *Neuroinformatics*, vol. 16, no. 1, pp. 51–63, Jan 2018.

[11] S. Jain, D. M. Sima, A. Ribbens, M. Cambron, A. Maertens, W. Van Hecke, J. De Mey, F. Barkhof, M. D. Steenwijk, M. Daams, F. Maes, S. Van Huffel, H. Vrenken, and D. Smeets, "Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images," *NeuroImage : Clinical*, vol. 8, pp. 367–375, May 2015.

[12] A. Galimzianova, Z. Lesjak, B. Likar, F. Pernus, and Z. Spiclin, "Locally adaptive MR intensity models and MRF-based segmentation of multiple sclerosis lesions," *Proc. SPIE Int. Soc. Opt. Eng.*, vol. 9413, p. 94133G, 20 Mar. 2015.

[13] A. Galimzianova, F. Pernus, B. Likar, and Z. Spiclin, "Stratified mixture modeling for segmentation of white-matter lesions in brain MR images," *NeuroImage*, vol. 124, no. Pt A, pp. 1031–1043, Jan. 2016.

[14] T. Jerman, A. Galimzianova, F. Pernus, B. Likar, and Z. Spiclin, "Combining unsupervised and supervised methods for lesion segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, ser. Lecture Notes in Computer Science, A. Crimi, B. Menze, O. Maier, M. Reyes, and H. Handels, Eds. Springer International Publishing, 2016, no. 9556, pp. 45–56.

[15] T. W. Schmitz, R. N. Spreng, M. W. Weiner, P. Aisen, R. Petersen, C. R. Jack, W. Jagust, J. Q. Trojanowki, A. W. Toga, L. Beckett *et al.*, "Basal forebrain degeneration precedes and predicts the cortical spread of alzheimer's pathology," *Nature communications*, vol. 7, p. 13249, 2016.

[16] N. J. Ray, S. Bradburn, C. Murgatroyd, U. Toseeb, P. Mir, G. K. Kountouriotis, S. J. Teipel, and M. J. Grothe, "In vivo cholinergic basal forebrain atrophy predicts cognitive decline in de novo Parkinson's disease," *Brain*, vol. 141, no. 1, pp. 165–176, Jan. 2018.

APPENDIX: THEOREM PROOFS

Theorem 1. *Let*

$$x'_p = \alpha_1 x_p + \alpha_0, \quad \alpha_0, \alpha_1 = const,\ \alpha_1 \neq 0 \quad \forall p \tag{34}$$

then there exist b'_{km} *such that*

$$\sum_{k=0}^{K} b_{km} x_p^k = \sum_{k=0}^{K} b'_{km} x_p'^k \quad \forall m \tag{35}$$

and thus

$$f(Y \mid B, \Sigma, \boldsymbol{x}) = f(Y \mid B', \Sigma, \boldsymbol{x}') \tag{36}$$

Proof. We drop the m subscript in this proof for brevity. Substitution of (34) into (35) gives

$$\sum_{k=0}^{K} b_k x_p^k = \sum_{k=0}^{K} b'_k (\alpha_1 x_p + \alpha_0)^k \tag{37}$$

Equating coefficients for the k-th power of x_p

$$b_K = b'_K \alpha_1^K \tag{38}$$

and

$$b'_K = \frac{b_K}{\alpha_1^K} \tag{39}$$

For the $(K-1)$-th power of x_p:

$$b_{K-1} = b'_{K-1} \alpha_1^{K-1} + b'_K \binom{K}{K-1} \alpha_0 \alpha_1^{K-1} \tag{40}$$

where $\binom{\bullet}{\bullet}$ is binomial coefficient. From (39)

$$b'_{K-1} = \frac{b_{K-1}}{\alpha_1^{K-1}} - b'_K \binom{K}{K-1} \alpha_0 \tag{41}$$

Proceeding similarly for lower powers we find that for all $k = K - l, l = 0..K$

$$b_{K-l} = b'_{K-l} \alpha_1^{K-l} + \sum_{j=K-(l-1)}^{K} \binom{j}{K-l} b'_j \alpha_0^{j-k} \alpha_1^{K-l} \tag{42}$$

$$b_k = b'_k \alpha_1^k + \sum_{j=k+1}^{K} \binom{j}{k} b'_j \alpha_0^{j-k} \alpha_1^k \tag{43}$$

and

$$b'_k = \frac{b_k}{\alpha_1^k} - \sum_{j=k+1}^{K} \binom{j}{k} b'_j \alpha_0^{j-k} \tag{44}$$

thereby defining b'_k for all k. □

Theorem 2. *Let* h *be analytic on* $(\underline{x}, \overline{x})$ *and*

$$x'_p = h(x_p) \quad \forall p \tag{45}$$

Then (35) has a solution in b'_{km} *for general* b_{km} *only when*

$$h(x_p) = \alpha_1 x_p + \alpha_0, \quad \alpha_0, \alpha_1 = const,\ \alpha_1 \neq 0 \tag{46}$$

Proof. First, since LHS of (35) is a polynomial in powers of x_p, RHS must be a polynomial in powers of x_p as well, consequently $h(x_p)$ is a polynomial. Let J be its degree:

$$x'_p = \sum_{j=o}^{J} \alpha_j x_p^j, \quad \alpha_J \neq 0. \tag{47}$$

Then we have to prove that (35) generally has no solutions when $J > 1$. Substitution of (47) in (35) gives

$$\sum_{k=0}^{K} b_{km} x_p^k = \sum_{k=0}^{K} b'_{km} \left(\sum_{j=0}^{J} \alpha_j x_p^j \right)^k \tag{48}$$

LHS is a polynomial of power K while RHS is a polynomial of power JK, therefore all RHS coefficients except the lowest K are equal to zero. Consequently, for the subset of terms that include α_J in RHS:

$$b'_{km} \alpha_J^k = 0, \quad \forall k > K/J \tag{49}$$

From (47) $\alpha_J \neq 0$, so

$$b'_{km} = 0, \quad \forall k > K/J \tag{50}$$

Now what remains is to consider b'_{km} for $k \leqslant K/J$. For $k = \left[\frac{K}{J}\right]$, where $[\bullet]$ is the integer part operator, the coefficient of x^k must satisfy

$$b_{km} = b'_{km} \alpha_J^k \tag{51}$$

The coefficient of x^{k-1} must satisfy

$$b_{(k-1)m} = b'_{km} \left(\binom{k-1}{1, k-2} \alpha_J^{k-2} \alpha_1 \right) \tag{52}$$

where $\binom{\bullet}{\bullet,\bullet}$ is multinomial coefficient.

Eliminating b'_{km} from the above two equations:

$$\frac{b_{km}}{b_{(k-1)m}} = \frac{\alpha_J^k}{\binom{k-1}{1,k-2} \alpha_1 \alpha_J^{k-2}} \quad \forall m \tag{53}$$

These equations cannot be satisfied simultaneously for all m if $\frac{b_{km}}{b_{(k-1)m}}$ differ across m thus proving that generally we cannot have $J > 1$. Consequently, h is (at most) linear. □