Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

ILJA KLEBANOV, ALEXANDER SIKORSKI, CHRISTOF SCHÜTTE, SUSANNA RÖBLITZ

Empirical Bayes Methods, Reference Priors, Cross Entropy and the EM Algorithm [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, Reference Priors, Cross Entropy and the EM Algorithm

Ilja Klebanov, Alexander Sikorski, Christof Schütte, Susanna Röblitz

November 30, 2016

Abstract

When estimating a probability density within the empirical Bayes framework, the non-parametric maximum likelihood estimate (NPMLE) usually tends to overfit the data. This issue is usually taken care of by regularization – a penalization term is subtracted from the marginal log-likelihood before the maximization step, so that the estimate favors smooth solutions, resulting in the so-called maximum penalized likelihood estimation (MPLE). The majority of penalizations currently in use are rather arbitrary brute-force solutions, which lack invariance under transformation of the parameters (reparametrization) and measurements. This contradicts the principle that, if the underlying model has several equivalent formulations, the methods of inductive inference should lead to consistent results. Motivated by this principle and using an information-theoretic point of view, we suggest an entropy-based penalization term that guarantees this kind of invariance. The resulting density estimate can be seen as a generalization of reference priors. Using the reference prior as a hyperprior, on the other hand, is argued to be a poor choice for regularization.

We also present an insightful connection between the NPMLE, the cross entropy and the principle of minimum discrimination information suggesting another method of inference that contains the doubly-smoothed maximum likelihood estimation as a special case.

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

1 Introduction

Inferring a parameter $X \in \mathcal{X}$ from a measurement $Z \in \mathcal{Z}$ using Bayes' rule requires prior knowledge about X, which is not given in many applications. This has led to a lot of controversy in the statistical community and to harsh criticism concerning the objectivity of the Bayesian approach.

However, if independent measurements $Z_m \in \mathcal{Z}$, $m = 1, \dots, M$, are given for a large number M of "individuals" with individual parametrizations $X_m \in \mathcal{X}$, which is the case in many statistical studies, empirical Bayes methods can provide a solution to this controversy. Instead of applying Bayes' rule to each

measurement *separately*, these methods usually boil down to gathering all measurements in order to construct an informative prior as a first step and then using this prior for the Bayesian inference of the individual parametrizations in a second step.

A typical application is the retrieval of patient-specific parametrizations in large clinical studies, e.g. [14].

The first step is often performed by maximizing the marginal likelihood $L(\pi)$ of all measurements over the prior π, for which the EM algorithm [3] is the standard tool. This procedure can be viewed as an interplay of frequentist and Bayesian statistics: The prior is viewed as a *hyperparameter* and chosen by maximum likelihood estimation (MLE), the actual individual parametrizations are then inferred using Bayes' rule. Roughly speaking, there are two possible assumptions on the prior distribution:

- Parametric MLE (PMLE): The prior π has a parametric form with a finite number of parameters, e.g. $\pi \sim N(\mu, \Sigma)$ is a normal distribution, and only these parameters have to be estimated, which are then referred to as hyperparameters (in our example the mean μ and the covariance matrix Σ).

- Non-parametric MLE (NPMLE): No parametric form of the prior π is assumed, in which case the hyperparameter is the prior π itself and the hyperparameter estimation problem is infinite-dimensional (in the continuous case).

We will concentrate on the second scenario, where we have no information about the form of the prior. In this case, it can be proven that the marginal likelihood is maximized by a discrete distribution π_{NPMLE} with at most M nodes, see [16, Theorems 2-5] or [17, Theorem 21]. This typical "overconfidence" of the maximum likelihood estimator is often dealt with by subtracting a *roughness penalty* $\Phi(\pi)$ from the marginal log-likelihood function $\log L(\pi)$, such that "smooth" priors are favored, resulting in the so-called *maximum penalized likelihood estimation* (MPLE):

$$\pi_{\mathrm{MPLE}} = \arg\max_{\pi} \log L(\pi) - \Phi(\pi). \tag{1}$$

The EM algorithm can be adapted to this situation, see [18, Section 1.6].

This approach can be viewed from the Bayesian perspective as choosing a hyperprior $f(\pi) \propto e^{-\Phi(\pi)}$ for the hyperparameter $\Pi = \pi$ on the set of all considered priors and then performing a maximum a posteriori (MAP) estimation for π:

$$\pi_{\mathrm{MAP}} = \arg\max_{\pi} L(\pi)\, e^{-\Phi(\pi)} = \arg\max_{\pi} \log L(\pi) - \Phi(\pi) = \pi_{\mathrm{MPLE}}\,. \tag{2}$$

Favorable properties of the roughness penalty function Φ are:

(a) penalization of rough/peaked behavior

(b) non-informativity: Without any extra information about the parameter or the prior, we want to keep our assumptions to a minimum (in the sense of objective Bayes methods).

(c) invariance under (equivalent) reparametrizations

(d) invariance under transformations of the measurement space $\mathcal{Z}$

(e) convexity: Since $\log L(\pi)$ can be proven to be concave in the NPMLE case [17, Section 5.1.3], a convex penalty function $\Phi(\pi)$ would guarantee convergence of the (modified) EM algorithm to a *global* optimum.

(f) implementability and simplicity

(g) natural and intuitive justification

We will concentrate mainly on the properties (c) and (d), which are formalized in Definition 3, since they guarantee consistent results: If two statisticians use equivalent models to explain equivalent data, their results must be consistent. The penalty functions currently used are mostly ad hoc and rather brute force solutions that confine amplitudes (e.g. ridge regression [18, Section 1.6]) or derivatives (see e.g. [10]) of the prior π, which are neither invariant under reparametrization of X nor have a natural derivation.

In [9], Good suggests to use the entropy as a roughness penalty

$$\Phi(\pi) = -\gamma\, H_X(\pi) = \gamma \int_{\mathcal{X}} \pi(x) \log \pi(x)\, \mathrm{d}x,$$

where γ is a weight that balances the trade-off between goodness of fit and smoothness of the prior (in this context, the term smoothness should be replaced by uncertainty or non-informativity of the prior, since that is what entropy measures), which is a very natural approach from an information-theoretic point of view. However, though claiming to apply the principle of maximum entropy, Good does not derive the roughness penalty from such a principle. Also, the above Φ is not invariant under reparametrization of X, making it, as Good puts it, "somewhat arbitrary" [9, p. 912]:

> "It could be objected that, especially for a continuous distribution, entropy is somewhat arbitrary, since it is variant under a transformation of the independent variable."

The *relative* entropy (also called Kullback-Leibler divergence or information gain) on the other hand provides a possibility to overcome this obstacle, since its expected value, the mutual information $\mathcal{I}[X;Z]$ of X and Z, is invariant under transformations of both X and Z. In the case of no measurements, when the only information at hand is the likelihood model, this method coincides with the construction of reference priors and thereby provides a generalization of reference priors.

Remark 1. *The idea to use information-theoretic considerations in order to construct non-informative (hyper-) priors is not new. Jeffreys prior [13] and its generalization to reference priors [2, 1] by Bernardo and Berger has become a widely used tool in objective Bayesian analysis. However, though suitable for constructing priors, these choices seem unfit for hyperpriors, since they promote peaked behaviour of distributions instead of penalizing it, as shown in Example 7.*

Remark 2. *An alternative to MPLE is the method of sieves introduced in [11] and applied to NPMLE in [8]. This approach restricts the space of priors to a*

suitable subspace and the restriction is weakened with the number of measurements. For a general discussion of MPLE and similar methods, see [5] and [6].

After introducing the mathematical framework in Section 2, the discussion on penalty terms and hyperpriors is given in Section 3. Section 4 deals with a connection between the log-likelihood and the cross entropy, which motivates another approach to regularize the NPMLE that can be viewed as a regularization in the measurement space $\mathcal{Z}$ and is a generalization of the doubly-smoothed MLE (DS-MLE) introduced by Seo and Lindsay [21]. Both approaches, MPLE and DS-MLE, are applied to a simple one-dimensional toy example in Section 5, followed by a short conclusion in Section 6. In Appendix A, we show how the EM algorithm in the NPMLE framework is connected to gradient ascent.

In a companion paper [14] we apply these theoretical results to a high-dimensional real life problem.

2 Setup and Notation

The main aim is to infer the parameter $X \in \mathcal{X} \subseteq \mathbb{R}^d$ from a measurement $Z \in \mathcal{Z} \subseteq \mathbb{R}^n$ for several "individuals", where $\mathcal{X}$ and $\mathcal{Z}$ are assumed to be open and convex subsets of $\mathbb{R}^d$ and $\mathbb{R}^n$. Denoting the probability density of a random variable Y by ρ_Y and its density conditioned on an event A by $\rho_Y(\bullet|A)$, the likelihood model is given by the conditional probability densities,

$$\mathcal{R} = \{\rho_Z(\bullet|X = x) \mid x \in \mathcal{X}\},$$

which we will assume to be given.

However, the true density $\pi_{\text{true}} = \rho_X$ of X is unknown and therefore Bayes' rule has to be applied using some elicited or constructed prior $\pi \in \mathcal{M}_1(\mathcal{X})$, where

$$\mathcal{M}_1(\mathcal{X}) := \left\{\pi \in L^1(\mathcal{X}) \mid \pi \geqslant 0,\, \|\pi\|_{L^1} = 1\right\}$$

is the set of all probability densities on $\mathcal{X}$. As described in the introduction, empirical Bayes methods rely on several measurements $Z_1 = z_1, \ldots, Z_M = z_m$ for the construction of a prior π, which can be performed via MPLE with an appropriate penalty $\Phi(\pi)$. The parameter X will serve as a latent variable,

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

In other words, viewing the prior $\Pi = \pi$ as a hyperparameter, we assume our hyperparametric model

$$\left(\{\rho_Z(\bullet|\Pi = \pi) \mid \pi \in \mathcal{M}_1(\mathcal{X})\}, \mathbb{P}\right)$$

to be correctly specified, i.e. there exists $\pi_{\text{true}} \in \mathcal{M}_1(\mathcal{X})$ such that the data-generating distribution $\mathbb{P}$ has the probability density $\rho_Z(\bullet|\Pi = \pi_{\text{true}})$:

$$Z_m \overset{\text{i.i.d.}}{\sim} \rho_Z = \rho_Z(\bullet|\Pi = \pi_{\text{true}}), \qquad m = 1, \ldots, M.$$

We will also assume the hyperparametric model to be identifiable, see [26, Section 5.5], i.e.

$$\rho_Z(\bullet|\Pi = \pi) = \rho_Z(\bullet|\Pi = \pi_{\text{true}}) \quad \Longleftrightarrow \quad \pi = \pi_{\text{true}}, \tag{3}$$

since otherwise there would be no chance to recover the true distribution from no matter how many measurements. The marginal likelihood of the prior is given by

$$L(\pi) = \prod_{m=1}^{M} \rho_Z(z_m|\Pi = \pi). \tag{4}$$

Here,

$$\rho_Z(z|\Pi = \pi) := \int_{\mathcal{X}} \rho_Z(z|X = x)\, \pi(x)\, \mathrm{d}x$$

denotes the "would-be probability density" of Z, if π was the correct prior. The posterior density given the prior π and the measurement $z \in \mathcal{Z}$ will be denoted by

$$p_\pi^z(x) := \frac{\pi(x)\, \rho_Z(z \mid X = x)}{\rho_Z(z \mid \Pi = \pi)}\,.$$

Our aim is to approximate π_{true} given a large number of measurements. This will be realized by MPLE after choosing an appropriate penalty $\Phi(\pi)$ (or, equivalently, a suitable hyperprior $f(\pi) = e^{-\Phi(\pi)}$):

$$\pi_{\text{true}} \approx \pi_{\text{MPLE}} = \arg\max_{\pi}\ \log L(\pi) - \Phi(\pi).$$

Throughout this manuscript, we will slightly abuse notation and not distinguish between a probability distribution and its probability density.

3 Choosing the Penalty $\Phi(\pi)$

As motivated in the introduction, penalizing by means of entropy provides a natural approach to incorporate the idea of non-informativity about the parameter into the inference process. However, there is more than one type of entropy that can be considered in our setup. We choose to penalize by means of the expected Kullback-Leibler divergence (relative entropy), the *mutual information* $\mathcal{I}[X;Z]$ of X and Z, which we will view as a function of π:

$$\begin{aligned}
\mathcal{I}[X;Z](\pi) &= \mathbb{E}_{Z\sim\rho_Z(\bullet|\Pi=\pi)}\left[D_{\mathrm{KL}}\left(p_\pi^Z \,\|\, \pi\right)\right] \\
&= \int_{\mathcal{X}}\int_{\mathcal{Z}} \pi(x)\, \rho_Z(z \mid X = x)\, \log\left[\frac{\rho_Z(z \mid X = x)}{\rho_Z(z \mid \Pi = \pi)}\right] \mathrm{d}z\, \mathrm{d}x \\
&= H_Z(\pi) - H_{Z|X}(\pi),
\end{aligned} \tag{5}$$

where

$$\begin{aligned}
H_Z(\pi) &:= -\int_{\mathcal{Z}} \rho_Z(z \mid \Pi = \pi)\, \log\left[\rho_Z(z \mid \Pi = \pi)\right] \mathrm{d}z, \\
H_{Z|X}(\pi) &:= -\int_{\mathcal{X}} \pi(x) \int_{\mathcal{Z}} \rho_Z(z \mid X = x)\, \log\left[\rho_Z(z \mid X = x)\right] \mathrm{d}z\, \mathrm{d}x
\end{aligned}$$

are the entropy of $Z \in \mathcal{Z}$ and the conditional entropy of Z given X, respectively, and D_{KL} denotes the Kullback-Leibler divergence.

The main reason for this choice is its invariance under transformations of X and Z, see properties (c) and (d) from the introduction, which guarantees consistent results. Let us make this more precise:

Definition 3. *Let $\mathcal{R} = \{\rho_Z(\bullet \mid X = x) \mid x \in \mathcal{X}\}$ be a likelihood model.*

(i) We call a function $F = F[\mathcal{R}]\colon \mathcal{M}_1(\mathcal{X}) \to \mathbb{R}$ invariant under transformation of X *or* invariant under reparametrization*, if for any diffeomorphism $\varphi\colon \mathcal{X} \to \tilde{\mathcal{X}},\ x \mapsto \tilde{x}$ and any $\pi \in \mathcal{M}_1(\mathcal{X})$,*

$$F[\mathcal{R}](\pi) = F[\tilde{\mathcal{R}}](\tilde{\pi}),$$

where $\tilde{\mathcal{R}}$ is the transformed likelihood model and $\tilde{\pi} = \varphi_{\#}\pi$ is the pushforward of π under φ:

$$\begin{aligned} \rho_Z(z \mid \tilde{X} = \tilde{x}) &= \rho_Z(z \mid X = \varphi^{-1}(\tilde{x})), \\ \tilde{\pi}(\tilde{x}) &= \left|\det(D\varphi^{-1})(\tilde{x})\right| \cdot \pi(\varphi^{-1}(\tilde{x})). \end{aligned} \tag{6}$$

(ii) We call a function $F = F[\mathcal{R}]\colon \mathcal{M}_1(\mathcal{X}) \to \mathbb{R}$ invariant under transformation of Z*, if for any diffeomorphism $\psi\colon \mathcal{Z} \to \tilde{\mathcal{Z}},\ z \mapsto \tilde{z}$,*

$$F[\mathcal{R}](\pi) = F[\tilde{\mathcal{R}}](\pi),$$

where $\tilde{\mathcal{R}} = \psi_{\#}\mathcal{R}$ is the pushforward of the likelihood model under ψ:

$$\rho_{\tilde{Z}}(\tilde{z} \mid X = x) = \left|\det(D\psi^{-1})(\tilde{z})\right| \cdot \rho_Z(\psi^{-1}(\tilde{z}) \mid X = x),$$

Here, $D\chi$ denotes the Jacobian of a diffeomorphism χ.

We are now ready to formulate and prove the invariance property of the mutual information $\mathcal{I}[X;Z]$ of X and Z in a slightly more general setup.

Proposition 4. *Let $\mathcal{R} = \{\rho_Z(\bullet \mid X = x) \mid x \in \mathcal{X}\}$ be a likelihood model with measurements $Z_m = z_m \in \mathcal{Z},\ m = 1, \dots, M$ and*

$$\Phi_g(\pi) = \Phi_g[\mathcal{R}](\pi) = -\int_{\mathcal{X}}\int_{\mathcal{Z}} \pi(x)\, \rho_Z(z \,|\, X = x)\, g\left[\frac{\rho_Z(z \,|\, X = x)}{\rho_Z(z \,|\, \Pi = \pi)}\right] \mathrm{d}z\, \mathrm{d}x$$

for some measurable function $g\colon \mathbb{R} \to \mathbb{R}$, for which the integral is defined. Then Φ_g, and in particular $\Phi_{\log} = -\mathcal{I}[X;Z]$, is invariant under transformations of X and Z and the marginal likelihood $L = L[\mathcal{R}]$ defined by (4) is invariant under transformations of X and Z up to a constant (transformation-dependent) factor.

Proof. Let $\varphi\colon \mathcal{X} \to \tilde{\mathcal{X}},\ x \mapsto \tilde{x}$ be a diffeomorphism, $\tilde{\mathcal{R}}$ the transformed likelihood model and $\tilde{\pi} = \varphi_{\#}\pi$ the pushforward of π under φ. Then, using the change of variables formula for $\tilde{x} = \varphi(x)$,

$$\begin{aligned} L[\tilde{\mathcal{R}}](\tilde{\pi}) &= \prod_{m=1}^{M} \int_{\tilde{\mathcal{X}}} \rho_Z(z_m \,|\, \tilde{X} = \tilde{x})\, \pi(\tilde{x})\, \mathrm{d}\tilde{x} \\ &= \prod_{m=1}^{M} \int_{\mathcal{X}} \rho_Z(z_m \,|\, X = x)\, \pi(x)\, \mathrm{d}x = L[\mathcal{R}](\pi), \\ \Phi_g[\tilde{\mathcal{R}}](\tilde{\pi}) &= -\int_{\tilde{\mathcal{X}}}\int_{\mathcal{Z}} \tilde{\pi}(\tilde{x})\, \rho_Z(z \,|\, \tilde{X} = \tilde{x})\, g\left[\frac{\rho_Z(z \,|\, \tilde{X} = x)}{\rho_Z(z \,|\, \tilde{\Pi} = \tilde{\pi})}\right] \mathrm{d}z\, \mathrm{d}\tilde{x} \\ &= -\int_{\mathcal{X}}\int_{\mathcal{Z}} \pi(x)\, \rho_Z(z \,|\, X = x)\, g\left[\frac{\rho_Z(z \,|\, X = x)}{\rho_Z(z \,|\, \Pi = \pi)}\right] \mathrm{d}z\, \mathrm{d}x = \Phi_g[\mathcal{R}](\pi). \end{aligned}$$

Let $\psi\colon \mathcal{Z} \to \tilde{\mathcal{Z}},\ z \mapsto \tilde{z}$ be a diffeomorphism and $\tilde{\mathcal{R}} = \psi_{\#}\mathcal{R}$ the pushforward of the likelihood model under ψ. Then, using the change of variables formula for $\tilde{z} = \psi(z)$ and denoting $c_m := \left|\det(D\psi^{-1})(\tilde{z}_m)\right|$,

$$\begin{aligned}
L[\tilde{\mathcal{R}}](\pi) &= \prod_{m=1}^{M} \int_{\mathcal{X}} \rho_{\tilde{Z}}(\tilde{z}_m \,|\, X = x)\, \pi(x)\, \mathrm{d}x \\
&= \prod_{m=1}^{M} c_m \int_{\mathcal{X}} \rho_Z(z_m \,|\, X = x)\, \pi(x)\, \mathrm{d}x = \left(\prod_{m=1}^{M} c_m \right) L[\mathcal{R}](\pi), \\
\Phi_g[\tilde{\mathcal{R}}](\pi) &= -\int_{\mathcal{X}} \int_{\tilde{\mathcal{Z}}} \pi(x)\, \rho_{\tilde{Z}}(\tilde{z} \,|\, X = x)\, g\left[\frac{\rho_{\tilde{Z}}(\tilde{z} \,|\, X = x)}{\rho_{\tilde{Z}}(\tilde{z} \,|\, \Pi = \pi)} \right] \mathrm{d}\tilde{z}\, \mathrm{d}x \\
&= -\int_{\mathcal{X}} \int_{\mathcal{Z}} \pi(x)\, \rho_Z(z \,|\, X = x)\, g\left[\frac{\rho_Z(z \,|\, X = x)}{\rho_Z(z \,|\, \Pi = \pi)} \right] \mathrm{d}z\, \mathrm{d}x = \Phi_g[\mathcal{R}](\pi).
\end{aligned}$$

□

Corollary 5. *Let $\mathcal{R} = \{\rho_Z(\bullet \mid X = x) \mid x \in \mathcal{X}\}$ be a likelihood model and $Z_m = z_m \in \mathcal{Z},\ m = 1, \ldots, M$ and $\Phi = \Phi[\mathcal{R}] = -\mathcal{I}[X; Z]$. Then the maximum penalized likelihood estimator π_{MPLE} defined by* (1) *is invariant under transformations of X and Z, where the invariance of a prior density is defined by* (6)*.*

Proof. Since $\log L$ and Φ are invariant under transformations of X and Z up to an additive constant by Proposition 4, so is

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

□

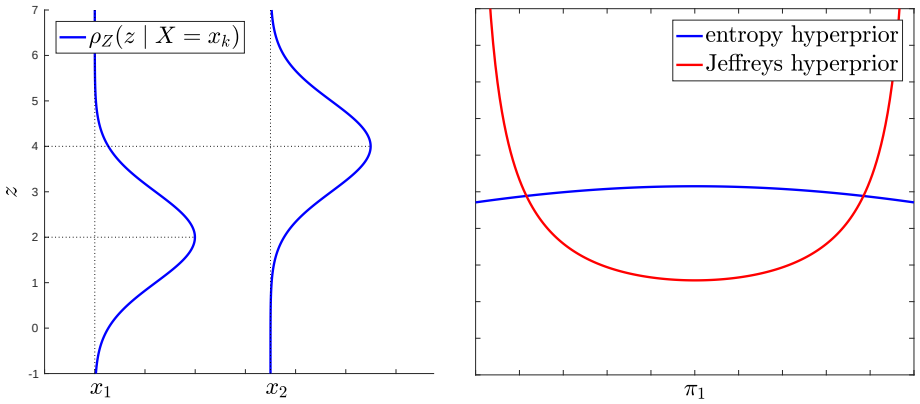
Definition 6. *For $\gamma > 0$, we will refer to $\Phi_{\mathcal{I},\gamma} = -\gamma\mathcal{I}[X; Z]$ as the* entropy penalty *and to the corresponding hyperprior $f_{\mathcal{I},\gamma}(\pi) = \exp(-\gamma\mathcal{I}[X; Z])$ as the* entropy hyperprior*.*

Penalizing with the mutual information $\mathcal{I}[X; Z]$ has a beautiful interpretation in the context of reference priors: If we have no measurements at hand, we want to maximize $\mathcal{I}[X; Z]$ in order to incorporate minimal informativity of the prior or, in other words, to maximize the expected information gain from some future measurement(s) [2, 1]. This results in the reference prior

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

If we do have measurements, we want to get a trade-off between goodness of fit and non-informativity of the prior, therefore we regularize the log-likelihood with $\mathcal{I}[X; Z]$.

One might also come up with the idea of using the concept of reference priors to construct hyperpriors. However, this appears to be a poor choice for regularization, since reference priors seem to favor peaked behaviour, as demonstrated in the following simple example:



(a) likelihood model (b) hyperpriors as functions of π_1

Figure 1: Jeffreys hyperprior tends to unregularize the prior, while the entropy hyperprior has a smoothing effect. Since $\pi_1 + \pi_2 = 1$, it is sufficient to plot the hyperprior over π_1.

Example 7. *Assume that $\mathcal{X} = \{x_1, x_2\}$ consists of only two points, $\mathcal{Z} = \mathbb{R}$ and that the likelihood model is given by normal distributions:*

$$\rho_Z(\bullet|X = x_1) = \mathcal{N}(2, 1), \qquad \rho_Z(\bullet|X = x_2) = \mathcal{N}(4, 1).$$

The hyperprior for the prior $\Pi = (\pi_1, \pi_2)$, where $\pi_j \geqslant 0$, $\pi_1 + \pi_2 = 1$, can be represented by a density over π_1. Figure 1 shows the results for the entropy hyperprior and for Jeffreys hyperprior. This example shows that Jeffreys hyperprior tends to unregularize the prior, since it favors the priors with high π_1 (and low π_2) as well as those with high π_2 (and low π_1) over the ones with balanced values of π_1, π_2. The entropy hyperprior, on the other hand, has its maximum at $\pi_1 = \pi_2 = 1/2$ and takes smaller values as the prior becomes more peaked.

In order to prove convexity of the entropy penalty $\Phi_{\mathcal{I},\gamma}(\pi)$, we will have to switch from $\mathcal{M}_1(\mathcal{X})$ to a convex subset K of the Banach space

$$\mathcal{B} = \{f \in L^1(\mathcal{X}) \mid \|f\|_1 = 0\}.$$

For this, we view $\mathcal{M}_1(\mathcal{X}) = \pi_* + K$ as a subset of an affine space, where π_* is any element of $\mathcal{M}_1(\mathcal{X})$ that acts as a displacement vector and

$$K = \{f \in \mathcal{B} \mid \pi_* + f \geqslant 0\}$$

is a convex subset of $\mathcal{B}$.

Proposition 8. *The entropy penalty $\Phi_{\mathcal{I},\gamma}(\pi)$ is convex in π.*

Proof. Since $\mathcal{I}[X; Z](\pi) = H_Z(\pi) - H_{Z|X}(\pi)$ by (5) and $H_{Z|X}(\pi)$ is linear in π, it is sufficient to consider the $H_Z(\pi)$. We transform it to a function $\Phi\colon K \to \mathbb{R}$ via

$$\Phi(f) := -H_Z(\pi_* + f).$$

Its Fréchet derivative $D\Phi\colon K \to B(K, \mathbb{R})$, where $B(K, \mathbb{R})$ denotes the set of bounded linear functionals on K, is given by

$$[D\Phi(f)](g) = \int_{\mathcal{Z}} \int_{\mathcal{X}} \rho_Z(z \mid X = x)\, g(x)\, \mathrm{d}x\, \log \rho_Z(z \mid \Pi = \pi_* + f)\, \mathrm{d}z.$$

Since

$$\Phi(g) - \Phi(f) - [D\Phi(f)](g-f) = \int \rho_Z(z \,|\, \pi_* + g) \, \log\left[\frac{\rho_Z(z \,|\, \Pi = \pi_* + g)}{\rho_Z(z \,|\, \Pi = \pi_* + f)}\right] \mathrm{d}z$$
$$= D_{\mathrm{KL}}\big(\rho_Z(z \,|\, \pi_* + g) \parallel \rho_Z(z \,|\, \pi_* + f)\big) \geqslant 0,$$

the claim follows from Kachurovskii's theorem [22, Proposition 7.4]. □

Corollary 9. *Since the marginal log-likelihood* $\log L(\pi)$ *is concave in* π *(see [17, Section 5.1.3]) and* $\Phi_{\mathcal{I},\gamma}(\pi)$ *is convex in* π*, the optimization problem*

$$\pi_{\mathrm{MPLE}} = \arg\max_{\pi} \ \log L(\pi) - \Phi_{\mathcal{I},\gamma}(\pi)$$

is concave in π*. Therefore, the (modified) EM algorithm converges globally.*

3.1 Derivation of the entropy penalty from a maximum entropy principle

Many Bayesian inverse problems are of the form [23]

$$Z = G(X) + E, \qquad E \sim \rho_E, \tag{7}$$

where Z is a noisy observation of $G(X)$, $G\colon \mathcal{X} \to \mathcal{Z}$ is the underlying model and E is an additive (often normally distributed) error term. In this case, the likelihood model consists of shifted versions of ρ_E,

$$\rho_Z(z \mid X = x) = \rho_E(z - G(x))$$

and hence the conditional entropy

$$H_{Z|X}(\pi) := -\int_{\mathcal{X}} \pi(x) \int_{\mathcal{Z}} \rho_Z(z \mid X = x) \, \log\left[\rho_Z(z \mid X = x)\right] \mathrm{d}z \, \mathrm{d}x$$
$$= \int_{\mathcal{Z}} \rho_E(z) \, \log\left[\rho_E(z)\right] \mathrm{d}z$$

is constant in π. Since $\mathcal{I}[X;Z](\pi) = H_Z(\pi) - H_{Z|X}(\pi)$ by (5), the entropy penalty $\Phi_{\mathcal{I},\gamma}$ is equivalent to the "Z-entropy penalty" $\Phi_{H_Z,\gamma} := -\gamma H_Z$.

We conjecture that this penalty can be derived from a hyperprior stemming from a maximum entropy principle, this time of the whole system (Π, Z) (remember that, when inferring the density Π from measurements Z, X serves only as a latent variable). So far, we were only able to formulate and prove this statement for discrete parameter spaces $\mathcal{X} = \{x_1, \dots, x_K\}$, where the hyperprior can be expressed by a probability density $f(\pi)$ on $\mathcal{M}_1(\mathcal{X}) \subseteq \mathbb{R}^K_{\geqslant 0}$. Maximizing the entropy of the whole model (Π, Z), which is given by

$$H_{(\Pi,Z)}(f) = H_{Z|\Pi}(f) + H_\Pi(f) = \int_{\mathcal{M}_1(\mathcal{X})} f(\pi) \, \big(H_Z(\pi) - \log f(\pi)\big) \, \mathrm{d}\pi,$$

leads to the following hyperprior:

Proposition 10. *If* f_E *is a (possibly improper) prior on* $\mathcal{M}_1(\mathcal{X})$ *that maximizes the total entropy* $H_{(\Pi,Z)}(f)$*, then*

$$f_E(\pi) \propto \exp(H_Z(\pi)).$$

Proof. Differentiation of $H_{(\Pi,Z)}(f)$ with respect to f yields

$$\partial_f H_{(\Pi,Z)}(f) = H_Z(\bullet) - \log f(\bullet),$$

which proves the claim. □

Corollary 11. *If the Bayesian inverse problem takes the form* (7)*, then, for* $\gamma = 1$*, the penalty* $\Phi_{\mathcal{I},\gamma} = \Phi_{H_Z,\gamma}$ *corresponds to a hyperprior* f_E *that maximizes the total entropy* $H_{\Pi,Z}$*.*

Proof. By Proposition 10, f_E is given by $f_E \propto \exp[-\gamma\Phi_{H_Z,1}]$. □

4 Cross entropy and the EM algorithm

In this section, we will discuss an interesting connection between the marginal likelihood $L(\pi)$ and the cross entropy $H^{\text{cross}}\big(\rho_Z, \rho_Z(\bullet|\Pi = \pi)\big)$ between the true measurement distribution $\rho_Z = \rho_Z(\bullet \mid \Pi = \pi_{\text{true}})$ and the one induced by π, $\rho_Z(\bullet \mid \Pi = \pi)$. As we shall see, the latter can be viewed as a (negative) *infinite data log-likelihood* relying on the knowledge of the entire distribution ρ_Z of Z instead of just a finite number of measurements $Z_1, \ldots, Z_M \sim \rho_Z$. This connection suggests an alternative approach to MPLE that is a generalization of the doubly-smoothed MLE [21, 20].

Further, we will generalize the EM algorithm to the infinite data scenario. Throughout this section, we will not restrict ourselves to the non-parametric case, therefore $\Pi = \pi$ can denote either a parametrization of the prior density, $\rho_X(\bullet \mid \Pi = \pi)$, in the PMLE setting or the prior density itself in the NPMLE setting, in which case $\rho_X(\bullet \mid \Pi = \pi) = \pi$.

Our generalization requires only the following two adaptations, where $\pi(x)$ is replaced by the more general formulation $\rho_X(x \mid \Pi = \pi)$:

$$\rho_Z(z|\Pi = \pi) := \int_{\mathcal{X}} \rho_Z(z|X = x)\, \rho_X(x \mid \Pi = \pi)\, \mathrm{d}x$$

and

$$p^z_\pi(x) := \frac{\rho_X(x \mid \Pi = \pi)\, \rho_Z(z \mid X = x)}{\rho_Z(z \mid \Pi = \pi)}\,.$$

The application of the EM algorithms in the NPMLE setting and the connection between the resulting fixed point iterations are discussed in the next section.

Let us first recall the usual EM algorithm, which maximizes the marginal likelihood $L(\pi)$ by iterations constructed in the following way:

- Formulate the *complete data likelihood function*

$$L^{\text{comp}}(\mathbf{x}, \mathbf{z} \mid \Pi = \pi) = \prod_{m=1}^{M} \rho_X(x_m \mid \Pi = \pi)\, \rho_Z(z_m \mid X = x_m),$$

  where the term "complete data" refers to the knowledge of both $\mathbf{X} = \mathbf{x} = (x_1, \ldots, x_M)$ and $\mathbf{Z} = \mathbf{z} = (z_1, \ldots, z_M)$.

- Formulate the *expected complete data log-likelihood* $Q_{\mathrm{cd}}(\pi \mid \pi_n)$ (the indices c and d are explained in Appendix A) with respect to the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$ under the current estimate π_n of Π:

$$\begin{aligned} Q_{\mathrm{cd}}(\pi \mid \pi_n) :=& \frac{1}{M}\,\mathbb{E}\Big[\log L^{\mathrm{comp}}\big([\mathbf{X} \mid \mathbf{Z} = \mathbf{z}, \Pi = \pi_n], \mathbf{z} \mid \Pi = \pi\big)\Big] \\ =& \frac{1}{M}\sum_{m=1}^{M} \int_{\mathcal{X}} p_{\pi_n}^{z}(x)\, \log\big(\rho_X(x \mid \Pi = \pi)\, \rho_Z(z_m \mid X = x)\big)\,\mathrm{d}x \end{aligned}$$

  The scaling factor $1/M$ is usually left out in the definition of Q_{cd}, but it will slightly simplify our notation.

- In each iteration step, maximize $Q_{\mathrm{cd}}(\pi \mid \pi_n)$ over π:

$$\pi_{n+1} = \arg\max_{\pi}\, Q_{\mathrm{cd}}(\pi \mid \pi_n).$$

It is more convenient to work with the marginal log-likelihood (scaled by $1/M$),

$$\mathcal{L}_{\mathrm{cd}}(\pi) := \frac{1}{M}\log L(\pi),$$

instead of the likelihood $L(\pi)$ itself (obviously, the local and global maxima of these two quantities coincide). It is well-known [3] that the above iteration increases $\mathcal{L}_{\mathrm{cd}}$ in each step,

$$\mathcal{L}_{\mathrm{cd}}(\pi_{n+1}) \geqslant \mathcal{L}_{\mathrm{cd}}(\pi_n), \qquad n \in \mathbb{N}, \tag{8}$$

and therefore converges to a local maximum of $\mathcal{L}_{\mathrm{cd}}$ (and even to a global one in the NPMLE setting, in which $\mathcal{L}_{\mathrm{cd}}$ is concave).

In order to prove the analogous statements in the infinite data scenario, let us introduce the analogues of $\mathcal{L}_{\mathrm{cd}}$ and Q_{cd}:

Definition 12. *We define the* infinite data log-likelihood *as*

$$\mathcal{L}_{\mathrm{cc}}(\pi) := -H^{\mathrm{cross}}\big(\rho_Z, \rho_Z(\bullet|\Pi = \pi)\big) = \int_{\mathcal{Z}} \rho_Z(z)\log \rho_Z(z|\Pi = \pi)\,\mathrm{d}z$$

and the infinite data expected complete data log-likelihood *as*

$$Q_{\mathrm{cc}}(\pi \mid \pi_n) := \int_{\mathcal{Z}} \rho_Z(z) \int_{\mathcal{X}} p_{\pi_n}^{z}(x)\, \log\big(\pi(x)\, \rho_Z(z \mid X = x)\big)\,\mathrm{d}x\,\mathrm{d}z.$$

These analogues are meaningful, since they appear as limits of $\mathcal{L}_{\mathrm{cd}}$ and Q_{cd}, if we let the number M of measurements go to infinity (therefore the term "infinite data"):

$$\mathcal{L}_{\mathrm{cd}} \xrightarrow{M\to\infty}_{\mathbb{P}} \mathcal{L}_{\mathrm{cc}} \quad \text{and} \quad Q_{\mathrm{cd}} \xrightarrow{M\to\infty}_{\mathbb{P}} Q_{\mathrm{cc}}, \tag{9}$$

since the measurements $Z_1, \ldots, Z_M$ are independent and ρ_Z-distributed (the index $\mathbb{P}$ denotes convergence in probability).

We will now prove the analogous statement to (8) in the infinite data scenario:

Proposition 13. *The EM algorithm given by*

$$\pi_{n+1} = \arg\max_{\pi} Q_{\mathrm{cc}}(\pi \mid \pi_n) \tag{10}$$

has the property

$$\mathcal{L}_{\mathrm{cc}}(\pi_{n+1}) \geqslant \mathcal{L}_{\mathrm{cc}}(\pi_n) \quad \textit{for each } n \in \mathbb{N}.$$

Proof. First note that Jensen's inequality implies

$$\begin{aligned}
\mathcal{L}_{\mathrm{cc}}(\pi) &= \int_{\mathcal{Z}} \rho_Z(z)\, \log\left(\int_{\mathcal{X}} \rho_Z(z \mid X = x)\, \pi(x)\, \frac{p^z_{\pi_n}(x)}{p^z_{\pi_n}(x)}\, \mathrm{d}x \right) \mathrm{d}z \\
&\geqslant \int_{\mathcal{Z}} \rho_Z(z) \int_{\mathcal{X}} p^z_{\pi_n}(x)\, \log\left(\frac{\rho_Z(z \mid X = x)\, \pi(x)}{p^z_{\pi_n}(x)}\, \frac{\rho_Z(z \mid \Pi = \pi_n)}{\rho_Z(z \mid \Pi = \pi_n)} \right) \mathrm{d}x\, \mathrm{d}z \\
&= \underbrace{\int_{\mathcal{Z}} \rho_Z(z)\, \log \rho_Z(z \mid \Pi = \pi_n)\, \mathrm{d}z}_{= \mathcal{L}_{\mathrm{cc}}(\pi_n)} + \underbrace{\int_{\mathcal{Z}} \rho_Z(z) \int_{\mathcal{X}} p^z_{\pi_n}(x)\, \log\left(\frac{\pi(x)}{\pi_n(x)} \right) \mathrm{d}x\, \mathrm{d}z}_{=: \ell(\pi \mid \pi_n)}.
\end{aligned}$$

Since $\ell(\pi \mid \pi_n) = 0$ for $\pi = \pi_n$ and

$$\arg\max_{\pi} \ell(\pi \mid \pi_n) = \arg\max_{\pi} Q_{cc}(\pi \mid \pi_n),$$

we have $\ell(\pi_{n+1} \mid \pi_n) \geqslant 0$ and therefore $\mathcal{L}_{\mathrm{cc}}(\pi_{n+1}) \geqslant \mathcal{L}_{\mathrm{cc}}(\pi_n)$. □

Remark 14. *Note that for $\pi_n = \pi_{\mathrm{true}}$ the quantity $Q_{cc}(\pi \mid \pi_n)$ is equal to the cross entropy between the true distribution of the pair (X, Z) and the one induced by π:*

$$Q_{cc}(\pi \mid \pi_{\mathrm{true}}) = -H^{\mathrm{cross}}\big(\rho_{(X,Z)}, \rho_{(X,Z)}(\bullet|\Pi = \pi)\big).$$

This proves that π_{true} is a fixed point of the EM algorithm (10).

Summing things up, we would like to apply the EM algorithm to maximize $\mathcal{L}_{\mathrm{cc}}$ and recover the true prior π_{true}. This corresponds to minimizing the cross entropy and is thereby equivalent to the principle of minimum discrimination information (or "Minxent"), see [15, 9]. Note that the cross entropy $H^{\mathrm{cross}}\big(\rho_Z, \rho_Z(\bullet|\Pi = \pi)\big)$ is minimal if and only if the two distributions ρ_Z and $\rho_Z(\bullet|\Pi = \pi)$ coincide, which is equivalent to $\pi = \pi_{\mathrm{true}}$ in the identifiable case (3).

However, since we have a *finite* number of data points $Z_1, \ldots, Z_M$, we can only apply the EM algorithm to maximize $\mathcal{L}_{\mathrm{cd}}$, which in general will not recover π_{true} exactly (see Section A). Alternatively, the desire to maximize $\mathcal{L}_{\mathrm{cc}}$ gives rise to another approach for the inference of Π:

(A) Approximate ρ_Z from $Z_1, \ldots, Z_M$, $\rho_Z^{\mathrm{appr}} \approx \rho_Z$, by your favorite density estimation method (e.g. Gaussian mixtures or kernel density estimation).

(B) Maximize $\mathcal{L}_{\mathrm{cc}}^{\mathrm{appr}}(\pi) := -H^{\mathrm{cross}}\big(\rho_Z^{\mathrm{appr}}, \rho_Z(\bullet|\Pi = \pi)\big)$.

This can be viewed as performing the regularization in the measurement space $\mathcal{Z}$ before the application of MLE.

Remark 15. *For this method to be invariant under transformations of Z, the density estimation $\rho_Z^{\mathrm{appr}} \approx \rho_Z$ has to be performed by a transformation invariant method. We will not discuss this issue here.*

4.1 Kernel density estimation and doubly-smoothed MLE

We will now discuss the case where the approximation $\rho_Z^{\text{appr}} \approx \rho_Z$ in step (A) is performed by a kernel density estimation,

$$\rho_Z^{\text{appr}}(z) = \frac{1}{M} \sum_{m=1}^{M} K_h(z - z_m), \tag{11}$$

where K_h denotes the kernel and h its bandwidth. In this case, it is meaningful to adjust the likelihood model by replacing $\rho_Z(z \mid \Pi = \pi)$ by

$$\tilde{\rho}_Z(\bullet \mid X = x) = \rho_Z(\bullet \mid X = x) * K_h$$

in the computation of $\rho_Z(\bullet|\Pi = \pi)$, since (for fixed h) we have an additional smoothing by the kernel:

$$\rho_Z^{\text{appr}} \xrightarrow{M\to\infty} \tilde{\rho}_Z := \rho_Z * K_h = \int \tilde{\rho}_Z(\bullet \mid X = x)\, \pi_{\text{true}}(x)\, \mathrm{d}x.$$

Alternatively, one could replace $\rho_Z(\bullet \mid \Pi = \pi)$ directly by

$$\tilde{\rho}_Z(\bullet \mid \Pi = \pi) := \rho_Z(\bullet \mid \Pi = \pi) * K_h.$$

This method was discussed by Seo and Lindsay [21], who called it the *doubly-smoothed MLE* (DS-MLE) due to the additional smoothing by the kernel density estimation and proved universal consistency under weak assumptions on the kernel and the likelihood model. The resulting density estimate is given by

$$\pi_{\text{DS-MLE}} := \arg\max_{\pi} \ \mathcal{L}_{\text{cc}}^{\text{appr}}(\pi).$$

If the numerical computation of $\mathcal{L}_{\text{cc}}^{\text{appr}}(\pi)$ is performed by Monte Carlo approximation,

$$\mathcal{L}_{\text{cc}}^{\text{appr}}(\pi) = \int_{\mathcal{Z}} \rho_Z^{\text{appr}}(z)\, \log \tilde{\rho}_Z(z|\Pi = \pi)\, \mathrm{d}z \approx \frac{1}{J} \sum_{j=1}^{J} \log \tilde{\rho}_Z(\zeta_j|\Pi = \pi) \tag{12}$$

where the points $\zeta_1, \ldots, \zeta_J \overset{\text{i.i.d.}}{\sim} \rho_Z^{\text{appr}}$ are samples of the density ρ_Z^{appr} (usually $J \gg M$), this approach retrieves the form of the standard NPMLE with measurements ζ_j instead of z_m (compare the right-hand side of the above expression with $\mathcal{L}_{\text{cd}}$). In other words, performing the kernel density estimation (11) and the Monte Carlo approximation (12) is equivalent to first "inflating" or "augmenting" each measurement z_m to a sampling $\zeta_1^{(m)}, \ldots, \zeta_\iota^{(m)} \overset{\text{i.i.d.}}{\sim} K_h(\bullet - z_m)$ and then applying NPMLE to the union of these samples $\{\zeta_1, \ldots, \zeta_J\} = \bigcup_{m=1}^{M} \{\zeta_1^{(m)}, \ldots, \zeta_\iota^{(m)}\}$ (and adjusted likelihood model), see [20].

5 Toy Example

We will illustrate the performance of the discussed methods on the FitzHugh-Nagumo model [7, 19], which is a simple model for the activation dynamics of a spiking neuron given by the following system of ODEs:

$$\begin{cases} \dot{v} &= v - \frac{v^3}{3} - w + I_{\text{ext}}, & v_0 = 1 \\ \tau \dot{w} &= v + a - bw, & w_0 = 0.1 \end{cases}$$

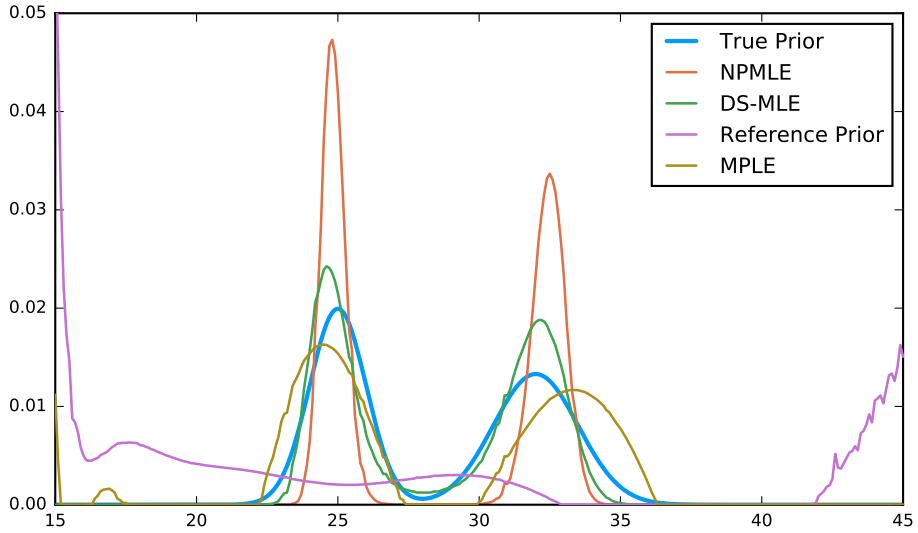


Figure 2: The prior estimates discussed in Sections 3 and 4. π_{NPMLE} and $\pi_{\text{DS-MLE}}$ were computed with the EM algorithm, while for the reference prior π_{ref} and π_{MPLE} a gradient ascent was used (with $\gamma = 19$). 200 iterations were performed. Note that for a higher number of iterations π_{NPMLE} and $\pi_{\text{DS-MLE}}$ (due to the discretization method discussed in Section 4.1) will not stop peaking.

We will assume the parameters $a = 0.7$, $b = 0.8$, $I_{\text{ext}} = 0.5$ to be given and only $X = \tau \in \mathcal{X} = [15, 45]$ to be inferred from a perturbed measurement Z of v at time $t^* = 40$:

$$Z = \phi(X) + E, \qquad \phi(X) = v(t^*), \qquad E \sim \mathcal{N}(0, 0.1^2).$$

We assume the true distribution of X to be the sum of two Gaussians:

$$\pi_{\text{true}} = \rho_X = \frac{1}{2}\left[\mathcal{N}\left(25, 1^2\right) + \mathcal{N}\left(32, 1.5^2\right)\right].$$

The resulting prior estimates π_{NPMLE}, π_{MPLE} and $\pi_{\text{DS-MLE}}$ as well as the true prior and the reference prior are shown in Figure 2.

6 Conclusion

We have introduced a penalty term for MPLE that is invariant under transformations in parameter and measurement space and thereby provides a consistent method of inference. It can be seen as a generalization of reference priors, since the two coincide in the case when no measurements are given.

Further, we have shown that the marginal log-likelihood and the EM algorithm have insightful analogues in the infinite data scenario. More precisely, if the number of measurements tends to infinity, the marginal log-likelihood converges to the negative cross entropy $H^{\text{cross}}\big(\rho_Z, \rho_Z(\bullet \mid \Pi = \pi)\big)$. The analogue of the EM algorithm can therefore be viewed as resulting from the principle of minimum cross entropy. The connections between the different log-likelihoods

and the corresponding EM algorithms is discussed more deeply in Appendix A and visualized as an demonstrative commutative diagram in Figure 3.

Since in practice only finitely many measurements are available, the minimization of the above cross entropy can not be implemented directly (ρ_Z is not accessible). However, it motivates another approach for density estimation by replacing ρ_Z by an approximation ρ_Z^{appr} and minimizing $H^{\text{cross}}\big(\rho_Z^{\text{appr}}, \rho_Z(\bullet \mid \Pi = \pi)\big)$ instead. As discussed in Section 4.1, this methodology contains the DS-MLE as a special case.

All methods were implemented for a simple one-dimensional toy example. In a companion paper [14], they were applied to a high-dimensional real life problem.

A EM algorithm and NPMLE

We now return to the NPMLE setting, where the hyperparameter $\Pi = \pi$ is the prior density itself, $\rho_X(\bullet \mid \Pi = \pi) = \pi$. We show how the NPMLE can be computed by a Monte Carlo discretization $\{x_1, \dots, x_K\} \subseteq \mathcal{X}$ of the space $\mathcal{X}$ and, correspondingly, of the prior π,

$$\pi = \sum_{k=1}^{K} w_k \, \delta_{x_k}, \quad w \in \mathcal{W} := \Big\{ w \in \mathbb{R}^K \mid w_k \geqslant 0 \, \forall k, \, \sum_{k=1}^{K} w_k = 1 \Big\}, \qquad \text{(X-MC)}$$

(of course, this also works if $\mathcal{X}$ is already discrete).

Here, the weights $W = w$ take the place of the hyperparameter and we have an analogous statement to (9) for the corresponding log-likelihoods and expected complete data log-likelihoods:

$$\begin{aligned}
\mathcal{L}_{\text{dd}}(w) &= \frac{1}{M} \sum_{m=1}^{M} \log \rho_Z(z_m | W = w) \\
&\xrightarrow{M\to\infty}_{\mathbb{P}} \int_{\mathcal{Z}} \rho_Z(z) \log \rho_Z(z | W = w) \, \mathrm{d}z \\
&=: \mathcal{L}_{\text{dc}}(w), \\
Q_{\text{dd}}(w \mid w_n) &= \frac{1}{M} \, \mathbb{E}\Big[\log L^{\text{comp}}\big([\mathbf{X} \mid \mathbf{Z} = \mathbf{z}, W = w_n], \mathbf{z} \mid W = w\big) \Big] \\
&= \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \big[\beta^z_{w_n}\big]_k \log\big(w_k \, \rho_Z(z_m \mid X = x_k)\big) \\
&\xrightarrow{M\to\infty}_{\mathbb{P}} \int_{\mathcal{Z}} \rho_Z(z) \sum_{k=1}^{K} \big[\beta^z_{w_n}\big]_k \log\big(w_k \, \rho_Z(z \mid X = x_k)\big) \, \mathrm{d}z \\
&=: Q_{\text{dc}}(w \mid w_n),
\end{aligned}$$

where

$$\rho_Z(z|W = w) := \sum_{k=1}^{K} w_k \, \rho_Z(z|X = x_k) \quad \text{and} \quad [\beta^z_w]_k := \frac{w_k \, \rho_Z(z|X = x_k)}{\rho_Z(z|W = w)}.$$

The indices c and d denote whether we consider the parameter space $\mathcal{X}$ (first index) and the measurement space $\mathcal{Z}$ (second index) to be discrete or continuous.

While the discretization in the parameter space $\mathcal{X}$ is just the Monte Carlo discretization (X-MC), the discretization in the measurement space $\mathcal{Z}$,

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

traces back to our lack of knowledge – instead of ρ_Z, we only have M ρ_Z-distributed measurements $Z_1, \ldots, Z_M$.

It is easy to see that, in the NPMLE setting, the application of the EM algorithms

$$\pi_{n+1} = \arg\max_{\pi} Q_{\mathrm{cd}}(\pi \mid \pi_n), \qquad w_{n+1} = \arg\max_{w} Q_{\mathrm{dd}}(w \mid w_n)$$

results in the fixed point iterations $\pi_{n+1} = \Psi_{\mathrm{cd}}\pi_n$ and $w_{n+1} = \Psi_{\mathrm{dd}}w_n$, respectively [16], where

$$\Psi_{\mathrm{cd}}\pi\,(x) = \frac{\pi(x)}{M} \sum_{m=1}^{M} \frac{\rho_Z(z_m|X=x)}{\rho_Z(z_m|\Pi=\pi)}\,, \tag{13}$$

$$[\Psi_{\mathrm{dd}}w]_k = \frac{w_k}{M} \sum_{m=1}^{M} \frac{\rho_Z(z_m|X=x_k)}{\rho_Z(z_m|W=w)}\,. \tag{14}$$

These iterations were first formulated by Turnbull [24, 25] and are usually referred to as self-consistency algorithms, since they fulfill the self-consistency principle introduced by Efron in 1967 [4].

We will now formulate an analogous result for the two cases with continous measurement space $\mathcal{Z}$:

Proposition 16. *In the NPMLE setting, the EM algorithms*

$$\pi_{n+1} = \arg\max_{\pi} Q_{\mathrm{cc}}(\pi \mid \pi_n), \qquad \pi_{n+1} = \arg\max_{\pi} Q_{\mathrm{dc}}(\pi \mid \pi_n)$$

are given by the fixed point iteration $\pi_{n+1} = \Psi_{\mathrm{cc}}\pi_n$ and $\pi_{n+1} = \Psi_{\mathrm{dc}}\pi_n$, respectively, where

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

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

Proof. For the first iteration, this follows from the fact that the derivative of $Q_{\mathrm{cc}}(\pi \mid \pi_n)$ with respect to π,

$$\partial_\pi Q_{\mathrm{cc}}(\pi \mid \pi_n) = \int_{\mathcal{Z}} \rho_Z(z)\, \frac{p^z_{\pi_n}}{\pi}\,\mathrm{d}z,$$

has to be a constant for the maximizer $\pi = \pi_{n+1}$. The second case goes analogously. □

Putting things together, we end up with the commutative diagram presented in Figure 3.

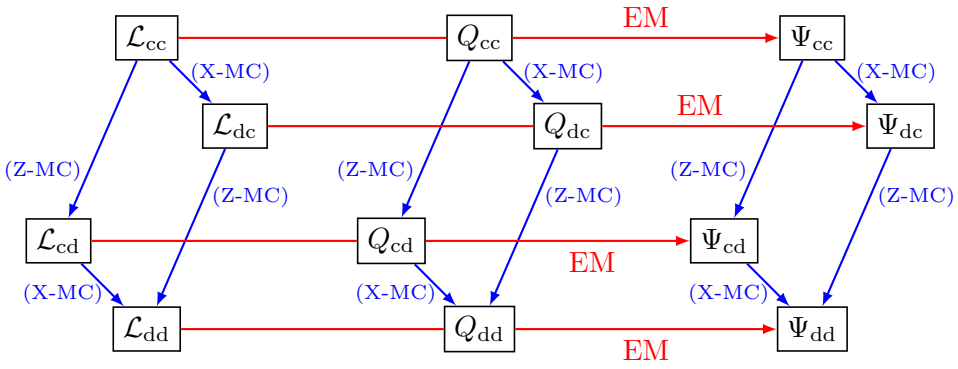


Figure 3: The relations between the log-likelihoods $\mathcal{L}$, the expected complete data log-likelihoods Q and the fixed point iterations Ψ resulting from the application of the EM algorithm can be summarized in a commutative diagram.

A.1 Connection to gradient ascent

In this section, we will establish an insightful connection between the EM algorithm in the non-parametric framework (NPMLE) and gradient ascent for the maximization of the log-likelihood $\mathcal{L}_{\mathrm{dd}}$. Analogous statements hold also for $\mathcal{L}_{\mathrm{cc}}$, $\mathcal{L}_{\mathrm{dc}}$ and $\mathcal{L}_{\mathrm{cd}}$. A natural way to maximize $\mathcal{L}_{\mathrm{dd}}$ is by moving stepwise in the direction of its gradient

$$\nabla\mathcal{L}_{\mathrm{dd}}(w) = \left(\frac{1}{M}\sum_{m=1}^{M}\frac{\rho_Z(z_m|X=x_k)}{\rho_Z(z_m|W=w)}\right)_{k=1,\dots,K}$$

after projecting it orthogonally onto the subspace $\mathcal{U} := \{u \in \mathbb{R}^K \mid \sum_k u_k = 0\}$:

$$w_{n+1} = w_n + \tau\, S_{\mathrm{orth}} \cdot \nabla\mathcal{L}_{\mathrm{dd}}(w_n), \qquad S_{\mathrm{orth}} := \frac{1}{K}\begin{pmatrix} K{-}1 & -1 & \cdots & -1 \\ -1 & K{-}1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ -1 & \cdots & -1 & K{-}1 \end{pmatrix},$$

$\tau > 0$ being the step size. The projection is necessary to guarantee the condition $\sum_k w_{n+1,k} = 1$. However, w_{n+1} might violate the condition $w_{n+1,k} \geqslant 0$, $k = 1, \dots, K$, and one would have to adjust the result after each iteration step in some way in order to push it into the simplex $\mathcal{W}$.

An alternative to this adjustment is to use a projection-like (and w-dependent) map $S_w \colon \mathbb{R}^K \to \mathcal{U}$ instead of an orthogonal projection, which guarantees that w_{n+1} lies in the correct simplex $\mathcal{W}$ (for $\tau = 1$):

$$w_{n+1} = \Psi_S\, w_n, \qquad \Psi_S\, w := w + S_w \cdot \nabla\mathcal{L}_{\mathrm{dd}}(w), \tag{15}$$

where

$$S_w := \mathrm{diag}(w) - ww^{\intercal} = \begin{pmatrix} w_1(1-w_1) & -w_1w_2 & \cdots & -w_1w_K \\ -w_2w_1 & w_2(1-w_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & -w_{K-1}w_K \\ -w_Kw_1 & \cdots & -w_Kw_{K-1} & w_K(1-w_K) \end{pmatrix}. \tag{16}$$

Proposition 17. *For all* $w \in \mathcal{W} = \Big\{ w \in \mathbb{R}^K \mid w_k \geqslant 0 \, \forall k, \ \sum_{k=1}^K w_k = 1 \Big\}$,

(i) S_w *maps into* $\mathcal{U} := \{u \in \mathbb{R}^K \mid \sum_k u_k = 0\}$,

(ii) S_w *is positive semi-definite,*

(iii) $\Psi_S \, w \in \mathcal{W}$ *and the fixed point iteration* (15) *coincides with* Ψ_{dd}.

Proof. (i) follows directly from

$$(1, \ldots, 1) \cdot S_w = w^\intercal - \Big(\sum_{k=1}^K w_k \Big) w^\intercal = 0,$$

while (ii) is a straightforward application of the Gershgorin circle theorem, see e.g. [12, Theorem 6.1.1].

A simple computation shows:

$$\begin{aligned}
\left[\Psi_S \, w\right]_k &= w_k + \frac{w_k}{M} \left(\sum_{m=1}^M \frac{\rho_Z(z_m|X = x_k)}{\rho_Z(z_m|W = w)} - \sum_{j=1}^K w_j \sum_{m=1}^M \frac{\rho_Z(z_m|X = x_j)}{\rho_Z(z_m|W = w)} \right) \\
&= w_k + \frac{w_k}{M} \sum_{m=1}^M \frac{\rho_Z(z_m|X = x_k)}{\rho_Z(z_m|W = w)} - \frac{w_k}{M} \sum_{m=1}^M \frac{\rho_Z(z_m|W = w)}{\rho_Z(z_m|W = w)} \\
&= \frac{w_k}{M} \sum_{m=1}^M \frac{\rho_Z(z_m|X = x_k)}{\rho_Z(z_m|W = w)} = \left[\Psi_{\mathrm{dd}} \, w\right]_k .
\end{aligned}$$

Since $S_w \left[\mathbb{R}^K\right] \subseteq \mathcal{U}$ by (1) and $\rho_Z(z|X = x) \geqslant 0$ for all $x \in \mathcal{X}$, $z \in \mathcal{Z}$, this proves (iii). □

Remark 18. *Since* $\langle v, S_w v \rangle \geqslant 0$ *for all* $v \in \mathbb{R}^K$ *by Proposition 17,* $S_w \nabla \mathcal{L}_{\mathrm{dd}}(w)$ *points in a non-descending direction of* $\mathcal{L}_{\mathrm{dd}}$*, though not in the steepest possible one. Therefore, the fixed point iteration* (15) *can be viewed as a nearly-gradient ascent. The classical results on the EM algorithm (see e.g. [3, Theorems 1 and 2]) combined with the equivalence of* (14) *and* (15) *(Proposition 17(iii)) prove that it actually yields a non-decreasing sequence and converges to a local maximum, which in the case of NPMLE is even a global one.*

The multiplication of S_w *with a vector* $v \in \mathbb{R}^K$ *can be performed in a simple way, since* $\mathrm{diag}(w) \cdot v$ *is just the pointwise multiplication and* $(ww^\intercal)v = w(w^\intercal v)$.

Acknowledgement

We would like to thank Shirin Riazy and Ingmar Schuster for carefully proof-reading this manuscript. We also express our gratitude to Tim Sullivan for a many insightful discussions.

References

[1] James Berger, José Bernardo, et al. On the development of reference priors. *Bayesian statistics*, 4(4):35–60, 1992.

[2] José Bernardo. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 113–147, 1979.

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

[4] 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.

[5] Paulus Petrus Bernardus Eggermont and Vincent LaRiccia. *Maximum penalized likelihood estimation*, volume 1. Springer, 2001.

[6] Paulus Petrus Bernardus Eggermont and Vincent LaRiccia. *Maximum penalized likelihood estimation*, volume 2. Springer, 2009.

[7] Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1(6):445, 1961.

[8] Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, pages 401–414, 1982.

[9] Irving Good. Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, pages 911–934, 1963.

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

[11] Ulf Grenander. *Abstract inference*. John Wiley & Sons, 1981.

[12] Roger Horn and Charles Johnson. *Matrix analysis*. Cambridge university press, 2012.

[13] Harold Jeffreys. An invariant form for the prior probability in estimation problems. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 186, pages 453–461. The Royal Society, 1946.

[14] I. Klebanov, A. Sikorski, C. Schütte, and S. Röblitz. Empirical Bayes methods for prior estimation in systems medicine. ZIB-Report 16-57, Zuse Institute Berlin, 2016.

[15] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

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

[17] Bruce Lindsay. Mixture models: theory, geometry and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–163. JSTOR, 1995.

[18] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[19] Jinichi Nagumo, Suguru Arimoto, and Shuji Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070, 1962.

[20] 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.

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

[22] Ralph Edwin Showalter. *Monotone operators in Banach space and nonlinear partial differential equations*, volume 49. American Mathematical Soc., 2013.

[23] Andrew M Stuart. Inverse problems: a bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[24] Bruce Turnbull. Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association*, 69(345):169–173, 1974.

[25] Bruce 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.

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