Master's Thesis

An Information-Theoretic Empirical Bayes Method and its Application to a Systems Biology Model

Free University of Berlin
Department of Mathematics and Computer Science

Alexander Sikorski

Supervisor: Prof. Susanna Röblitz

Berlin 2017

Declaration

I hereby declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institute. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Alexander Sikorski

Berlin, March 14th, 2017

Contents

1. Introduction 1

2. Empirical Bayes Methods 3
   - 2.1. The Bayesian formalism . . . 3
   - 2.2. The likelihood model . . . 4
   - 2.3. The empirical Bayes model . . . 5

3. Regularization 7
   - 3.1. Regularization via a penalization term . . . 8
   - 3.2. The mutual-information penalty . . . 8
   - 3.3. Regularization by smoothing of the data . . . 13

4. Numerical schemes 14
   - 4.1. Monte Carlo approximations . . . 14
   - 4.2. EM algorithm . . . 16
   - 4.3. Optimization for MPLE . . . 18
   - 4.4. Markov Chain Monte Carlo . . . 19

5. Application 21
   - 5.1. The problem . . . 22
   - 5.2. Sampling . . . 24
   - 5.3. Prior estimation . . . 26
   - 5.4. Results . . . 27

6. Conclusion 28

A. References 29

B. Implementation 31

1. Introduction

This thesis will cover the development and application of an empirical Bayes method to the problem of parameter estimation in systems biology. The goal is to provide a general and practical solution to the Bayesian inverse problem in the case of high dimensional parameter spaces making use of present cohort-data.

Regarding it's application to systems biology or medicine a quantification of uncertainty of the results is of utmost importance to any practitioner facing decisions such as whether to apply a specific treatment or release a new drug.

Although the classical frequentist approach to statistics offers answers to this in terms of confidence intervals and hypothesis tests, these techniques, aiming for point estimates, have a hard time dealing with ill-posed problems incorporating uncertainties and unidentifiabilities, often appearing in applications with large nonlinear models.

We will henceforth adopt the Bayesian view which, contrary to the point estimates in the frequentist approach, naturally confines the treatment of uncertainty by its description in terms of distributions.

As we will see it also allows for a natural incorporation of data, a circumstance of ever greater importance in a time of progressing digitalization of our lives, the medicine and the sciences coining the term big data.

We will furthermore tackle one of the main points of criticism on the Bayesian approach, namely its subjectivity in the choice of the prior: Two scientists, given the same data and working with the same model, can come up with different results (the posterior) imposing different a priori knowledge about the parameters in questions (the prior).

This critique lead to the school of objective Bayesian analysis (c.f. [1]) with the goal to provide methods ensuring consistent results in repeated usage by scientists.

One approach is the choice of so called *non-informative priors*, using information theoretic considerations to formalize the notion of non-informativity, giving rise to the *Jeffrey's prior* [12] and it's generalization to more general spaces, the *reference priors* [3].

An alternative approach is given by the empirical Bayes methods:

> The empirical Bayes approach to statistical decision problems is applicable when the same decision problem presents itself repeatedly and independently with a fixed but unknown a priori distribution of the parameter - Herbert Robbins [18].

Here repeated measurements of individuals from the population in question, called co-

hort data, are used to construct a prior representing that population. The first major contribution was by Robbins[17], deriving explicit formulas for the Bayes estimators of nonparametric priors for different families of likelihood functions giving rise to the *nonparametric maximum likelihood estimator* (NPMLE). For a while these ideas got largely neglected, probably due to its computational cost, until they was brought back to attention in the parametric case by Efron and Morris [5] in 1973.

Unfortunately the NPMLE approach applied to finite data results in discrete distributions. In Section 3 we will therefore discuss its regularization. We start by treating the topic of penalization of unsmooth solution, leading to the *maximal penalized likelihood estimator* (MPLE), and will set this into relation with Bayesian hyperpriors, i.e. prior assumption on the prior.

We then continue by introducing the *mutual information penalty*, a specific choice for penalizing unsmooth priors based on information theoretic considerations. We end up with a nonparametric, hence generally applicable, method combining the assumptions of least information from the non-informative priors with the usage of cohort data from the empirical Bayes approach combining their strengths. The suggested prior can as well be seen as a generalization of the reference priors to the cohort data regime.

We additionally discuss the *doubly-smoothed maximum likelihood estimator* (DS-MLE) estimate by Seo and Lindsay [22], an alternative approach with the idea of tackling the problem of finite data at its root by smoothing the data itself followed by the NPMLE.

In section 4 we will then address the question of how to compute the introduced prior estimates numerically. We will work with a sample driven Monte Carlo approach, expressing the priors as importance samplings. We then derive the explicit formulas for the EM algorithm [4] for the approximation of the NPMLE and the DS-MLE and provide the Jacobian for the optimization of the MPLE with the *mutual information penalty*.

We conclude the section with a digression to Markov Chain Monte Carlo sampling used for the computation of the Monte Carlo discretization.

Finally we will apply the developed methods and algorithms to a high-dimensional model from systems biology and discuss the results in section 5.

The essence of this thesis, including the general presentation and lots of the notation, originated in close relation to the articles [14, 13], to which we hence refer for further reading.

2. Empirical Bayes Methods

2.1. The Bayesian formalism

We start by laying out the basic formal tools in the Bayesian setting.

Definition 1. Let $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ denote continuous random variables with joint probability density $\rho_{X,Y}(x, y)$. The *conditional probability* density of Y given the event $X = x$ (i.e. $\rho_X(x) > 0$) X is defined as

$$\rho_{Y|X}(y|x) := \frac{\rho_{X,Y}(x, y)}{\rho_X(x)}, \tag{2.1}$$

where $\rho_X(x)$ is the *marginal density* of X, i.e. the joint density $\rho(x, y)$ marginalized over all possible y:

$$\rho_X(x) := \int_y \rho(x, y)\,\mathrm{d}y. \tag{2.2}$$

Throughout this thesis, we will slightly abuse notation and not distinguish between a probability distribution X and its density ρ_X. We will furthermore, whenever conditioning on an event $X = x$ directly denote the condition in the subscript, i.e.

$$\rho_{Y|x}(y) := \rho_{Y|X}(y \mid x),$$

whenever the corresponding random variable is clear from the context.

The above equations already imply the heart of Bayesian statistics, *Bayes' theorem:*

Theorem 2. *Let X and Y be as above. For all $y \in \mathcal{Y}$, such that $\rho_Y(y) > 0$ holds:*

$$\rho_{X|y}(x) = \frac{\rho_{Y|x}(y)\,\rho_X(x)}{\rho_Y(y)} = \frac{\rho_{Y|x}(y)\,\rho_X(x)}{\int_{\mathcal{X}} \rho_{Y|x'}(y)\,\rho_X(x')\,\mathrm{d}x'}. \tag{2.3}$$

Proof. This follows by successive insertion of (2.1) and (2.2):

$$\rho_{X|y}(x) \overset{(2.1)}{:=} \frac{\rho_{X,Y}(x, y)}{\rho_Y(y)} \overset{(2.1)}{=} \frac{\rho_{Y|x}(y)\,\rho_X(x)}{\rho_Y(y)} \overset{(2.2)}{=} \frac{\rho_{Y|x}(y)\,\rho_X(x)}{\int_{\mathcal{X}} \rho_{Y|x'}(y)\,\rho_X(x')\,\mathrm{d}x'}.$$

□

This formula, obtained by successive insertion of (2.1) and (2.2), constitutes the heart of Bayesian statistics. It tells us how to reconstruct the *posterior distribution* $\rho_{X|y}$ of the

unknown parameter x given data y, using the *likelihood* $\rho_{Y|x}(y)$ of x given y as well as the *prior* $\rho_X(x)$ reflecting our prior assumptions on the density of X.

Note that for fixed y the posterior $\rho_{X|y}$ and prior ρ_X both are probability densities in $\mathcal{X}$. The likelihood on the other hand is not a probability density in $\mathcal{X}$ (it is one one in $\mathcal{Y}$ for fixed x). For this reason it is also called the *likelihood function*

$$L(x \mid y) := \rho_{Y|x}(y).$$

2.2. The likelihood model

Our inference bases on the combination of a deterministic physical model, our description of the reality, with a stochastic measurement error and the formalism of Bayes'.

Definition 3. The *physical model* is a map $\Phi : \mathcal{X} \subseteq \mathbb{R}^n \to \mathcal{Y} \subseteq \mathbb{R}^m$, mapping some *parameter* $x \in \mathcal{X}$ to a resulting *state* $y \in \mathcal{Y}$. The *parameters* $x \in \mathcal{X}$ are distributed according to some prior $X : \Omega \to \mathcal{X} \sim \rho_X$, with $\rho_X(x) > 0 \forall x \in \mathcal{X}$.

The data generating *measurement process* $Z : \Omega \to \mathcal{Z} \subseteq \mathbb{R}^m$ is modeled as an independent Gaussian perturbation with prescribed covariance matrix Σ of the state conditioned on the parameter:

$$\rho_{Z|X}(z \mid x) = \Phi(x) + E, \quad E \sim \mathcal{N}(0, \Sigma), \tag{2.4}$$

where $\mathcal{N}(y, \Sigma)$ denotes the normal distribution with mean y and covariance matrix Σ.

This *likelihood model* gives us the means to compute the probability of measuring a single measurement z, given the underlying parameter x.

Assuming the *prior distribution* on X was known, this would enable us to compute the *posterior* $\rho_{X|z}$ given some measurement $z \in \mathcal{Z}$ by straightforward application of the Bayes' theorem (2.3).

Remark 4. Note that whilst this model is extensible to inference based on multiple measurements $\boldsymbol{z}^M := \left(z_i \overset{\text{i.i.d.}}{\sim} \rho_{Z|x}(x)\right)_{i=1}^M$ using the product-rule,

$$\rho_{X|\boldsymbol{z}^M}(x) = \frac{\prod_{m=1}^M \rho_{Z|x}(z_i)\, \rho_X(x)}{\prod_{m=1}^M \rho_Z(z_i)}$$

the assumption of identically distributed z_i implies that they all depend on the same parameter $X = x$. This is the right inference if all measurements come from the same

realization of X, e.g. the same subject, but is wrong assuming different measurements correspond to independent draws from the prior ρ_X, e.g. multiple subjects. We will therefore proceed to a more general framework allowing for the inference in the latter case.

2.3. The empirical Bayes model

Since in general the *prior* X cannot be assumed to be known, a number of different methods have been established for estimating this prior based on empirical cohort data, giving rise to so=called *empirical Bayes methods.* In that context we extend the model by parametrizing the prior ρ_X on a hyperparameter Π, the prior on the set of priors if one likes, resulting in the hierarchical model $\Pi \to X \to Z$, which we will refer to as the *hyperparametric model.*

Most literature confines itself to (finite dimensional) parametric empirical Bayes methods, characterized by considering parametrized families of distributions for the priors, e.g.

$$\begin{aligned} &\pi \in \Pi : \Omega \to = \mathbb{R}^k, \\ &\rho_{X|\pi} \sim \mathcal{N}(\pi, I) \end{aligned}$$

since these may admit explicit formulas for prior point estimates if the likelihood model and prior families admit simple forms, as well as circumventing regularization issues. We aim at a more general solution to the inference problem by allowing arbitrary distributions as priors, i.e.

$$\begin{aligned} &, \\ &\Pi : \Omega \to \mathcal{M}_1(\mathcal{X}) := \left\{\rho \in L^1(\mathcal{X}) \mid \rho \geq 0,\ \|\rho\|_{L^1} = 1\right\}, \\ &\rho_{X|\pi}(x) = \pi(x) \end{aligned}$$

resulting in *nonparametric empirical Bayes* methods.

The marginal likelihood of a prior $\Pi = \pi$ given a single measurement z is then given by

$$L(\pi \mid z) = \rho_{Z|\pi}(z) = \int_{\mathcal{X}} \rho_{Z|x}(z)\, \pi(x)\, \mathrm{d}x.$$

Since this likelihood, in contrast to the basic Bayesian model above, does not depend on a specific realization of the latent variable X anymore, this allows us to handle multiple measurements coming from independent samples of X correctly using the product distribution:

Definition 5. For finite data $\boldsymbol{z}^M = (z_m \in \mathcal{Z})_{m=1,\dots,M}$ we define the likelihood of a prior π as

$$L\left(\pi \mid \boldsymbol{z}^M\right) := \prod_{m=1}^{M} \rho_{Z|\pi}\left(z_m\right) \tag{2.5}$$

or alternatively in its renormalized logarithmic form as *finite data log-likelihood*

$$\mathcal{L}\left(\pi \mid \boldsymbol{z}^M\right) := \frac{1}{M} \log \rho\left(\boldsymbol{z}^M | \pi\right) = \frac{1}{M} \sum_{m=1}^{M} \log \rho_{Z|\pi}\left(z_m\right).$$

In the case of "infinite data", represented by the probability density ρ_Z of the data-generating random variable Z we define analogously

Definition 6. Let ρ_Z be a probability density on $\mathcal{Z}$. The corresponding *infinite data log-likelihood* is defined as

$$\mathcal{L}^{\infty}\left(\pi \mid \rho_Z\right) := \int_{\mathcal{Z}} \rho_Z\left(z\right) \log \rho_{Z|\pi}\left(z\right) \mathrm{d}z.$$

This definition, also referred to as the cross entropy between ρ_Z and $\rho_{Z|\pi}$, follows from the former in the limit for $m \to \infty$ from the law of large numbers assuming that ρ_Z is indeed the data generating distribution:

$$z_m \overset{i.i.d.}{\sim} \rho_Z \Rightarrow \mathcal{L}\left(\pi \mid \boldsymbol{z}^M\right) \underset{M\to\infty}{\overset{\text{a.s.}}{\longrightarrow}} \mathcal{L}^{\infty}\left(\pi \mid \rho_Z\right). \tag{2.6}$$

The following proposition demonstrates how we can recover the data underlying "true prior" π^* in the infinite data regime by maximizing the corresponding likelihood functional $\mathcal{L}^{\infty}$.

Proposition 7. *Let the hyperparametric model be* well specified, *i.e.* $\exists \pi^* \in \mathcal{M}_1\left(\mathcal{X}\right)$*:*

$$\rho_Z = \rho_{Z|\pi^*}$$

and identifiable *[25, chapter 5], i.e.* $\forall \pi \in \mathcal{M}_1$:

$$\rho_{Z|\pi} = \rho_{Z|\pi^*} \Rightarrow \pi = \pi^*.$$

Then π^* *is the unique maximizer of* $\mathcal{L}^{\infty}$

$$\pi^* = \underset{\pi \in \mathcal{M}_1(\mathcal{X})}{\arg\max} \mathcal{L}^{\infty}\left(\pi\right).$$

Proof. Gibbs' inequality says that

$$\int_{\mathcal{Z}} \rho_Z(z) \log \rho_Z(z) \, \mathrm{d}z \geq \int_{\mathcal{Z}} \rho_Z(z) \log q(z) \, \mathrm{d}z$$

for any probability density q, with equality if and only if $q = \rho_Z$. The claim follows from the assumptions. □

Both assumptions arise rather naturally. If the model is not well specified this merely means the measured data ρ_Z cannot be explained by any prior, thus resulting in an ill-posed problem. If on the other hand the model is not identifiable, which by definition corresponds to the injectivity of the marginal likelihood function $\rho_{Z|\pi}$ as a function of π, there exists another $\pi' \neq \pi^*$ inducing the same measurement distribution so we cannot hope to recover the true prior from the data.

The latter problem can be retracted through lifting the inference problem to equivalence classes of priors leading to the same measurements,

$$\pi \sim \pi' : \Longleftrightarrow \left\| \rho_{Z|\pi} - \rho_{Z|\pi'} \right\|_{L^1(\mathcal{Z})} = 0, \tag{2.7}$$

which, in the non-identifiable case, is all we can hope for.

Though in practice usually only finite data is available, and even though the limiting property (2.6) might give hope that the NPMLE estimate

$$\pi_{ML} := \operatorname*{arg\,max}_{\pi' \in \mathcal{M}_1(\mathcal{X})} \mathcal{L}\left(\pi \mid \boldsymbol{z}^M\right)$$

might approximate π^* properly, one can prove [15, Theorem 21] that the maximizer of $\mathcal{L}$ is a discrete distribution with at most M nodes.

In the field of machine learning this phenomenon, commonly occurring for insufficient data, is referred to as *overfitting* and usually approached by regularization techniques, methods to enforce more regular and smooth solutions.

3. Regularization

To address this problem of irregularity we will introduce two regularization methods, the *maximum penalized likelihood estimation* (MPLE), introducing a penalization term to the former optimization problem, and the *doubly smoothed maximum likelihood estimation* (DS-MLE), based on applying NPMLE after smoothing the data.

3.1. Regularization via a penalization term

For this section, let us fix the data data $\boldsymbol{z}^M$ in Definition(5), denoting $L(\pi) := L(\pi \mid \boldsymbol{z}^M)$.

Definition 8. For a given *roughness penalty* (or *regularization term*) $\phi : \mathcal{M}_1 \to \mathbb{R}$, responsible for penalizing unsmooth or unwanted solutions with high values, the MPLE estimate π_ϕ admits the form

$$\pi_\phi = \arg\max_\pi \log L(\pi) - \phi(\pi). \tag{3.1}$$

This approach also allows for an interpretation in the context of the Bayesian hyperparametric model by identifying the penalty ϕ with the hyperprior ρ_Π on Π via $\rho_\Pi \propto e^{-\phi}$. The posterior for Π is then

$$\rho_{\Pi|Z} \propto L(\pi)\, e^{-\phi(\pi)}$$

and thus the *maximum a posteriori* estimate π_{MAP} for the hyperparametric model corresponds to the MPLE estimate:

$$\pi_{MAP} = \arg\max_\pi L(\pi)\, e^{-\phi(\pi)} = \arg\max_\pi \log L(\pi) - \phi(\pi) = \pi_\phi.$$

One now might might argue that we started with the question of finding the correct prior and just complicated the situation by transferring this problem to the question of the correct hyperprior. While this may be true we argue that the latter can be tackled from a rather abstract, problem independent standpoint, hence leading to a more general answer.

3.2. The mutual-information penalty

Many of the common penalty functions currently in use, penalizing either large amplitudes (e.g. ridge regression [16, section 1.6]) or derivatives (c.f. [9]) of the prior, are not invariant under reparametrizations of the parameter space $\mathcal{X}$ and are rather ad-hoc without a natural derivation. Following a more information-theoretic view Good [8] suggested the use of the differential entropy for the penalty

$$\phi_{H_X}(\pi) := \gamma \int_{\mathcal{X}} \rho_{X|\pi}(x) \log \rho_{X|\pi}(x)\, \mathrm{d}x,$$

with $\gamma \in \mathbb{R}^+$, the regularization constant, determining the degree of smoothing due to this penalty. This prior however is still variant under reparametrizations of $\mathcal{X}$. This means that if two scientists estimate the prior using equivalent models, using different systems of units, they could end up with different estimates. Hence this penalty does not rectify the

problem of subjectivity in the Bayesian method. We therefore look for a penalty which is invariant under coordinate transformations.

Embracing the information theoretic approach we therefore propose the use of the *mutual information* instead of the entropy for the penalty:

Definition 9. Let $A : \Omega \to \mathcal{A}$ and $B : \Omega \to \mathcal{B}$ be two continuous random variables, $\rho_{A,B}$ their joint probability density and ρ_A, ρ_B their respective marginal densities. Their *mutual information* is defined as

$$\begin{aligned}\mathcal{I}(A;B) :&= \int_{\mathcal{A}}\int_{\mathcal{B}} \rho_{A,B}(a,b) \log\left(\frac{\rho_{A,B}(a,b)}{\rho_A(a)\,\rho_B(b)}\right) \mathrm{d}a\mathrm{d}b \\ &= \mathbb{E}_{b\sim B}\left[D_{KL}\left(\rho_{A|b} \parallel \rho_A\right)\right] \\ &= H(B) - H(B;A),\end{aligned}$$

with D_{KL} being the *Kullback-Leibler divergence* from ρ_A to $\rho_{A|b}$

$$D_{KL}\left(\rho_{A|b} \parallel \rho_A\right) := \int_{\mathcal{A}} \rho_{A|b}(a) \log \frac{\rho_{A|b}(a)}{\rho_A(a)} \mathrm{d}a,$$

and $H(B)$, $H(B;A)$ being the *differential entropy* of B respectively the *conditional differential* entropy of B given A

$$\begin{aligned}H(B) &:= -\int_{\mathcal{B}} \rho_B(b) \log(\rho_B)\,\mathrm{d}b \\ H(B;A) &:= \int_{\mathcal{A}} \rho_A(a)\, H(B \mid a)\,\mathrm{d}a \\ &= -\int_{\mathcal{A}} \rho_A(a) \int_{\mathcal{B}} \rho_{B|a}(b) \log\left(\rho_{B|a}\right) \mathrm{d}b\mathrm{d}a.\end{aligned}$$

The mutual information quantifies the "amount of information" that one random variable shares with the respective other, expressed by the information content of the their joint distribution ($\rho_{A,B}$) relative to their joint distribution if they were independent ($\rho_A\rho_B$), weighted by their joint distribution.

We can gain further insights into its meaning by expressing it in terms of another fundamental information-theoretic quantity, the Kullback-Leibler divergence $D_{KL}(A \parallel B)$ from B to A (also called *information gain* or *relative entropy*). It is a measure for the loss of information when considering B as an approximation to A, or consequently in the Bayesian context it is the gain of information revising one's beliefs from the prior B to the posterior A.

Hence the mutual information of A and B corresponds to the expected information gain from the prior to the posterior over A when the measurements are B-distributed. This interpretation gives rise to the following definition:

Definition 10. For $\gamma > 0$ constant we define the *mutual information penalty*

$$\begin{aligned} \phi_I^{\gamma}(\pi): &= -\gamma \mathcal{I}(X \mid \pi; Z \mid \pi) \\ &= -\gamma \int_{\mathcal{X}} \rho_{X|\pi}(x) \int_{\mathcal{Z}} \rho_{Z|x}(z) \log\left(\frac{\rho_{Z|x}(z)}{\rho_{Z|\pi}(z)}\right) \mathrm{d}z\mathrm{d}x. \end{aligned} \tag{3.2}$$

Minimizing this penalty then corresponds to maximizing the amount of shared information between the prior prediction $Z \mid \pi$ and the prior $X \mid \pi$ itself (where we understand these random variables as restrictions of Z respectively X onto the event $\Pi = \pi$). In terms of the Kullback-Leibler formulation this means priors π are rewarded by the amount of information gain expected from their hypothetically induced measurements, hence encoding a notion of non-informativity.

Fortunately the mutual information is furthermore transformation invariant, thus allowing for the application of the mutual information penalty independent of the models parameterization:

Lemma 11. *Let X, Y be two random variables and $\varphi^{-1}: \mathcal{X} \to \tilde{\mathcal{X}}$, $\psi^{-1}: \mathcal{Z} \to \tilde{\mathcal{Z}}$ be diffeomorphisms defining coordinate transformations and corresponding transformed random variables $\tilde{X}$, $\tilde{Z}$ with the densities*

$$\rho_{\tilde{X},\tilde{Z}}(\tilde{x},\tilde{z}) := \rho_{X,Z}(\varphi(\tilde{x}), \psi(\tilde{z})) \left|D\varphi(\tilde{x})\right| \left|D\psi(\tilde{z})\right|$$

$$\rho_{\tilde{X}}(\tilde{x}) := \rho_X(\varphi(\tilde{x})) \left|D\varphi(\tilde{x})\right|, \quad \rho_{\tilde{Z}}(\tilde{z}) := \rho_Z(\psi(\tilde{z})) \left|D\psi(\tilde{z})\right|.$$

Let $\tilde{\pi}(\tilde{x}) := \pi(\varphi(\tilde{x})) \left|D\varphi(\tilde{x})\right|$ define the corresponding pullback onto π and $\tilde{\boldsymbol{z}}^M := (\tilde{z}_m)_{m=1}^M$, $\tilde{z}_m = \psi^{-1}(z_m)$, $m = 1, ..., M$ the transformed data.

1. The mutual information $\mathcal{I}(X;Z)$ is invariant under these coordinate transformations of $\mathcal{X}$ and $\mathcal{Z}$:
$$\mathcal{I}(X;Z) = \mathcal{I}(\tilde{X};\tilde{Z}).$$

2. The likelihood $L\left(\pi \mid \boldsymbol{z}^M\right)$ (2.5) is invariant under these coordinate transformations up to a constant factor $c = \prod_{m=1}^M \left|D\psi(\tilde{\boldsymbol{z}}_m)\right|$.
$$L\left(\pi \mid \boldsymbol{z}^M\right) = cL\left(\tilde{\pi} \mid \tilde{\boldsymbol{z}}^M\right)$$

Proof. According to the change of variables formula

$$\begin{aligned}
&\mathcal{I}(X;Z)\\
=&\int_{\mathcal{Z}}\int_{\mathcal{X}}\rho_{X,Z}(x,z)\log\left(\frac{\rho_{X,Z}(x,z)}{\rho_X(x)\,\rho_z(z)}\right)\mathrm{d}x\mathrm{d}z.\\
=&\int_{\tilde{\mathcal{Z}}}\int_{\tilde{\mathcal{X}}}\rho_{X,Z}(\varphi(\tilde{x}),\psi(\tilde{z}))\log\left(\frac{\rho_{X,Z}(\varphi(\tilde{x}),\psi(\tilde{z}))}{\rho_X(\varphi(\tilde{x}))\,\rho_Z(\psi(\tilde{z}))}\right)|D\varphi(\tilde{x})|\,|D\psi(\tilde{z})|\,\mathrm{d}\tilde{x}\mathrm{d}\tilde{z}\\
=&\int_{\tilde{\mathcal{Z}}}\int_{\tilde{\mathcal{X}}}\rho_{\tilde{X},\tilde{Z}}(\tilde{x},\tilde{z})\log\left(\frac{\rho_{\tilde{X},\tilde{Z}}(\tilde{x},\tilde{z})}{\rho_{\tilde{X}}(\tilde{x})\,\rho_{\tilde{Z}}(\tilde{z})}\right)\mathrm{d}\tilde{x}\mathrm{d}\tilde{z}\\
=&\mathcal{I}(\tilde{X};\tilde{Z}).
\end{aligned}$$

Analogously

$$\begin{aligned}
&L\left(\pi\mid\boldsymbol{z}^M\right)\\
=&\prod_{m=1}^{M}\int_{\mathcal{X}}\rho_{Z|x}(z_m)\,\pi(x)\,\mathrm{d}x\\
=&\prod_{m=1}^{M}\int_{\mathcal{X}}\rho_{Z|\tilde{x}}(\tilde{z}_m)\,\pi(\varphi(\tilde{x}))\,|D\varphi(\tilde{x})|\,|D\psi(\tilde{\boldsymbol{z}}_m)|\,\mathrm{d}\tilde{x}\\
=&\prod_{m=1}^{M}|D\psi(\tilde{\boldsymbol{z}}_m)|\int_{\mathcal{X}}\rho_{Z|\tilde{x}}(\tilde{z_m})\,\tilde{\pi}(\tilde{x})\,\mathrm{d}\tilde{x}\\
=&cL\left(\tilde{\pi}\mid\tilde{\boldsymbol{z}}^M\right).
\end{aligned}$$

□

Corollary 12. *The maximum penalized likelihood estimator $\pi_{\phi_I^\gamma}$ (3.1) with the mutual information penalty ϕ_I^γ (3.2) is invariant under the transformations of $\mathcal{X}$ and $\mathcal{Y}$ as specified in Lemma 11.*

Proof. This follow immediately from the above Lemma, recognizing that the multiplicative constant turns into an additive shift due to the logarithm, which does not change the arg max. □

In the case of an additive measurement error we can simplify the MPLE estimator by only computing the entropy of the prior predictive distribution:

Theorem 13. *For models with additive measurement error $E\sim\rho_E$*

$$Z=\Phi(X)+E$$

the MPLE estimate $\pi_{\phi^\gamma_{H_Z}}$*, penalized by the* Z-entropy

$$\phi^\gamma_{H_Z}(\pi) := -\gamma H(Z \mid \pi),$$

and the mutual information penalty coincide:

$$\pi_{\phi^\gamma_{H_Z}} = \pi_{\phi^\gamma_I}.$$

Proof. In the case of additive measurement error $\rho_{Z|x}$ consists of shifts of ρ_E

$$\rho_{Z|x}(z) = \rho_E(z - \Phi(x))$$

and thus

$$\begin{aligned} H(Z \mid x) &= \int_{\mathcal{Z}} \rho_{Z|x}(z) \log\left(\rho_{Z|x}(z)\right) \mathrm{d}z \\ &= \int_{\mathcal{Z}} \rho_E(z) \log\left(\rho_E(z)\right) \mathrm{d}z \\ &= H(E). \end{aligned}$$

Hence the conditional entropy part is constant and both penalties agree up to an additive constant

$$\begin{aligned} \phi^\gamma_I(\pi) &= -\gamma\left(H(Z \mid \pi) - H(Z; X \mid \pi)\right) \\ &= -\gamma\left(H(Z \mid \pi) - \int_{\mathcal{X}} \rho_{X|\pi}(x) H(Z \mid x) \mathrm{d}x\right) \\ &= -\gamma\left(H(Z \mid \pi) + H(E)\right) \\ &= \phi^\gamma_{H_Z} + \gamma H(E). \end{aligned}$$

Therefore their maxima agree and the estimates are the same. □

Remark 14. For the mentioned additive error model and discrete parameter space $\mathcal{X}$, Klebanov et al. [14, p. 3.1] show that the hyperprior $\rho_\Pi(\pi) :\propto \exp H(Z \mid \pi)$ maximizes the total entropy $H(Z, \Pi)$ of the whole model and furthermore conjecture this to hold as well for continuous $\mathcal{X}$. This derivation from a maximum entropy principle is a further justification for the choice of $\phi^\gamma_{H_Z}$ as a meaningful penalty.

Remark 15. In the case of no data $\boldsymbol{z}^M$ the log-likelihood term of $\pi_{\phi^\gamma_I}$ in (3.1) vanishes. It then matches the definition of reference priors [2] by Berger et al. Therefore this estimation

routine can be seen as an extension of reference priors from a purely non-informative to a cohort-data based empirical Bayes approach.

Remark 16. It can be shown that the mutual information penalty is convex in π, with strict convexity in the identifiable case, as well as that the likelihood function L is convex[14]. The MPLE (3.1) becomes a concave optimization problem for $\phi = \phi_I^\gamma$.

3.3. Regularization by smoothing of the data

Instead of relying on the coarse approximation of the data generating distribution by the empirical distribution $\rho_{\boldsymbol{z}^M} := \frac{1}{M}\sum_{m=1}^M \delta_{z_m} \approx \rho_Z$ and then penalizing overconfident priors, we may as well address the issue of overfitting by using a smooth approximation $\tilde{\rho}_{\boldsymbol{z}^M}$ to $\rho_{\boldsymbol{z}^M}$.

In the following we will apply the idea of smoothing the data by a kernel convolution, introduced by Seo and Lindsay [22] as the DS-MLE, to the empirical Bayes setting.

Definition 17. Let $K : \mathcal{Z} \to \mathbb{R}$ be a kernel density function (i.e. $K \geq 0$, $\int_{\mathcal{Z}} K(z)\,\mathrm{d}z = 1$) and $\boldsymbol{z}^M = \left(z_m \overset{i.i.d.}{\sim} \rho_Z\right)_{m=1}^M$ be M measurements. The *smoothed data density* is then defined as

$$\tilde{\rho}_{\boldsymbol{z}^M}(z) = (\rho_{\boldsymbol{z}^M} * K)(z) := \frac{1}{M}\sum_{m=1}^M K(z - z_m).$$

Note that when smoothing the data we also have to smooth the model (hence *doubly smoothed*) to amount for the additional uncertainty in the data and stay consistent.

That is in the identifiable case with data-generating prior π^* (c.f. Proposition 7) we want

$$\tilde{\rho}_{\boldsymbol{z}^M} \overset{M\to\infty}{\longrightarrow} \rho_Z * K =: \tilde{\rho}_{Z|\pi^*} \neq \rho_{Z|\pi^*},$$

to hold.

Hence the corresponding *smoothed likelihood model* is given by

$$\begin{aligned} \tilde{\rho}_{Z|x} &= \rho_{Z|x} * K \\ \Rightarrow \tilde{\rho}_{Z|\pi} &= \rho_{Z|\pi} * K \end{aligned}$$

The resulting DS-MLE then takes the form

$$\pi_{DS} := \underset{\pi\in\mathcal{M}_1(\mathcal{X})}{\arg\max}\, \tilde{\mathcal{L}}^\infty(\pi \mid \tilde{\rho}_{\boldsymbol{z}^M}),$$

with $\tilde{\mathcal{L}}$ as in Definition (5) with adjusted likelihood.

This estimator is proven to be consistent under weak assumptions on the kernel and likelihood model (c.f. [22]).

Note however that the choice of a kernel K leaves space for debate. Furthermore this procedure is not invariant under reparametrizations of the measurement space $\mathcal{Z}$ for fixed kernels. Hence this approach, although rather natural and simple does not remedy the problem of subjectivity in the Bayesian inference.

4. Numerical schemes

4.1. Monte Carlo approximations

Since the arising integrals are in general not tractable analytically, we will make use of sample based discretization of the continuous spaces $\mathcal{X}$, $\mathcal{Z}$ and use Monte Carlo integration for the corresponding integrals.

1. Given M measurements $\boldsymbol{z}^M = (z_i)_{i=1}^M$ sampled across the population, these are distributed across the marginal measurement distribution $z_i \sim \rho_Z$ by construction of the model. We can hence approximate

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

2. In the case of the parameter space $\mathcal{X}$ we start with an arbitrary sampling $\boldsymbol{x} = (x_k \in \mathcal{X})_{k=1}^K$ distributed according to some density $x_i \sim \rho_A$. We can now approximate any other density distribution ρ_B on $\mathcal{X}$ as an importance sampling with weights

$$\boldsymbol{w} \in \mathcal{W} := \mathcal{M}_1\left(\{1, ..., K\}\right)$$

such that $w_i \propto \frac{\rho_B(x_i)}{\rho_A(x_i)}$. We then have

$$\rho_B \approx \sum_{k=1}^{K} w_k \delta_{x_k}.$$

Let us express the prior π in terms of its weights $\boldsymbol{w}$,

$$\pi \approx \rho_{X|w} := \sum_{k=1}^{K} w_k \delta_{x_k}. \tag{4.1}$$

For ease of notation we will, by slightly abusing it, refer to the discretization w of a continuous density π by its latter name, where appropriate.

We can now approximate the expectation value of any measurable function g under π via

$$\mathbb{E}_{x\sim\pi}\left[g\left(x\right)\right] = \int_{\mathcal{X}} g\left(x\right)\pi\left(x\right)\mathrm{d}x \approx \sum_{k=1}^{K} w_k g\left(x_k\right).$$

For the prior predictive distribution this leads to

$$\begin{aligned}\rho_{Z|\pi}\left(z\right) &= \int_{\mathcal{X}} \rho_{Z|x}\left(z\right)\pi\left(x\right)\mathrm{d}x \\ &\approx \rho_{Z|\boldsymbol{w}}\left(z\right) := \sum_{k=1}^{K} w_k \rho_{Z|x_k}\left(z\right).\end{aligned}$$

Inserting this into the marginal likelihood yields

$$L\left(\pi \mid \boldsymbol{z}^M\right) \approx L\left(\boldsymbol{w} \mid \boldsymbol{z}^M\right) := \prod_{m=1}^{M}\sum_{k=1}^{K} w_k \rho_{Z|x_k}\left(z_m\right).$$

In order to integrate over the density $\rho_{Z|\pi}$ for the entropy hyperprior, we approximate its density by an additional weighted sampling consisting of $\bar{K} > K$ $\mathcal{Z}$-samples generated from the given $\mathcal{X}$-sampling by

$$\bar{\boldsymbol{z}} := \left(\bar{z}_j \sim \rho_{Z|x_{J(j)}}\right)_{j=1}^{\bar{K}}$$

with corresponding weights

$$\bar{w_j} := \frac{w_{J(j)}}{\#J^{-1}\left(J\left(j\right)\right)}.$$

Here $J : \{1, 2, ..., \bar{K}\} \to \{1, 2, ..., K\}$ denotes a surjective index mapping function, mapping from the $\mathcal{Z}$- to the corresponding $\mathcal{X}$-samples indices. The normalizing factor in the weights amounts for the inflation by multiple $\mathcal{Z}$-samples $\bar{z}_i$, $\bar{z}_j$ from a single $\mathcal{X}$-sample in the case of $J\left(i\right) = J\left(j\right)$.

The Monte Carlo approximation to the $\mathcal{Z}$-entropy then takes the form

$$H\left(Z \mid \pi\right) \approx H\left(Z \mid \boldsymbol{w}\right) := -\sum_{j=1}^{\bar{K}} \bar{w}_j \log\left(\sum_{k=1}^{K} w_k \rho_{Z|x_k}\left(\bar{z}_j\right)\right).$$

4.2. EM algorithm

We will first show how to apply the well known *Expectation Maximization* (EM) algorithm for the NPMLE and DS-MLE.

Herefore let us first recapitulate the EM-algorithm following the classic paper of Dempster, Laird and Rubin [4].

We start by defining the *complete data likelihood functio*n

$$L^c(\pi \mid \boldsymbol{x}, \boldsymbol{z}) := \prod_{m=1}^{M} \rho_{X|\pi}(x_m)\, \rho_{Z|x_m}(z_m),$$

the likelihood of a specific prior represented by w given the measurements $\boldsymbol{z} = (z_m)_{m=1}^M$ with corresponding parameters $\boldsymbol{x} = (x_m)_{m=1}^M$, where the completeness is meant in the context of knowing all involved variables, including $\boldsymbol{x}$.

Based on this we define the *expected complete data log-likelihood* of π as the expectation over the posterior $\boldsymbol{x}_n := \left(x_{n,m} \overset{\text{ind.}}{\sim} \rho_{X|\pi_n,z_m}\right)_{m=1}^M$ under the current estimate π_n of the logarithm of the complete data likelihood conditioned on π and the observed cohort-data $\boldsymbol{z}^M$:

$$\begin{aligned} Q(\pi \mid \pi_n) :=& \mathbb{E}_{\boldsymbol{x}_n}\left[\log\left(L^c\left(\pi \mid \boldsymbol{x}_n, \boldsymbol{z}^M\right)\right)\right] \\ =& \mathbb{E}_{\boldsymbol{x}_n}\left[\sum_{m=1}^{m} \log\left(\rho_{X|\pi}(x_{n,m})\, \rho_{Z|\pi}(z_m)\right)\right] \\ =& \sum_{m=1}^{M} \mathbb{E}_{x\sim\rho_{X|\pi_n,z_m}}\left[\log\left(\rho_{X|\pi}(x)\, \rho_{Z|x}(z_m)\right)\right]. \end{aligned} \tag{4.2}$$

The second step follows from the insight that after exchanging integration and summation the log term for the m'ths summand depends only on a single $x_{n,m}$ and thus the other $x_{n,m'}$, $m' \neq m$, get marginalized out.

The EM algorithm works by iteratively maximizing the expected complete-data log-likelihood under the current estimate π_n:

$$\pi_{n+1} := \arg\max_{\pi\in\mathcal{W}} Q(\pi \mid \pi_n).$$

For a proof of uniqueness and convergence in the case of strictly convex likelihoods we refer to [4] and [26].

Let us compute the corresponding formulas for our Monte Carlo approximations.

For the approximations (4.1) of π and π_n in terms of $\boldsymbol{w}$ and $\boldsymbol{w}_n$, respectively,

$$\begin{aligned}\pi &\approx \sum_{k=1}^{K} w_k \delta_{x_k},\\ \pi_n &\approx \sum_{k=1}^{K} w_{n,k} \delta_{x_k},\end{aligned}$$

we have for any measurable function f

$$\begin{aligned}\mathbb{E}_{x\sim\rho_{X|\pi_n,z_m}}\left[f\left(x\right)\right] &= \int_{\mathcal{X}} \rho_{X|\pi_n,z_m}\left(x\right) f\left(x\right) \mathrm{d}x\\ &\approx \sum_{k=1}^{K} w_{n,k} \frac{\rho_{Z|x_k}\left(z_m\right)}{\rho_{Z|\boldsymbol{w_n}}\left(z_m\right)} f\left(x_k\right).\end{aligned}$$

Therefore we can approximate (4.2) by

$$\begin{aligned}Q\left(\pi \mid \pi_n\right) \approx Q\left(\boldsymbol{w} \mid \boldsymbol{w}_n\right) &:= \sum_{m=1}^{M}\sum_{k=1}^{K} w_{n,k} \frac{\rho_{Z|x_k}\left(z_m\right)}{\rho_{Z|\boldsymbol{w_n}}\left(z_m\right)} \log\left(\rho_{X|\boldsymbol{w}}\left(x_k\right)\rho_{Z|x_k}\left(z_m\right)\right)\\ &= \sum_{m=1}^{M}\sum_{k=1}^{K} w_{n,k} \frac{\rho_{Z|x_k}\left(z_m\right)}{\rho_{Z|\boldsymbol{w_n}}\left(z_m\right)} \log\left(w_k \rho_{Z|x_k}\left(z_m\right)\right).\end{aligned}$$

We can furthermore explicitly compute the maximizer to the arising optimization problem

$$\boldsymbol{w}_{n+1} := \arg\max_{\boldsymbol{w}\in\mathcal{W}} Q\left(\boldsymbol{w} \mid \boldsymbol{w}_n\right).$$

A necessary condition is that the gradient of $Q(\cdot \mid \boldsymbol{w}_n)$at $\boldsymbol{w}$ is orthogonal to the tangent

space $T_{\boldsymbol{w}}\mathcal{W}$ of $\mathcal{W}$, which means that all components of the gradient are equal:

$$\begin{aligned}
&\frac{\partial Q\left(\boldsymbol{w} \mid \boldsymbol{w}_n\right)}{\partial w_k} \perp T_{\boldsymbol{w}}\mathcal{W} \\
\Rightarrow &\exists c \in \mathbb{R}\, \forall k = 1, ..., K: \\
&\frac{\partial Q\left(\boldsymbol{w} \mid \boldsymbol{w}_n\right)}{\partial w_k} = \sum_{m=1}^{M} \frac{w_{n,k}}{w_k} \frac{\rho_{Z|x_k}\left(z_m\right)}{\rho_{Z|w_n}\left(z_m\right)} = c \\
\Rightarrow &w_k = \frac{w_{n,k}}{c} \sum_{m=1}^{M} \frac{\rho_{Z|x_k}\left(z_m\right)}{\rho_{Z|w_n}\left(z_m\right)}.
\end{aligned}$$

Since $\sum_k w_k = 1$ we conclude $c = 1/M$ and hence end up with the explicit EM step $\boldsymbol{w}_{n+1} = \psi\left(\boldsymbol{w}_n\right)$:

$$\psi\left(\boldsymbol{w}_n\right)_k := \frac{w_{n,k}}{M} \sum_{m=1}^{M} \frac{\rho_{Z|x_k}\left(z_m\right)}{\rho_{Z|w_n}\left(z_m\right)}. \tag{4.3}$$

Iterative application of this formula henceforth converges to the weights approximating the desired NPMLE estimate π_{ML}.

Applying the Monte Carlo discretization to the DS-MLE setting ([22]) we end up with the same EM-algorithm of the NPMLE applied to the augmented data points $\left(z_m \overset{i.i.d}{\sim} \tilde{\rho}_{\boldsymbol{z}^M}\right)_{m=1}^{M^{\text{aug}}}$ and the smoothed likelihoods (c.f. Section 3.3). Since M data points result in maximally M peaks in the NPMLE [15], a necessary condition for strictly positive weights is that the number of augmented data points is at least that of the parameter-space nodes, $M^{\text{aug}} > K$.

4.3. Optimization for MPLE

In the case of an additive measurement error the Monte Carlo discretized version of the MPLE (3.1) with the mutual information penalty takes the form

$$w_{\phi_I^\gamma} = \arg\max_{\boldsymbol{w} \in \mathcal{W}} O\left(\boldsymbol{w}\right)$$

with objective function

$$\begin{aligned}
O\left(w\right) &= \log L\left(\boldsymbol{w} \mid \boldsymbol{z}^M\right) + \gamma H\left(Z \mid w\right) \\
&= \sum_{m=1}^{M} \log \sum_{k=1}^{K} w_k \rho_{Z|x_k}\left(z_m\right) - \gamma \sum_{j=1}^{\bar{K}} \bar{w}_j \log\left(\sum_{k=1}^{K} w_k \rho_{Z|x_k}\left(\bar{z}_j\right)\right).
\end{aligned}$$

Being a composition of linear and concave functions the objective O is easily shown to be concave, allowing for the use of constraint concave/convex optimization routines.

The high dimensionality of $\boldsymbol{w}$ suggests the usage of derivative based optimization techniques. The derivative of the objective is given by

$$\begin{aligned}\frac{\partial O}{\partial w_k}(w) = &\sum_{m=1}^{M} \frac{\rho_{Z|x_k}(z_m)}{\sum_{k'=1}^{K} w_{k'} \rho_{Z|x_{k'}}(z_m)} \\ &+ \gamma \sum_{j=1}^{\bar{K}} \frac{\bar{w}_j \rho_{Z|x_k}(\bar{z}_j)}{\sum_{k'=1}^{K} w_{k'} \rho_{Z|x_{k'}}(\bar{z}_j)} \\ &+ \gamma \sum_{j \in J^{-1}(k)} \frac{\log\left(\sum_{k'=1}^{K} w_{k'} \rho_{Z|x_{k'}}(\bar{z}_j)\right)}{\#J^{-1}(k)}.\end{aligned}$$

Remark. Instead of first discretizing and then deriving, one might as well try to work with the discretization of the derivative (of the corresponding continuous objective). Whilst this promises to better approximate the gradient of the underlying continuous problem, the gradient does not fit to the objective anymore and may therefore lead to problems with optimization routines relying on correctness of the gradient.

4.4. Markov Chain Monte Carlo

The quality of the Monte Carlo approximations greatly depends on the choice of the importance samples $\boldsymbol{x}$. Whilst an equidistant grid may work well for low dimensional parameter spaces $\mathcal{X}$, its number of grid points increases exponentially with the dimension of $\mathcal{X}$. This so called curse of dimensionality suggests the use of Markov Chain Monte Carlo (MCMC) sampling methods, a popular sampling scheme for probability densities f defined over high-dimensional spaces.

The basic idea of MCMC methods revolves around constructing an ergodic Markov Chain, a stochastic process whose conditional probability for future states depends only on the current state, which has the desired target density f as stationary density. In the limit of infinite sample sizes the samples from the Markov Chain then are distributed according to f.

The probably most common MCMC scheme is the Metropolis–Hastings (MH) algorithm. It works by iteratively sampling a proposal $x_n^{'} \in \mathcal{X}$ "around" the last MCMC sample x_{n-1} according to a prescribed *proposal density* $Q\left(x_n^{'} \mid x_{n-1}\right)$ and accepting or rejecting that proposal in such a way that the resulting MCMCs stationary distribution is f.

Let us define the corresponding Markov process in terms of its *transition probabilities* $P(x_{n+1} \mid x_n)$, starting from the detailed balance condition

$$\begin{aligned} & P(x' \mid x) P(x) = P(x \mid x') P(x') \\ \Leftrightarrow \quad & \frac{P(x)}{P(x')} = \frac{P(x \mid x')}{P(x' \mid x)} \end{aligned}$$

which ensures the existence of a stationary distribution $\pi = P\pi$.

Taking the ansatz of splitting the transition probability into a proposal distribution g and an acceptance distribution A

$$P(x' \mid x) = g(x' \mid x) A(x' \mid x),$$

we end up with

$$\frac{A(x' \mid x)}{A(x \mid x')} = \frac{P(x')}{P(x)} \frac{g(x \mid x')}{g(x' \mid x)}. \tag{4.4}$$

The Metropolis-Hastings choice for the acceptance distribution

$$A(x' \mid x) := \min\left(1, \frac{P(x')}{P(x)} \frac{g(x \mid x')}{g(x' \mid x)}\right)$$

satisfies this equation, since either $A(x' \mid x)$ or $A(x \mid x')$ is 1, while the other equals the desired right hand side of (4.4) (or its inverse).

This choice allows for the formulation of the following theorem.

Theorem 18. *Let $f, g \in \mathcal{M}_1(\mathcal{X})$ with $f(x) > 0$, $g(x) > 0 \forall x \in \mathcal{X}$. Then the Markov chain defined by*

$$P(x' \mid x) = g(x' \mid x) \min\left(1, \frac{f(x')}{f(x)} \frac{g(x \mid x')}{g(x' \mid x)}\right)$$

admits f as unique stationary distribution

$$f = Pf.$$

Proof. The fact that f is indeed a stationary distribution follows by construction, uniqueness follows from irreducibility and aperiodicity which is given due to positivity of P. □

Hence we can use this Markov process to sample a Markov Chain according to the MH-

MCMC algorithm:

1. Start from an arbitrary point x_0
2. Sample a proposal state $x'_{n+1} \sim g\left(x_{n+1} \mid x_n\right)$
3. With probability $A\left(x'_{n+1} \mid x_n\right)$ accept the proposal, i.e. set $x_{n+1} := x'_{n+1}$, otherwise reject it, i.e. set $x_{n+1} := x_n$
4. Increase n by 1 and resume with 2 until sufficiently many states were generated

The speed of convergence to the stationary distribution is strongly influenced by the choice of the proposal distribution. A common choice for the proposal distribution is the normal distribution centered at the current state x_n, but even here the choice of the covariance is vital for rapid mixing of the resulting Markov Chain.

To circumvent the obstacle of choosing a proper proposal covariance we decided to use the adaptive Metropolis (AM) algorithm by Haario et. al [10][10] which tunes the proposal covariance matrix online based on the current samples, according to

$$
\begin{aligned}
g_n(\cdot \mid x_n) &= \mathcal{N}(x, \Sigma_0) && \text{if } n \leq 2d, \\
g_n(\cdot \mid x_n) &= (1-\beta)\mathcal{N}\left(x, \frac{2.38^2}{d}\Sigma_n\right) + \beta\mathcal{N}(x, \Sigma_0) && \text{if } n > 2d,
\end{aligned}
$$

with d being the dimensionality of the sampling space $\mathcal{X}$ and $\frac{2.38^2}{d}$ a scaling constant considered to be optimal for high dimensions [19]. Σ_n is the covariance matrix estimate based on the previous samples $(x_i)_{i=0}^n$, which can be computed recursively, and Σ_0 an initially chosen positive definite covariance matrix. The linear combination with the fixed normal by $0 < \beta < 1$ ensures that the resulting proposal covariance stays positive definite even in the case of singular Σ_n.

The acceptance step remains the same as with the standard MH-MCMC. This Markov chain still admits the desired target density as stationary distribution, assuming it is log-concave outside of some arbitrary region (for a proof see [20]).

5. Application

We will now discuss the application of the developed empirical Bayes method on the basis of a high-dimensional ordinary differential equation model accompanied by real-life measurements. We will therefore introduce the model and obtain a discretization of $\mathcal{X}$ by sampling from the individual Bayesian posteriors by means of MCMC sampling. We will then compute the NPMLE, MPLE and DS-MLE prior estimates and discuss the results.

5.1. The problem

Our physical model, consisting of a system of 33 ordinary differential equations (ODE), models the feedback mechanisms of the prevalent hormones in the female menstrual cycle with a focus on GnRH-receptor binding and was derived by Röblitz et al. [21]. For the defining equations we refer to the original paper.

This system is parametrized by 114 parameters, out of which 21 (the Hill parameters) are considered fixed for the following survey. In [21] the authors provide point estimates for the parameters, although concluding that only 52 were identifiable, as well as initial conditions. We will denote these as *nominal parameters* θ^{nom} and *initial conditions* y_0^{nom} and use them as initial conditions for the Markov chain as well as for the prior computation.

Let us denote the forward solution of the ODE at time t, given parameters $\theta \in \mathbb{R}^{82}$ and initial conditions $y_0 \in \mathbb{R}^{33}$, as

$$\phi\left(t; \theta, y_0\right) \in \mathbb{R}^{33}.$$

Our data consists of blood concentrations of follicle-stimulating hormone (FSH), luteinizing hormone (LH), estradiol (E2) and progesterone (P4) measured from 53 healthy women over thirty days, roughly every second day. This data was collected in the context of PAEON, a collaborative European research project on eHealth. Denote the set of measurements for a single patient m

$$z^m := \left(z_{t,i}^m\right)_{(t,i)\in I_z^m}$$

with $z_{t,i}^m$ being the single measurement of species $i = 1, ..., 4$ at time t and I_z^m denoting the index set of available measurements.

To impose the condition of periodicity of the data onto our inference process we further augment the data by a copy of itself shifted in time by an additional latent parameter representing the period length $\tau > 0$:

$$z^{m,\tau} := \left(\hat{z}_{t,i}^m\right)_{(t,i)\in I_z^{m,\tau}},$$

with augmented index set

$$I_z^{m,\tau} = \bigcup_{(t,i\in I_z^m)} \left\{(t,i), (t+\tau, i)\right\}$$

and augmented measurements

$$\hat{z}^m_{t,i} := \begin{cases} z^m_{t,i} & \text{if } (t,i) \in I^m_z \\ z^m_{t-\tau,i} & \text{otherwise} \end{cases}.$$

Subsuming the latent model parameters θ (82), the initial conditions y_0 (33) and the period length τ we end up with 116 parameters

$$x = (\theta, y_0, \tau) \in X := (\Theta, Y_0, T)$$

for the Bayesian inference.

We model each single as afflicted by a independent Gaussian measurement error with independent componentwise standard deviations of 10% of their respective order of magnitude, estimated from the nominal solution $\phi^{nom}(t) := \phi(t; y_0^{\text{nom}}, \theta^{\text{nom}})$:

$$\sigma_1^{\text{meas}} = 12,\ \sigma_2^{\text{meas}} = 1,\ \sigma_3^{\text{meas}} = 40,\ \sigma_4^{\text{meas}} = 1.5.$$

We end up with the following likelihood function for the parameters x given a single person's data z_m:

$$L(x \mid z^m) = L(\theta, y_0, \tau \mid z^m) = \prod_{(t,i)\in I_z^{m,\tau}} (2\pi)^{-\frac{1}{2}} (\sigma_i^{\text{meas}})^{-2} \exp\left(-\frac{1}{2}\left(\frac{\phi(t;\theta,y_0)_i - z^{m,\tau}_{t,i}}{\sigma_i^{\text{meas}}}\right)^2\right).$$

Remark 19. This likelihood correctly reflects the amount of information available dependent on the number of measurements. A higher number of measurements results in sharper specified likelihood function. This will allow us to correctly treat different patients/measurements together in the upcoming empirical Bayes analysis.

Remark 20. One might argue that this model manipulates the data and hence is no more of the form (2.4). But it is equivalent to the model obtained by just duplicating the data and subsuming the shift operation into a new forward solution operator, hence Theorem 13 still applies.

For the initial Bayesian sampling procedure, necessary for arrival at a discretization of $\mathcal{X}$, we furthermore need to specify an initial prior π_0 on our parameters $X := (\theta, y_0, \tau)$.

Since we did not want to imply any knowledge on θ but their order of magnitude we chose

the uniform prior bounded by $\alpha := 5$ times the multiple of its corresponding nominal parameter value

$$\Theta_i \mid \pi_0 = \mathcal{U}\left(0, \alpha\theta_i^{\text{nom}}\right), \quad i = 1, ..., 82.$$

The prior for the initial conditions $y_{0,i} = \theta_{82+i}$, $i = 1, ..., 33$ is constructed as a mixture of Gaussians centered at the trajectories of the nominal solution

$$Y_0 \mid \pi_0 := \frac{1}{31}\sum_{t=0}^{30} \mathcal{N}\left(\phi^{\text{nom}}(t), \Sigma\right), \quad i = 1, ..., 33,$$

and the covariance Σ being a diagonal matrix with the component-wise covariance estimates

$$\Sigma_{ii} := \text{Cov}\left(\left(\phi_i^{\text{nom}}(t)\right)_{t=0}^{30}\right), \quad i = 1, ..., 33.$$

The prior for the period length θ_{116} was chosen to be Gaussian with mean 28.9 days and a standard deviation of 3.4 days (c.f. [6]),

$$T \mid \pi_0 := \mathcal{N}\left(28.9, 3.4^2\right).$$

5.2. Sampling

We will now compute the individual posterior samples

$$\boldsymbol{x}^m := (x_i^m)_{i=1}^N \overset{\text{i.i.d}}{\sim} \rho_{X|z^m} = \rho_{X|\pi_0} L\left(\theta \mid z^m\right)$$

for each patient m.

Since all sampled parameters $(x_i)_{i=1}^{116}$ are restricted to $\mathbb{R}^+$ but the used AM sampler uses normal proposal densities with global support, we first rescale the original parameters using $\log : \mathbb{R}^+ \to \mathbb{R}$:

$$\tilde{x}_i = \log\left(x_i\right), \quad i = 1, ..., 116.$$

The normal proposals in the log-space now correspond to lognormal proposals in the original parameter space.

However undergoing this transformation we also have to adjust the likelihood function

according to the change of variables formula:

$$\tilde{L}\left(\tilde{x} \mid z\right) = L\left(\exp\left(\tilde{x}\right) \mid z\right) \prod_{i=1}^{116} \tilde{x}_i.$$

Choosing the initial value for the Markov chain according to our nominal values

$$\begin{aligned} x_{0,i} &:= \log \theta_i^{\text{nom}}\,, \quad i = 1, .., 82 \\ x_{0,82+i} &:= \log y_{0,i}^{\text{nom}}, \quad i = 1, .., 33 \\ x_{0,116} &:= \log 28.9, \end{aligned}$$

we may hope to start in a region of high density which henceforth is already representative for the target density and thus expect a relatively short burn-in phase.

In the first runs the initial covariance matrix Σ_0 of the proposal density for the AM sampler was chosen to be small value uniform in each direction. Upon later runs we reused the covariance structure Σ_N of present samplings according to

$$\Sigma_0 := \frac{2.38^2}{d} \Sigma_N,$$

to speed up the adaption process.

We computed 50.000.000 samples $\boldsymbol{x}^m$ of the individual posteriors $\rho_{X|z^m}$ for each patient m at an average speed of around one million samples per hour and core, of which we rejected the first 10.000.000 as a burn-in phase.

We applied the convergence diagnostics by Gelman and Rubin[7] and concluded convergence of a parameter if the potential scale reduction factors 0.975 quantiles were estimated below 1.2. This lead to on average 74 converged parameters (varying per patient). The Heidelberger and Welch diagnostic [11] furthermore assessed componentwise stationarity of the Markov Chains for on average 109 parameters.

Keeping in mind that many of the parameters are probably unidentifiable (hence random walking in 100 dimensional space) we consider these as good results.

Remark 21. Since the MCMC sampling, consuming most of the computational time of the here proposed prior estimation procedure, is easily parallelizable over the individual persons, the proposed scheme is well suited for the use in e.g. clinical studies, where lots of data is available (c.f. [13], Section 4).

5.3. Prior estimation

After thinning the individual posterior samples $(\boldsymbol{x}^m)_{m=1}^{53}$ for computational feasibility, we combined these in order to obtain a grid for the prior estimation.

$$\boldsymbol{x}^M := \bigcup_{m=1}^{M} \boldsymbol{x}^m.$$

The results below were computed with $\#\boldsymbol{x}^M = 2596$.

Note that these samples, which conform to the average density of the individual posterior densities, are $\pi_1 := \psi(\pi_0)$ distributed. Our experiments have shown that this first EM step (4.3) is already quite close to the NPMLE. Since furthermore the MPLE is expected to be a smoothed version of the NPMLE, these points serve as a good basis for the importance sampling:

$$\boldsymbol{x}^M \sim \psi(\pi_0) \approx \pi_{ML} \approx \pi_{\phi_I^\gamma}.$$

We computed the NPMLE and DS-MLE estimates using 100 iterations of the EM algorithm, starting with uniform weights.

In case of the DS-MLE model we smoothed each data point z_i by 50 additional samples. The kernel K was chosen to be a multivariate Gaussian with diagonal covariance Σ^K, computed by Silverman's rule of thumb[23] for kernel density bandwidth estimation, based on the individual data components:

$$\Sigma^K_{i,i} = 0.9 M^{-\frac{1}{5}} \operatorname{Cov}(z_i^m)_{m=1}^M.$$

For the optimization of the MPLE we used the Method of Moving Asymptotes algorithm [24], a globally-convergent, gradient-based optimizer. To guarantee the constraint $\sum_{k=1}^K w_k = 1$ we used the built in Augmented Lagrangian algorithm. This basically amends a penalty to the objective, whose impact is scaled based on the current misfit of the constraint. We furthermore set the upper bounds for each component to 1, and the lower bounds to 10^{-12} to avoid eventual divisions by zero.

The penalty constant γ is chosen by trial and error, until achieving desired smoothing results. It would be interesting to find a rule for its choice, e.g. by information theoretic reasons or cross validation.

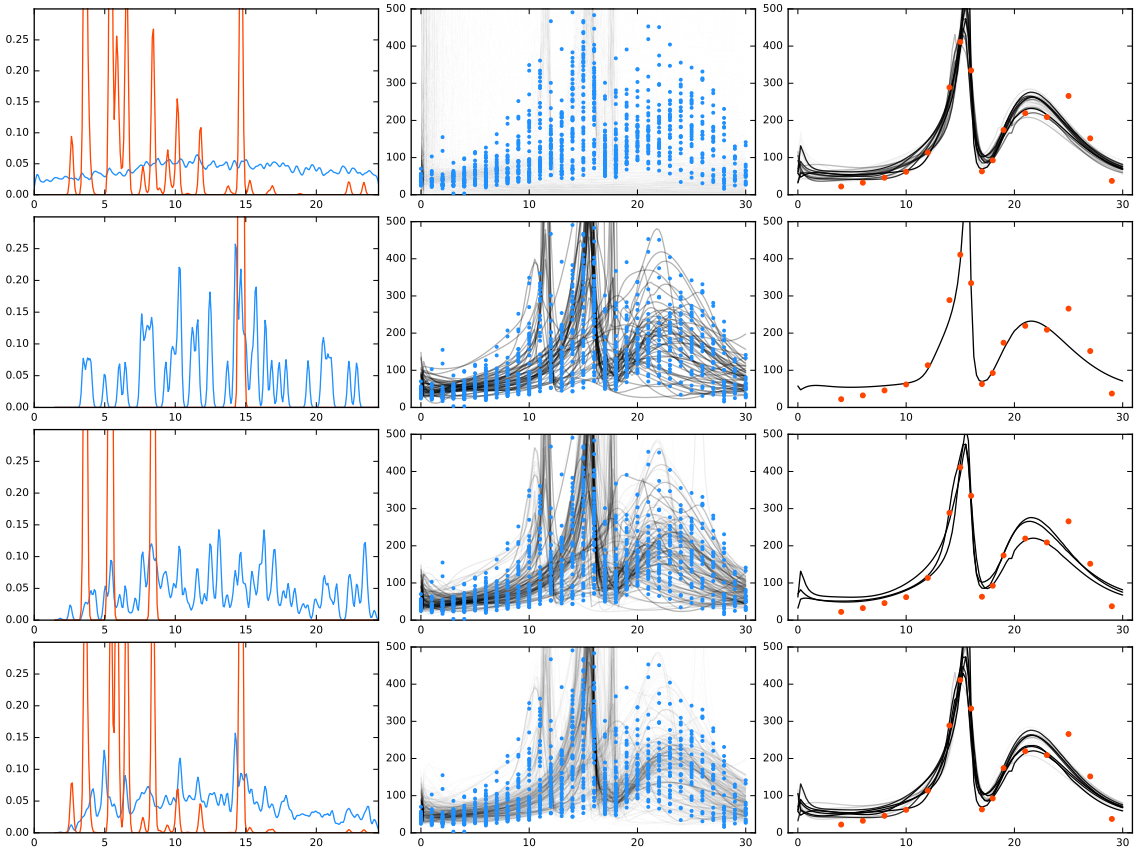

Figure 5.1: π_0, NPMLE, DS-MLE and MPLE (top to bottom) "transition rate constant from PrA2 to SeF1"-marginals of the respective priors (blue) with posteriors (red) for a single subject (left), trajectories sampled from the prior together with the cohort data (middle) and the subject data with trajectories sampled from the posterior (right).

5.4. Results

We will now compare the initial prior π_0 with the NPMLE π_{ML}, the DS-MLE π_{DS} and the information penalized MPLE $\pi_{\phi_I^\gamma}$ for $\gamma = 100$.

Figure 5.1 shows the marginal distributions of the transition rate constant from PrA2 to SeF1, plotted as a kernel density estimate (KDE) with a bandwidth of 0.1, of the individual estimated priors and an individual subjects ($m = 14$) posteriors. It furthermore illustrates the prior respectively posterior predictive distribution as density plots with the cohort- respectively individual data.

For the prior π_0 we combined the MCMC samples $\boldsymbol{x}^M \sim \pi_1$ with the same amount of independent samples from the true density π_0 and then reweighted this sampling with the

inverse of a KDE (with the bandwidths chosen according to Silverman's rule of thumb[23]) at each of the grid-points to approximate a true sampling of π_0, which itself did not lead to any interesting plots for the given, relative low sample size. The first row can therefore only be seen as an approximation to π_0. We can still recognize a lot of unfeasible solutions in the first few days but also a diverse posterior plot.

Looking at the NPMLE (second row), we can observe, keeping in mind that we are looking at a KDE, that the estimated prior consists of a few, relatively high peaks ($\max_i w_i = 0.023$). This expected behavior can also be seen in the prior predictive density plot (middle), where the individual trajectories of the prior approximating the cohort-data are distinctly visible. The corresponding posterior exhibits a single peak ($\max_i w_i = 0.999$) for the most likely solution found.

The DS-MLE (third row) provides a slightly more diverse view with smaller peaks ($\max_i w_i = 0.019$) then the NPMLE and a smoother prior predictive density plot, but still exhibiting a rather erratic variation. The posterior is dominated by three peaks ($\max_i w_i = 0.439$),

Finally the MPLE (last row) provides the smoothest of the estimated priors ($\max_i w_i = 0.014$), while still sharing the informative peak at around $x = 15$ with the NPMLE. The prior predictive density plot is also more densely covered in the feasible region, indicating a higher explanatory power. The individual posterior (right) shares the mode of the NPMLE but also offers a lot of additional trajectories as explanation ($\max_i w_i = 0.258$).

We henceforth conclude that the MPLE with the mutual information penalty indeed leads to satisfactory results, improving on the overconfidence of the MPLE, and being at the very least capable of keeping up with other contemporary methods such as DS-MLE.

6. Conclusion

We have shown that the MPLE with the information penalty ϕ_I^γ is an attractive approach to the empirical Bayes method. It bases on natural, information-theoretic considerations and admits the desirable property of transformation invariance, generalizing the notion of the reference priors to the empirical Bayes framework. Due to its concavity the objective is computationally feasible and its Monte Carlo approximation enables it's application to high-dimensional problems eluding the curse of dimensionality.

We furthermore showed how to apply the developed methods to a real world problem by the means of MCMC sampling, affirming its proficiency in a practical scenario.

A. References

[1] J. Berger et al. "The case for objective Bayesian analysis". In: *Bayesian analysis* 1.3 (2006), pp. 385–402.

[2] J. O. Berger, J. M. Bernardo, and D. Sun. "The formal definition of reference priors". In: *The Annals of Statistics* (2009), pp. 905–938.

[3] J. M. Bernardo. "Reference posterior distributions for Bayesian inference". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1979), pp. 113–147.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the royal statistical society. Series B (methodological)* (1977), pp. 1–38.

[5] B. Efron and C. Morris. "Stein's estimation rule and its competitors-an empirical Bayes approach". In: *Journal of the American Statistical Association* 68.341 (1973), pp. 117–130.

[6] R. J. Fehring, M. Schneider, and K. Raviele. "Variability in the phases of the menstrual cycle". In: *Journal of Obstetric, Gynecologic, & Neonatal Nursing* 35.3 (2006), pp. 376–384.

[7] A. Gelman and D. B. Rubin. "Inference from iterative simulation using multiple sequences". In: *Statistical science* (1992), pp. 457–472.

[8] I. J. Good. "Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables". In: *The Annals of Mathematical Statistics* (1963), pp. 911–934.

[9] I. J. Good and R. A. Gaskins. "Nonparametric roughness penalties for probability densities". In: *Biometrika* (1971), pp. 255–277.

[10] H. Haario, E. Saksman, and J. Tamminen. "An adaptive Metropolis algorithm". In: *Bernoulli* (2001), pp. 223–242.

[11] P. Heidelberger and P. D. Welch. "Simulation run length control in the presence of an initial transient". In: *Operations Research* 31.6 (1983), pp. 1109–1144.

[12] H. 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*. Vol. 186. 1007. The Royal Society. 1946, pp. 453–461.

[13] I. Klebanov, A. Sikorski, C. Schütte, and S. Röblitz. "Empirical Bayes methods for prior estimation in systems medicine". In: (2016). arXiv: 1612.01403.

[14] I. Klebanov, A. Sikorski, C. Schütte, and S. Röblitz. "Empirical Bayes Methods, Reference Priors, Cross Entropy and the EM Algorithm". In: (2016). arXiv: 1612.00064.

[15] B. Lindsay. *Mixture Models: Theory, Geometry, and Applications*. Conference Board of the Mathematical Sciences: NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics, 1995. ISBN: 9780940600324.

[16] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Vol. 382. John Wiley & Sons, 2007.

[17] H. Robbins. "An Empirical Bayes Approach to Statistics". In: (1956), pp. 157–163. URL: http://projecteuclid.org/euclid.bsmsp/1200501653.

[18] H. Robbins. "The empirical Bayes approach to statistical decision problems". In: *The Annals of Mathematical Statistics* 35.1 (1964), pp. 1–20.

[19] G. O. Roberts, A. Gelman, W. R. Gilks, et al. "Weak convergence and optimal scaling of random walk Metropolis algorithms". In: *The annals of applied probability* 7.1 (1997), pp. 110–120.

[20] G. O. Roberts and J. S. Rosenthal. "Examples of Adaptive MCMC". In: *Journal of Computational and Graphical Statistics* 18.2 (2009), pp. 349–367.

[21] S. Röblitz et al. "A mathematical model of the human menstrual cycle for the administration of GnRH analogues". In: *Journal of Theoretical Biology* 321 (2013), pp. 8–27. ISSN: 0022-5193. DOI: 10.1016/j.jtbi.2012.11.020.

[22] B. Seo and B. G. Lindsay. "A universally consistent modification of maximum likelihood". In: *Statistica Sinica* (2013), pp. 467–487.

[23] B. W. Silverman. "Algorithm AS 176: Kernel Density Estimation Using the Fast Fourier Transform". In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31.1 (1982), pp. 93–99. ISSN: 00359254, 14679876. URL: http://www.jstor.org/stable/2347084.

[24] K. Svanberg. "A class of globally convergent optimization methods based on conservative convex separable approximations". In: *SIAM journal on optimization* 12.2 (2002), pp. 555–573.

[25] A. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000. ISBN: 9780521784504.

[26] C. J. Wu. "On the convergence properties of the EM algorithm". In: *The Annals of statistics* (1983), pp. 95–103.

B. Implementation

The introduced algorithms were implemented in Julia, a modern programming language for numerical computing, and bundled in the open-source package *GynC.jl*, available at www.github.com/axsk/GynC.jl.

For installation and usage run the following commands in the Julia REPL:

```
Pkg.clone(https://github.com/axsk/GynC.jl)
using GynC
```

The package was build in a modular way trying to separate the generic empirical Bayes methods and the GynC model specifics in their respective own submodules and should hence be easily extensible to other additive-error models.

It also provides a series of Jupyter-notebooks, containing the code which was ran to generate the results from Section 5.

In the following we will give a short overview over the core components and methods of the software and mention some implementation details.

Empirical Bayes

The code to this module can be found in the `src/eb` directory.

```
LikelihoodModel(xs, ys, zs, datas, measerr, zsampledistr)
```

This is the base-type of the empirical Bayes module, providing the data necessary for the empirical Bayes methods. It holds the following fields:

- `xs`: vector of $\mathcal{X}$-samples (x_k) constituting the grid of the discretization
- `ys`: vector of corresponding pushforward under Φ
- `zs`: vector of additional $\mathcal{Z}$-samples for the $\mathcal{Z}$-entropy computation (c.f. Section 4.1)
- `datas`: vector containing the cohort-data
- `measerr`: specification of the measurement error for likelihood computation. Whilst basic support is added for generic `Distributions.Distribution`, we also provide the type `MatrixNormalCentered` used for the computation of the likelihoods in our application.

- `zsampledistr`: specification of the distribution used for generating samples for the $\mathcal{Z}$-entropy computation

Based on this type the following convenience functions are available

```
em(m::LikelihoodModel, w0, niter)
```

Perform `niter` iterations of the EM-algorithm on model `m`, starting with initial weights `w0`.

```
mple(m::LikelihoodModel, w0, niter, reg, h)
```

Perform a constraint gradient ascent search to optimize the MPLE with $\gamma =$ `reg`, initial weights `w0`, stepsize `h` and `niter` iterations.

```
optimmple(m, reg, w0)
```

Perform a constraint optimization using the Method of Moving Asymptotes algorithm [24].

```
hz(m,w), dhz(m,w), logl(m,w), dlogl(m,w)
```

Compute the log-likelihood, $\mathcal{Z}$-entropy and their derivatives for the given model `m` at weights `w`.

```
smoothedmodel(m, mult)
```

Perform the smoothing of the data and likelihood functions for the DS-MLE.

```
likelihoodmat(xs::Vector,ys::Vector,d::Distribution)
```

Compute the pairwise likelihoods based for a distribution `d`. In the case that `d::MatrixNormalCentered` an optimized version making use of the binomial decomposition and renormalization of the likelihoods for stability is implemented. It also handles incomplete data by excluding NaNs of the computations.

```
gyncmodel(n)
```

Generate a GynC-likelihoodmodel with `n` presampled x-samples.

```
gyncmodel(xs, datas; zmult, sigma)
```

Generate a GynC-likelihoodmodel for the given samples `xs` and `datas` with `zmult` z samples (for the entropy) per x sample and measurement error standard deviation `sigma` times the components magnitude.

```
samplepi0(m), samplepi1(m), subsample(m)
```

Convenience functions for generating respectively loading (precomputed) samples corresponding to the prior and posterior distributions, as well as for subsampling.

GynC

The code to this module can be found in the `src/gync` directory.

`Config`

Base-type managing the configuration for the MCMC sampling (patient data, measurement error, initial proposal variance for MCMC, adaptivity of MCMC, thinning, initial values and priors). For more information refer to `src/gync/model.jl` .

`logprior(c, x), logpost(c, x), llh(c,x)`

Log prior, posterior and likelihood values for given config `c` and sample `x`.

`sample(config, iters)`

Sample `iters` samples using the the AMM sampler from the Mamba.jl package

`batch(configs, iters; dir)`

Sample the given vector of `configs` in parallel on a Slurm cluster and store them in the `dir` directory whenever any of the iterations in the vector `iters` is reached.

`alldatas()`

Return all present cohort-data $\boldsymbol{z}^M$.

`gync(y0, p, ts)`

Compute a trajectory for given initial values `y0`, parameters `p` and times `ts` using the Sundials CVode solver.