NONPARAMETRIC PRIOR ESTIMATION FROM COHORT DATA AND ITS APPLICATION TO SYSTEMS BIOLOGY

1. Introduction

This thesis will cover the development and application of an empirical Bayes method to cohort data.

Since the invention of the computer Bayesian methods, beforehand often untractable to compute, have gained a great amount of importance in the field of reverse problems.

Now, with the Internet of things coming and the progressing digitalisation of scienes, healthcare and *, ever greater amounts of data is beeing gathered, coining the term big data.

This poses new challenges for storage and performance, but also introduces problems of erroneous, incomplete and diverse/cohort * data raising the demand for adjusted analytical methods.

By its nature the Bayesian formalism is very well suited to handle uncertainty and missing information and the here presented method fits (cohort parallelisation..*)

Taking a nonparametric, sampling based approch allows the application of this method to a wide class of problems.

The Bayesian's strength but also weakness is the need to incorporate prior belief/knowledge about the system in question. Whilst the choice of the prior remains a question of debate, we will use cohort data, e.g. measurements of several persons, to improve our prior knowledge by incorporating the individual posteriors into a new, informative, prior.

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 and Y denote continuous random variables and $\rho_{X,Y}(x,y)$ their joint probability density. The *conditional probability* density of Y given the value x for X is

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

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

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

Succesive insertion of these identities leads to the probability density form of *Bayes' theorem*

$$\rho_{X,Y}(x|y) := \frac{\rho_{X,Y}(x,y)}{\rho_Y(y)} = \frac{\rho_{Y|X}(y|x)\,\rho_X(x)}{\rho_Y(y)} = \frac{\rho_{Y|X}(y|x)\,\rho_X(x)}{\int_x \rho_{Y|X}(y|x')\,\rho_X(x')\,\mathrm{d}x'}. \tag{2.1}$$

This formula, constituting the heart of Bayesian statistics, tells us how to reconstruct the *posterior distribution* $\rho_{X|Y}(x|y)$ of the unknown parameter X given data Y, using the *likelihood* $\rho_{Y|X}(y|x)$ 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 prior $\rho_X(x)$ and posterior $\rho_{X|Y}(x|y)$ both are probability densities in x, whilst the likelihood $\rho_{Y|X}(y|x)$ would be in y but is not in x, which is why it is often called the *likelihood function* $L(x \mid y)$ for emphasis.

Also note that the denominator, the *evidence,* does not depend on x and thus is merely a scaling constant, preserving the probability density property of having measure one. This will come in handy later for Markov Chain Monte Carlo sampling, since we will be able to omit it in crucial calculations.

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'. The phyiscal model is represented as 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 $y \in \mathcal{Y}$. We furthermore model the the data generating measurement process Z as

an independent Gaussian perturbation with prescribed covariance Σ of that state:

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

or shorthand

$$Z \mid X \sim \mathcal{N}(\Phi(X), \Sigma),$$

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 probabilty of measuring a single measurement z, given the underlying parameter x.

Assuming the *prior distribution* X was known, this would enable us to compute the *posterior* $X(x|z)$ given some measurement z by straightforward application of the Bayes' theorem equation (2.1).

Note also that whilst this model is extensible to multiple measurements using the product of likelihoods as likelihood, this is only valid under the assumption of identically distributed likelihoods, which corresponds to the assumption that the same parameter x is underlying the different measurements. This would be the right model if all measurements result from the same realization of X (the same subject), but is wrong assuming different measurements correspond to independant draws from the prior X (multiple subjects).

2.3. **The empirical Bayes model.** robbins efron emalgo

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 the so called *empirical Bayes methods.* In that context we extend the model by conditioning the prior ρ_X on a hyperparater Π, the prior 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.

$$\rho_{X|\pi} = \mathcal{N}(\pi, I), \pi \in \mathrm{Im}(\Pi) = \mathbb{R}^n,$$

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

$$\rho_{X|\pi} = \pi,\, \pi \in \mathrm{Im}\left(\Pi\right) = \mathcal{M}_1\left(\mathcal{X}\right) := \left\{\rho \in L^1\left(\mathcal{X}\right) \mid \rho \geq 0,\, \|\rho\|_{L^1} = 1\right\},$$

resulting in a *nonparametric empirical Bayes* method.

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

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

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 different samples of X correctly using the product distribution:

Definition 2. For finite data/measurements $\boldsymbol{z}^M = \left(z_m \in \mathcal{Z}\right)_{m=1,\ldots,M}$ we define the likelihood of a prior π as

$$\rho\left(\boldsymbol{z}^M|\pi\right) := \prod_{m=1}^{M} \rho_Z\left(z_m \mid \Pi = \pi\right)$$

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

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

For "infinite data/measurements", represented in the form of a data-generating probabilty distribution p_Z, we define the corresponding *infinite data log-likelihood*

$$\mathcal{L}_{cc}\left(\pi \mid \rho_Z\right) := \int \rho_Z\left(z\right)\log\rho_Z\left(z \mid \pi\right)\mathrm{d}z.$$

The latter definition follows from the former in the limit for $m \to \infty$ assuming that p_Z is indeed the data generating distribution:

$$z_m \overset{i.i.d.}{\sim} \rho_Z \Rightarrow \mathcal{L}_{cd}\left(\pi \mid \boldsymbol{z}^M\right) \overset{M\to\infty}{\longrightarrow} \mathcal{L}_{cc}\left(\pi \mid \rho_Z\right)$$

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 L_{cc}.

Proposition 3. *Let the hyperparametric model be* well specified, *i.e.*

$$\exists \pi^* \in \mathcal{M}_1(\mathcal{X}) : \rho_Z = \rho_{Z|\pi^*}$$

and identifiable [12, chapter 5]

$$\rho_{Z|\pi'} = \rho_{Z|\pi^*} \Rightarrow \pi' = \pi^* \quad \forall \pi' \in \mathcal{M}_1.$$

We then have that

$$\pi^* = \underset{\pi' \in \mathcal{M}_1(\mathcal{X})}{\arg\max}\, L_{cc}(\pi').$$

Proof. □

Both of the assumptions arise nather 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 ρ_Z, there exists another $\pi' \neq \pi^*$ inducing the same measurements so we cannot hope to recover the right prior candidate from the data.

This can be retracted though by lifting the inference problem to equivalence classes of priors leading to the same measurements,

$$\pi \sim \pi' : \Longleftrightarrow \left\| \rho_{Z|\pi} - \rho_{Z|\pi'} \right\|_{L^1(\mathcal{Z})} = 0, \tag{2.2}$$

which is all we can hope for.

3. Regularization

In practice though usually only finite data is available, and even though the limiting property might give hope that maximization of L_{cd} might approximate π^* properly, one can prove [6, Theorem 21] that the maximizer of L_{cd} is a discrete distribution with at most M nodes.

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

3.1. **Regularization via a penalization term.** To address this problem we will introduce two such regularization methods, the *maximum penalized likelihood estimation* (MPLE), introducing a penalization term to the former optimization problem, and the *doubly smoothed maximum likelihood estimation* (DSMLE), lifting the problem from the finite to the infinite data regime by smoothing the measurements and thus approximating the continuous data-generating distribution ρ_Z.

For a given likelihood function L and a *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 admits the form

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

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

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

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

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

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. **Choice of the penalty.** Many of the common penalty functions currently in use, penalizing either large amplitudes (e.g. ridge regression [7, section 1.6]) or derivatives (c.f. [4]) 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 [3] suggested the use of the differential entropy for the penalty

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

with $\gamma \in \mathbb{R}^+$ being a parameter determining the degree of smoothing due to this penalty. This prior however is still variant under reparametrizations of $\mathcal{X}$ due to the log-nonlinearity. This means that if two scientists estimate the prior using equivalent models, e.g. by using different systems of units, they would 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 4. 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} beeing 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 $H(B)$, $H(B;A)$ beeing the differential entropy of B and the conditional differential entropy of B given A

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

Lemma 5. *Let A, B as above, and $\varphi^{-1} : \mathcal{A} \to \tilde{\mathcal{A}}$, $\psi^{-1} : \mathcal{B} \to \tilde{\mathcal{B}}$ be diffeomorphisms defining the coordinate transformations and corresponding transformed random variables $\tilde{A}$, $\tilde{B}$ with the densities*

$$\rho_{\tilde{A},\tilde{B}}\left(\tilde{a},\tilde{b}\right) := \rho_{A,B}\left(\varphi\left(\tilde{a}\right),\psi\left(\tilde{b}\right)\right)\left|D\varphi\left(\tilde{a}\right)\right|\left|D\psi\left(\tilde{b}\right)\right|$$

$$\rho_{\tilde{A}}\left(\tilde{a}\right) := \rho_{A}\left(\varphi\left(\tilde{a}\right)\right)\left|D\varphi\left(\tilde{a}\right)\right|, \quad \rho_{\tilde{B}}\left(\tilde{b}\right) := \rho_{B}\left(\psi\left(\tilde{b}\right)\right)\left|D\psi\left(\tilde{b}\right)\right|.$$

The mutual information $\mathcal{I}(A;B)$ is then invariant under these coordinate transformations of $\mathcal{A}$ and $\mathcal{B}$:

$$\mathcal{I}(A;B) = \mathcal{I}(\tilde{A};\tilde{B}).$$

Proof. According to the change of variables formula

$$\begin{aligned}
&\mathcal{I}(A;B)\\
&= \int_{\mathcal{B}}\int_{\mathcal{A}} \rho_{A,B}(a,b)\log\left(\frac{\rho_{A,B}(a,b)}{\rho_A(a)\,\rho_B(b)}\right)\mathrm{d}a\mathrm{d}b.\\
&= \int_{\tilde{\mathcal{B}}}\int_{\tilde{\mathcal{A}}} \rho_{A,B}\left(\varphi\left(\tilde{a}\right),\psi\left(\tilde{b}\right)\right)\log\left(\frac{\rho_{A,B}\left(\varphi\left(\tilde{a}\right),\psi\left(\tilde{b}\right)\right)}{\rho_A\left(\varphi\left(\tilde{a}\right)\right)\rho_B\left(\psi\left(\tilde{b}\right)\right)}\right)\left|D\varphi\left(\tilde{a}\right)\right|\left|D\psi\left(\tilde{b}\right)\right|\mathrm{d}\tilde{a}\mathrm{d}\tilde{b}\\
&= \int_{\tilde{\mathcal{B}}}\int_{\tilde{\mathcal{A}}} \rho_{\tilde{A},\tilde{B}}\left(\tilde{a},\tilde{b}\right)\log\left(\frac{\rho_{\tilde{A},\tilde{B}}\left(\tilde{a},\tilde{b}\right)}{\rho_{\tilde{A}}\left(\tilde{a}\right)\rho_{\tilde{B}}\left(\tilde{b}\right)}\right)\mathrm{d}\tilde{a}\mathrm{d}\tilde{b}\\
&= \mathcal{I}(\tilde{A};\tilde{B})
\end{aligned}$$

□

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

We can gain further insights into its meaning by expressing it in terms of another fundamental information-theoretic quantity, the Kullback-Leibler divergence.

The Kullbach-Leibler divergence (also called *information gain* or *relative entropy*) is a measure for the loss of information when considering B as a 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.

that is the mutual information of X and Y corresponds to the expected information gain from the prior to the posterior over $\mathcal{X}$ when the measurements are Y-distributed.

Definition 6. For $\gamma > 0$ constant we define the *information penalty*

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

Minimizing this penalty then corresponds to maximizing the amount of shared information between the prior-predictive distribution $\rho_{Z|\pi}$ and the prior $\rho_{X|\pi}$ itself. In terms of the Kullback-Leibler formulation this means priors π are rewarded by the amount of information gain expected from their induced measurements hence leading to non-informativity.

Corollary 7. *The maximum penalized likelihood estimator π_{MPLE} 3.1 with the information penalty Φ_I 3.2 is invariant under transformations of $\mathcal{X}$ and $\mathcal{Y}$ as in 5.*

The

In the case of an additive measurement error we can simplify the MPLE estimator using its formulation in the entropy form

Theorem 8. *For models with additive measurement error $E \sim \rho_E$*

$$Z = G(X) + E$$

the MPLE estimates with the Z-entropy penalty

$$\Phi_{E_Z}(\pi) := -\gamma H(Z \mid \pi)$$

and the information penalty coincide:

$$\pi_{\Phi_I} = \pi_{\Phi_{E_Z}}.$$

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

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

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

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

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

Remark 9. For the mentioned additive error model and discrete parameter space $\mathcal{X}$, [?, 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. This derivation from a maximum entropy principle, which they conjecture to hold as well in the more general setting with continious $\mathcal{X}$, back the choice of Φ_{H_Z} and hence its generalization $\Phi_{\mathcal{I}}$ as a meaningful penalty.

3.3. **Regularization by smoothing of the data.** Instead of relying on the coarse approximation of the data generating distribution by the empiricial distribution $\rho_{\boldsymbol{z}^M} \approx \rho_Z$ and then penalizing overconfident priors, we may as well adress the issue of overfitting by providing a smooth approximation ρ_Z^{appr} to ρ_Z.
Seo and Lindsay [11] introduced the *doubly-smoothed maximum likelihood estimator* (DS-MLE) based on a kernel density estimate of ρ_Z.

Definition 10. Let $K : \mathcal{Z} \to \mathbb{R}$ be a kernel density function and $z_m \overset{i.i.d.}{\sim} \rho_Z$, $m = 1, ..., M$ be M measurements. The *smoothed data density* is then defined as

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

The corresponding *smoothed likelihood model* is given by

$$\tilde{\rho}_{Z|x} = \rho_{Z,x} * K,\ \tilde{\rho}_{Z|\pi} = \rho_{Z|\pi} * K.$$

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:

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

where π^* is the data-generating prior as in 3.

The resulting DS-MLE then takes the form

$$\pi_{DS} := \operatorname*{arg\,max}_{\pi\in\mathcal{M}_1(\mathcal{X})} L_{cc}\left(\pi \mid \tilde{\rho}_{\boldsymbol{z}^M}\right)$$

and is proved to be consistent under weak assumptions on the kernel and likelihood model (c.f. [11]).

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

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 continious spaces $\mathcal{X}$, $\mathcal{Z}$ and use Monte Carlo integration for the corresponding integrals.

(1) Given M measurements $\boldsymbol{z} = (z_i)_{i=1}^M$ 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

$$\rho_Z \approx \frac{1}{\#\boldsymbol{z}} \sum_{z\in\boldsymbol{z}} \delta_z.$$

(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 a density $x_i \sim \rho_X$. We can now approximate any other density distribution ρ_Y on $\mathcal{X}$ as an importance sampling with weights $\boldsymbol{w} = (w_k)_{k=1}^K, \quad w_k \geq 0, \quad \sum_{k=1}^K w_k = 1$ such that

$w_i \propto \frac{\rho_Y(x_i)}{\rho_X(x_i)}$. We then have

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

Inserting these approximations into above integrals of interest we end up with the following Monte Carlo approximations.

Let

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

The prior predictive distribution can be approximated by

$$\rho_{Z|\pi}(z) = \int_{\mathcal{X}} \rho_{Z|x}(z)\,\pi(x)\,\mathrm{d}x \approx \rho_{Z|w}(z) := \sum_{k=1}^{K} w_k \rho_{Z|x_k}(z)$$

Inserting this into the marginal likelihood leads to

$$L\left(\pi \mid \boldsymbol{z}^M\right) \approx L\left(w \mid \boldsymbol{z}^M\right) := \prod_{m=1}^{M} \sum_{k=1}^{K} w_k L_{mk}.$$

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

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

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 $J(i) = J(j)$.

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

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

4.2. **EM algorithm for NPMLE and DS-MLE.** Equipped with these formulas we can now tackle the problem of the Maximum Likelihood estimation.

Herefore we introduce the *Expectation Maximization* (EM) algorithm following the classic paper of Dempster, Laird and Rubin [1].

We start by defining the *complete data likelihood function*

$$\begin{aligned} L^c(\boldsymbol{x}, \boldsymbol{z} \mid w) :&= \prod_{m=1}^{M} \rho_{X|w}(x_m)\, \rho_{Z|w}(z_m) \\ &= \prod_{m=1}^{M} w_m \sum_{k=1}^{K} w_k \rho_{Z|x_k}(z_m) \end{aligned}$$

the likelihood of a specific prior represented by w given the measurements $\boldsymbol{z} = (z_m)_{m=1}^M$ as well as the 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 with respect to the estimate π_n, the expectation over the log of the compete data likelihood conditioned on the current estimate π, i.e.

$$\begin{aligned} Q(w \mid w_n) :&= \mathbb{E}_{\boldsymbol{X}_n}\left[\log\left(L^c(\boldsymbol{X}, \boldsymbol{z} \mid w)\right)\right], \\ &= \mathbb{E}_{\boldsymbol{X}_n}\left[\sum_{m=1}^{m} \log\left(\rho_{X|w}(x_m)\, \rho_{Z|w}(z_m)\right)\right] \\ &= \sum_{m=1}^{M} \mathbb{E}_{x \sim \rho_{X|w_n, z_m}}\left[\log\left(\rho_{X|w}(x)\, \rho_{Z|w}(z_m)\right)\right] \\ &= \sum_{m=1}^{M} \sum_{k=1}^{K} \frac{w_{n,k} \rho_{Z|x_k}(z_m)}{\rho_{Z|w_n}(z_m)} \log\left(w_k \rho_{Z|x_k}(z_m)\right) \end{aligned}$$

where $\boldsymbol{X}_n = \left(x_m \overset{\text{i.i.d.}}{\sim} \rho_{X|w_n, z_m}\right)_{m=1}^M$. The second step follows from the insight that after exchanging integration and summation each log term depends only on a single x_m and hence the other x_m's get marginalized out.

The EM algorithm works by iteratively maximizing the expected complete-data log-likelihood under the current estimate w_n over the space of admissible weightings $\mathcal{W} := w^* \in \mathcal{M}_1(\{x_1, ..., x_K\})$

$$w_{n+1} := \arg\max_{w^* \in \mathcal{W}} Q(w^* \mid w_n).$$

Proof of convergence and uniqueness.

We can furthermore explicitly compute the maximizer w^*:

A necessary condition is that the gradient at w^* is orthogonal to the tangent space of $\mathcal{W}$, which means that all components of the gradient are the same:

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

Since $\sum_k w_k^* = 1$ we conclude $c = 1/M$ and hence end up with the explicit EM step:

$$w_{n+1,k} = \frac{w_{n,k}}{M} \sum_{m=1}^{M} \frac{\rho_{Z|x_k}\left(z_m\right)}{\rho_{Z|w_n}\left(z_m\right)}.$$

4.3. **Optimization for MPLE.**

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 small dimensional parameter spaces $\mathcal{X}$, its number of samples/gridpoints increases exponentially with the dimension of $\mathcal{X}$. This so called curse of dimensionality suggests the use of Markov Chain Monte Carlo (MCMC) sampling, 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 taget 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 p*roposal 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 Markoc process in terms of its *transition probabilities* $P\left(x_{n+1} \mid x_n\right)$, starting from the detailed balance condition

$$\begin{aligned} P\left(x' \mid x\right) P\left(x\right) = & \qquad P\left(x \mid x'\right) P\left(x'\right) \\ \Leftrightarrow \frac{P\left(x\right)}{P\left(x'\right)} = & \qquad \frac{P\left(x \mid x'\right)}{P\left(x' \mid x\right)}, \end{aligned}$$

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

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

we end up with

$$\frac{A\left(x' \mid x\right)}{A\left(x \mid x'\right)} = \frac{P\left(x'\right)}{P\left(x\right)} \frac{g\left(x \mid x'\right)}{g\left(x' \mid x\right)}. \tag{4.1}$$

The Metropolis-Hastings choice for the acceptance distribution

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

satisfies this equation, since either $A\left(x' \mid x\right)$ or $A\left(x \mid x'\right)$ is 1, while the other equals the desired right hand side of 4.1.

This choice allows for the formulation of the following theorem.

Theorem 11. *Let* $f, g \in \mathcal{M}_1\left(\mathcal{X}\right)$ *with* $f\left(x\right) > 0, g\left(x\right) > 0 \, \forall x \in \mathcal{X}$. *Then the Markov Process defined by*

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

admits f *as unique stationary distribution*

$$f = Pf.$$

Proof. That f indeed is a stationary distribution follows by construction, uniqueness follows from irreducibility which is given due to positivity of P. □

We hence 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 $A\left(x'_{n+1} \mid x_n\right)$ set $x_{n+1} := x'_{n+1}$, otherwise set $x_{n+1} := x_n$
(4) Resume with 2 until sufficient states were generated

The speed of convergence to the stationary distribution is strongly influenced by the choice of the proposal distribution. While sampling proposals according to the target distribution would lead to the best results, this was not possible firsthand which is why we resorted to MCMC. A common choice for the proposal distribution is the normal distribution, but even here the choice of the covariance is vital for rapid mixing of the resulting Markov Chain.

To circument the obstacle of choosing a proper proposal covariance we decided to use the adaptive Metropolis (AM) algorithm by Haario et. al [5] which tunes the proposal covariance online based on the current samples.

Specifically we chose the version by Roberts and Rosenthal [9] defining the proposal density as

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

with d beeing the dimensionality of the sampling space $\mathcal{X}$, $\frac{2.38^2}{d}$ a scaling constant considered to be optimal for large dimensions [8]. Σ_n is the covariance estimate based on the previous samples $(x_i)_{i=0}^n$ and Σ_0 an initially chosen positive definite covariance. The mixture of the nonrandom normal by $\beta > 0$ 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 [9]).

5. Application

We will now discuss the application of the developed empirical Bayes method on the basis of a large-dimensional ordinary differential model accompanied by real-life measurements. We will hence introduce the model, discuss the MCMC sampling procedure, compute the MLE, 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, 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. [10]. For the 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 [10] the authors furthermore give a point estimate of parameters and initial conditions, which we will denote as *nominal parameters* θ^{nom} and *initial conditions* y_0^{nom}, fitting the model on average to a dataset of 13 women.

Our data consists of blood concentrations of follicle-stimulating hormone (FSH), luteinizing hormone (LH), estradiol (E2) and progesterone (P4) measured from 45 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.

Since we know that the healthy menstrual cycle is periodic, we impose this condition onto our inference process by introducing a further latent parameter p for the period length in days, augmenting the data by a copy of itself after that time.

Since furthermore only 4 of the involved 33 species are measured we cannot assume a specific initial condidition and are thus estimating the initial conditions as well.

Subsuming the latent parameters ($i = 1, ..., 82$), the initial conditions ($i = 83, ..., 115$) and the period length ($i = 116$) we end up with 116 parameters $\theta = (\theta_i)_{i=1}^{116}$ for the estimation.

We furthermore have to specify the priors. For we did not want to imply any knowlege on the parameters but their order of magnitude we chose the uniform prior bounded by the $\alpha := 5$ multiple of its corresponding nominal parameter

$$P(\theta_i) = \mathcal{U}(0, \alpha\theta_i^{\text{nom}}) \quad i = 1, ..., 82.$$

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

$$P(\theta_{82+i}) := \frac{1}{31} \sum_{t=0}^{30} \mathcal{N}(\phi_t^{\text{nom}}, \Sigma) \quad i = 1, ..., 33$$

with reference solutions

$$\phi_t^{nom} := \phi(t; y_0^{\text{nom}}, \theta^{\text{nom}}), \quad t = 0, ..., 30$$

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

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

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. [2])

$$P(\theta_{116}) = \mathcal{N}\left(28.9, 3.4^2\right).$$

Denote the set of measurements for a patient/woman/subject

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

with $z_{t,i}$ beeing the single measurement of species i at time t and I_z denoting the index set of available measurements.

We model each single measurement as independent on the others and afflicted by a Gaussian measurement error with independent componentwise standard deviations of 10% of their respective order of magnitude

$$\sigma_1^{\text{meas}} = 12,\ \sigma_2^{\text{meas}} = 1^2,\ \sigma_3^{\text{meas}} = 40^2,\ \sigma_4^{\text{meas}} = 1.5^2.$$

We end up with the following likelihood function for the parameters θ given a single persons data z

$$L(\theta \mid z) = \prod_{(t,i)\in I_z} \mathcal{N}\left(\phi(t;\theta)_i, \sigma_i^{\text{meas}}\right)(z_{t,i})$$

or explicitly

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

Note that 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 in the upcomming empirical Bayes analysis.

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

$$\tilde{\Theta}_i = \log(\Theta_i), \quad i = 1, ..., 116.$$

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

Undergoing this transformation however we also have to adjust the likelihood function according to the change of variables formula:

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

Chosing 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 burnin phase.

In the first runs the initial covariance Σ_0 of the proposal density for the AM sampler was chosen to be uniform. 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.

5.3. **Prior estimation.** Given the individual samplings

5.4. **Results.** plot mcmc chain

paperplot

2d marginals?

6. Conclusion

References

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

[2] Richard J Fehring, Mary Schneider, and Kathleen Raviele. Variability in the phases of the menstrual cycle. *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, 35(3):376–384, 2006.

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

[4] Irving J Good and Ray A Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, pages 255–277, 1971.

[5] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, pages 223–242, 2001.

[6] B.G. 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.

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

[8] Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. Weak convergence and optimal scaling of random walk metropolis algorithms. *The annals of applied probability*, 7(1):110–120, 1997.

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

[10] Susanna RÃ¶blitz, Claudia StÃ¶tzel, Peter Deuflhard, Hannah M. Jones, David-Olivier Azulay, Piet H. van der Graaf, and Steven W. Martin. A mathematical model of the human menstrual cycle for the administration of gnrh analogues. *Journal of Theoretical Biology*, 321:8 – 27, 2013.

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

[12] A.W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2000.

Appendix

Implementation.