On Hyper-Parameter Estimation in Empirical Bayes: A Revisit of the MacKay Algorithm

Chune Li[1], **Yongyi Mao**[2], **Richong Zhang**[1] and **Jinpeng Huai**[1] *
[1]School of Computer Science and Engineering, Beihang University, Beijing, P. R. China 100191
[2]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, ON K1N 6N5, Canada

Abstract

An iterative procedure introduced in MacKay's evidence framework is often used for estimating the hyper-parameter in empirical Bayes. Despite its effectiveness, the procedure has stayed primarily as a heuristic to date. This paper formally investigates the mathematical nature of this procedure and justifies it as a well-principled algorithm framework. This framework, which we call the MacKay algorithm, is shown to be closely related to the EM algorithm under certain Gaussian assumption.

1 INTRODUCTION

As a bridge between full Bayesian models and completely frequentist models, the empirical Bayesian method (also known as empirical Bayes in short, or type-II maximum likelihood) has been applied to many learning, inference or prediction applications (see. e.g. [Schäfer and Strimmer, 2005, Efron, 2012, Heskes, 2000, Yang et al., 2004, Frost and Savarino, 1986, DuMouchel and Pregibon, 2001]). The generic setup of empirical Bayes consists the observed data $\mathbf{D}$, the model parameter $\mathbf{z}$ that parametrizes the data likelihood function $p(\mathbf{D}|\mathbf{z})$, and the prior distribution $p(\mathbf{z})$ of the model parameter. In the parametric version of empirical Bayes (Figure 1), the prior distribution is parameterized by certain hyper-parameter $\boldsymbol{\alpha}$, namely, as $p(\mathbf{z}|\boldsymbol{\alpha})$, and the philosophy of empirical Bayes is to estimate the hyper-parameter $\boldsymbol{\alpha}$ from the observed data $\mathbf{D}$.

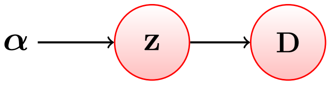


Figure 1: The generic model of empirical Bayesian method

As empirical Bayes treats the model parameter $\mathbf{z}$ as a latent random variable, the estimation of the hyper-parameter $\boldsymbol{\alpha}$ naturally fits in the framework of the EM algorithm [Dempster et al., 1977, Carlin and Louis, 1997], and the EM-based solutions have been developed to solve this problem in various application domains (see, e.g., [Inoue and Tanaka, 2001, Clyde and George, 2000]). Among other approaches to this problem, a technique introduced by MacKay is also widely adopted in practice[Bishop, 1999, Tipping, 2001, Tipping et al., 2003, Wipf and Nagarajan, 2008, Tan and Févotte, 2009].

In his "evidence framework" [MacKay, 1992b, MacKay and Neal, 1994, MacKay, 1995], MacKay considers a hierarchical Bayesian model similar to that in Figure 1 but with one distinction: an additional hyper-prior $p(\boldsymbol{\alpha}|\mathcal{H})$, which depends on the choice $\mathcal{H}$ of model, is placed on the hyper-parameter $\boldsymbol{\alpha}$. In this setting, the evidence framework addresses three levels of inference problems: 1) given the hyper-parameter $\boldsymbol{\alpha}$, inferring $\mathbf{z}$, 2) given the model $\mathcal{H}$, inferring $\boldsymbol{\alpha}$, and 3) evaluating model $\mathcal{H}$. MacKay shows [MacKay, 1995] that the three levels of inference may be combined for prediction and for automatic shrinkage of parameter spaces (namely, Automatic Relevance Determination, or ARD) for neural network regression models. The second-level inference in the evidence framework is closely related to empirical Bayes. In particular, when a flat hyper-prior $p(\boldsymbol{\alpha}|\mathcal{H})$ is placed on $\boldsymbol{\alpha}$, the objective of the second-level inference coincides with the objective of empirical Bayes. For the second-level inference, MacKay introduces a procedure that alternates between inferring $\mathbf{z}$ given $\boldsymbol{\alpha}$ (first-level inference) and inferring $\boldsymbol{\alpha}$ given $\mathbf{z}$. This procedure, although well appreciated in some classical papers (e.g., [Bishop, 1999]) and highly cited in ARD related literature (e.g., [Bishop, 1999, Tipping, 2001, Tan and Févotte, 2009]), is called the *MacKay algorithm* in this paper.

*This work is supported partly by China 973 program (No. 2014CB340305), by the National Natural Science Foundation of China (No. 61300070, 61421003), and by the Beijing Advanced Innovation Center for Big Data and Brain Computing. Richong Zhang is the corresponding author of this work (email: zhangrc@act.buaa.edu.cn)

Since its birth, the MacKay algorithm has been applied to various empirical Bayes models and its performance is often compared with the EM algorithm. For example, the MacKay algorithm is applied to the Bayesian PCA model [Bishop, 1999] and a non-negative matrix factorization model [Tan and Févotte, 2009] for automated shrinkage of the latent-space dimensions. In [Tipping, 2001], the MacKay algorithm is applied to SVM regression models for promoting sparsity, and it is shown to converge faster than the EM algorithm.

Despite its effectiveness, the mathematical principle and optimization objective of the MacKay algorithm are however not well characterized in the literature to date. In MacKay's original exposition [MacKay and Neal, 1994, MacKay, 1995], the second-level inference task is clearly stated, but the justification of the iterative procedure (i.e., the MacKay algorithm) is mainly heuristic. In addition, since the MacKay's algorithm is often implemented with a particular update procedure, known as the fixed-point iteration [Solomon, 2015, Hyvärinen, 1999], or the "MacKay update", the boundary between the *framework* of the MacKay algorithm and MacKay's fixed-point *update rule* is often blurred in the literature. This makes the MacKay algorithm often understood in a narrow sense as this specific fixed-point update rule, rather than as an *algorithm framework*.

In this paper, under a generic formulation of the empirical Bayes model (Figure 1), we re-formulate the MacKay algorithm as a coordinate-ascent procedure for solving a well-defined optimization problem. This optimization problem shares some similarity with the optimization problem underlying the EM algorithm: its objective function is a lower bound of the true objective function defining the optimization objective of empirical Bayes. Also similar to the EM algorithm, this lower bound is not "far" from the true objective function and one of the two update steps in the coordinate-ascent procedure guarantees to make the lower bound meet the true objective function. This understanding justifies the MacKay algorithm (whether or not implemented with the MacKay update) as a well-principled algorithm framework, juxtaposed on equal footing with the EM framework.

Under a specific linear regression model, it has been observed that the MacKay update and the EM algorithm are closely reated [Murphy, 2012]. It is then curious to investigate the relationship between the two algorithms in a more general setting. To that end, we show that as long as the the posterior distribution $p(\mathbf{z}|\mathbf{D}, \boldsymbol{\alpha})$ is a Gaussian distribution, the objective function for the MacKay algorithm is simply a restriction of the EM objective function where two of the three variables are restricted on a curve. Under this Gaussian condition, we show that the MacKay optimization problem is a relaxation of the original optimization problem in empirical Bayes, and that the EM optimization problem is a relaxation of the MacKay optimization problem. In addition, the three problems attain their optimum at the same configuration of the hyper-parameter $\boldsymbol{\alpha}$. These understandings then help to explain why the MacKay algorithm converge faster than the EM algorithm.

The objective of this paper is to rigorously formulate the MacKay algorithm and to investigate its connection to the EM algorithm. We have made an effort to be pedagogical in our presentation. In particular, we use a linear regression model and the Bayesian PCA model as running examples throughout the paper.

2 SETUP

The generic model for empirical Bayes is given in Figure 1, where $\mathbf{D}$ is the observed data, $\mathbf{z}$ is the model parameter, and $\boldsymbol{\alpha}$ is the hyper-parameter. We note that both $\mathbf{z}$ and $\boldsymbol{\alpha}$ can be a scalar, a vector, a matrix or of an arbitrary form. The model is specified by the likelihood function $p(\mathbf{D}|\mathbf{z})$ and the prior distribution $p(\mathbf{z}|\boldsymbol{\alpha})$. The objective of empirical Bayes is then to estimate the hyper-parameter $\boldsymbol{\alpha}$ from the data $\mathbf{D}$.

Let

$$l(\boldsymbol{\alpha}) := \log p(\mathbf{D}|\boldsymbol{\alpha}) = \log \int p(\mathbf{D}|\mathbf{z}) p(\mathbf{z}|\boldsymbol{\alpha}) d\mathbf{z} \qquad (1)$$

be the log-marginal likelihood or the "log-evidence" [MacKay, 1992b] of the hyper-parameter $\boldsymbol{\alpha}$. Then the estimation of $\boldsymbol{\alpha}$ can be naturally formulated as solving the following optimization problem.

Opt-I

Find $\boldsymbol{\alpha}$ that maximizes $l(\boldsymbol{\alpha})$.

We now use the examples of linear regression and Bayesian PCA[Bishop, 1999] to illustrate this. Throughout the paper, we will use $\mathcal{N}(x; \mu, \Lambda)$ to denote the Gaussian density function with variable x, mean μ and covariance matrix Λ, and we will use I_d to denote the $d \times d$ identity matrix, $\mathrm{Tr}(\cdot)$ to denote the trace operator, $\mathrm{Det}(\cdot)$ to denote the determinant operator, $\|\cdot\|$ to denote L2 norm, and $\mathbb{E}_q[\cdot]$ to denote expectation under distribution q.

Linear Regression Example–1 *Let* $\mathbf{D} := \{(x_{(i)}, y_{(i)}) : i = 1, 2, \ldots, n\}$ *be the observed data, where each* $x_{(i)}$ *is a vector in* $\mathbb{R}^d$*, and each* $y_{(i)}$ *is a scalar in* $\mathbb{R}$*. The dependency of* $y_{(i)}$ *on* $x_{(i)}$ *is modelled as*

$$y_{(i)} = \mathbf{z}^T x_{(i)} + \epsilon_{(i)}.$$

Here $\epsilon_{(i)}$ *is a zero-mean Gaussian noise with variance* σ^2*, and* $\mathbf{z}$ *is the model parameter, which is modelled as a d-dimensional spherical Gaussian variable with zero mean and variance* $1/\boldsymbol{\alpha}$*. For simplicity, we assume that the parameter* σ^2 *is known and the objective of empirical Bayes* *is to estimate the hyper-parameter $\boldsymbol{\alpha}$. Then the objective function in Opt-l is:*

$$l(\boldsymbol{\alpha})=\log\int\mathcal{N}(\mathbf{z};0,\frac{1}{\boldsymbol{\alpha}}I_d)\prod_{i=1}^{n}\mathcal{N}(y_{(i)};\mathbf{z}^T x_{(i)},\sigma^2)d\mathbf{z}$$

Bayesian PCA Example–1 *Following [Bishop, 1999], let $\mathbf{D} := \{t_{(i)} : i = 1, 2, \ldots, n\}$ be the observed data, where each $t_{(i)}$ is a vector in $\mathbb{R}^m$. Each observed vector $t_{(i)}$ depends on a latent variable $x_{(i)} \in \mathbb{R}^d$ via*

$$t_{(i)} = \mathbf{z}x_{(i)} + \epsilon_{(i)}.$$

Here $x_{(i)}$ is a zero-mean Gaussian variable with covariance I_d, $\epsilon_{(i)}$ is a spherical Gaussian noise with zero mean and known variance σ^2, and parameter $\mathbf{z} \in \mathbb{R}^{m\times d}$ ($d < m$) is modelled as a random matrix whose kth column $\mathbf{z}_k$ is drawn from a spherical Gaussian distribution with zero mean and variance $1/\boldsymbol{\alpha}_k$. Let $\boldsymbol{\alpha} := \{\boldsymbol{\alpha}_k : k = 1, 2, \ldots, d\}$. Then $\boldsymbol{\alpha}$ is the hyper-parameter on $\mathbf{z}$ and Opt-l has the following objective function:

$$l(\boldsymbol{\alpha}) = \log\left(\int\prod_{k=1}^{d}\mathcal{N}(\mathbf{z}_k;0,\frac{1}{\boldsymbol{\alpha}_k}I_m)\right.$$
$$\left.\prod_{i=1}^{n}\int\mathcal{N}(x_{(i)};0,I_d)\mathcal{N}(t_{(i)};\mathbf{z}x_{(i)},\sigma^2 I_m)dx_{(i)}\right)d\mathbf{z}$$

The optimization problem Opt-l can sometimes be solved easily, for instance, in the above linear regression setting. In practice, however, this problem is usually difficult and requires special algorithmic techniques. The above Bayesian PCA setting is one such example.

The EM approach to Opt-l is well-known. In the remainder of this paper, we develop the MacKay algorithm for this problem. To compare and relate to EM, we also present the EM algorithm in parallel. The above linear regression and Bayesian PCA settings will be carried along our development as illustrative examples.

3 THE TWO ALGORITHMS

In this section, we will show that both the EM algorithm and the MacKay algorithm can be formulated as optimizing a lower bound of the objective function $l(\boldsymbol{\alpha})$ via coordinate ascent. While this is well known for the EM algorithm, it has been quite obscure for the MacKay algorithm.

3.1 THE EM ALGORITHM

The Expectation-Maximization (EM) algorithm [Dempster et al., 1977] is a classical method for maximizing the log-likelihood function or log-posterior density function in which certain latent variables have been integrated over. When applied to the optimization problem Opt-l, the EM algorithm implicitly constructs a lower bound $\mathcal{F}_{\mathrm{EM}}$ of the objective function $l(\boldsymbol{\alpha})$.

$$\begin{aligned}\mathcal{F}_{\mathrm{EM}}(q,\boldsymbol{\alpha}) &:= \mathbb{E}_q\left[\log\frac{p(\mathbf{D}|\mathbf{z})p(\mathbf{z}|\boldsymbol{\alpha})}{q(\mathbf{z})}\right] \\ &= \int q(\mathbf{z})\log\frac{p(\mathbf{D}|\mathbf{z})p(\mathbf{z}|\boldsymbol{\alpha})}{q(\mathbf{z})}d\mathbf{z}\end{aligned} \quad (2)$$

where $q(\cdot)$ is an arbitrary probability distribution on the space of $\mathbf{z}$. By the Jensen's Inequality[Jensen, 1906], the follow result is well-known in the literature of the EM algorithm [Dempster et al., 1977].

Lemma 1. *$\mathcal{F}_{\mathrm{EM}}(q,\boldsymbol{\alpha}) \leq l(\boldsymbol{\alpha})$, where the equality is achieved if and only if $q(\mathbf{z})$ is the posterior distribution $p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha})$.*

Instead of optimizing the original objective function $l(\boldsymbol{\alpha})$, we now define an alternative optimization problem.

OptEM

Find $\boldsymbol{\alpha}$ and a distribution q that maximize $\mathcal{F}_{\mathrm{EM}}(q,\boldsymbol{\alpha})$.

The EM algorithm is then the coordinate ascent solver for OptEM. More precisely, the update rule at the t^{th} iteration of the coordinate ascent is given below.

EM Algorithm

E-Step:

$$\begin{aligned} q^{(t)} &:= \arg\max_q \mathcal{F}_{\mathrm{EM}}(q,\boldsymbol{\alpha}^{(t)}) \\ &= p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha}^{(t)})\end{aligned}$$

M-Step:

$$\begin{aligned} \boldsymbol{\alpha}^{(t+1)} &:= \arg\max_{\boldsymbol{\alpha}} \mathcal{F}_{\mathrm{EM}}(q^{(t)},\boldsymbol{\alpha}) \\ &= \arg\max_{\boldsymbol{\alpha}} \mathbb{E}_{q^{(t)}}\left[\log p(\mathbf{z}|\boldsymbol{\alpha})\right]\end{aligned}$$

We note that by Lemma 1, at the end of E-Step,

$$\mathcal{F}_{\mathrm{EM}}(q^{(t)},\boldsymbol{\alpha}^{(t)}) = l(\boldsymbol{\alpha}^{(t)}). \quad (3)$$

This gives rise to the following lemma [Dempster et al., 1977].

Lemma 2. *The iteration of the EM algorithm continuously increases the log-evidence function $l(\boldsymbol{\alpha})$ and therefore is guaranteed to converge.*

Linear Regression Example–2 *For the linear regression model, let $[x_{(1)}, x_{(2)}, \ldots, x_{(n)}]$ be denoted by a matrix $X \in \mathbb{R}^{d\times n}$ and $[y_{(1)}, y_{(2)}, \ldots, y_{(n)}]^T$ denoted by a vector*

$Y \in \mathbb{R}^n$. The objective function (2) is

$$\begin{aligned}\mathcal{F}_{\mathrm{EM}}(q,\boldsymbol{\alpha}) = {} & \frac{1}{\sigma^2}\sum_{i=1}^{n} y_{(i)}x_{(i)}^T\mathbb{E}_q[\mathbf{z}] \\ & - \frac{1}{2\sigma^2}\mathbb{E}_q\left[\mathbf{z}^T\left(\sum_{i=1}^{n} x_{(i)}x_{(i)}^T + \boldsymbol{\alpha}\sigma^2 I_d\right)\mathbf{z}\right] \\ & - \mathbb{E}_q\left[\log q(\mathbf{z})\right] + \frac{n+d}{2}\log 2\pi + \frac{n}{2}\log\sigma^2 \\ & + \frac{d}{2}\log\boldsymbol{\alpha} - \frac{1}{2\sigma^2}\sum_{i=1}^{n} y_{(i)}^2\end{aligned}$$

It is easy to verify that the posterior distribution of $\mathbf{z}$ *is also Gaussian and the E-Step update becomes*

$$q^{(t)}(\mathbf{z}) = \mathcal{N}\left(\mathbf{z};\mu_{\mathrm{LR}}(\boldsymbol{\alpha}^{(t)}),K_{\mathrm{LR}}(\boldsymbol{\alpha}^{(t)})\right) \qquad (4)$$

where

$$\mu_{\mathrm{LR}}(\boldsymbol{\alpha}) := \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}XX^T + \boldsymbol{\alpha}I_d\right)^{-1}XY, \qquad (5)$$

and

$$K_{\mathrm{LR}}(\boldsymbol{\alpha}) := \left(\frac{1}{\sigma^2}XX^T + \boldsymbol{\alpha}I_d\right)^{-1}. \qquad (6)$$

For the M-Step update, noting that

$$\begin{aligned}\mathcal{F}_{\mathrm{EM}}(q^{(t)},\boldsymbol{\alpha}) = {} & -\frac{\boldsymbol{\alpha}}{2}\left\|\mu_{\mathrm{LR}}(\boldsymbol{\alpha}^{(t)})\right\|^2 \\ & - \frac{\boldsymbol{\alpha}}{2}\mathrm{Tr}\left(K_{\mathrm{LR}}(\boldsymbol{\alpha}^{(t)})\right) + \frac{d}{2}\log\boldsymbol{\alpha} + const,\end{aligned}$$

it is possible to express the maximizing $\boldsymbol{\alpha}$ *for this function directly in terms of* $\boldsymbol{\alpha}^{(t)}$ *as:*

$$\boldsymbol{\alpha}^{(t+1)} = \frac{d}{\left\|\mu_{\mathrm{LR}}(\boldsymbol{\alpha}^{(t)})\right\|^2 + \mathrm{Tr}\left(K_{\mathrm{LR}}(\boldsymbol{\alpha}^{(t)})\right)} \qquad (7)$$

That is, the updates in E-Step and M-Step can be combined into the single update equation (7).

Bayesian PCA Example–2 *For the BPCA model, the objective function (2) is*

$$\begin{aligned}\mathcal{F}_{\mathrm{EM}}(q,\boldsymbol{\alpha}) = {} & -\frac{n}{2}\mathbb{E}_q\left[\log\mathrm{Det}(\mathbf{z}\mathbf{z}^T + \sigma^2 I_m)\right] \\ & - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_q\left[t_{(i)}^T(\mathbf{z}\mathbf{z}^T + \sigma^2 I_m)^{-1}t_{(i)}\right] \\ & - \frac{1}{2}\sum_{k=1}^{d}\boldsymbol{\alpha}_k\mathbb{E}_q\left[\left\|\mathbf{z}_k\right\|^2\right] - \mathbb{E}_q\left[\log q(\mathbf{z})\right] \\ & + \sum_{k=1}^{d}\frac{m}{2}\log\boldsymbol{\alpha}_k - \frac{dn+dm}{2}\log 2\pi.\end{aligned}$$

The M-Step update is then

$$\boldsymbol{\alpha}_k^{(t+1)} = \frac{m}{\mathbb{E}_{q^{(t)}}\left[\left\|\mathbf{z}_k\right\|^2\right]}. \qquad (8)$$

However, the E-Step update of $q^{(t)}$ *can not be expressed in explicit forms and one usually relies on various approximation techniques. For example, a sampling approach [Neal, 1993] may be used for this purpose. Later in this paper, we will discuss the approach that approximates the posterior as a Gaussian.*

3.2 MACKAY ALGORITHM

In [MacKay, 1992a, MacKay, 1992c, MacKay, 1992b, MacKay, 1995], MacKay presented the influential evidence framework that addresses inference as three levels. A heuristic iterative procedure is introduced for the second-level inference, namely, for inferring the hyperparameter. Here, we re-formulate the this procedure as a well-principled algorithm framework, and call it the *MacKay algorithm.*

To begin, note that the objective function $l(\boldsymbol{\alpha})$ can be expressed as

$$l(\boldsymbol{\alpha}) = \log p(\mathbf{D},\mathbf{z}|\boldsymbol{\alpha}) - \log p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha}),$$

for any $\mathbf{z}$ (with $p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha})$ non-zero). Define

$$\mathcal{F}_{\mathrm{MacKay}}(\mathbf{z},\boldsymbol{\alpha}) := \log p(\mathbf{D},\mathbf{z}|\boldsymbol{\alpha}) - \max_{\mathbf{z}'}\log p(\mathbf{z}'|\mathbf{D},\boldsymbol{\alpha}). \qquad (9)$$

The following lemma is easy to verify.

Lemma 3. $\mathcal{F}_{\mathrm{MacKay}}(\mathbf{z},\boldsymbol{\alpha}) \leq l(\boldsymbol{\alpha})$, *where the equality is achieved if and only if* $\mathbf{z} = \arg\max_{\mathbf{z}}\log p(\mathbf{D},\mathbf{z}|\boldsymbol{\alpha})$.

We now introduce another optimization problem.

OptMacKay

Find $\boldsymbol{\alpha}$ and $\mathbf{z}$ that maximize $\mathcal{F}_{\mathrm{MacKay}}(\mathbf{z},\boldsymbol{\alpha})$.

The MacKay algorithm is then defined as the following coordinate ascent procedure for optimizing $\mathcal{F}_{\mathrm{MacKay}}$.

MacKay Algorithm

$\mathbf{z}$-Step:

$$\begin{aligned}\mathbf{z}^{(t)} & := \arg\max_{\mathbf{z}}\mathcal{F}_{\mathrm{MacKay}}(\mathbf{z},\boldsymbol{\alpha}^{(t)}) \\ & = \arg\max_{\mathbf{z}}\log p(\mathbf{D},\mathbf{z}|\boldsymbol{\alpha}^{(t)})\end{aligned}$$

$\boldsymbol{\alpha}$-Step:

$$\boldsymbol{\alpha}^{(t+1)} := \arg\max_{\boldsymbol{\alpha}}\mathcal{F}_{\mathrm{MacKay}}(\mathbf{z}^{(t)},\boldsymbol{\alpha}).$$

At the end of $\mathbf{z}$-Step, by Lemma 3,

$$\mathcal{F}_{\mathrm{MacKay}}(\mathbf{z}^{(t)},\boldsymbol{\alpha}^{(t)}) = l(\boldsymbol{\alpha}^{(t)}). \qquad (10)$$

This property clearly parallels Equation (3) of the EM algorithm. That is, although the MacKay algorithm maximizes the lower bound $\mathcal{F}_{\text{MacKay}}$ of the true objective function $l(\boldsymbol{\alpha})$, the lower bound $\mathcal{F}_{\text{MacKay}}$ is in fact "not far" below $l(\boldsymbol{\alpha})$ and at the end of each $\mathbf{z}$-Step update, the lower-bound meets $l(\boldsymbol{\alpha})$. Then by the coordinate-ascent nature of the MacKay algorithm, we have the following lemma, parallel to Lemma 2 of the EM algorithm.

Lemma 4. *The iteration of the MacKay algorithm continuously increases the log-evidence function $l(\boldsymbol{\alpha})$ and therefore is guaranteed to converge.*

In the MacKay algorithm, it is worth noting that the update in the $\boldsymbol{\alpha}$-Step is usually performed with a "fixed-point iteration" procedure [Solomon, 2015, MacKay, 1992a, Bishop, 1999, Murphy, 2012], which we describe next for self-containedness.

Fixed-Point Iteration Suppose that the equation

$$\partial\mathcal{F}_{\text{MacKay}}(\mathbf{z},\boldsymbol{\alpha})/\partial\boldsymbol{\alpha} = 0$$

can be reduced to the form $\boldsymbol{\alpha} = h(\boldsymbol{\alpha},\mathbf{z})$. The fixed-point iteration approach for the α-Step update in the MacKay algorithm is the following update rule.

$$\boldsymbol{\alpha}^{(t+1)} = h(\boldsymbol{\alpha}^{(t)},\mathbf{z}^{(t)}).$$

Linear Regression Example–3 *In the linear regression model, the lower bound $\mathcal{F}_{\text{MacKay}}$ is*

$$\begin{aligned}\mathcal{F}_{\text{MacKay}}(\mathbf{z},\boldsymbol{\alpha}) =& \frac{1}{\sigma^2}\sum_{i=1}^{n} y_{(i)}x_{(i)}^T\mathbf{z} - \frac{1}{2\sigma^2}\sum_{i=1}^{n} y_{(i)}^2 \\ &- \frac{1}{2\sigma^2}\mathbf{z}^T\left(\sum_{i=1}^{n} x_{(i)}x_{(i)}^T + \boldsymbol{\alpha}\sigma^2 I_d\right)\mathbf{z} \\ &+ \frac{n+d}{2}\log 2\pi + \frac{n}{2}\log\sigma^2 + \frac{d}{2}\log\boldsymbol{\alpha} \\ &+ \frac{1}{2}\log\text{Det}\left(2\pi K_{\text{LR}}(\boldsymbol{\alpha})\right)\end{aligned} \tag{11}$$

The $\mathbf{z}$-Step turns out to be

$$\mathbf{z}^{(t)} = \mu_{\text{LR}}(\boldsymbol{\alpha}^{(t)}).$$

Note

$$\frac{\partial\mathcal{F}_{\text{MacKay}}(\mathbf{z},\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} = \frac{d}{2\boldsymbol{\alpha}} - \frac{1}{2}\|\mathbf{z}\|^2 - \frac{1}{2}\text{Tr}\left(K_{\text{LR}}(\boldsymbol{\alpha})\right).$$

When setting $\frac{\partial\mathcal{F}_{\text{MacKay}}(\mathbf{z},\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} = 0$, we obtain

$$\boldsymbol{\alpha} = \frac{d - \boldsymbol{\alpha}\text{Tr}\left(K_{\text{LR}}(\boldsymbol{\alpha})\right)}{\|\mathbf{z}\|^2}$$

This gives rise to the fixed-point iteration of the $\boldsymbol{\alpha}$-Step:

$$\boldsymbol{\alpha}^{(t+1)} = \frac{d - \boldsymbol{\alpha}^{(t)}\text{Tr}\left(K_{\text{LR}}(\boldsymbol{\alpha}^{(t)})\right)}{\left\|\mathbf{z}^{(t)}\right\|^2}.$$

Bayesian PCA Example–3 *For the BPCA model, the objective function $\mathcal{F}_{\text{MacKay}}$ is*

$$\begin{aligned}\mathcal{F}_{\text{MacKay}}(\mathbf{z},\boldsymbol{\alpha}) =& -\frac{n}{2}\log\text{Det}\left(\mathbf{z}\mathbf{z}^T + \sigma^2 I_m\right) \\ &- \frac{1}{2}\sum_{i=1}^{n} t_{(i)}^T(\mathbf{z}\mathbf{z}^T + \sigma^2 I_m)^{-1}t_{(i)} \\ &- \frac{1}{2}\sum_{k=1}^{d}\alpha_k\|\mathbf{z}_k\|^2 - \max_{\mathbf{z}'} p(\mathbf{z}'|\mathbf{D},\boldsymbol{\alpha}) \\ &+ \sum_{k=1}^{d}\frac{m}{2}\log\boldsymbol{\alpha}_k - \frac{dn+dm}{2}\log 2\pi\end{aligned}$$

The $\mathbf{z}$-Step and $\boldsymbol{\alpha}$-Step updates then become

$$\begin{aligned}\mathbf{z}^{(t)} &= \arg\max_{\mathbf{z}} p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha}^{(t)}) \\ \boldsymbol{\alpha}^{(t+1)} &= \arg\max_{\boldsymbol{\alpha}}\left[-\frac{1}{2}\sum_{k=1}^{d}\boldsymbol{\alpha}_k\left\|\mathbf{z}_k^{(t)}\right\|^2\right. \\ &\qquad\left. - \max_{\mathbf{z}'} p(\mathbf{z}'|\mathbf{D},\boldsymbol{\alpha}) + \sum_{k=1}^{d}\frac{m}{2}\log\boldsymbol{\alpha}_k\right]\end{aligned}$$

Since in general there does not exist closed-form solution for the $\mathbf{z}$-Step update, the two update equations can not be further expressed. In practice, a Gaussian approximation is applied to the posterior function $p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha}^{(t)})$ in order to derive these update equations (see next section).

It is perhaps worth noting that the z-step update of the MacKay algorithm resembles the E-step update of an approximate version of the EM algorithm, known as "Hard EM" (or "Viterbi-EM" in the context of Hidden Markov Models)[Allahverdyan and Galstyan, 2011]. However, the M-step of Hard EM/Viterbi-EM is different from the $\boldsymbol{\alpha}$-step of the MacKay algorithm, due to the fact the OptEM and OptMacKay have different objective functions. It is not clear whether there is a more direct connection between Hard EM and the MacKay algorithm bypassing the generic EM algorithm, although we suspect that the answer is "no".

4 GAUSSIAN APPROXIMATION

As seen above, in both the EM algorithm and the MacKay algorithm, it is desirable to compute the posterior distribution of model parameter, namely, $p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha})$. In the case of EM, this is for updating q in the E-Step and in the case of MacKay, this is for updating $\mathbf{z}$ in the $\mathbf{z}$-Step. For some models, such Bayesian PCA, it is difficult to carry out explicit computation of the posterior. A commonly used technique is to approximate the posterior as a Gaussian density function, namely,

$$p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha}) \approx \mathcal{N}(\mathbf{z};\mu,K),\ \text{for some}\ \mu,\ K. \tag{12}$$

Clearly, the mean vector μ and the covariance matrix K of the Gaussian density depend on the hyper-parameter $\boldsymbol{\alpha}$ and will be denoted by $\mu(\boldsymbol{\alpha})$ and $K(\boldsymbol{\alpha})$ respectively.

When $p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha})$ is a continuous function of $\mathbf{z}$, a common technique for obtain such an approximation (12) is the following [MacKay, 1995].

Let $\widehat{\mathbf{z}}$ maximizes $p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha})$. By Taylor-expanding $\log p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha})$ at $\mathbf{z}=\widehat{\mathbf{z}}$, up to the second-order terms, it is easy to see that

$$\log p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha}) \approx \log p\left(\widehat{\mathbf{z}}|\mathbf{D},\boldsymbol{\alpha}\right)+\frac{1}{2}\left(\mathbf{z}-\widehat{\mathbf{z}}\right)^{T}H(\boldsymbol{\alpha})\left(\mathbf{z}-\widehat{\mathbf{z}}\right)$$

where $H(\boldsymbol{\alpha})$ denotes the Hessian matrix of function $\log p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha})$ at $\mathbf{z}=\widehat{\mathbf{z}}$. This gives the customary approximation of $p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha})$ as [MacKay, 1995]

$$p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha}) \approx \mathcal{N}\left(\mathbf{z};\widehat{\mathbf{z}},-H(\boldsymbol{\alpha})^{-1}\right)$$

Then $\mu(\boldsymbol{\alpha})$ and $K(\boldsymbol{\alpha})$ in the Gaussian approximation (12) can be taken as

$$\mu(\boldsymbol{\alpha})=\widehat{\mathbf{z}},\ K(\boldsymbol{\alpha})=-H(\boldsymbol{\alpha})^{-1}. \tag{13}$$

We note that in (12), the approximation is sometimes accurate, namely, that the strict equality is satisfied. In such cases, the Gaussian approximation as stated in (12) and (13) in fact holds precisely.

Linear Regression Example–4 *As seen in (4), (5) and (6), the posterior distribution $p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha})$ is indeed a Gaussian density function. That is, the Gaussian approximation (12) holds with equality, where $\mu(\boldsymbol{\alpha})=\mu_{\rm LR}(\boldsymbol{\alpha})$ and $K(\boldsymbol{\alpha})=K_{\rm LR}(\boldsymbol{\alpha})$ (defined in (5) and (6) respectively).*

Bayesian PCA Example–4 *Let $\mathbf{Z}$ be the vector representation of matrix $\mathbf{z}$, namely, $\mathbf{Z}$ is a length-md vector obtained by stacking columns of the matrix $\mathbf{z}$. That is, $\mathbf{Z}:=(\mathbf{z}_1^T,\mathbf{z}_2^T,\ldots,\mathbf{z}_d^T)^T$. Let $\widehat{\mathbf{Z}}$ denote the maximizing configuration for function $p(\mathbf{Z}|\mathbf{D},\boldsymbol{\alpha})$, and similarly let $H(\boldsymbol{\alpha})$ denote the Hessian of $\log p(\mathbf{Z}|\mathbf{D},\boldsymbol{\alpha})$ at $\mathbf{Z}=\widehat{\mathbf{Z}}$. The Gaussian approximation (12) then becomes*

$$p(\mathbf{Z}|\mathbf{D},\boldsymbol{\alpha})\approx\mathcal{N}(\mathbf{Z};\mu_{\rm BPCA}(\boldsymbol{\alpha}),K_{\rm BPCA}(\boldsymbol{\alpha})) \tag{14}$$

where

$$\mu_{\rm BPCA}(\boldsymbol{\alpha}):=\widehat{\mathbf{Z}},\ K_{\rm BPCA}(\boldsymbol{\alpha}):=-H(\boldsymbol{\alpha})^{-1}.$$

We note that in this case, (12) is only an approximation. In addition, since $\widehat{\mathbf{Z}}$ and $H(\boldsymbol{\alpha})^{-1}$ are difficult to compute analytically, numerical solutions are usually sought.

In the remainder of this section, we assume that (12) holds with equality and further investigate the optimization problems in the EM and MacKay algorithms.

4.1 EM

Recall that with the EM algorithm, the objective function in the optimization problem is $\mathcal{F}_{\rm EM}$ in (2). Since the optimizing distribution q for any given $\boldsymbol{\alpha}$ is the posterior $p(\mathbf{z}|D,\boldsymbol{\alpha})$, this, under the Gaussian assumption (12) of the posterior, allows us to restrict q to the form $\mathcal{N}(\mathbf{z};u,S)$ parametrized by mean vector u and covariance S. The the objective function $\mathcal{F}_{\rm EM}(q,\boldsymbol{\alpha})$ can then be re-expressed as $\mathcal{F}_{\rm EM}(u,S,\boldsymbol{\alpha})$. That is,

$$\begin{aligned}\mathcal{F}_{\rm EM}(u,S,\boldsymbol{\alpha}) &= \mathbb{E}_{\mathcal{N}(\mathbf{z};u,S)}\left[\log\frac{p(\mathbf{D}|\mathbf{z})p(\mathbf{z}|\boldsymbol{\alpha})}{\mathcal{N}(\mathbf{z};u,S)}\right]\\ &= \mathbb{E}_{\mathcal{N}(\mathbf{z};u,S)}\left[\log p(\mathbf{D}|\mathbf{z})p(\mathbf{z}|\boldsymbol{\alpha})\right] \qquad (15)\\ &\quad+\frac{J}{2}\log 2\pi+\frac{1}{2}\log\mathrm{Det}(S)+\frac{J}{2}\end{aligned}$$

where J is the length of the vector $\mathbf{z}$. The following lemma is established by noting the following two-way factorization of $p(\mathbf{D},\mathbf{z}|\boldsymbol{\alpha})$.

$$p(\mathbf{D},\mathbf{z}|\boldsymbol{\alpha})=p(\mathbf{D}|\mathbf{z})p(\mathbf{z}|\boldsymbol{\alpha})=p(\mathbf{D}|\boldsymbol{\alpha})p(\mathbf{z}|\mathbf{D},\boldsymbol{\alpha}) \tag{16}$$

Lemma 5. *When the Gaussian approximation (12) holds with equality, the function $\mathcal{F}_{\rm EM}$ can be re-expressed as*

$$\begin{aligned}\mathcal{F}_{\rm EM}(u,S,\boldsymbol{\alpha}) &= \log\mathcal{N}(u;\mu(\boldsymbol{\alpha}),K(\boldsymbol{\alpha}))-\frac{1}{2}\mathrm{Tr}(K^{-1}(\boldsymbol{\alpha})S)\\ &\quad+\log p(\mathbf{D}|\boldsymbol{\alpha})+\frac{J}{2}\log 2\pi+\frac{1}{2}\log\mathrm{Det}(S)+\frac{J}{2}\end{aligned}$$

To derive the update rule for the EM algorithm under the Gaussian approximation (12), we prove the following results.

Lemma 6. *For any given S and $\boldsymbol{\alpha}$,*

$$\arg\max_{u}\mathcal{F}_{\rm EM}(u,S,\boldsymbol{\alpha})=\mu(\boldsymbol{\alpha}).$$

Proof: Based on Lemma 5, we express $\mathcal{F}_{\rm EM}(u,S,\boldsymbol{\alpha})$ further.

$$\begin{aligned}\mathcal{F}_{\rm EM}(u,S,\boldsymbol{\alpha}) &= -\frac{J}{2}\log 2\pi-\frac{1}{2}\log\mathrm{Det}(K(\boldsymbol{\alpha}))\\ &\quad-\frac{1}{2}\left(u-\mu(\boldsymbol{\alpha})\right)^{T}K^{-1}(\boldsymbol{\alpha})(u-\mu(\boldsymbol{\alpha}))\\ &\quad-\frac{1}{2}\mathrm{Tr}\left(K^{-1}(\boldsymbol{\alpha})S\right)+\log p(\mathbf{D}|\boldsymbol{\alpha})\\ &\quad+\frac{J}{2}\log 2\pi+\frac{1}{2}\log\mathrm{Det}(S)+\frac{J}{2}.\end{aligned}$$

$$\frac{\partial\mathcal{F}_{\rm EM}}{\partial u}=-\frac{1}{2}\left(K^{-1}(\boldsymbol{\alpha})+(K^{-1}(\boldsymbol{\alpha}))^{T}\right)(u-\mu(\boldsymbol{\alpha}))$$

By setting $\frac{\partial\mathcal{F}_{\rm EM}}{\partial u}$ to zero, we prove the result. □

Lemma 7. *For any u and $\boldsymbol{\alpha}$,*

$$\arg\max_{S}\mathcal{F}_{\rm EM}(u,S,\boldsymbol{\alpha})=K(\boldsymbol{\alpha}).$$

Proof: By the expression of $\mathcal{F}_{\rm EM}(u, S, \boldsymbol{\alpha})$ in Lemma 5,

$$\frac{\partial \mathcal{F}_{\rm EM}}{\partial S} = -\frac{1}{2}\frac{\partial}{\partial S}{\rm Tr}(K^{-1}(\boldsymbol{\alpha})S) + \frac{1}{2}\frac{\partial}{\partial S}\log {\rm Det}(S).$$

By ${\rm Tr}(AB) = \sum_i \sum_j A_{ij}B_{ji}$, we have

$$\begin{aligned}{\rm Tr}(K^{-1}(\boldsymbol{\alpha})S) &= \sum_i\sum_j (K^{-1}(\boldsymbol{\alpha}))_{ij}S_{ji};\\ \frac{\partial}{\partial S_{ij}}{\rm Tr}(K^{-1}(\boldsymbol{\alpha})S) &= (K^{-1}(\boldsymbol{\alpha}))_{ji}\\ \frac{\partial}{\partial S}{\rm Tr}(K^{-1}(\boldsymbol{\alpha})S) &= (K^{-1}(\boldsymbol{\alpha}))^T.\end{aligned}$$

On the other hand, since $\frac{\partial {\rm Det}(S)}{\partial S_{ij}} = {\rm Det}(S)(S^{-1})_{ji}$, we have

$$\begin{aligned}\frac{\partial}{\partial S_{ij}}\log {\rm Det}(S) &= (S^{-1})_{ji}\\ \frac{\partial}{\partial S}\log {\rm Det}(S) &= (S^{-1})^T.\end{aligned}$$

Then

$$\frac{\partial \mathcal{F}_{\rm EM}}{\partial S} = -\frac{1}{2}(K^{-1}(\boldsymbol{\alpha}))^T + \frac{1}{2}(S^{-1})^T.$$

The lemma is then proved by setting this derivative to zero. □

As a corollary of Lemma 6, Lemma 7 and (15), the update rule of the EM algorithm can be established.

Lemma 8. *When the Gaussian approximation (12) holds with equality, the EM algorithm becomes the following EM-Gauss Procedure.*

EM-Gauss Procedure:

E-Step:

$$\begin{aligned}u^{(t)} &:= \mu(\boldsymbol{\alpha}^{(t)})\\ S^{(t)} &:= K(\boldsymbol{\alpha}^{(t)})\end{aligned}$$

M-Step:

$$\boldsymbol{\alpha}^{(t+1)} := \arg\max_{\boldsymbol{\alpha}} \mathbb{E}_{\mathcal{N}(\mathbf{z};u^{(t)},S^{(t)})}\left[\log p(\mathbf{z}|\boldsymbol{\alpha})\right]$$

Linear Regression Example–5 *In the linear regression model, the Gaussian assumption (12) holds true. The E-Step then reduces to (4), which can be integrated into the M-Step. The EM update can then be expressed as a single update equation (7), the same as that in Linear Regression Example-2.*

Bayesian PCA Example–5 *Note that $\mu_{\rm BPCA}(\boldsymbol{\alpha})$ is a vector of length md and $K_{\rm BPCA}(\boldsymbol{\alpha})$ is an $md\times md$ matrix. Let $\mu_{{\rm BPCA},k}(\boldsymbol{\alpha})$ denote the component of $\mu_{\rm BPCA}(\boldsymbol{\alpha})$ corresponding to $\mathbf{z}_k$ component of $\widehat{\mathbf{Z}}$, and $K_{{\rm BPCA},k}(\boldsymbol{\alpha})$ denote the sub-matrix of $K_{\rm BPCA}(\boldsymbol{\alpha})$ that serves as the covariance matrix of $\mathbf{z}_k$. With the Gassian approximation (14) holds with equality, the update (8) of $\boldsymbol{\alpha}$ becomes*

$$\boldsymbol{\alpha}_k^{(t+1)} = \frac{m}{\left\|\mu_{{\rm BPCA},k}(\boldsymbol{\alpha}^{(t)})\right\|^2 + {\rm Tr}\left(K_{{\rm BPCA},k}(\boldsymbol{\alpha}^{(t)})\right)}.$$

4.2 MACKAY

Lemma 9. *When the Gaussian approximation (12) holds with equality, the function $\mathcal{F}_{\rm MacKay}$ in (9) becomes*

$$\mathcal{F}_{\rm MacKay}(\mathbf{z}, \boldsymbol{\alpha}) = \log p(\mathbf{D}, \mathbf{z}|\boldsymbol{\alpha}) + \frac{1}{2}\log {\rm Det}(2\pi K(\boldsymbol{\alpha})).$$

Proof: This lemma follows from $\max_{\mathbf{z}'}\log p(\mathbf{z}'|\mathbf{D}, \boldsymbol{\alpha}) = -\frac{1}{2}\log {\rm Det}(2\pi K(\boldsymbol{\alpha}))$. □

Lemma 10. *When the Gaussian approximation (12) holds with equality, for any $\boldsymbol{\alpha}$,*

$$\arg\max_{\mathbf{z}} \log p(\mathbf{D}, \mathbf{z}|\boldsymbol{\alpha}) = \mu(\boldsymbol{\alpha}).$$

The proof of this lemma follows the same line as that of Lemma 6. The MacKay algorithm under the Gaussian approximation (12) can then be established from Lemma 10 and Lemma 9.

Lemma 11. *When the Gaussian approximation (12) holds with equality, the MacKay algorithm becomes the following MacKay-Gauss Procedure.*

MacKay-Gauss Procedure:

$\mathbf{z}$-Step:

$$\mathbf{z}^{(t)} := \mu(\boldsymbol{\alpha}^{(t)})$$

$\boldsymbol{\alpha}$-Step:

$$\boldsymbol{\alpha}^{(t+1)} := \arg\max_{\boldsymbol{\alpha}}\left[\log p(\mathbf{z}^{(t)}|\boldsymbol{\alpha}) + \frac{1}{2}\log {\rm Det}(K(\boldsymbol{\alpha}))\right]$$

Linear Regression Example–6 *Since the Gaussian assumption (12) holds true in linear regression, the updates of the MacKay algorithm are those in Linear Regression Example–3.*

Bayesian PCA Example–6 *Assuming that the Gassian approximation (12) holds with equality, the $\mathbf{z}$-Step and $\boldsymbol{\alpha}$-Step updates become*

$$\begin{aligned}\mathbf{Z}^{(t)} &= \mu_{\rm BPCA}(\boldsymbol{\alpha}^{(t)})\\ \boldsymbol{\alpha}^{(t+1)} &= \arg\max_{\boldsymbol{\alpha}}\Bigg[-\frac{1}{2}\sum_{k=1}^{d}\boldsymbol{\alpha}_k\left\|\mathbf{z}_k^{(t)}\right\|^2\\ &\quad + \frac{1}{2}\log {\rm Det}\left(K_{\rm BPCA}(\boldsymbol{\alpha})\right) + \sum_{k=1}^{d}\frac{m}{2}\log \boldsymbol{\alpha}_k\Bigg]\end{aligned}$$

When setting $\frac{\partial \mathcal{F}_{\text{MacKay}}(\mathbf{z},\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}_k} = 0$*, we can obtain*

$$\boldsymbol{\alpha}_k = h(\boldsymbol{\alpha}_k, \mathbf{z}) = \frac{m - \boldsymbol{\alpha}_k \mathrm{Tr}\left(K_{\mathrm{BPCA},k}(\boldsymbol{\alpha})\right)}{\|\mathbf{z}_k\|^2}.$$

This gives rise to the fixed-point iteration of the $\boldsymbol{\alpha}$*-Step*

$$\boldsymbol{\alpha}_k^{(t+1)} = \frac{m - \boldsymbol{\alpha}_k^{(t)} \mathrm{Tr}\left(K_{\mathrm{BPCA},k}(\boldsymbol{\alpha}^{(t)})\right)}{\left\|\mathbf{z}_k^{(t)}\right\|^2}$$

4.3 THE RELATIONSHIP BETWEEN MAKCAY AND EM

As is shown earlier, the MacKay and EM algorithms correspond to solving two different optimization problems. However, we will show next that when the Gaussian approximation (12) holds exactly, the two algorithms are closely related.

First note that $\mathcal{F}_{\mathrm{EM}}$ is a trivariate function whereas $\mathcal{F}_{\mathrm{MacKay}}$ is a bivariate function. The theorem below suggests that if the Gaussian approximation of the posterior distribution $p(\mathbf{z}|\mathbf{D}, \boldsymbol{\alpha})$ is exact, then by setting its covariance variable S of $\mathcal{F}_{\mathrm{EM}}$ to the covariance matrix of the posterior, the function $\mathcal{F}_{\mathrm{EM}}$ reduces $\mathcal{F}_{\mathrm{MacKay}}$.

Theorem 1. *When the Gaussian approximation (12) holds with equality,* $\mathcal{F}_{\mathrm{EM}}(u, K(\boldsymbol{\alpha}), \boldsymbol{\alpha}) = \mathcal{F}_{\mathrm{MacKay}}(u, \boldsymbol{\alpha})$*, where* $K(\boldsymbol{\alpha})$ *is defined in (13).*

Proof: Suppose that the Gaussian approximation (12) holds with equality. Invoking (16), we have

$$\begin{aligned}\mathcal{F}_{\mathrm{MacKay}}(\mathbf{z}, \boldsymbol{\alpha}) &= \log p(\mathbf{D}, \mathbf{z}|\boldsymbol{\alpha}) + \frac{1}{2}\log \mathrm{Det}(2\pi K(\boldsymbol{\alpha}))\\ &= \log \mathcal{N}(\mathbf{z}; \mu(\boldsymbol{\alpha}), K(\boldsymbol{\alpha})) + \log p(\mathbf{D}|\boldsymbol{\alpha})\\ &\quad + \frac{1}{2}\log \mathrm{Det}(2\pi K(\boldsymbol{\alpha}))\end{aligned}$$

But by Lemma 5, we have

$$\begin{aligned}\mathcal{F}_{\mathrm{EM}}(\mathbf{z}, K(\boldsymbol{\alpha}), \boldsymbol{\alpha}) &= \log \mathcal{N}(\mathbf{z}; \mu(\boldsymbol{\alpha}), K(\boldsymbol{\alpha}))\\ &\quad - \frac{1}{2}\mathrm{Tr}\left(K^{-1}(\boldsymbol{\alpha})K(\boldsymbol{\alpha})\right) + \log p(\mathbf{D}|\boldsymbol{\alpha})\\ &\quad + \frac{J}{2}\log 2\pi + \frac{1}{2}\log \mathrm{Det}(K(\boldsymbol{\alpha})) + \frac{J}{2}\\ &= \log \mathcal{N}(\mathbf{z}; \mu(\boldsymbol{\alpha}), K(\boldsymbol{\alpha})) + \log p(\mathbf{D}|\boldsymbol{\alpha})\\ &\quad + \frac{1}{2}\log \mathrm{Det}(2\pi K(\boldsymbol{\alpha}))\\ &= \mathcal{F}_{\mathrm{MacKay}}(\mathbf{z}, \boldsymbol{\alpha})\end{aligned}$$

This proves the theorem. □

Theorem 1 suggests that the function $\mathcal{F}_{\mathrm{MacKay}}$ is a restriction of function $\mathcal{F}_{\mathrm{EM}}$. Denote by $\mathfrak{C}$ the set of all $(S, \boldsymbol{\alpha})$ configurations with $S = K(\boldsymbol{\alpha})$. That is, $\mathfrak{C}$ is the curve on the $(S, \boldsymbol{\alpha})$ plane specified by $S = K(\boldsymbol{\alpha})$. Under this notation, $\mathcal{F}_{\mathrm{MacKay}}$ is the function $\mathcal{F}_{\mathrm{EM}}$ with variables $(S, \boldsymbol{\alpha})$ restricted on the curve $\mathfrak{C}$.

Theorem 2. *When the Gaussian approximation (12) holds with equality,*

$$\mathcal{F}_{\mathrm{MacKay}}(u, \boldsymbol{\alpha}) = \max_S \mathcal{F}_{\mathrm{EM}}(u, S, \boldsymbol{\alpha}), \qquad (17)$$

$$l(\boldsymbol{\alpha}) = \max_u \mathcal{F}_{\mathrm{MacKay}}(u, \boldsymbol{\alpha}). \qquad (18)$$

Proof: Denote $S^* := \arg\max_S \mathcal{F}_{\mathrm{EM}}(u, S, \boldsymbol{\alpha}) = K(\boldsymbol{\alpha})$. Thus

$$\max_S \mathcal{F}_{\mathrm{EM}}(u, S, \boldsymbol{\alpha}) = \mathcal{F}_{\mathrm{EM}}(u, S^*, \boldsymbol{\alpha}) = \mathcal{F}_{\mathrm{EM}}(u, K(\alpha), \boldsymbol{\alpha}).$$

Then the equation (17) holds by Theorem 1, On the other hand, by Lemma 3, $\mathcal{F}_{\mathrm{MacKay}}(\mathbf{z}, \boldsymbol{\alpha}) \leq l(\boldsymbol{\alpha})$ and the equality can be achieved. We thus obtain the equation (18). □.

The following result follows immediately from Theorem 2.

Corollary 1. *The optimizing configurations for* Opt-l*,* OptEM *and* OptMacKay *are identical in* $\boldsymbol{\alpha}$*.*

Theorem 2 and Corollary 1 essentially suggest that OptMacKay is a relaxation of Opt-l, that OptEM is a relaxation of OptMacKay, and that such successive relaxations do not alter the solution of the original problem Opt-l.

Lemma 12. *The EM-Gauss Procedure is identical to the 3-way coordinate ascent on* $\mathcal{F}_{\mathrm{EM}}$*, namely, iterating over the following three steps.*

$$\begin{aligned}u^{(t)} &:= \arg\max_u \mathcal{F}_{\mathrm{EM}}(u, S^{(t-1)}, \boldsymbol{\alpha}^{(t)})\\ S^{(t)} &:= \arg\max_S \mathcal{F}_{\mathrm{EM}}(u^{(t)}, S, \boldsymbol{\alpha}^{(t)})\\ \boldsymbol{\alpha}^{(t+1)} &:= \arg\max_{\boldsymbol{\alpha}} \mathcal{F}_{\mathrm{EM}}(u^{(t)}, S^{(t)}, \boldsymbol{\alpha})\end{aligned}$$

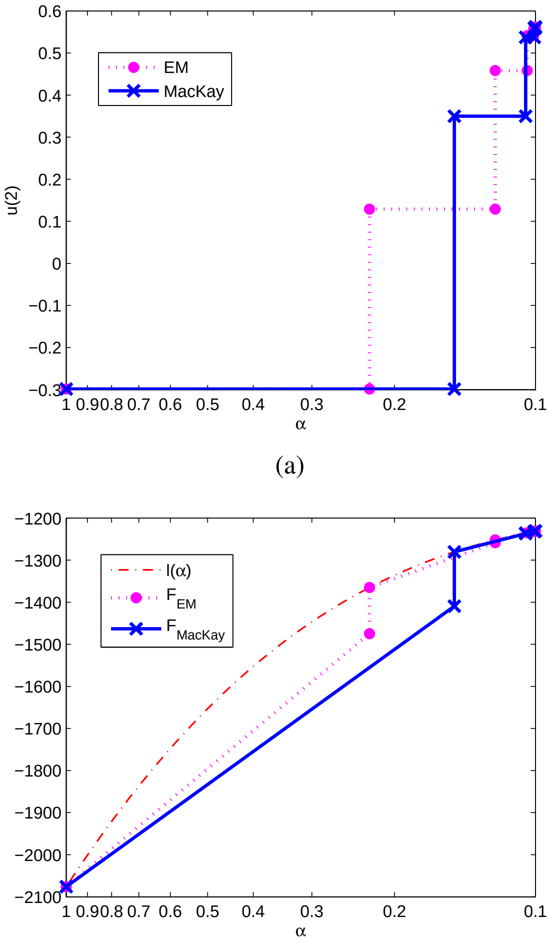
Proof: This follows from the fact that in the EM-Gauss Procedure, the update of u is independent of S and the update of S is independent of u. □

Since OptEM is a relaxation of OptMacKay, it has higher degrees of freedom during optimization. This extra degree of freedom is fully explored in the three-way coordinate descent of EM-Gauss, making its convergence slower than that of MacKay-Gauss. This slower convergence of EM-Gauss can also be understood from another perspective, in which MacKay-Gauss and EM-Gauss are both considered as optimizing the function $\mathcal{F}_{\mathrm{EM}}$.

Lemma 13. *The MacKay-Gauss Procedure is equivalent to the following two-way coordinate-ascent on* $\mathcal{F}_{\mathrm{EM}}$*.*

$$\begin{aligned}u^{(t)} &:= \arg\max_u \mathcal{F}_{\mathrm{EM}}(u, S^{(t-1)}, \boldsymbol{\alpha}^{(t-1)})\\ \left(S^{(t)}, \boldsymbol{\alpha}^{(t)}\right) &:= \arg\max_{(S,\boldsymbol{\alpha})\in\mathfrak{C}} \mathcal{F}_{\mathrm{EM}}(u^{(t)}, S, \boldsymbol{\alpha}).\end{aligned}$$

Following directly from the Lemma 11 and Theorem 1, this lemma suggests that MacKay-Gauss can be viewed as optimizing the same objective function $\mathcal{F}_{\mathrm{EM}}$ as EM-Gauss, but taking a particular coordinate-ascent path, namely, alternating between maximization over u and maximization over $(S, \boldsymbol{\alpha})$ *along the curve* $\mathfrak{C}$. This allows MacKay-Gauss to take a "shorter-cut" than EM-Gauss.



(a)

(b)

Figure 2: The convergence of both the EM algorithm and the MacKay algorithm from the initial configuration $\boldsymbol{\alpha} = 1$ to the final configuration $\boldsymbol{\alpha} \approx 0.1$. (a) The $(\boldsymbol{\alpha}, u)$-trajectories of the EM and MacKay algorithms, where an arbitrary component (in this case, the second component $u(2)$) of vector u is taken as a representative for u. (b) the function $\mathcal{F}_{\mathrm{EM}}$ evaluated along the EM trajectory and the function $\mathcal{F}_{\mathrm{MacKay}}$ evaluated along the MacKay trajectory, and the log-evidence function $l(\boldsymbol{\alpha})$ plotted using its closed-form expression.

4.4 Experiments

Experiments are performed to study the dynamics of the MacKay algorithm and the EM algorithm. We generate a simulated dataset $\mathbf{D}$ for a linear regression model according to the setup in Linear Regression Example–1 with $n = 300, d = 200, \sigma^2 = 10, \boldsymbol{\alpha} = 0.1$ where each each $x_{(i)}$ is drawn uniformly at random from the open interval $(0, 1)$. Both the EM algorithm (in Linear Regression Example-2) and the MacKay algorithm (in Linear Regression Example-3) are used to estimate $\boldsymbol{\alpha}$ from $\mathbf{D}$. For both algorithms, $\boldsymbol{\alpha}$ is initialized to 1 and σ^2 is treated as known.

The optimization trajectories of the two algorithms in Figure 2 (a) show that the MacKay algorithm converges faster towards the fixed point (top right corner) than the EM algorithm. In Figure 2 (b), we see that both the MacKay algorithm and the EM algorithm increase their respective objective functions along their optimization paths, but MacKay achieves higher value of the log-evidence function $l(\boldsymbol{\alpha})$ than EM at every iteration step.

5 Concluding Remarks

In his influential evidence framework, MacKay presented practical Bayesian methods for inference at the parameter level, at the hyper-parameter level and at the model level. For inference at the hyper-parameter level, MacKay introduced a heuristic procedure that iterates between estimating the parameter for a given hyper-parameter setting and estimating the hyper-parameter for the previous parameter setting. Although this procedure is widely adopted in empirical Bayesian methods, its mathematical principle had not been carefully explored prior to this work. In this paper, we formulate this procedure as a well-principled algorithmic framework, and call it the MacKay algorithm.

We show that the MacKay algorithm, like the EM algorithm, can be understood as a coordinate-ascent solution to optimizing a lower bound of the objective function in empirical Bayes. Although this lower bound is different from the lower bound that is optimized by the EM algorithm, we show that as long as the posterior distribution of the parameter is a Gaussian density function, the two algorithms are closely related. In particular, the EM optimization problem, the MacKay optimization problem, and the original empirical Bayes optimization problem all have the same solution. In addition, the MacKay problem is a relaxation of the original problem, and the EM problem is a relaxation of the MacKay problem. This understanding provides and intuitive explanation as to why the MacKay algorithm converges faster than the EM algorithm.

We believe that the close relationship between the MacKay algorithm and the EM algorithm revealed in this paper strongly depends on the Gaussian condition. Although this paper does not show the necessity of this condition, we believe that, in case of non-Gaussian posterior or non-Gaussian models, this relationship will break down, and the MacKay algorithm will diverge from the EM algorithm towards its own optimization objective and along its own optimization path. This makes the MacKay algorithm a framework in its own right. We hope that this paper inspire more applications of the MacKay algorithm to more general models, a territory appearing completely unexplored.

References

[Allahverdyan and Galstyan, 2011] Allahverdyan, A. and Galstyan, A. (2011). Comparative analysis of viterbi training and maximum likelihood estimation for hmms. In *Advances in Neural Information Processing Systems*, pages 1674–1682.

[Bishop, 1999] Bishop, C. M. (1999). Bayesian PCA. *Advances in neural information processing systems*, pages 382–388.

[Carlin and Louis, 1997] Carlin, B. P. and Louis, T. A. (1997). Bayes and empirical Bayes methods for data analysis. *Statistics and Computing*, 7(2):153–154.

[Clyde and George, 2000] Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):681–698.

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

[DuMouchel and Pregibon, 2001] DuMouchel, W. and Pregibon, D. (2001). Empirical Bayes screening for multi-item associations. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 67–76. ACM.

[Efron, 2012] Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

[Frost and Savarino, 1986] Frost, P. A. and Savarino, J. E. (1986). An empirical Bayes approach to efficient portfolio selection. *Journal of Financial and Quantitative Analysis*, 21(03):293–305.

[Heskes, 2000] Heskes, T. (2000). Empirical bayes for learning to learn. In *ICML*, pages 367–374.

[Hyvärinen, 1999] Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *Neural Networks, IEEE Transactions on*, 10(3):626–634.

[Inoue and Tanaka, 2001] Inoue, J.-i. and Tanaka, K. (2001). Dynamics of the maximum marginal likelihood hyperparameter estimation in image restoration: Gradient descent versus expectation and maximization algorithm. *Phys. Rev. E*, 65:016125.

[Jensen, 1906] Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les ingalits entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193.

[MacKay, 1992a] MacKay, D. J. (1992a). Bayesian interpolation. *Neural computation*, 4(3):415–447.

[MacKay, 1992b] MacKay, D. J. (1992b). The evidence framework applied to classification networks. *Neural computation*, 4(5):720–736.

[MacKay, 1992c] MacKay, D. J. (1992c). A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472.

[MacKay, 1995] MacKay, D. J. (1995). Probable networks and plausible predictions: a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems*, 6(3):469–505.

[MacKay and Neal, 1994] MacKay, D. J. and Neal, R. M. (1994). Automatic relevance determination for neural networks. In *Technical Report in preparation*. Cambridge University.

[Murphy, 2012] Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

[Neal, 1993] Neal, R. (1993). Probabilistic inference using markov chain monte carlo methods. Technical Report CRG-TR -93-1, Dept. of Computer Science, University of Toronto.

[Schäfer and Strimmer, 2005] Schäfer, J. and Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21(6):754–764.

[Solomon, 2015] Solomon, J. (2015). *Numerical Algorithms: Methods for Computer Vision, Machine Learning, and Graphics*. CRC Press.

[Tan and Févotte, 2009] Tan, V. Y. and Févotte, C. (2009). Automatic relevance determination in nonnegative matrix factorization. In *SPARS'09-Signal Processing with Adaptive Sparse Structured Representations*.

[Tipping, 2001] Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *The journal of machine learning research*, 1:211–244.

[Tipping et al., 2003] Tipping, M. E., Faul, A. C., et al. (2003). Fast marginal likelihood maximisation for sparse bayesian models. In *AISTATS*.

[Wipf and Nagarajan, 2008] Wipf, D. P. and Nagarajan, S. S. (2008). A new view of automatic relevance determination. In *Advances in neural information processing systems*, pages 1625–1632.

[Yang et al., 2004] Yang, D., Zakharkin, S. O., Page, G. P., Brand, J. P., Edwards, J. W., Bartolucci, A. A., and Allison, D. B. (2004). Applications of Bayesian statistical methods in microarray data analysis. *American Journal of Pharmacogenomics*, 4(1):53–62.