Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

ILJA KLEBANOV, ALEXANDER SIKORSKI, CHRISTOF SCHÜTTE, SUSANNA RÖBLITZ

Empirical Bayes Methods for Prior Estimation in Systems Medicine [1]

[1]This research was carried out in the framework of MATHEON supported by Einstein Foundation Berlin.

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30-84185-0
Telefax: +49 30-84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

Empirical Bayes Methods for Prior Estimation in Systems Medicine

Ilja Klebanov, Alexander Sikorski, Christof Schütte, Susanna Röblitz

December 1, 2016

Abstract

One of the main goals of mathematical modelling in systems medicine related to medical applications is to obtain patient-specific parameterizations and model predictions. In clinical practice, however, the number of available measurements for single patients is usually limited due to time and cost restrictions. This hampers the process of making patient-specific predictions about the outcome of a treatment. On the other hand, data are often available for many patients, in particular if extensive clinical studies have been performed. Therefore, before applying Bayes' rule *separately* to the data of each patient (which is typically performed using a non-informative prior), it is meaningful to use empirical Bayes methods in order to construct an informative prior from all available data. We compare the performance of four priors – a non-informative prior and priors chosen by nonparametric maximum likelihood estimation (NPMLE), by maximum penalized lilelihood estimation (MPLE) and by doubly-smoothed maximum likelihood estimation (DS-MLE) – by applying them to a low-dimensional parameter estimation problem in a toy model as well as to a high-dimensional ODE model of the human menstrual cycle, which represents a typical example from systems biology modelling.

Keywords: Parameter estimation, Bayesian inference, Bayesian hierarchical modelling, hyperparameter, hyperprior, principle of maximum entropy, NPMLE, MPLE, DS-MLE, EM algorithm, Jeffreys prior, reference prior

1 Introduction

The estimation of a parameter $X \in \mathbb{R}^d$ from a measurement $Z \in \mathbb{R}^n$ given a model $\phi\colon \mathbb{R}^d \to \mathbb{R}^n$ can be an ill-posed problem, especially if ϕ is non-injective. In addition, measurements are usually defective and the measurement error should be taken into account. For these reasons, it is meaningful to view X, Z and the error term E as random variables with

$$Z = \phi(X) + E, \qquad E \sim \rho_E \ \text{ independent of } X, \tag{1}$$

where ρ_E denotes the density of the error term E. After establishing a *prior density* $\rho_X(x)$ of X, which reflects our initial knowledge about the parameter, Bayes' rule gives us the proper tool to update our knowledge when taking into

accout the measurement:

$$\rho_X(x \mid Z = z) = \frac{\rho_E(z - \phi(x))\, \rho_X(x)}{\rho_Z(z)},$$

where

$$\rho_Z(z) = \int_{\mathbb{R}^d} \rho_E(z - \phi(x))\, \rho_X(x)\, \mathrm{d}x.$$

In contrast to point estimators like the maximum likelihood estimator (MLE)

$$\hat{x}_{\mathrm{MLE}} = \arg\max_x \rho_E(z - \phi(x)),$$

which usually try to optimize the fit, the result of the Bayesian inference is a whole *distribution*, the "posterior", in the parameter space $\mathcal{X}$. Usually, especially in high dimensions, it is given by a (possibly weighted) sampling, since the density function is too costly to compute due to the integral $\rho_Z(z)$ in the denominator. A sampling can be produced without the knowledge of this denominator by e.g. the Metropolis-Hastings algorithm [8, 4] and is useful for the computation of certain expectation values with respect to the posterior.

However, in many applications no comprehensible prior can be assigned, which results in eliciting priors based on expert opinion and therefore in different posterior distributions depending on which expert was asked. This unsatisfactory lack of objectivity and consistency has led to many controversial discussions about the reasonability and trustability of Bayesian inference.

Empirical Bayes methods provide one possible solution to this issue by first estimating the prior distribution. This, naturally, will require further knowledge, in our case (independent, ρ_Z-distributed) measurements $Z_1 = z_1, \ldots, Z_M = z_M$ for *several* individuals, which exists in many statistical trials.

An important application is the prediction of patient-specific treatment success rates based on clinical measurement data and a mathematical model describing the underlying physiological processes. One example are hormonal treatments of the human menstrual cycle, as they are frequently performed in reproductive medicine. In this case, clinical data are available as well as a robust mathematical model, which allows to simulate the cyclic behavior under varying external conditions [10].

Typically, predictions are required for a specific patient in the daily clinical practice, where the number of measurements is limited due to time and cost restrictions. On the other hand, data are often available for many (hundreds or even thousands of) patients, in particular if extensive clinical studies have been performed for the approval of a drug. Using these population data, we propose an iterative algorithm for constructing an informative prior distribution, which then serves as the basis for computing patient-specific posteriors and obtaining individual predictions.

In the empirical Bayes framework, the prior $\Pi = \pi$ is considered as a hyperparameter with unknown true value $\pi_{\mathrm{true}} = \rho_X$. A corresponding likelihood function $L(\pi)$ can be derived and statistical inference (frequentist or Bayesian) can be applied for its estimation. If the prior has no prescribed parametric form, the resulting nonparametric maximum likelihood estimate (NPMLE)

$$\pi^{\mathrm{NPMLE}} = \arg\max_\pi \ \log L(\pi) \tag{2}$$

is given by a discrete distribution with at most M nodes, see [6, Theorems 2-5] or [7, Theorem 21]. In order to avoid this kind of "peaked" behavior (overfitting of the MLE) and to favor "smooth" priors, a penalty term $\Phi(\pi)$ is often subtracted from the marginal log-likelihood. The resulting maximum penalized likelihood estimate (MPLE),

$$\pi^{\mathrm{MPLE}} = \arg\max_{\pi} \ \log L(\pi) - \Phi(\pi),$$

is a trade-off between goodness of fit and smoothness of the prior.

Following the discussion in [5], we will use the negative mutual information $\mathcal{I}[X;Z](\pi)$ as our penalty, which is equivalent to using the negative "Z-entropy" $H_Z(\pi)$ (the entropy in measurement space) in the case of an additive error (1),

$$\begin{aligned} \Phi(\pi) = -\gamma\, H_Z(\pi) &:= \gamma \int \rho_Z(z|\Pi = \pi)\ \log \rho_Z(z|\Pi = \pi)\, \mathrm{d}z \\ &= -\gamma\, \mathcal{I}[X;Z](\pi) + \text{const.}, \end{aligned} \tag{3}$$

where the weight γ equilibrates the trade-off between smoothness of the prior and goodness of fit. The resulting prior estimate π^{MPLE} has a beautiful connection to reference priors – the two coincide in the case of no measurements. As an alternative approach, we will apply the doubly-smoothed MLE (DS-MLE) introduced by Seo and Lindsay in [12, 11] to our problem, which can be viewed as a regularization in the measurement space before the application of NPMLE.

This paper has a theorical counterpart [5], where the theory is explained in detail. Here, we will concentrate on applications.

We will introduce the notation in Section 2, set out the theory and derive the numerical scheme in Section 3. The resulting algorithm for two different scenarios and their implementation are discussed in Section 4. Corresponding numerical results for a low-dimensional toy example are presented in Section 5, together with an example that shows what can go wrong. Finally, in Section 6 the algorithm is applied to a high-dimensional parameter estimation problem in an ODE model of the human menstrual cycle, which represents a typical example from systems medicine modelling.

2 Setup and Notation

Throughout this manuscript, we will use the following notation:

(i) The probability density function of a random variable Y will be denoted by ρ_Y, while $\rho_Y(\bullet \mid A)$ will stand for its conditional density given an event A (typically $A = \{X = x\}$ or $A = \{\Pi = \pi\}$). Other probabilities will be denoted by π for "priors" and p for "posteriors", in particular,

$$p_\pi^z(x) := \frac{\rho_Z(z \mid X = x)\, \pi(x)}{\int \rho_Z(z \mid X = \tilde{x})\, \pi(\tilde{x})\, \mathrm{d}\tilde{x}}$$

denotes the posterior density of X given the measurement $Z = z$ and the prior π.

(ii) In the case of an additive error (1), which we will always assume, the likelihood model $\{\rho_Z(\bullet \mid X = x) \mid x \in \mathbb{R}^d\}$ is given by

$$\rho_Z(z \mid X = x) := \rho_E\left(z - \phi(x)\right),$$

where $\phi\colon \mathbb{R}^d \to \mathbb{R}^n$ is the (known) underlying model and ρ_E is the (known) probability density of the additive error term E.

(iii) Since we assume to have several patients with (independent and ρ_X-distributed, but unknown) parametrizations X_m and (known) measurements Z_m,

$$X_m \overset{\text{i.i.d.}}{\sim} \pi_{\text{true}} = \rho_X, \qquad Z_m \overset{\text{indep.}}{\sim} \rho_Z(\bullet \mid X = x_m), \qquad m = 1, \dots, M,$$

our likelihood model $\{\rho_Z(\bullet \mid \Pi = \pi) \mid \pi \in \mathcal{M}_1(\mathbb{R}^d)\}$ for the hyperparameter $\Pi = \pi$, where $\mathcal{M}_1(\mathbb{R}^d)$ denotes the set of all probability densities on $\mathbb{R}^d$, is given by

$$\rho_Z(\bullet \mid \Pi = \pi) = \int \rho_Z(z \mid X = x)\, \pi(x)\, \mathrm{d}x, \tag{4}$$

which is the "would-be probability density" of Z, if π was the true prior (so, the true density of Z is given by $\rho_Z = \rho_Z(\bullet \mid \Pi = \pi_{\text{true}})$). The (marginal) likelihood $L(\pi)$ is then given by

$$L(\pi) = \prod_{m=1}^{M} \rho_Z(z_m \mid \Pi = \pi).$$

We will call the likelihood model identifiable (see e.g. [14, Section 5.5]), if

$$\rho_Z(\bullet \mid \Pi = \pi) = \rho_Z(\bullet \mid \Pi = \pi_{\text{true}}) \iff \pi = \pi_{\text{true}}. \tag{5}$$

We will slightly abuse notation by using (4) in the more general case $\pi \in L^1(\mathbb{R}^d)$ and by utilizing the same notation for probability densities and the corresponding probability distributions.

3 Theory

In order to reproduce the probability density ρ_X of the parameters from the measurements $z_1, \dots, z_M$, we will recursively apply an approximation to the fixed point iteration

$$\pi_{n+1}(x) = (\Psi \pi_n)(x), \quad \text{where} \tag{6}$$

$$(\Psi\pi)(x) := \int p_\pi^z(x)\, \rho_Z(z)\, \mathrm{d}z = \pi(x) \int \frac{\rho_Z(z \,|\, X = x)}{\rho_Z(z \mid \Pi = \pi)}\, \rho_Z(z)\, \mathrm{d}z. \tag{7}$$

This iteration is motivated by the observation that the true prior density $\pi_{\text{true}} = \rho_X$ of X is a fixed point of Ψ,

$$(\Psi\pi_{\text{true}})(x) = \pi_{\text{true}}(x) \int \frac{\rho_Z(z \,|\, X = x)}{\rho_Z(z)}\, \rho_Z(z)\, \mathrm{d}z = \pi_{\text{true}}(x), \tag{8}$$

and can be seen as an analogon to the EM algorithm applied to the "infinite data" log-likelihood,

$$\mathcal{L}_{\text{cc}}(\pi) := -H^{\text{cross}}\big(\rho_Z,\, \rho_Z(z \mid \Pi = \pi)\big) = \int \rho_Z\, \log \rho_Z(z \mid \Pi = \pi)\, \mathrm{d}z, \tag{9}$$

where H^{cross} denotes the cross entropy. It can be shown that $\mathcal{L}_{\text{cc}}$ is concave in π and that its value is increased in each iteration step (see [5]):

$$\mathcal{L}_{\text{cc}}(\Psi\pi) \geqslant \mathcal{L}_{\text{cc}}(\pi) \quad \text{for all } \pi \in \mathcal{M}_1(\mathbb{R}^d).$$

More precisely, the following statement holds:

Proposition 1. *Let $\pi \in \mathcal{M}_1(\mathbb{R}^d)$ be a globally supported probability density function. Then the following two statements are equivalent:*

(i) $\rho_Z(\bullet \mid \Pi = \pi) = \rho_Z$,

(ii) $\Psi\pi = \pi$,

(iii) π *maximizes* $\mathcal{L}_{\text{cc}}(\pi)$.

Proof. The proof for (i)⇒(ii) goes analogously to (8). For (ii)⇒(i) we will use the abbreviation

$$\rho_{Z,f} := \rho_Z(\bullet \mid \Pi = f), \qquad f \in L^1(\mathbb{R}^d).$$

We define the subspace

$$\mathcal{E} = \big\{\rho_{Z,f} \mid f \in L^1(\mathbb{R}^d)\big\} \subseteq L^1(\mathbb{R}^n)$$

with weighted L^2 inner product

$$\big\langle \rho_{Z,f_1}, \rho_{Z,f_2} \big\rangle_\pi := \int_{\mathbb{R}^n} \frac{\rho_{Z,f_1}(z)\,\rho_{Z,f_2}(z)}{\rho_{Z,\pi}(z)}\,\mathrm{d}z.$$

We can formulate the following chain of implications:

$$\begin{aligned}
\text{(ii)} &\implies \forall x: \quad \int \frac{\rho_Z(z)}{\rho_{Z,\pi}(z)}\,\rho_Z(z \mid X = x)\,\mathrm{d}z = 1 \\
&\implies \forall x: \quad \int \left(1 - \frac{\rho_Z(z)}{\rho_{Z,\pi}(z)}\right)\rho_Z(z \mid X = x)\,\mathrm{d}z = 0 \\
&\implies \int (\pi - \rho_X)(x) \int \frac{\rho_{Z,\pi}(z) - \rho_Z(z)}{\rho_{Z,\pi}(z)}\,\rho_Z(z \mid X = x)\,\mathrm{d}z\,\mathrm{d}x = 0 \\
&\implies \int \frac{\rho_{Z,\pi}(z) - \rho_Z(z)}{\rho_{Z,\pi}(z)}\,\rho_{Z,(\pi-\rho_X)}(z)\,\mathrm{d}z = 0 \\
&\implies \big\langle \rho_{Z,\pi} - \rho_Z, \rho_{Z,\pi} - \rho_Z \big\rangle_\pi = 0,
\end{aligned}$$

which implies (i) by the positive definiteness of the inner product.

The equivalence of (i) and (iii) is given by the fact that the cross entropy of two densities is minimal if and only if the two densities agree. □

So, in the identifiable case (5), π_{true} is the only fixed point of the iteration (6) and one can prove convergence.

As discussed in [5], Proposition 1 suggests yet another estimation approach for the density $\pi_{\text{true}} = \rho_X$: Compute an approximation ρ_Z^{appr} of the density ρ_Z using the measurements $Z_1, \ldots, Z_M$ and then minimize the cross entropy between ρ_Z^{appr} and $\rho_Z(z \mid \Pi = \pi)$:

$$\pi_* = \arg\max_\pi \mathcal{L}_{\text{cc}}^{\text{appr}}(\pi), \qquad \mathcal{L}_{\text{cc}}^{\text{appr}}(\pi) := -H^{\text{cross}}\big(\rho_Z^{\text{appr}},\, \rho_Z(z \mid \Pi = \pi)\big).$$

If the approximation of ρ_Z is performed by kernel density estimation, the resulting method is the so-called doubly-smoothed maximum likelihood estimation (DS-MLE) introduced by Seo and Lindsay in [12, 11]. Note that, in this case, due to the additional smoothing by the kernel, the likelihood model has to be smoothed as well in order to get consistent results. We will denote the resulting density estimate π_* by $\pi_{\text{DS-MLE}}$.

3.1 Numerical realization

The numerical approximation of the fixed point iteration (6)–7 will be realized by two discretization steps:

(i) The first discretization in $\mathbb{R}^n$ is due to the fact that we have only finitely many ρ_Z-distributed measurements $Z_1, \dots, Z_M$ instead of the density ρ_Z appearing in (7). We will use the following Monte-Carlo approximation:

$$\rho_Z \approx \frac{1}{M} \sum_{m=1}^{M} \delta_{z_m}. \tag{Z-MC}$$

(ii) In order to compute the high-dimensional integrals $\rho_Z(z_m \mid \Pi = \pi)$, a second discretization in $\mathbb{R}^d$ is necessary, which will be realized by another Monte-Carlo approximation for prior densities π (and the hyperparameter $\Pi = \pi$ will be replaced by $W = w$):

$$\pi \approx \sum_{k=1}^{K} w_k \delta_{x_k}, \quad x_k \in \mathbb{R}^d,\ w \in \mathcal{W} := \Big\{ w \in \mathbb{R}^K \mid w_k \geqslant 0 \,\forall k,\ \sum_{k=1}^{K} w_k = 1 \Big\}. \tag{X-MC}$$

The application of these discretizations to the infinite data log-likelihood $\mathcal{L}_{\text{cc}}$ defined by (9) and to the fixed point iteration $\Psi_{\text{cc}} := \Psi$ defined by (6) – (7) results in the commutative diagram displayed in Figure 1, which is discussed in detail in [5].

Here, the indices c and d denote whether the parameter space (first index) and the measurement space (second index) are considered continuous or discretized in the above sense. The corresponding log-likelihoods and fixed point iterations are given by

$$\mathcal{L}_{\text{cc}}(\pi) = \int_{\mathcal{Z}} \rho_Z(z) \log \rho_Z(z|\Pi = \pi)\, \mathrm{d}z, \qquad \Psi_{\text{cc}}\pi\,(x) = \pi(x) \int_{\mathcal{Z}} \rho_Z(z)\, \frac{\rho_Z(z|X = x)}{\rho_Z(z|\Pi = \pi)}\, \mathrm{d}z,$$

$$\mathcal{L}_{\text{cd}}(\pi) = \frac{1}{M} \sum_{m=1}^{M} \log \rho_Z(z_m|\Pi = \pi), \qquad \Psi_{\text{cd}}\pi\,(x) = \frac{\pi(x)}{M} \sum_{m=1}^{M} \frac{\rho_Z(z_m|X = x)}{\rho_Z(z_m|\Pi = \pi)},$$

$$\mathcal{L}_{\text{dc}}(w) = \int_{\mathcal{Z}} \rho_Z(z) \log \rho_Z(z|W = w)\, \mathrm{d}z, \qquad [\Psi_{\text{dc}}w]_k = w_k \int_{\mathcal{Z}} \rho_Z(z)\, \frac{\rho_Z(z|X = x_k)}{\rho_Z(z|W = w)}\, \mathrm{d}z,$$

$$\mathcal{L}_{\text{dd}}(w) = \frac{1}{M} \sum_{m=1}^{M} \log \rho_Z(z_m|W = w), \qquad [\Psi_{\text{dd}}w]_k = \frac{w_k}{M} \sum_{m=1}^{M} \frac{\rho_Z(z_m|X = x_k)}{\rho_Z(z_m|W = w)}.$$

Remark 2. *Note that $\mathcal{L}_{\text{cd}}$ and $\mathcal{L}_{\text{dd}}$ are the log-likelihood functions of the hyperparameter $\Pi = \pi$, $W = w$ respectively (devided by M). The fixed point iterations Ψ_{cd} and Ψ_{dd} resulting from the application of the EM algorithm have*

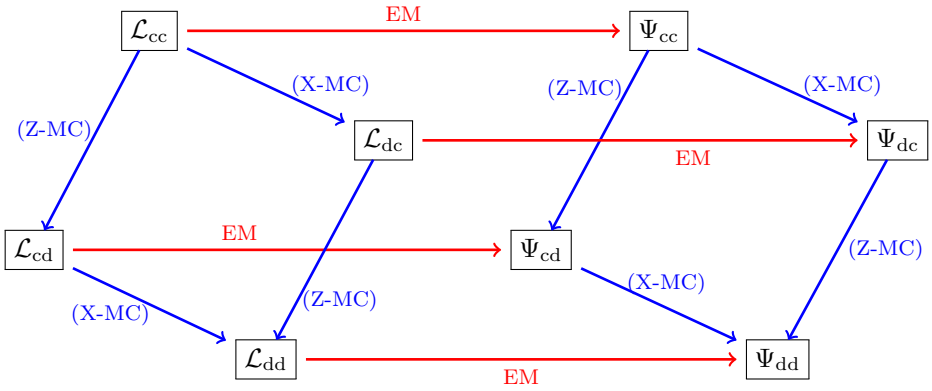


Figure 1: The relations between the log-likelihoods $\mathcal{L}$ and the fixed point iterations Ψ resulting from the application of the EM algorithm summarized by a commutative diagram.

been studied before, see e.g. [13, 6]. They are often referred to as the "self-consistency algorithm" since they fulfill the self-consistency principle introduced by Efron [2]. Therefore, the stepwise increase of $\mathcal{L}_{\mathrm{cd}}$ and $\mathcal{L}_{\mathrm{dd}}$ when applying the iterations Ψ_{cd} and Ψ_{dd}, respectively, follows from the theory on the EM algorithm [1]. The proof of an analogous statement for Ψ_{cc} and Ψ_{dc} is given in [5].

In order to guarantee that the nodes $x_1, \dots, x_K$ from the discretization (X-MC) lie in an area of high probability of the iterates π_n, they will be chosen π_1-distributed, since our experiments have shown that π_1 already gives a considerable improvement over π_0 in approximating π_{true}. This will be realized by producing a $p_{\pi_0}^{z_m}$-distributed sampling $(x_1^{(m)}, \dots, x_L^{(m)})$ for each $m = 1, \dots, M$ by the Metropolis-Hastings algorithm and merging these to get a π_1-distributed sampling

$$\mathcal{X} = \{x_1, \dots, x_K\} = \bigcup_{m=1}^{M} \left\{x_1^{(m)}, \dots, x_L^{(m)}\right\}$$

with weights $w_{1,k} = 1/K$, $k = 1, \dots, K$.

The numerical realization of DS-MLE goes analogously to NPMLE but with an augmented set of measurements $\zeta_1, \dots, \zeta_{\tilde{M}}$ instead of $Z_1, \dots, Z_M$, which are samples from the kernel density estimate ρ_Z^{appr}, see the description in [5] or [11].

Unfortunately, even though $\mathcal{L}_{\mathrm{cd}}$ is an approximation of $\mathcal{L}_{\mathrm{cc}}$ and $\mathcal{L}_{\mathrm{cc}}$ is maximized by the true prior π_{true} of $\mathcal{L}_{\mathrm{cc}}$, $\mathcal{L}_{\mathrm{cd}}$ can be shown to be maximized by a discrete distribution with at most M nodes, see [6, Theorems 2-5] or [7, Theorem 21]. This originates from the typical "overconfidence" (or "overfitting") of maximum likelihood estimators and results in a poor approximation of the true prior:

$$\pi_{\mathrm{cd},\infty} = \arg\max_{\pi} \mathcal{L}_{\mathrm{cd}}(\pi) \not\approx \arg\max_{\pi} \mathcal{L}_{\mathrm{cc}}(\pi) = \pi_{\mathrm{true}}.$$

As discussed in the introduction, one way to deal with this issue is to subtract a penalty term from $\mathcal{L}_{\mathrm{cd}}$ before maximizing it, which we will choose to be the

negative entropy $-H_Z(\pi)$ in the measurement space defined by (3):

$$\pi^{\mathrm{MPLE}} = \arg\max_{\pi} \; M\mathcal{L}_{\mathrm{cd}}(\pi) + \gamma H_Z(\pi). \tag{10}$$

The numerical implementaion of (10) will be realized by a gradient ascent of the discretized functional

$$\mathcal{A} := M\mathcal{L}_{\mathrm{dd}}(w) + \gamma H_Z(w),$$

where

$$H_Z(w) := -\int \rho_Z(z \mid W = w) \, \log \rho_Z(z \mid W = w) \, \mathrm{d}z.$$

In order for the iterates w_n to stay in the simplex $\mathcal{W}$ as defined in (X-MC), an additional bestapproximation is necessary in each step.

The integral appearing in the gradient of the functional $\mathcal{A}$,

$$\nabla\mathcal{A} = \left(\sum_{m=1}^{M} \frac{\rho_Z(z_m \mid X = x_k)}{\rho_Z(z_m \mid W = w)} - \gamma \underbrace{\int \rho_Z(z \mid X = x_k) \, \log(\rho_Z(z \mid W = w)) \, \mathrm{d}z}_{=: I} - \gamma \right)_{k=1}^{K},$$

is computed using an importance sampling,

$$\begin{aligned} I &= \int \frac{\rho_Z(z \mid X = x_k)}{\rho_Z(z \mid W = w)} \, \log(\rho_Z(z \mid W = w)) \, \rho_Z(z \mid W = w) \, \mathrm{d}z \\ &\approx \sum_{j=1}^{J} \frac{\rho_Z(z_j \mid X = x_k)}{\rho_Z(z \mid W = w)} \, \log(\rho_Z(z_j \mid W = w)), \end{aligned}$$

where the points z_j are chosen $\rho_Z(\bullet \mid W = w)$-distributed.

3.2 Non-identifiable case

If the identifiability assumption (5) is not fulfilled, we cannot expect the fixed point iteration (6)–(7) to converge to the true prior $\pi_{\mathrm{true}} = \rho_X$. The best we can hope for is to get close to a prior π such that $\rho_Z(\bullet \mid \Pi = \pi) = \rho_Z$. Therefore, one way to enforce convergence is to restrict ourselves to equivalence classes of densities with respect to the equivalence relation

$$\pi \sim \pi' \iff \left\| \rho_Z(\bullet \mid \Pi = \pi) - \rho_Z(\bullet \mid \Pi = \pi') \right\|_{L^1(\mathbb{R}^n)} = 0.$$

Note that the set of equivalence classes $L^1(\mathbb{R}^d)/\!\sim$ is, in fact, the quotient space $L^1(\mathbb{R}^d)/\ker(\psi)$ emerging from the linear map

$$\psi \colon L^1(\mathbb{R}^d) \to L^1(\mathbb{R}^n), \qquad \pi \mapsto \rho_Z(\bullet \mid \Pi = \pi) = \int \rho_Z(\bullet \mid X = x) \, \pi(x) \, \mathrm{d}x.$$

Therefore, $L^1(\mathbb{R}^d)/\!\sim$ inherits the L^1-norm and L^1-distance via

$$\|[\pi]\|_{L^1} = \inf_{\pi' \in [\pi]} \|\pi'\|_{L^1}, \qquad \|[\pi_1] - [\pi_2]\|_{L^1} = \inf_{\substack{\pi_1' \in [\pi_1] \\ \pi_2' \in [\pi_2]}} \|\pi_1' - \pi_2'\|_{L^1}$$

and we can choose from the following two definitions for the convergence of π_n to π_∞:

$$\pi_n \xrightarrow[Z]{n\to\infty} \pi_\infty \;:\Longleftrightarrow\; \|[\pi_n] - [\pi_\infty]\|_{L^1} \xrightarrow{n\to\infty} 0$$

or

$$\pi_n \xrightarrow[Z]{n\to\infty} \pi_\infty \;:\Longleftrightarrow\; \|\rho_Z(\bullet \mid \Pi = \pi_n) - \rho_Z(\bullet \mid \Pi = \pi_\infty)\|_{L^1} \xrightarrow{n\to\infty} 0.$$

Both definitions are meaningful but the second definition yields a weaker form of convergence as stated by the following proposition:

Proposition 3. *Let $\pi \in L^1(\mathbb{R}^d)$. Then, in the above notation,*

$$\|\rho_Z(\bullet \mid \Pi = \pi)\|_{L^1} \leqslant \|[\pi]\|_{L^1} \ .$$

Proof. For each $\pi' \in [\pi]$ we have

$$\begin{aligned} \|\rho_Z(\bullet \mid \Pi = \pi)\|_{L^1} \;&=\; \|\rho_Z(\bullet \mid \Pi = \pi')\|_{L^1} \;=\; \int \left| \int \rho_Z(z \mid X = x)\, \pi'(x)\, \mathrm{d}x \right| \mathrm{d}z \\ &\leqslant \int |\pi'(x)| \underbrace{\int \rho_Z(z \mid X = x)\, \mathrm{d}z}_{=1}\, \mathrm{d}x \;=\; \|\pi'\|_{L^1} . \end{aligned}$$

□

4 Resulting Algorithm

Given the data $\mathcal{Z} = \{z_1, \ldots, z_M\} \subseteq \mathbb{R}^n$ of M patients, we will discuss two scenarios:

(A) No diagnoses have been made.

(B) The patients have been diagnosed with diseases/sicknesses $s_1, \ldots, s_L$ (for simplicity, we will assume that each patient has exactly one disease), resulting in a partition of the data set, where $Z^{(l)}$ denotes the data of the patients with disease s_l:

$$\mathcal{Z} = \{z_1, \ldots, z_M\} = \bigsqcup_{l=1}^{L} \mathcal{Z}^{(l)}, \qquad \mathcal{Z}^{(l)} = \{z_1^{(l)}, \ldots, z_{M_l}^{(l)}\} \qquad (M = \sum_{l=1}^{L} M_l).$$

Approach for scenario (A):

- Starting with a non-informative prior π_0, we construct informative priors π_{NPMLE}, $\pi_{\text{DS-MLE}}$ and π_{MPLE} by the fixed point iterations discussed in Section 3.1. All four priors, $\pi = \pi_0$, $\pi_{\text{DS-MLE}}$, π_{NPMLE}, π_{MPLE}, will be given by the same sampling $\mathcal{X} = \{x_1, \ldots, x_K\}$ but with weights $w = w_0$, w_{NPMLE}, $w_{\text{DS-MLE}}$, w_{MPLE} and their performance will be compared in the next steps.

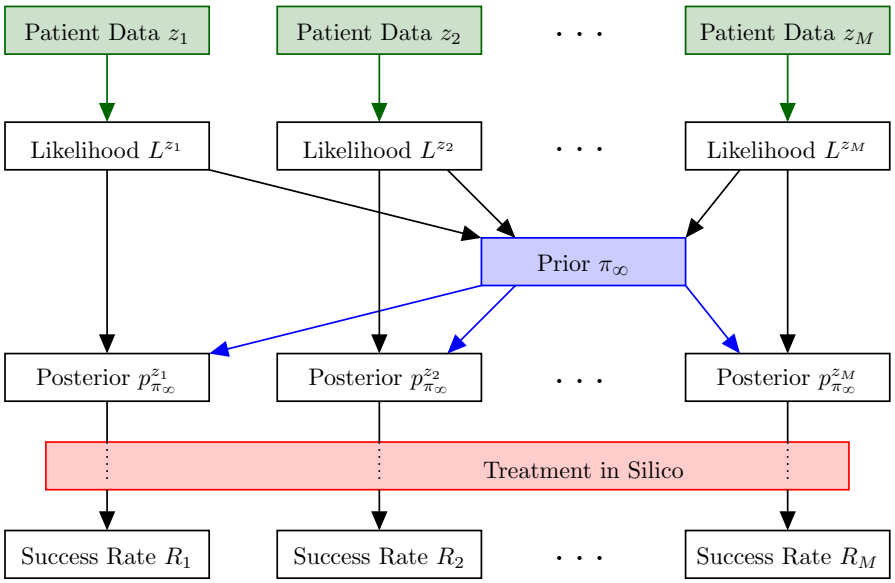


Figure 2: Algorithmic scheme for the computation of patient-specific parametrizations and predictions of individual treatment success rates.

- Given a patient with data $Z = z_*$, we compute the individual posteriors $p^{z_*}_\pi$ with respect to these priors, which will be given by new (individual) weights,

$$v_k^* = \frac{w_k \, \rho_Z(z^* \mid X = x_k)}{\sum_{j=1}^K w_j \, \rho_Z(z^* \mid X = x_j)} \, .$$

- The success rate R_* can now be approximated via

$$R_* = \int r(x) \, p^{z_*}_\pi(x) \, \mathrm{d}x \approx \frac{1}{K} \sum_{k=1}^K v_k^* \, r(x_k),$$

where

$$r(x) = \begin{cases} 1 & \text{if the treatment, given the parameters } x, \text{ is successful,} \\ 0 & \text{otherwise.} \end{cases}$$

Approach for scenario (B):
If the patients are diagnosed with diseases $s_1, \ldots, s_L$ and the number of patients M_l is large for each disease s_l, then this extra information can be used by applying the procedure described in (A) to each subset $\mathcal{Z}^{(l)}$ separately in order to obtain more precise results.

5 Toy Example

We will start with an easy to grasp low-dimensional non-identifiable mechanical example for scenario (A), where the patients will be represented by springs with

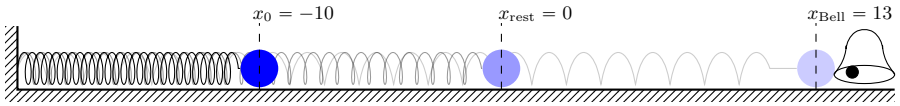


Figure 3: Experimental arrangement for the toy example described above.

different stiffness values as possible parameter values. Different stiffness values result in different system responses when a certain force is applied to the springs, which represents a treatment of the patients. The example demonstrates how our algorithm can be applied to predict success rates of such a treatment.

We buy two boxes (representing two diseases) of springs, the first containing 150 springs with stiffness $K_1 = 15\,N/m$, the second containing 150 springs with stiffness $K_2 = 30\,N/m$. The springs are of a low quality and their actual stiffness varies from the nominal value with a standard deviation of 15% (we assume a normal distribution for each box).

Once we arrive at home, we realize that the boxes are not labeled and that we already forgot the values K_1 and K_2 as well as the standard deviations. As described above, there are two possible scenarios and we will only treat scenario (A), since (B) goes analogously:

(A) We mix up the springs by putting all of them into one big box (no diagnosis for each spring).

(B) We keep them in the two separate boxes (the springs are diagnosed with diseases s_1 or s_2, depending on the box they come from).

In order to determine the stiffness of a single spring, we perform the following experiment (harmonic oscillator, see Figure 3):

- We fix one end of the spring and put a mass $m = 700\text{g}$ to the other end.
- After compressing it by 10cm, we let it swing. Applying Hooke's law this results in the following ODE

$$x''(t) = -\frac{K}{m}\,x(t), \qquad x(0) = -10\text{cm}. \tag{11}$$

- We measure its amplitude at time $t^* = 1\text{s}$. Therefore, the model $\phi : \mathbb{R} \to \mathbb{R}$ and the measurement $Z \in \mathbb{R}$ are given by

$$\phi(K) = x(t^*), \qquad Z = \phi(K) + E,$$

  where the measurement error $E \sim \rho_E = \mathcal{N}(0, \sigma_E^2)$ is assumed to be standard normal distributed with mean 0 and a standard deviation of $\sigma_E = 1\text{cm}$.

We implemented the fixed point iterations for NPMLE, DS-MLE and MPLE discussed in Section 3 (for the latter $\gamma = 49$ appeared adequate), starting with a "non-informative" prior π_0, which we chose as the uniform distribution on $[1, 50]$ (in kg/s^2), as well as the computation of the corresponding posteriors for one of the springs. The results are shown in Figure 4. Since the model is highly

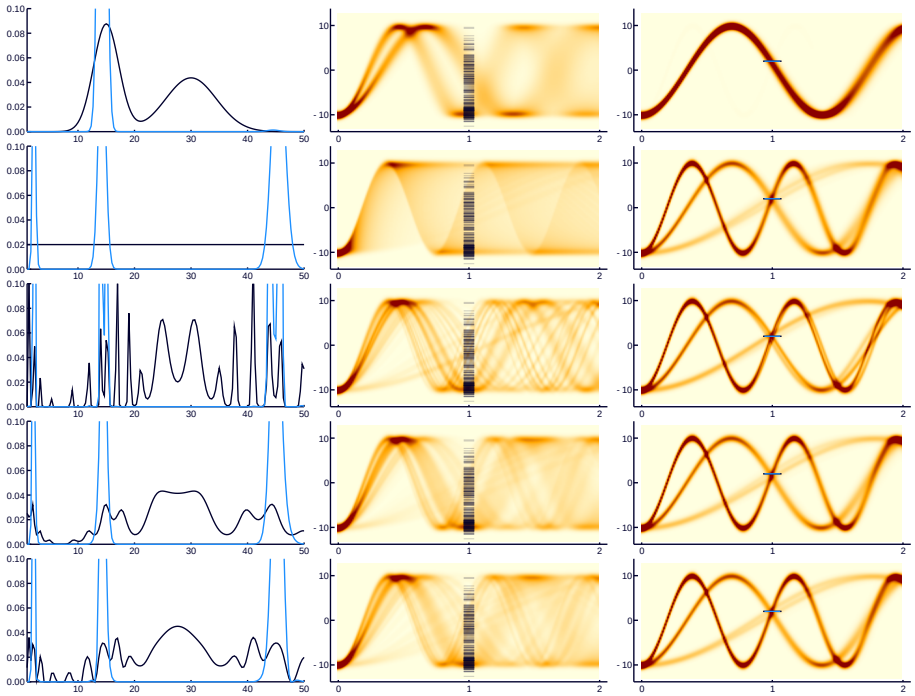


Figure 4: From top to bottom: priors π_{true}, π_0, π_{NPMLE}, $\pi_{\text{DS-MLE}}$, π_{MPLE}.
From left to right: the prior (black) and the corresponding posterior for one specific spring (blue), distribution of trajectories sampled from that prior plotted over time and all measurements, measurement of that spring and corresponding posterior distribution of trajectories.
500 iterations have been performed. Note that for a higher number of iterations π_{NPMLE} and $\pi_{\text{DS-MLE}}$ (due to the discretization method discussed in Section 3.1) will not stop peaking.

non-identifiable (for each point $x^* \in [-10, 10]$ there are several stiffness values K for which the trajectory of (11) goes through (t^*, x^*), i.e. $\phi(K) = x^*$), the prior estimates are rather non-informative. Therefore, the effect on the posterior is barely visible, while the smoothing of the prior is evident.

The "treatment procedure" will be modeled by hitting the mass in positive x-direction with several pulses at certain times given by the force $F(t)$ plotted in Figure 5, which results in the following perturbed ODE:

$$x''(t) = -\frac{K}{m}x(t) + F(t), \qquad x(0) = -10\text{cm}.$$

The treatment will be considered successful, if the mass hits the bell located at $x_{\text{Bell}} = 13$cm within ten seconds.

For each $\Xi \in \{\text{true}, 0, \text{NPMLE}, \text{DS-MLE}, \text{MPLE}\}$, we computed the success rates R_m^Ξ, $m = 1, \dots, M$ with respect to π_Ξ and their (empirical) standard deviations from the true success rates:

$$\sigma_R^\Xi = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M} |R_m^\Xi - R_m^{\text{true}}|^2}.$$

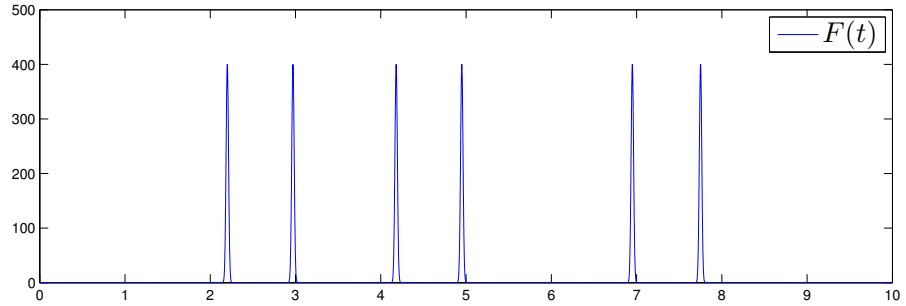


Figure 5: The force F used for modelling the treatment procedure plotted over time t.

The results are given in Table 1.

Ξ	0	NPMLE	DS-MLE	MPLE
σ_R^{Ξ}	0.126	0.135	0.106	0.097

Table 1: The standard deviations of the predicted success rates for different prior distributions π_{Ξ}.

5.1 Example of an unsmooth reference prior

By the following example, we want to make the reader aware of the fact that regularizing by means of the mutual information $\mathcal{I}[X;Z]$ not always results in a smooth prior. Even if we ignore the likelihood in (2) and maximize $\mathcal{I}[X;Z]$ alone, the resulting prior, which is the reference prior

$$\pi_{\text{ref}} = \arg\max_{\pi} \ \mathcal{I}[X;Z](\pi),$$

can be very irregular.

If, in the above example, we perform two measurements at times $t_1^* = 1$s and $t_2^* = 1.7$s instead of only one, the model $\phi : \mathbb{R} \to \mathbb{R}^2$ and the measurement $Z \in \mathbb{R}^2$ are given by

$$\phi(K) = (x(t_1), x(t_2))^\intercal, \qquad Z = \phi(K) + E,$$

with error

$$E = (E_1, E_2)^\intercal \ \sim \ \rho_E = \mathcal{N}(0, \sigma_E^2) \otimes \mathcal{N}(0, \sigma_E^2).$$

The resulting reference prior shown in Figure 6 is very unsmooth. Let us analyze the reason for this!

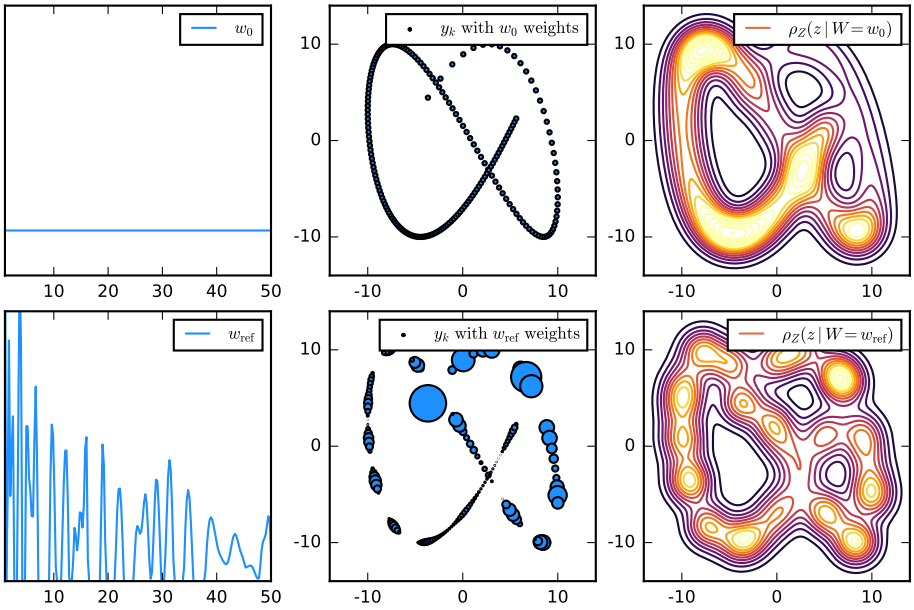


Figure 6: From left to right: Two different priors (uniform and reference prior), resulting weights in the linear combination (12), resulting densities $\rho_Z(z \mid W = w)$ in measurement space. One can clearly see that the reference prior results in a higher entropy $H_Z(\pi)$ and thereby in a higher mutual information $\mathcal{I}[X;Z](\pi)$, which it was optimized for. However, it is very irregular.

If we discretize the parameter space $\mathcal{X} = [1, 50]$ by 200 equidistant points x_k and define $y_k := \phi(x_k)$, the density $\rho_Z(z \mid W = w)$ consists of a linear combination of shifted versions of ρ_E,

$$\rho_Z(z \mid W = w) = \sum_{k=1}^{K} w_k \, \rho_E(z - y_k), \tag{12}$$

in our case a Gaussian mixture with centers y_k. The aim is to choose $w = (w_k)_k$ in such a way that it maximizes the entropy $H_Z(w)$, i.e. to make $\rho_Z(z \mid W = w)$ rather "spread" and "flat".

Therefore, we get huge weights w_k for those k, where $\rho_E(z - y_k)$ covers a large region that cannot be covered by any other $\rho_E(z - y_j)$. Low weights w_k are chosen small if the corresponding y_k lie in a regions that are already covered by many $\rho_E(z - y_j)$ (otherwise the density $\rho_Z(z \mid W = w)$ would get too high around y_k). This way, we arrive at the peaked prior π_{ref} shown in Figure 6.

The unsmoothness of the reference prior π_{ref} (and of our MPLE) is not necessarily a downside, but it is worthwile to be aware of such situations when using these methods.

6 Parameter estimation in a large ODE model

As model system, we consider a model for the human menstrual cycle, named GynCycle [10]. This model is given by a system of 33 ordinary differential equations (ODEs) and 114 parameters. It has been calibrated previously with time-series data of blood concentrations for four hormones from 12 patients during the unperturbed cycle and during treatment (dose-response experiments). Using deterministic, local optimization (an error-oriented Gauss-Newton method), only 63 out of 114 parameters could be identified from the given data. The remaining parameters kept their values from previous versions of the model. In the following, we will denote these parameter values as nominal values. The model is currently used to make patient-specific predictions about the outcome of treatment strategies in reproductive medicine[1]. Hence, quantification of uncertainty in these predictions is of utmost importance.

We got access to additional measurement values of 36 woman for the four hormones LH, FSH, E2 and P4 during normal cycles[2]. These data are sparse or incomplete in the sense that measurements were not taken on all cycle days, resulting in about 15 measurement time points per patient and hormone. Our approach, however, is flexible enough to handle such a data situation.

Based on these data, our aim is to estimate the prior distribution for 82 out of the 114 parameters Θ (Hill exponents have been fixed), denoted by π_0^Θ, as well as for the initial values Y_0 of the 33 model species, $\pi_0^{Y_0}$, resulting in a total of 115 dimensions: $X = (Y_0, \Theta) \in \mathbb{R}^{115}$.

As an initial guess for the prior π_0^Θ of model parameters, we have chosen uniform distributions on the intervals between zero and five times the nominal parameter values as non-informative approach. For $\pi_0^{Y_0}$, we chose a mixture distribution of independent normals around daily values of a reference cycle computed with the nominal parameters, i.e.

$$\pi_0^{Y_0} = \frac{1}{31} \sum_{t=0}^{30} G\left[y_{\text{ref}}(t), \Sigma\right],$$

where $G[m, C]$ denotes the Gaussian density with mean m and covariance matrix C, $y_{\text{ref}} : \mathbb{R} \to \mathbb{R}^{33}$ is the reference solution over one menstrual cycle and Σ is a diagonal matrix consisting of the squared standard deviations of each species, respectively,

$$\Sigma = \text{diag}(\sigma_1^1, \ldots, \sigma_{33}^2), \qquad \sigma_j^2 = \frac{1}{30} \sum_{t=0}^{30} |y_{\text{ref},j}(t) - \overline{y}_{\text{ref},j}|^2.$$

The total prior $\pi_0 \colon \mathbb{R}^{115} \to \mathbb{R}$ is build up from $\pi_0^{Y_0}$ and π_0^Θ under the assumption that Y_0 and Θ are independent,

$$\pi_0(y_0, \theta) := \pi_0^{Y_0}(y_0)\, \pi_0^\Theta(\theta).$$

The likelihood for specific measurements $z \in \mathbb{R}^{4\times 31}$ is chosen normally distributed with a (relative) standard deviation of $\sigma = 10\%$,

$$\rho_Z(z \,|\, X = x) \propto \exp\left(-\frac{d(\phi(x), z)^2}{2\sigma^2}\right),$$

[1] EU project PAEON-Model Driven Computation of Treatments for Infertility Related Endocrinological Diseases, project number 600773.

[2] Courtesy of Dorothea Wunder, University Hospital of Lausanne.

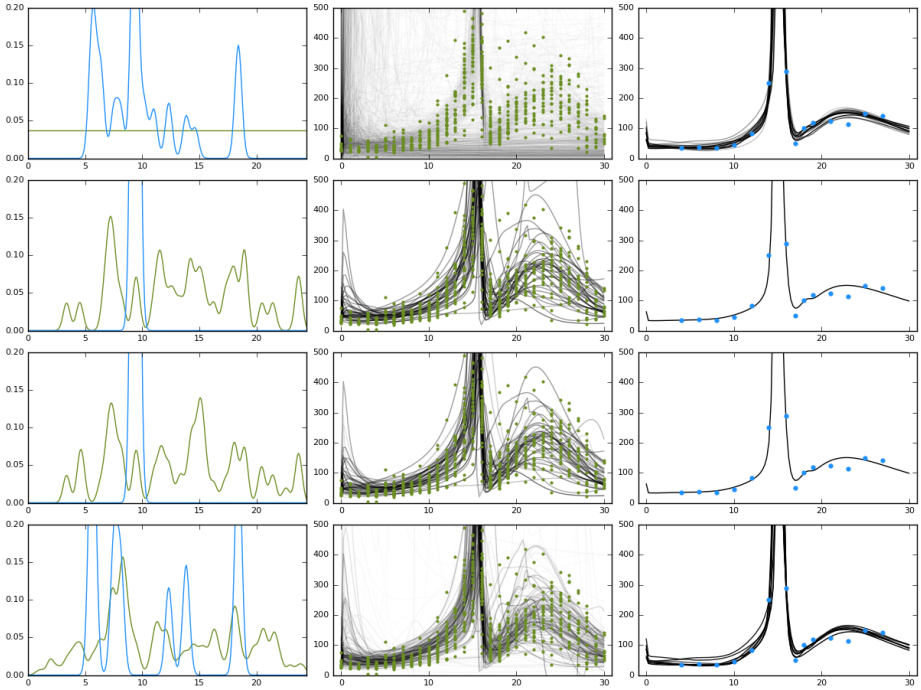


Figure 7: From top to bottom: prior given by π_0, π_{NPMLE}, $\pi_{\text{DS-MLE}}$, π_{MPLE}. From left to right: marginal distribution of the prior (black) for one parameter (the transition rate from the late primary follicle PrA2 to the early secondary follicle SeF1) and the corresponding posterior for one specific patient (blue), trajectories sampled from that prior and all measurements, measurements of that patient and trajectories sampled from the posterior.
300 iterations have been performed. Note that for a higher number of iterations π_{NPMLE} and $\pi_{\text{DS-MLE}}$ (due to the discretization method discussed in Section 3.1) will not stop peaking.

where

$$d(u,v)^2 = \sum_{j=1}^{4} \sum_{t=0}^{30} \left| \frac{u_j(t) - v_j(t)}{c_j} \right|^2$$

is the relative distance between simulated and measured data, c_j are suitably chosen constants of the magnitude of the measurements $(v_j(t))_{t=0,\dots,30}$,

$$\phi(x) = \phi(y_0, \theta) = (\text{proj}_4(y(t)))_{t=0,\dots,30}\,,$$

proj_4 denotes the projection onto the four measurable components, and $y(t)$ is the solution of the GynCycle model with initial values y_0 and parameters θ.

Remark 4. *As mentioned above, the measurements for most women were not taken daily, resulting in incomplete data. In this case, ϕ and d have to be chosen separately for each woman, restricting them to measured components. This does not influence the applicability of our algorithm.*

We sampled the posteriors for each patient (in order to get π_1-distributed samples as explained in Section 3.1) using the adaptive mixture Metropolis algo-

rithm by Roberts and Rosenthal [9], which is basically a multivariate Metropolis-Hastings algorithm tuning its Gaussian proposal density for the current sample based on the covariance of the former ones. As the computation is independent for each patient, this problem is well-scalable in the number of patients and we were thus able to compute 10 million samples for each patient. The Raftery-Lewis diagnostic suggests around 7 million samples for convergence and the Gelman and Rubin criterion confirms this in our case with potential scale reduction factors smaller than 1.05.

Once the π_1-distributed samples had been computed, we implemented the fixed point iterations for NPMLE, DS-MLE and MPLE discussed in Section 3 (for the latter $\gamma = 19$ appeared adequate), as well as the corresponding posteriors for one of the springs. The results are shown in Figure 7. From the plots it seems counterintuitive that the posterior $p_{\pi_0}^{z_m}$ stemming from the prior π_0 is so much less informative then those coming from the estimated priors. However, this is less surprising if one keeps in mind that we only consider *marginal* densities and therefore the priors π_{NPMLE}, $\pi_{\text{DS-MLE}}$ and π_{MPLE} can be much more informative (compared to π_0) than they look for each marginalization (e.g., they might be concentrated around a submanifold of $\mathcal{X}$). This demonstrates the importance and strength of empirical Bayes methods.

Our next step will be to compute individual success rates for treatments frequently used in reproductive medicine (modelled by a perturbed ODE) and to compare the results with clinical outcomes.

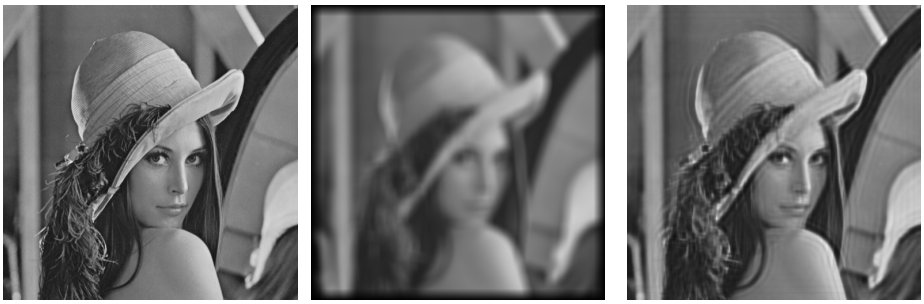
7 Conclusion

We have introduced a method that estimates the prior distribution in the empirical Bayes framework, when measurements for a large number of individuals are available. We discussed the issue of convergence in the identifiable case and also what happens in the non-identifiable case.

A detailed scheme for the numerical realization of the method has been elaborated, see Sections 3.1 and 4. The numerical approximation to the fixed point iteration has to be applied with caution, since its convergence properties differ from the ones of the exact iteration, see the discussion in the introduction and in Section 3.1. A transformation-invariant regularization approach has been applied in order to deal with this situation.

The method has been applied to a toy example in low dimensions to confirm our theoretical results, as well as to a high-dimensional real life problem. As a byproduct, the method can be applied to deconvolve blurred images, as discussed in Appendix A.

As demonstrated and explained in Section 5.1, the resulting π_{MPLE} can be rather unsmooth. As stated by Good and Gaskins in [3], "continuous distributions [...] could have violent small ripples with little effect on the entropy". They advise against the use of entropy as a roughness penalty in the continuous case. From our point of view, "small ripples" in the density are not a criterion for the exclusion of a distribution. Not only can the true distribution itself have ripples, but, even if not so, a distribution with small ripples can be a good approximation to it. However, if the aim is to get a smooth prior, this approach might be inadequate. It is worth mentioning that in this case the reference prior encounters the same problem. A possible solution is given by DS-MLE, which



(a) original image (b) blurred image (c) deconvolved, 50 iter.

(d) deconvolved, 1500 iter. (e) deconvolved, 10000 iter. (f) deconvolved, 50000 iter.

Figure 8: Deconvolution of an artificially blurred image (b) using the fixed point iteration (6)–(7).

was also discussed and implemented for both problems.

We wish to extend the application in Section 6 to data bases of thousands of patients (instead of just 36), which are available by now at the University Hospitals in Zurich and Basel.

A Deconvolution of Blurred Images

The fixed point iteration (6)–(7) can readily be applied for the deconvolution of blurred images with known point spread function (non-blind deconvolution).

For this, the probability densities π_n are viewed as blurred versions of the true density $\pi_{\text{true}} = \rho_X$, whereby ρ_X is smoothed by a convolution kernel G_n:

$$\pi_{n+1}(x) = \Psi\pi_n(x) = \int p^z_{\pi_n}(x)\,\rho_Z(z)\,\mathrm{d}z = \int \rho_X(\tilde{x}) \underbrace{\int p^z_{\pi_n}(x)\,\rho_Z(z \,|\, X = \tilde{x})\,\mathrm{d}z}_{=:G_{n+1}(x,\tilde{x})}\,\mathrm{d}\tilde{x}.$$

With growing number n the iterates become less smoothed, converging to ρ_X. Therefore, the fixed point iteration (6)–(7) results in a deconvolution process of π_0 to the original prior ρ_X.

In fact, if we choose $\phi = \mathrm{Id}$ and the error density ρ_E as the point spread function, we can view $\rho_X : \mathbb{R}^2 \to \mathbb{R}_{\geqslant 0}$ as the original image (without loss of generality, we can assume that it is normalized and given by gray scale values)

and the evidence

$$\rho_Z(z) = \int \rho_X(x)\, \rho_E(z - \phi(x))\, \mathrm{d}x = (\rho_X * \rho_E)(z)$$

as the blurred image. In this setup, our algorithm provides a method for the restoration of the original image from the blurred image, as demonstrated in Figure 8.

References

[1] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B 39*, pages 1–38, 1977.

[2] Bradley Efron. The two sample problem with censored data. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 4, pages 831–853, 1967.

[3] IJ Good and Ray A Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.

[4] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[5] I. Klebanov and A. Sikorski. Empirical Bayes methods, reference priors, cross entropy and the EM algorithm. ZIB-Report 16-56, Zuse Institute Berlin, 2016.

[6] Nan Laird. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*, 73(364):805–811, 1978.

[7] Bruce G Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*. JSTOR, 1995.

[8] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[9] Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

[10] S. Röblitz, C. Stötzel, P. Deuflhard, H.M. Jones, D.-O. Azulay, P. van der Graaf, and S.W. Martin. A mathematical model of the human menstrual cycle for the administration of GnRH analogues. *J. Theoretical Biology*, 321:8–27, 2013.

[11] Byungtae Seo and Bruce G Lindsay. A computational strategy for doubly smoothed MLE exemplified in the normal mixture model. *Computational Statistics & Data Analysis*, 54(8):1930–1941, 2010.

[12] Byungtae Seo and Bruce G Lindsay. A universally consistent modification of maximum likelihood. *Statistica Sinica*, pages 467–487, 2013.

[13] Bruce W Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 290–295, 1976.

[14] Aad Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.