BAYESIAN LINEAR BASIS MODELS WITH GAUSSIAN PRIORS FOR MUSICAL EXPRESSION

CARLOS EDUARDO CANCINO CHACÓN, MAARTEN GRACHTEN, AND GERHARD WIDMER

Austrian Research Institute for Artificial Intelligence

OFAI-TR-2014-12 Version 1.0

Abstract. We present a probabilistic linear basis model for musical expression. The model is an extension of prior work by Grachten and Widmer [Grachten and Widmer, 2012] and is a generalization of the work by Grachten et al. [Grachten et al., 2014]. By assuming the prior distribution of the model parameters to be Gaussian with arbitrary mean and covariance, this model allows for specifying musical knowledge, and for modeling multiple distinct performances of the same piece. We show that in its current state, the model performs at least on par of the original approach.

Keywords: Musical expression, probabilistic linear basis modeling, Bayesian linear regression

1. Introduction

Performances of (notated) musical pieces typically show large variations in expressive patters like tempo, dynamics, articulation, etc. Musicians use such variations to transmit an expressive interpretation of the music to the listener. This interpretation may contain elements related to the musical structure, and also affective elements [Clarke, 1988, Palmer, 1997].

Models for musical expression using the machine learning paradigm can have different aims. On the one hand, such models can be used to to produce results that maximally resemble patterns of observed (expert) performances through a data-driven approach [Grachten, 2002, Teramura et al., 2008, Ohishi et al., 2014]. On the other hand, a machine learning approach can be used to analyze musical performances [Grachten, 2002, Widmer, 2003]. This would result in models that describe principles that emerge from performance data as precise and compact as possible [Grachten and Widmer, 2012].

The Linear Basis Model (LBM) for musical expression is a framework proposed in [Grachten and Widmer, 2011] for expressive dynamics and in [Grachten and Widmer, 2012] for general expressive parameters that follows the machine learning paradigm. Such a framework models the effect of annotated expressive markings on music performances by means of *basis functions* that capture some structural aspect of a musical score, and

express the relation of each score note to that aspect as a real number. This framework can be used for both predictive and analytical purposes [Grachten and Widmer, 2012, Grachten et al., 2014].

The original formulation of the LBM used a least squares (LS) regression to compute the optimal model parameters. A probabilistic LBM using the Bayesian linear regression assuming zero mean Gaussian priors with isotropic covariance was presented in [Grachten et al., 2014]. In this paper we generalize this framework to consider Gaussian priors with arbitrary mean and covariance.

The outline of this paper is as follows: In Section 2, the LBM framework for musical expression is described. In Section 3, a mathematical formulation of the LBM is provided. In Section 4, the optimal model parameters of the LBM in the Bayesian linear regression sense are derived. A discussion on the interpretation of the parameters of the Bayesian linear regression is presented in Section 5. A proof-of-concept experiment for testing the predictive accuracy of the proposed framework is described in Section 6, while its results are discussed in Section 7. Conclusions are presented in Section 8.

2. Basis Functions

Musical expression consists of a number of individual factors that jointly determine the way a musical piece is rendered [Palmer, 1996]. Under this assumption, expressive information from human performances can be decomposed into components.

In this paper, a *musical score* is defined as a sequence of elements that hold information, and may also refer to other elements [Grachten and Widmer, 2012]. These elements may include note elements (e.g. pitch, duration) and non-note elements (e.g. *piano*, *forte*, *crescendo*). If we denote the set of all note elements in a score by $\mathcal{X}$, a basis function is a real valued mapping in the form $\varphi\colon \mathcal{X} \mapsto [0, 1]$.

The LBM framework relies on two simplifying assumptions. First, it is considered that each expressive parameter is only dependent on score information. This assumption implies that mutual dependencies between expressive parameters, as well as temporal dependencies within parameters are not explicitly modeled. Second, each expressive parameter is modeled as depending linearly on score information.

As an example, let us take the use of three different kinds of basis functions to represent different categories of dynamics annotations. Figure 1 illustrates the idea modeling loudness as a weighted sum of basis functions schematically. Dynamics annotations that specify a constant loudness level that is in effect until another such directive occurs, such as *piano* (relatively low loudness level) and *forte* (relatively high loudness level) can be represented with a step-like function. Gradual increases and decreases of loudness, such as *crescendo* and *decrescendo*, can be modeled using ramp-like functions. A third class of annotations, concerning only a single position in time (e.g. *sforzato*, accents) can be represented with impulsive functions.

In [Grachten and Widmer, 2012], several types of basis functions for representing different aspects of musical expression are described. One of the most important issues

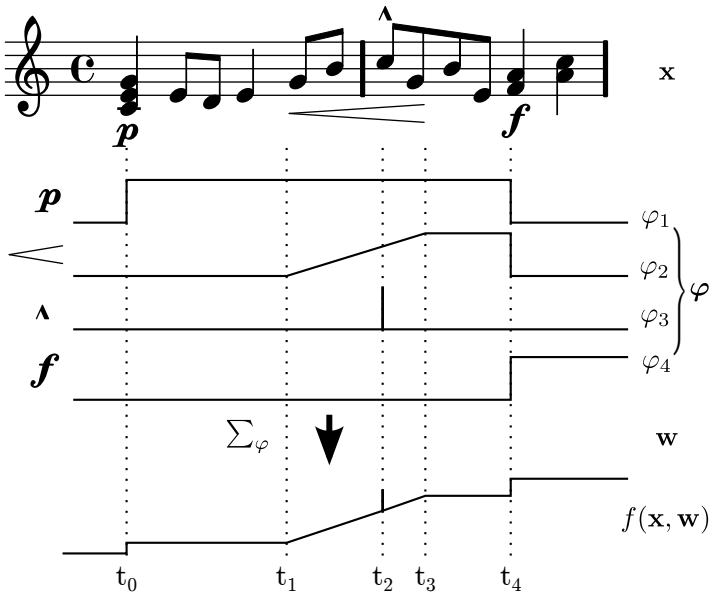

FIGURE 1. Schematic view of note dynamics as a weighted sum $f(\mathbf{x}, \mathbf{w})$ of basis functions φ, representing dynamic annotations

regarding the design of basis functions is whether a basis function corresponds to a feature in general, or to a single instance of that feature. This leads to the definition of *global* basis functions and *local* basis functions, respectively. In this paper, only global basis functions are considered, and the extension of the methods derived here to local basis functions will be the subject of a future study.

3. Linear Basis Model

In this section we provide a mathematical formulation of the LBM framework. Let $\mathbf{x} = (x_1, \ldots, x_N)^T \in \mathbb{R}^N$ be a vector representing a set of N notes in a musical score and $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^N$ be a vector representing expressive parameters (e.g. dynamics [Grachten and Widmer, 2011], expressive tempo [Krebs and Grachten, 2012]). In order to be consistent with the machine learning literature, in this paper we also denominate $\mathbf{y}$ as *targets* [Bishop, 2006, Grachten and Widmer, 2012]. Let $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_M)^T \in \mathbb{R}^M$ be a vector whose elements represent the basis functions. The influence of a basis function φ_i is expressed by a weight w_i. Hence, we can model the expressive parameters $\mathbf{y}$ as weighted sum of the basis functions, i.e.

$$\mathbf{y} = f(\mathbf{x}, \mathbf{w}) + \boldsymbol{\epsilon} = \boldsymbol{\Phi}\mathbf{w} + \boldsymbol{\epsilon}, \tag{1}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{N \times M}$ is a matrix with elements $\Phi_{ij} = \varphi_j(x_i)$, $\mathbf{w} = (w_1, \ldots, w_M)^T \in \mathbb{R}^M$ is a vector of weights and $\boldsymbol{\epsilon}$ is a zero mean Gaussian noise with covariance $\beta^{-1}\mathbf{E}$, with $\mathbf{E}$ an identity matrix. It follows that the conditional distribution of $\mathbf{y}$ given the weights $\mathbf{w}$

is given by

$$p(\mathbf{y} \mid \mathbf{w}) = \mathcal{N}(\mathbf{y} \mid \mathbf{\Phi w}, \beta^{-1}\mathbf{E}). \tag{2}$$

4. Bayesian Linear Regression with Gaussian priors

Given a set of performances as training data in form $(\mathbf{x}_t, \mathbf{y}_t)$, we are interested in using the model from Eq. (1) to estimate the optimal set of weights $\mathbf{w}$. This approach represents an optimization problem [Boyd and Vandenberghe, 2004]. In previous work [Grachten and Widmer, 2011], an LS regression was used to compute the weights. Nevertheless, this solution has several shortcomings. These limitations include the computation of only point (deterministic) estimates of the expressive parameters $\mathbf{y}$ over the weights $\mathbf{w}$, instead of probability distributions; and the inability of the model to combine prior information on the weights with weight estimates based on empirical data [Grachten and Widmer, 2012]. Another drawback of the LS solution is that interactions between input variables are not modeled. A detailed derivation of this result can be found in Appendix A.

An alternative solution, that allows the model to include prior information of the weights and interactions between basis functions is Bayesian linear regression [Bishop, 2006]. This approach can be understood as probabilistic Bayesian estimation of the weights using the Maximum a Posteriori (MAP) paradigm. As stated in the Introduction, in this paper we generalize the work in [Grachten et al., 2014] by relaxing the assumption that the prior distribution of the model parameters is a zero mean Gaussian distribution with isotropic covariance.

Following this Bayesian interpretation, we assume that the prior probability distribution of the weights is given by

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0), \tag{3}$$

where $\mathbf{m}_0$ is the prior mean, and $\mathbf{S}_0$ is the prior covariance matrix. Using Bayes' theorem, the posterior conditional probability of the weights given the training targets $\mathbf{y}_t$ can be expressed as

$$p(\mathbf{w} \mid \mathbf{y}_t) = \frac{p(\mathbf{y}_t \mid \mathbf{w})p(\mathbf{w})}{\int_{\mathbf{w}} p(\mathbf{y}_t \mid \mathbf{w})p(\mathbf{w})d\mathbf{w}}. \tag{4}$$

Substituting Eq. (2) and Eq. (3) in the above equation, the MAP estimation of the weights results in

$$\begin{aligned} \mathbf{w}_{map} &= \underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathbf{w} \mid \mathbf{y}_t) \\ &= \left(\beta \mathbf{\Phi}_t^T \mathbf{\Phi}_t + \mathbf{S}_0^{-1}\right)^{-1} \left(\beta \mathbf{\Phi}_t^T \mathbf{y}_t + \mathbf{S}_0^{-1}\mathbf{m}_0\right). \end{aligned} \tag{5}$$

A detailed derivation of this result is provided in Appendix B. A method for estimating the prior parameters β, $\mathbf{m}_0$ and $\mathbf{S}_0$ from training data using the Expectation-Maximization (EM) algorithm is provided in Appendix C.

5. Interpretation of the prior parameters

The formulation proposed in the previous section allows for inclusion of prior information in the form of the prior mean $\mathbf{m}_0$, and models interactions between basis functions in the form of the prior covariance matrix $\mathbf{S}_0$, addressing directly the shortcomings of the LS-based LBM framework. As a predictive model for expressive performance, this approach allows for two methods for generating the performance of a musical piece. On the one hand, a performance can be obtained directly from the model of Eq. (1), by substituting $\mathbf{w}_{map}$ for $\mathbf{w}$. On the other hand, we can use the posterior conditional probability of the weights given some training targets to compute a predictive distribution [Bishop, 2006]. In order to generate a new performance, we draw a sample from such a distribution.

As an analysis tool, estimates of the prior parameters $\mathbf{m}_0$ and $\mathbf{S}_0$ and the precision β can be used to analyze the effect of the score annotations in musical performance by a musician. By estimating the parameters from performance data of different musicians, or from different eras, they might provide a starting point for modeling performance styles.

6. Experiments

In this section, we present a proof-of-concept experiment to show how the presented model is able to account for aspects of expressive dynamics.

6.1. **Data set.** A set of random scores with 200 notes is generated. We denote $\mathbf{x}_i = (x_1, \ldots, x_{200})$ the notes in the i-th score. Such scores are structured as follows: Every score has a random number of sections, ranging from 2 to 5. Each section has only one constant dynamics indication, namely *piano*, $mezzoforte$ or $forte$. Each section is then divided into 2 or 3 subsections. We pick randomly one of these subsections per section for a gradual change of loudness, namely, *crescendo* or *decrescendo*. Finally, a random number of notes in the score are picked to have accents >. As described in Section 2, for these dynamics annotations we have three types of basis functions. For p, mf and f, defined in range $[x_s, x_e]$ the corresponding function is given by

$$\varphi_{\{p,mf,f\}}(x_i) = \begin{cases} 1 & \text{if } s \leq i \leq e, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

For *cresc.* and *decresc.*, defined in range $[x_s, x_e]$, the basis function is defined as

$$\varphi_{\{cresc.,decresc.\}}(x_i) = \begin{cases} \frac{i-s}{e-s} & \text{if } s \leq i \leq e, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

For an accent > in note x_s, the basis function is given by

$$\varphi_{>}(x_i) = \begin{cases} 1 & \text{if } i = s, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

In order to generate a performance, we fix the parameters of the prior distribution of the weights. The prior mean is set to

$$\mathbf{m}_0 = (p \mapsto 50, mf \mapsto 70, f \mapsto 100, cresc. \mapsto 20, decresc. \mapsto -20, > \mapsto 15)^T. \tag{9}$$

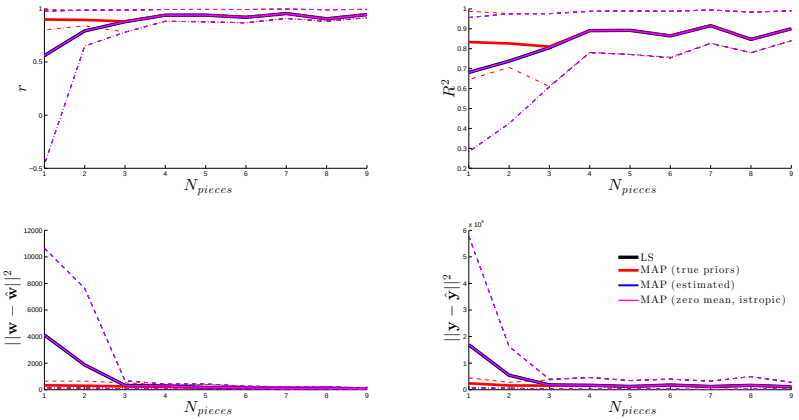

FIGURE 2. Average predictive accuracy in a leave-one-out scenario over performance of 50 trials. See Section 6 for abbreviations.

The prior covariance is calculated as $\mathbf{S}_0 = \mathbf{A}\mathbf{A}^T$, with $\mathbf{A} \in \mathbb{R}^{6\times 6}$ a random matrix sampled from the uniform distribution $\mathcal{U}(0, 1)$. We set $\beta = 1$. Given $\mathbf{\Phi}_i$, the matrix of basis functions for the i-th random piece, a set of target dynamics $\mathbf{y}_i$ is sampled from[1]

$$p(\mathbf{y}_i) = \mathcal{N}\left(\mathbf{y}_i \mid \mathbf{\Phi}_i\mathbf{m}_0, \beta^{-1}\mathbf{E} + \mathbf{\Phi}_i\mathbf{S}_0\mathbf{\Phi}_i^T\right). \tag{10}$$

6.2. **Predictive accuracy.** To evaluate the accuracy of the predictions of the proposed framework, a leave-one-out cross validation was performed. Each model was trained with N_{pieces} pieces, with N_{pieces} ranging from 1 to 9, and it was used to predict the dynamics of a new piece. As a measure of the quality of the predictive accuracy, we use r, the Pearson correlation coefficient, R^2, the coefficient of determination, $\|\mathbf{w} - \hat{\mathbf{w}}\|^2$, the squared error of the weights, and $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$, the squared error of the prediction. The correlation coefficient denotes how strongly the observed dynamics and the dynamics proposed by the model correlate, while R^2 expresses the proportion of variance explained by the model. The squared error of the weights indicates the deviation of the estimated weights to the ground truth, while the squared error of the prediction denotes the how different are the predicted dynamics to the true dynamics.

Figure 2 shows the average and standard deviation over the 50 trials of the above discussed quality measures of the predictive accuracy for different models for every N_{pieces}. In this table, LS denotes the old least squares approach from [Grachten and Widmer, 2012], and MAP the proposed Bayesian method. With (true prior) we denote the Bayesian linear regression using the prior parameters above discussed, while (estimated) refers to the estimation of these parameters using the EM algorithm from Appendix C, and (zero mean, isotropic) refers to the previous approach from [Grachten et al., 2014].

[1] see Eq. (104) in Appendix D.2

7. Discussion

The results show that the proposed Bayesian approach with true prior information performs better than the LS approach for a limited amount of training data. For example, the Bayesian LBM with true prior information has an increase of ca. 23% in the correlation coefficient and ca. 12% in the explained variance, while achieving the lowest squared error, ca. 3.5 times smaller than the other approaches for training with 2 pieces. These results suggest that incorporating prior knowledge into the model increases its ability to generalize from a training data that is not representative.

The Bayesian model with zero mean and isotropic Gaussian priors performs on par with the LS regression. This is consistent with the results in reported in [Grachten et al., 2014]. Furthermore, in case of estimation of the prior parameters from the data, the new approach performs on par to LS regression.

8. Conclusions and Future Work

In this paper, a fully probabilistic version of the LBM for musical expression using Bayesian Linear regression with Gaussian priors was presented. This model addresses some limitations of the previous LS solution, and is a generalization of the work presented in [Grachten et al., 2014]. Experimental results suggest that the predictive accuracy of the Bayesian model performs on par of the LS regression, with an improvement over LS in case of limited training data.

The framework presented in this paper only considers the case of global basis functions. In order to extend this work to local basis functions, future work will explore a joint clustering + Bayesian linear regression solution. Such a method could be interpreted in terms of the more general assumption of the prior probability distribution of the weights being a mixture of Gaussians.

Acknowledgments

This work is supported by European Union Seventh Framework Programme, through the Lrn2Cre8 project (FET grant agreement no. 610859).

References

[Bishop, 2006] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer Verlag, Microsoft Research Ltd.

[Boyd and Vandenberghe, 2004] Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

[Burden and Faires, 2010] Burden, R. L. and Faires, J. D. (2010). *Numerical Analysis*. Brooks/Cole, 9 edition.

[Clarke, 1988] Clarke, E. F. (1988). Generative principles in music. In Sloboda, J., editor, *Generative Processes in Music: The Psychology of Performance, Improvisation, and Composition*. Oxford University Press.

[Fong and Saunders, 2011] Fong, D. C.-L. and Saunders, M. (2011). LSMR: An Iterative Algorithm for Sparse Least-Squares Problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971.

[Grachten, 2002] Grachten, M. (2002). Summary of the Music Performance Panel, MOSART Workshop 2001, Barcelona. In *MOSART Workshop*, pages 1–17.

[Grachten et al., 2014] Grachten, M., Cancino Chacón, C. E., and Widmer, G. (2014). Analysis and prediction of expressive dynamics using Bayesian linear models. In *Proceedings of the 1st international workshop on computer and robotic Systems for Automatic Music Performance*, pages 545–552.

[Grachten and Widmer, 2011] Grachten, M. and Widmer, G. (2011). Explaining musical expression as a mixture of basis functions. In *Proceedings of the Eigth Sound and Music Computing Conference*, pages 1–7, Padua, Italy.

[Grachten and Widmer, 2012] Grachten, M. and Widmer, G. (2012). Linear Basis Models for Prediction and Analysis of Musical Expression. *Journal of New Music Research*, 41(4):311–322.

[Kay, 2009] Kay, S. M. (2009). *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall Signal Processing Series. Prentice Hall, University of Rhode Island.

[Krebs and Grachten, 2012] Krebs, F. and Grachten, M. (2012). Combining score and filter based models to predict tempo fluctuations in expressive music performances. In *Proceedings of the Ninth Sound and Music Computing Conference (SMC)*, Copenhagen, Denmark.

[Ohishi et al., 2014] Ohishi, Y., Mochihashi, D., Kameoka, H., and Kashino, K. (2014). Mixture of Gaussian process experts for predicting sung melodic contour with expressive dynamic fluctuations. In *ICASSP 2014 - 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3714–3718. IEEE.

[Paige and Saunders, 1982] Paige, C. C. and Saunders, M. A. (1982). LSQR, An algorithm for sparse linear equations and sparse least squares. *ACM Trans. Math. Software*, 8(1):43–71.

[Palmer, 1996] Palmer, C. (1996). Anatomy of a performance: Sources of musical expression. *Music Perception*, 13(3):433–453.

[Palmer, 1997] Palmer, C. (1997). Music performance. *Annual Review of Psychology*, 48:115–138.

[Petersen and Pedersen, 2012] Petersen, K. B. and Pedersen, M. S. (2012). *The Matrix Cookbook* . Technical University of Denmark.

[Spiegel and Liu, 1999] Spiegel, M. R. and Liu, J. M. (1999). *Mathematical Handbook of Formulas and Tables*. Schaum's outlines. McGraw-Hill Companies.

[Teramura et al., 2008] Teramura, K., Okuma, H., and Taniguchi, Y. (2008). Gaussian process regression for rendering music performance. *Proceedings of the 10th International Conference on Music Perception and Cognition*.

[Widmer, 2003] Widmer, G. (2003). Discovering simple rules in complex data: A meta-learning algorithm and some surprising musical discoveries. *Artificial Intelligence*, 146(2):129–148.

Appendix A. Least Squares regression, Maximum Likelihood estimation and Minimum Variance Unbiased estimator

In this appendix, a detailed derivations of the computation of the weights for linear regression in the LS sense and maximum likelihood, as well as the minimum variance unbiased estimator of the weights is provided.

The weights computed in the LS sense are given by

$$\mathbf{w}_{ls} = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathcal{L}_{ls}(\mathbf{w}), \tag{11}$$

where $\mathcal{L}_{ls}$, the squared error, is a cost function that can be written as

$$\begin{aligned} \mathcal{L}_{ls}(\mathbf{w}) &= \|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|^2 \\ &= (\mathbf{y} - \mathbf{\Phi}\mathbf{w})^T(\mathbf{y} - \mathbf{\Phi}\mathbf{w}). \end{aligned} \tag{12}$$

Computing the gradient of the cost function with respect to the weights results in

$$\nabla_{\mathbf{w}}\mathcal{L}_{ls}(\mathbf{w}) = -2\mathbf{\Phi}^T\mathbf{y} + 2\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}. \tag{13}$$

The weights that minimize the cost function are given by the solutions to $\nabla_{\mathbf{w}}\mathcal{L}_{ls}(\mathbf{w}) = \mathbf{0}$, i.e.

$$\mathbf{w}_{ls} = \mathbf{\Phi}^{\dagger}\mathbf{y}, \tag{14}$$

where

$$\mathbf{\Phi}^{\dagger} = \left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T \tag{15}$$

is known as the *Moore-Penrose pseudo-inverse* of the matrix $\mathbf{\Phi}$ [Bishop, 2006]. The computation of $\mathbf{\Phi}^{\dagger}$ results in numerical instabilities for ill conditioned problems. Furthermore, for large training data sets, the above described LS technique can be computationally expensive [Bishop, 2006]. More efficient alternatives for computing the LS regression include iterative approaches such as the Least Mean Squares algorithm, which uses the technique known as stochastic gradient descent [Bishop, 2006], and the LSQR and LSMR algorithms for large sparse systems [Paige and Saunders, 1982, Fong and Saunders, 2011].

Under the assumption of a Gaussian conditional distribution of $\mathbf{y}$ given $\mathbf{w}$ (as in Eq. (2)), the LS regression is analytically equivalent to ML estimation of $\mathbf{w}$. The conditional log-likelihood of $\mathbf{y}$ given $\mathbf{w}$ is given by

$$\log p(\mathbf{y} \mid \mathbf{w}) = \frac{N}{2}\log(\beta) - \frac{N}{2}\log(2\pi) - \frac{\beta}{2}(\mathbf{y} - \mathbf{\Phi}\mathbf{w})^T(\mathbf{y} - \mathbf{\Phi}\mathbf{w}). \tag{16}$$

The gradient of the above log-likelihood is proportional to the gradient of the LS cost function, i.e.

$$\nabla_{\mathbf{w}}\log p(\mathbf{y} \mid \mathbf{w}) = \beta\mathbf{\Phi}^T\mathbf{y} - \beta\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} = -\frac{\beta}{2}\nabla_{\mathbf{w}}\mathcal{L}_{ls}(\mathbf{w}). \tag{17}$$

Therefore, computing the solution that maximizes the conditional log-likelihood is equivalent to computing the solution that minimizes the squared error. Hence, the weights computed using ML are given by

$$\mathbf{w}_{ml} = \mathbf{w}_{ls} = \mathbf{\Phi}^{\dagger}\mathbf{y} \tag{18}$$

We can show that the ML estimation of the weights is an MVUE [Kay, 2009]. According to the Cramér-Rao lower bound (CRLB) theorem, the covariance of $\mathbf{w}_{ml}$ is bounded by

$$\text{cov}(\mathbf{w}_{ml}) \geq \mathbf{I}(\mathbf{w})^{-1}, \tag{19}$$

where $\mathbf{I}(\mathbf{w})$ is the *Fisher information matrix*, given by

$$\mathbf{I}(\mathbf{w}) = -\mathbb{E}\left\{\nabla_{\mathbf{w}} \nabla_{\mathbf{w}} \log p(\mathbf{y} \mid \mathbf{w})\right\}. \tag{20}$$

Using Eq. (17), we can rewrite the above equation as

$$\begin{aligned}\mathbf{I}(\mathbf{w}) &= -\mathbb{E}\left\{\nabla_{\mathbf{w}} \left(\beta \mathbf{\Phi}^T \mathbf{y} - \beta \mathbf{\Phi}^T \mathbf{\Phi}\right)\right\} \\ &= -\mathbb{E}\left\{-\beta \mathbf{\Phi}^T \mathbf{\Phi}\right\} \\ &= \beta \mathbf{\Phi}^T \mathbf{\Phi}. \end{aligned} \tag{21}$$

Therefore, the CRLB results in

$$\text{cov}(\mathbf{w}_{ml}) \geq \left(\beta \mathbf{\Phi}^T \mathbf{\Phi}\right)^{-1}. \tag{22}$$

Furthermore, we can factorize $\nabla_{\mathbf{w}} \log p(\mathbf{y} \mid \mathbf{w})$ as

$$\begin{aligned}\nabla_{\mathbf{w}} \log p(\mathbf{y} \mid \mathbf{w}) &= \beta \mathbf{\Phi}^T \mathbf{\Phi} \left[\mathbf{\Phi}^{\dagger}\mathbf{y} - \mathbf{w}\right] \\ &= \mathbf{I}(\mathbf{w}) \left(\mathbf{\Phi}^{\dagger}\mathbf{y} - \mathbf{w}\right). \end{aligned} \tag{23}$$

By the CRLB theorem, this means that $\mathbf{\Phi}^{\dagger}\mathbf{y}$ is an MVUE estimator [Kay, 2009].

Appendix B. Derivation of the weights for Bayesian linear regression using a Gaussian prior

In this appendix, we provide a detailed derivation of the weights for Bayesian linear regression described in Section 4.

The joint probability distribution of the targets $\mathbf{y}$ and the weights $\mathbf{w}$ given by

$$p(\mathbf{y}, \mathbf{w}) = p(\mathbf{y} \mid \mathbf{w})p(\mathbf{w}), \tag{24}$$

where $p(\mathbf{w})$ is the conjugate prior distribution of the weights [Bishop, 2006]. As previously stated in Section 4, we assume this prior probability distribution to be Gaussian, i.e.

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_0, \mathbf{S}_0), \tag{25}$$

where $\mathbf{m}_0$ is the prior mean and $\mathbf{S}_0$ is the prior covariance matrix.

Using Bayes' Theorem, the posterior probability distribution of $\mathbf{w}$ given $\mathbf{y}$ (in log-domain) is given by

$$\log p(\mathbf{w} \mid \mathbf{y}) = \log p(\mathbf{y} \mid \mathbf{w}) + \log p(\mathbf{w}) - \underbrace{\log \int_{\mathbf{w}} p(\mathbf{y} \mid \mathbf{w})p(\mathbf{w})d\mathbf{w}.}_{\text{does not depend on } \mathbf{w}} \tag{26}$$

The weights for Bayesian linear regression can be computed using MAP estimation, i.e.

$$\begin{aligned} \mathbf{w}_{map} &= \underset{\mathbf{w}}{\operatorname{argmax}} \left(\log p(\mathbf{w} \mid \mathbf{y})\right) \\ &= \underset{\mathbf{w}}{\operatorname{argmax}} \left(\underbrace{\log p(\mathbf{y} \mid \mathbf{w}) + \log p(\mathbf{w})}_{\mathcal{L}_{map}(\mathbf{w})} \right). \end{aligned} \tag{27}$$

Let us analyze every term of the cost function $\mathcal{L}_{map}$ separately. The log-posterior probability of the targets given the weights can be written as

$$\log p(\mathbf{y} \mid \mathbf{w}) = \frac{N}{2}\log(\beta) - \frac{N}{2}\log(2\pi) - \frac{\beta}{2}\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{\Phi}\mathbf{w} + \mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}\right). \tag{28}$$

The log-prior probability of the weights is given by

$$\log p(\mathbf{w}) = -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log(\det(\mathbf{S}_0)) - \frac{1}{2}\left((\mathbf{w} - \mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0)\right). \tag{29}$$

The gradient with respect to the weights of the log likelihood $\mathcal{L}_{map}(\mathbf{w})$ results in

$$\begin{aligned} \nabla_{\mathbf{w}}\mathcal{L}_{map}(\mathbf{w}) &= \beta\mathbf{y}^T\mathbf{\Phi} - \beta\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - \frac{1}{2}\left(\mathbf{S}_0 + \mathbf{S}_0^{-1^T}\right)(\mathbf{w} - \mathbf{m}_0) \\ &= \beta\mathbf{y}^T\mathbf{\Phi} - \beta\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w} - \mathbf{S}_0^{-1}(\mathbf{w} - \mathbf{m}_0). \end{aligned} \tag{30}$$

The weights that maximize the cost function are the solution to $\nabla_{\mathbf{w}}\mathcal{L}_{map}(\mathbf{w}) = \mathbf{0}$, i.e.

$$\mathbf{w}_{map} = \left(\beta\mathbf{\Phi}^T\mathbf{\Phi} + \mathbf{S}_0^{-1}\right)^{-1}\left(\beta\mathbf{\Phi}^T\mathbf{y} + \mathbf{S}_0^{-1}\mathbf{m}_0\right). \tag{31}$$

In case $\mathbf{m}_0 = \mathbf{0}$ and $\mathbf{S}_0 = \alpha^{-1}\mathbf{E}$, the solution simplifies to

$$\mathbf{w}_{map} = \left(\frac{\alpha}{\beta}\mathbf{E} + \mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^T\mathbf{y} \tag{32}$$

This case was considered in [Grachten et al., 2014]. From the above equation is easy to see that the prior distribution of $\mathbf{w}$ represents a regularization to the LS regression. Therefore, the Bayesian linear regression it is less prone to overfitting [Bishop, 2006]. In the case of $\|\mathbf{S}_0^{-1}\| \to 0$ (or $\alpha \to 0$), the weights $\mathbf{m}_B$ converge to the LS weights $\mathbf{w}_{ls}$. In such scenario, the prior probability distribution is said to be *non informative* [Bishop, 2006].

Appendix C. Estimation of prior parameters using the EM algorithm

This section describes the estimation of the mean value $\mathbf{m}_0$ and covariance matrix $\mathbf{S}_0$ of the prior distribution of the weights $p(\mathbf{w})$ and the noise precision β of the conditional probability distribution $p(\mathbf{y} \mid \mathbf{w})$ using the EM-Algorithm.

For the E-step we need to compute the conditional expectation of the log-likelihood $\mathcal{L}_{map}$, given as in Eq. (27), with respect to the posterior probability of the weights $\mathbf{w}$ given the targets $\mathbf{y}$. Using Bayes' theorem, the posterior probability of $\mathbf{w}$ given $\mathbf{y}$ is given by[2]

$$p(\mathbf{w} \mid \mathbf{y}) = \mathcal{N}(\mathbf{w} \mid \mathbf{m}_N, \mathbf{S}_N), \tag{33}$$

where $\mathbf{m}_N$ is the posterior mean and $\mathbf{S}_N$ is the posterior covariance matrix, calculated as

$$\mathbf{m}_N = \mathbf{S}_N \left(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\mathbf{\Phi}^T\mathbf{y}\right) \qquad \text{and} \qquad \mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\mathbf{\Phi}^T\mathbf{\Phi}. \tag{34}$$

The conditional expectation of the log-likelihood function results in

$$\mathbb{E}\left\{\mathcal{L}_{map}(\mathbf{w})\right\} = \mathbb{E}\left\{\log p(\mathbf{y} \mid \mathbf{w})\right\} + \mathbb{E}\left\{\log p(\mathbf{w})\right\} \tag{35}$$

Let us compute the conditional expected values from the above equation individually

$$\begin{aligned}\mathbb{E}\left\{\log p(\mathbf{y} \mid \mathbf{w})\right\} &= \int_{\mathbf{w}} \log p(\mathbf{y} \mid \mathbf{w}) p(\mathbf{w} \mid \mathbf{y}) d\mathbf{w} \\ &= \frac{N}{2}\log(\beta) - \frac{N}{2}\log(2\pi) - \frac{\beta}{2}\left(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{\Phi}\underbrace{\mathbb{E}\left\{\mathbf{w}\right\}}_{=\mathbf{m}_N}\right. \\ &\quad \left. + \underbrace{\mathbb{E}\left\{\mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}\right\}}_{**}\right).\end{aligned} \tag{36}$$

In order to compute $**$, we consider the eigenvalue decomposition of $\mathbf{\Phi}^T\mathbf{\Phi}$. Since this product is a symmetric matrix, we can rewrite it as

$$\mathbf{\Phi}^T\mathbf{\Phi} = \sum_{i=1}^{M} \nu_i \mathbf{u}_i \mathbf{u}_i^T, \tag{37}$$

[2]See Eq. (103) in Appendix D.2.

where ν_i is the i-th eigenvalue and $\mathbf{u}_i$ the i-th eigenvector of $\mathbf{\Phi}^T\mathbf{\Phi}$, respectively. Substituting the above equation in $**$ results in

$$\begin{aligned}
** &= \mathbb{E}\left\{\mathbf{w}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{w}\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^{M}\nu_i\mathbf{w}^T\mathbf{u}_i\mathbf{u}_i^T\mathbf{w}\right\} \\
&= \mathbb{E}\left\{\sum_{i=1}^{M}\nu_i\mathbf{u}_i^T\mathbf{w}\mathbf{w}^T\mathbf{u}_i\right\} \\
&= \sum_{i=1}^{M}\nu_i\mathbf{u}_i^T\underbrace{\mathbb{E}\left\{\mathbf{w}\mathbf{w}^T\right\}}_{=\mathbf{m}_N\mathbf{m}_N^T+\mathbf{S}_N}\mathbf{u}_i \\
&= \mathbf{m}_N^T\left(\sum_{i=1}^{M}\nu_i\mathbf{u}_i\mathbf{u}_i^T\right)\mathbf{m}_N + \sum_{i=1}^{M}\nu_i\mathbf{u}_i^T\mathbf{S}_N\mathbf{u}_i \\
&= \mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N + \underbrace{\sum_{i=1}^{M}\nu_i\mathbf{u}_i^T\mathbf{S}_N\mathbf{u}_i}_{***}.
\end{aligned} \tag{38}$$

We can use properties of the trace[3] to rewrite $***$ as

$$\begin{aligned}
*** &= \sum_{i=1}^{M}\nu_i\mathbf{u}_i^T\mathbf{S}_N\mathbf{u}_i \\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\nu_i\mathbf{u}_i^T\mathbf{S}_N\mathbf{u}_i\right) \\
&= \mathrm{Tr}\left(\sum_{i=1}^{M}\nu_i\mathbf{S}_N\mathbf{u}_i\mathbf{u}_i^T\right) \\
&= \mathrm{Tr}\left(\mathbf{S}_N\sum_{i=1}^{M}\nu_i\mathbf{u}_i\mathbf{u}_i^T\right) \\
&= \mathrm{Tr}\left(\mathbf{S}_N\mathbf{\Phi}^T\mathbf{\Phi}\right) \\
&= \mathrm{Tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{S}_N\right).
\end{aligned} \tag{39}$$

Substituting the above result in $**$, we get

$$** = \mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N + \mathrm{Tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{S}_N\right). \tag{40}$$

[3] See Eq. (92) and Eq. (93) in Appendix D.1.

Using the above equation, Eq. (36) can be rewritten as

$$\begin{aligned}\mathbb{E}\left\{\log p(\mathbf{y} \mid \mathbf{w})\right\} =& \frac{N}{2}\log(\beta) - \frac{N}{2}\log(2\pi) - \frac{\beta}{2}\mathbf{y}^T\mathbf{y} \\ &+ \beta\mathbf{y}^T\mathbf{\Phi}\mathbf{m}_N - \frac{\beta}{2}\left(\mathbf{m}_N\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N + \operatorname{Tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{S}_N\right)\right). \end{aligned} \tag{41}$$

In a similar fashion, the conditional expectation of the prior probability is given by

$$\begin{aligned}\mathbb{E}\left\{\log p(\mathbf{w})\right\} =& -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log(\det(\mathbf{S}_0)) \\ &- \frac{1}{2}\left(\underbrace{\mathbb{E}\left\{\mathbf{w}^T\mathbf{S}_0^{-1}\mathbf{w}\right\}}_{=\mathbf{m}_N^T\mathbf{S}_0^{-1}\mathbf{m}_N+\operatorname{Tr}\left(\mathbf{S}_0^{-1}\mathbf{S}_N\right)} - \mathbf{m}_0^T\left(\mathbf{S}_0^{-1^T} + \mathbf{S}_0^{-1}\right)\underbrace{\mathbb{E}\left\{\mathbf{w}\right\}}_{=\mathbf{m}_N} + \mathbf{m}_0^T\mathbf{S}_0^{-1}\mathbf{m}_0\right) \\ =& -\frac{M}{2}\log(2\pi) - \frac{1}{2}\log(\det(\mathbf{S}_0)) - \frac{1}{2}(\mathbf{m}_N - \mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{m}_N - \mathbf{m}_0) \\ &- \frac{1}{2}\operatorname{Tr}\left(\mathbf{S}_0^{-1}\mathbf{S}_N\right). \end{aligned} \tag{42}$$

Substituting in Eq. (41) and Eq. (42) in Eq. (35), the expectation of the cost function $\mathcal{L}_{map}$ results in

$$\begin{aligned}\mathbb{E}\left\{\mathcal{L}_{map}(\mathbf{w})\right\} =& \frac{N}{2}\log(\beta) - \frac{N+M}{2}\log(2\pi) - \frac{1}{2}\log(\det(\mathbf{S}_0)) \\ &- \frac{\beta}{2}\mathbf{y}^T\mathbf{y} + \beta\mathbf{y}^T\mathbf{\Phi}\mathbf{m}_N - \frac{\beta}{2}\operatorname{Tr}(\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{S}_N) - \frac{\beta}{2}\mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N \\ &- \frac{1}{2}(\mathbf{m}_N - \mathbf{m}_0)^T\mathbf{S}_0^{-1}(\mathbf{m}_N - \mathbf{m}_0) - \frac{1}{2}\operatorname{Tr}\left(\mathbf{S}_0^{-1}\mathbf{S}_N\right). \end{aligned} \tag{43}$$

We can proceed to the M-step. In order to avoid overfitting of the mean value $\mathbf{m}_0$ caused by $\|\mathbf{S}_0\| \to 0$ and $\beta \to \infty$, thus, rendering the priors non informative, we need to introduce a constraint on $\mathbf{S}_0$. A natural bound for the estimated covariance is the CRLB from Eq. (22) as

$$\|\mathbf{S}_0\|^2 \geq \left\|\left(\beta\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2. \tag{44}$$

Using the Frobenius norm[4], we can express this constraint as

$$-g(\beta, \mathbf{S}_0) = \operatorname{Tr}\left(\mathbf{S}_0\mathbf{S}_0^T\right) - \frac{1}{\beta^2}\operatorname{Tr}\left(\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1^T}\right) \geq 0. \tag{45}$$

[4]See Eq. (98) in Appendix D.1.

We proceed to maximize $\mathbb{E}\{\mathcal{L}_{map}\}$ subject to the above constraint using the method of Lagrange multipliers. The Lagrangian function is given by

$$\begin{aligned}\mathcal{L}_{em}(\beta, \mathbf{m}_0, \mathbf{S}_0, \kappa) =& \frac{N}{2}\log(\beta) - \frac{N+M}{2}\log(2\pi) - \frac{1}{2}\log\left(\det(\mathbf{S}_0)\right) \\ &- \frac{\beta}{2}\mathbf{y}^T\mathbf{y} + \beta\mathbf{y}^T\mathbf{\Phi}\mathbf{m}_N - \frac{\beta}{2}\operatorname{Tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{S}_N\right) - \frac{\beta}{2}\mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N \\ &- \frac{1}{2}\left(\mathbf{m}_N - \mathbf{m}_0\right)^T\mathbf{S}_0^{-1}(\mathbf{m}_N - \mathbf{m}_0) - \frac{1}{2}\operatorname{Tr}\left(\mathbf{S}_0^{-1}\mathbf{S}_N\right) \\ &+ \kappa\left(\operatorname{Tr}\left(\mathbf{S}_0\mathbf{S}_0^T\right) - \frac{1}{\beta^2}\operatorname{Tr}\left(\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1^T}\right)\right). \end{aligned} \tag{46}$$

Since this optimization problem has an inequality constraint, we consider the Karush-Kuhn-Tucker (KKT) conditions [Boyd and Vandenberghe, 2004]. These conditions can be stated as follows:

$$\begin{aligned} &\text{Stationarity:} && \nabla_{\beta,\mathbf{m}_0,\mathbf{S}_0}\mathcal{L}_{em} = \mathbf{0} \\ &\text{Primal feasibility:} && g(\beta, \mathbf{S}_0) \leq 0 \\ &\text{Dual feasibility:} && \kappa \geq 0 \\ &\text{Complimentary slackness:} && \kappa g(\beta, \mathbf{S}_0) = 0, \end{aligned} \tag{47}$$

where κ is the KKT multiplier.

Since the covariance matrix $\mathbf{S}_0$ must be symmetric and positive definite[5]. Therefore we can express $\mathbf{S}_0$ as

$$\mathbf{S}_0 = \mathbf{L}\mathbf{L}^T, \tag{48}$$

where $\mathbf{L} \in \mathbb{R}^{K\times K}$ is an invertible matrix[6]. We can rewrite the cost function in terms of $\mathbf{L}$ as

$$\begin{aligned}\mathcal{L}_{em}(\beta, \mathbf{m}_0, \mathbf{L}, \kappa) =& \frac{N}{2}\log(\beta) - \frac{N+M}{2}\log(2\pi) - \frac{1}{2}\log(\det(\mathbf{L}\mathbf{L}^T)) \\ &- \frac{\beta}{2}\mathbf{y}^T\mathbf{y} + \beta\mathbf{y}^T\mathbf{\Phi}\mathbf{m}_N - \frac{\beta}{2}\operatorname{Tr}(\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{S}_N) - \frac{\beta}{2}\mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N \\ &- \frac{1}{2}(\mathbf{m}_N - \mathbf{m}_0)^T\mathbf{L}^{-1^T}\mathbf{L}^{-1}(\mathbf{m}_N - \mathbf{m}_0) - \frac{1}{2}\operatorname{Tr}\left(\mathbf{L}^{-1^T}\mathbf{L}^{-1}\mathbf{S}_N\right) \\ &+ \kappa\left(\operatorname{Tr}\left(\mathbf{L}\mathbf{L}^T\mathbf{L}^T\mathbf{L}\right) - \frac{1}{\beta^2}\operatorname{Tr}\left(\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1^T}\right)\right) \end{aligned} \tag{49}$$

[5]By definition, covariance matrices are positive semi-definite. Furthermore, an invertible positive semidefinite matrix is positive definite [Burden and Faires, 2010], and therefore, since we assume that $\mathbf{S}_0$ is invertible, $\mathbf{S}_0$ is positive definite.

[6]If $\mathbf{L}$ is a lower triangular matrix with non-negative elements in the diagonal, this is the Cholesky decomposition [Burden and Faires, 2010], nevertheless a particular structure of matrix $\mathbf{L}$ is not required for $\mathbf{S}_0^{-1}$ to be positive definite (See Section 9.6.6 in [Petersen and Pedersen, 2012]).

The partial derivative of the cost function with respect to β is given by

$$\begin{aligned}\frac{\partial \mathcal{L}_{em}}{\partial \beta} =& \frac{N}{2\beta} - \frac{1}{2}\mathbf{y}^T\mathbf{y} + \mathbf{y}^T\mathbf{\Phi}\mathbf{m}_N - \frac{1}{2}\operatorname{Tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{S}_N\right) - \frac{1}{2}\mathbf{m}_N\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N \\ &+ \frac{2\kappa}{\beta^3}\operatorname{Tr}\left(\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1^T}\right). \end{aligned} \tag{50}$$

The gradient of the cost function with respect to $\mathbf{m}_0$ is given by

$$\frac{\partial \mathcal{L}_{em}}{\partial \mathbf{m}_0} = \mathbf{L}^{-1^T}\mathbf{L}^{-1}(\mathbf{m}_N - \mathbf{m}_0). \tag{51}$$

For the sake of clarity of the calculation of the gradient of the cost function with respect to $\mathbf{L}$, we compute individually the partial derivatives of the the every non constant term of the cost function that has a dependency in $\mathbf{L}$:

$$\begin{aligned}\frac{\partial}{\partial \mathbf{L}}\log(\det(\mathbf{L}\mathbf{L}^T)) &= \frac{\partial}{\partial \mathbf{L}}\log(\det(\mathbf{L})\det(\mathbf{L}^T)) \\ &= 2\frac{\partial}{\partial \mathbf{L}}\log((\det(\mathbf{L}))) \\ &= 2\mathbf{L}^{-1^T}, \end{aligned} \tag{52}$$

$$\begin{aligned}\frac{\partial}{\partial \mathbf{L}}(\mathbf{m}_N - \mathbf{m}_0)^T(\mathbf{L}\mathbf{L})^{-1}(\mathbf{m}_N - \mathbf{m}_0) &= -(\mathbf{L}\mathbf{L}^T)^{-1}\frac{\partial(\mathbf{m}_N - \mathbf{m}_0)^T\mathbf{L}\mathbf{L}^T(\mathbf{m}_N - \mathbf{m}_0)}{\partial \mathbf{L}}(\mathbf{L}\mathbf{L}^T)^{-1} \\ &= -2(\mathbf{m}_N - \mathbf{m}_0)^T(\mathbf{m}_N - \mathbf{m}_0)(\mathbf{L}\mathbf{L}^T)^{-1}\mathbf{L}(\mathbf{L}\mathbf{L}^T)^{-1}, \end{aligned} \tag{53}$$

$$\begin{aligned}\frac{\partial}{\partial \mathbf{L}}\operatorname{Tr}\left(\left(\mathbf{L}\mathbf{L}^T\right)^{-1}\mathbf{S}_N\right) &= \frac{\partial}{\partial \mathbf{L}}\operatorname{Tr}\left(\mathbf{S}_N\left(\mathbf{L}\mathbf{L}^T\right)^{-1^T}\right) \\ &= \frac{\partial}{\partial \mathbf{L}}\operatorname{Tr}\left(\mathbf{S}_N\left(\mathbf{L}^T\mathbf{L}\right)^{-1}\right) \\ &= \frac{\partial}{\partial \mathbf{L}}\operatorname{Tr}\left(\left(\mathbf{L}^T\mathbf{E}\mathbf{L}\right)^{-1}\mathbf{S}_N\right) \\ &= -2\left(\mathbf{E}\mathbf{L}\left(\mathbf{L}^T\mathbf{E}\mathbf{L}\right)^{-1}\right)\mathbf{S}_N\left(\mathbf{L}^T\mathbf{E}\mathbf{L}\right)^{-1} \\ &= -2\mathbf{L}^{-1^T}\mathbf{S}_N\left(\mathbf{L}^T\mathbf{L}\right)^{-1}, \end{aligned} \tag{54}$$

$$\begin{aligned}\frac{\partial}{\partial \mathbf{L}}\operatorname{Tr}\left(\mathbf{L}\mathbf{L}^T\mathbf{L}^T\mathbf{L}\right) &= \frac{\partial}{\partial \mathbf{L}}\operatorname{Tr}\left(\left(\mathbf{L}\mathbf{L}^T\right)^T\left(\mathbf{L}^T\mathbf{L}\right)\right) \\ &= \frac{\partial}{\partial \mathbf{L}}\operatorname{Tr}\left(\mathbf{E}\mathbf{L}^T\mathbf{E}\mathbf{L}\mathbf{L}^T\mathbf{E}\mathbf{L}\mathbf{E}\right) \\ &= \mathbf{E}\mathbf{L}\mathbf{L}^T\mathbf{E}\mathbf{L}\mathbf{E}\mathbf{E}^T + \mathbf{E}\mathbf{L}\mathbf{E}\mathbf{E}^T\mathbf{L}^T\mathbf{E}\mathbf{L} \\ &\quad + \mathbf{E}\mathbf{L}\mathbf{E}\mathbf{E}\mathbf{L}^T\mathbf{E}\mathbf{L} + \mathbf{E}^T\mathbf{L}\mathbf{L}^T\mathbf{E}^T\mathbf{L}\mathbf{E}\mathbf{E} \\ &= 3\mathbf{L}\mathbf{L}^T\mathbf{L}. \end{aligned} \tag{55}$$

Using the above results, the partial derivative of the cost function with respect to $\mathbf{L}$ is given by

$$\begin{aligned}\frac{\partial \mathcal{L}_{em}}{\partial \mathbf{L}} =& -\mathbf{L}^{-1^T} + (\mathbf{m}_N - \mathbf{m}_0)^T(\mathbf{m}_N - \mathbf{m}_0)(\mathbf{L}\mathbf{L}^T)^{-1}\mathbf{L}(\mathbf{L}\mathbf{L}^T)^{-1} \\ &+ \mathbf{L}^{-1^T}\mathbf{S}_N\left(\mathbf{L}^T\mathbf{L}\right)^{-1} + 3\kappa\mathbf{L}\mathbf{L}^T\mathbf{L} \end{aligned} \tag{56}$$

Substituting Equations (50), (51) and (56) in the KKT conditions from Eq. (47), maximizing the cost function is equivalent to solve the following system of equations for β, $\mathbf{m}_0$, $\mathbf{S}_0$ and κ:

$$a\beta^3 - N\beta^2 - 4\kappa\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2 = 0 \tag{57}$$

$$\mathbf{L}^{-1^T}\mathbf{L}^{-1}(\mathbf{m}_N - \mathbf{m}_0) = \mathbf{0} \tag{58}$$

$$\begin{aligned}&-\mathbf{L}^{-1^T} + \mathbf{L}^{-1^T}\mathbf{S}_N\left(\mathbf{L}^T\mathbf{L}\right)^{-1} + 3\kappa\mathbf{L}\mathbf{L}^T\mathbf{L} - \mathbf{L}^{-1^T} \\ &+(\mathbf{m}_N - \mathbf{m}_0)^T(\mathbf{m}_N - \mathbf{m}_0)(\mathbf{L}\mathbf{L}^T)^{-1}\mathbf{L}(\mathbf{L}\mathbf{L}^T)^{-1} = \mathbf{0}_M \end{aligned} \tag{59}$$

$$\kappa\left(\frac{1}{\beta^2}\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2 - \mathrm{Tr}\left(\mathbf{L}\mathbf{L}^T\mathbf{L}^T\mathbf{L}\right)\right) = 0 \tag{60}$$

$$\left\|\mathbf{L}\mathbf{L}^T\right\|^2 - \frac{1}{\beta^2}\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2 \geq 0 \tag{61}$$

$$\kappa \geq 0, \tag{62}$$

where $\mathbf{0}_M$ represents the null matrix in $\mathbb{R}^{M\times M}$, and a is constant factor given by

$$a = \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{\Phi}\mathbf{m}_N + \mathrm{Tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{S}_N\right) + \mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N. \tag{63}$$

In order to satisfy the KKT constraints (Eq. (61) and Eq. (62)) we have two alternatives, namely $\kappa = 0$ or $\kappa > 0$ and $\left\|\mathbf{L}\mathbf{L}^T\right\|^2 = \frac{1}{\beta^2}\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2$. In the following, we consider the two cases individually:

(1) $\kappa = 0$

By substituting κ in the KKT conditions, it is easy to see that the parameters that maximizes the cost function are given by

$$\beta = \frac{N}{a} \tag{64}$$

$$\mathbf{m}_0 = \mathbf{m}_N \tag{65}$$

$$\mathbf{L}\mathbf{L}^T = \mathbf{S}_N. \tag{66}$$

(2) $\kappa > 0$.

Since Eq. (57) is a cubic polynomial equation in β, we can find the roots using the general formula [Spiegel and Liu, 1999] as

$$\beta_k = -\frac{1}{3a}\left(-N + u_k C + \frac{\Delta_0}{u_k C}\right), \tag{67}$$

where

$$u_1 = 1, \qquad u_2 = \frac{-1 + j\sqrt{3}}{2}, \qquad u_3 = \frac{-1 - j\sqrt{3}}{2}, \tag{68}$$

and

$$C = \sqrt[3]{\frac{\Delta_1 + \sqrt{\Delta_1^2 - 4\Delta_0^3}}{2}}, \tag{69}$$

with

$$\begin{aligned} \Delta_0 &= N^2 \\ \Delta_1 &= -2N^3 - 108a^2\kappa\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2. \end{aligned} \tag{70}$$

The discriminant for this polynomial equation is given as

$$\Delta = -16\kappa\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2\left(N^3 + 27a^2\kappa\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2\right) < 0, \tag{71}$$

which means that there is only one real solution [Spiegel and Liu, 1999]. Since β is the precision of the noise $\boldsymbol{\epsilon}$, it must be real valued and positive. Therefore, the only meaningful solution for β is taking u_1, which results in

$$\beta = \frac{1}{3a}\left(N + \frac{N^2 + 1}{\sqrt[3]{N^3 + 54a^2\kappa\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2 + 6\sqrt{3}\sqrt{a^2N^3\kappa\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2 + 27a^4\kappa^2\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^4}}}\right) \tag{72}$$

From Eq. (58), it follows that

$$\mathbf{m}_0 = \mathbf{m}_N. \tag{73}$$

Substituting the above result in Eq. (59), results in

$$\mathbf{L}\mathbf{L}^T = \mathbf{S}_N + 3\kappa\mathbf{L}\mathbf{L}^T\mathbf{L}\mathbf{L}^T\mathbf{L}\mathbf{L}^T. \tag{74}$$

Taking the trace of the above equation we can obtain κ as

$$\begin{aligned}
\operatorname{Tr}\left(3\kappa\left(\mathbf{L}\mathbf{L}^T\right)^3\right) &= \operatorname{Tr}\left(\mathbf{L}\mathbf{L}^T - \mathbf{S}_N\right) \\
3\kappa \operatorname{Tr}\left(\left(\mathbf{L}\mathbf{L}^T\right)^3\right) &= \operatorname{Tr}\left(\mathbf{L}\mathbf{L}^T - \mathbf{S}_N\right) \\
\kappa &= \frac{\operatorname{Tr}\left(\mathbf{L}\mathbf{L}^T - \mathbf{S}_N\right)}{3\operatorname{Tr}\left(\left(\mathbf{L}\mathbf{L}^T\right)^3\right)}.
\end{aligned} \tag{75}$$

Using the Woodbury identity[7] we can rewrite $\mathbf{S}_N$ from Eq. (34) as

$$\begin{aligned}
\mathbf{S}_N &= \mathbf{S}_0 - \mathbf{S}_0\mathbf{\Phi}^T\left(\frac{1}{\beta}\mathbf{E} + \mathbf{\Phi}\mathbf{S}_0\mathbf{\Phi}^T\right)^{-1}\mathbf{\Phi}\mathbf{S}_0 \\
&= \mathbf{L}\mathbf{L}^T - \mathbf{L}\mathbf{L}^T\mathbf{\Phi}^T\left(\frac{1}{\beta}\mathbf{E} + \mathbf{\Phi}\mathbf{L}\mathbf{L}^T\mathbf{\Phi}^T\right)^{-1}\mathbf{\Phi}\mathbf{L}\mathbf{L}^T.
\end{aligned} \tag{76}$$

Substituting the above equation in Eq. (75) we get

$$\kappa = \frac{\operatorname{Tr}\left(\mathbf{L}\mathbf{L}^T\mathbf{\Phi}^T\left(\frac{1}{\beta}\mathbf{E} + \mathbf{\Phi}\mathbf{L}\mathbf{L}^T\mathbf{\Phi}^T\right)^{-1}\mathbf{\Phi}\mathbf{L}\mathbf{L}^T\right)}{3\operatorname{Tr}\left(\left(\mathbf{L}\mathbf{L}^T\right)^3\right)}, \tag{77}$$

which is strictly positive, since $\mathbf{L}\mathbf{L}^T\mathbf{\Phi}^T\left(\frac{1}{\beta}\mathbf{E} + \mathbf{\Phi}\mathbf{L}\mathbf{L}^T\mathbf{\Phi}^T\right)^{-1}\mathbf{\Phi}\mathbf{L}\mathbf{L}^T$ is the product of positive definite matrices.

Both solutions satisfy the KKT conditions, nevertheless, the solution for $\kappa > 0$ enforces that the norm of the estimated covariance to attain the norm of the CRLB and bounds the value of β.

[7] See Eq. (99) in Appendix D.1

In Algorithm 1, the computation of the parameters of the prior distribution of the weights and the precision of the noise using the EM algorithm is summarized.

Data: $\mathbf{\Phi} \in \mathbb{R}^{N\times K}$, $\mathbf{y} \in \mathbb{R}^{N\times 1}$, $\varepsilon \in \mathbb{R}_{\geq 0}$: tolerance

Initialize $\mathbf{m}_0^{(old)}$, $\mathbf{L}^{(old)} = \text{Cholesky}\left(\mathbf{S}_0^{(old)}\right)$, $\beta^{(old)}$, $\kappa^{(old)}$, $\mathcal{L}^{(old)} := \mathcal{L}_{em}(\beta^{(old)}, \mathbf{m}_0^{(old)}, \mathbf{L}^{(old)}, \kappa^{(old)})$;

Set $Convergence := False$;

while $Convergence\ is\ False$ **do**

Compute new parameters

$$\mathbf{S}_N := \left(\left(\mathbf{L}^{(old)}\mathbf{L}^{(old)^T}\right)^{-1} + \beta^{(old)}\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1} \tag{78}$$

$$\mathbf{m}_N := \mathbf{S}_N\left(\left(\mathbf{L}^{(old)}\mathbf{L}^{(old)^T}\right)^{-1}\mathbf{m}_0^{(old)} + \beta^{(old)}\mathbf{\Phi}^T\mathbf{y}\right) \tag{79}$$

$$a := \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{\Phi}\mathbf{m}_N + \text{Tr}\left(\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{S}_N\right) + \mathbf{m}_N^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{m}_N \tag{80}$$

$$\mathbf{m}_0^{(new)} := \mathbf{m}_N \tag{81}$$

$$\mathbf{L}^{(new)} := \text{Cholesky}\left(\mathbf{S}_N + \kappa^{(old)}\left(\mathbf{L}^{(old)}\mathbf{L}^{(old)^T}\right)^3\right) \tag{82}$$

$$\beta^{(new)} := \frac{1}{3a}\left(N + \frac{N^2+1}{\sqrt[3]{N^3 + 54a^2\kappa^{(old)}\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2 + 6\sqrt{3}\sqrt{a^2N^3\kappa^{(old)}\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^2 + 27a^4\kappa^{(old)^2}\left\|\left(\mathbf{\Phi}^T\mathbf{\Phi}\right)^{-1}\right\|^4}}}\right) \tag{83}$$

$$\kappa^{(new)} := \frac{\text{Tr}\left(\mathbf{L}^{(old)}\mathbf{L}^{(old)^T} - \mathbf{S}_N\right)}{3\,\text{Tr}\left((\mathbf{L}\mathbf{L}^T)^3\right)} \tag{84}$$

$$\mathcal{L}^{(new)} := \mathcal{L}_{em}(\beta^{(new)}, \mathbf{m}_0^{(new)}, \mathbf{S}_0^{(new)}, \kappa^{(new)}) \tag{85}$$

if $|\mathcal{L}^{(old)} - \mathcal{L}^{(new)}| \leq \varepsilon$ **then**

Set $Convergence := True$;

else

Update parameters

$$\beta^{(old)} := \beta^{(new)} \qquad \mathbf{m}_0^{(old)} := \mathbf{m}_0^{(new)} \qquad \mathbf{L}^{(old)} := \mathbf{L}^{(new)} \qquad \kappa^{(old)} := \kappa^{(new)} \qquad \mathcal{L}^{(old)} := \mathcal{L}^{(new)} \tag{86}$$

end

end

return

$$\mathbf{m}_0 = \mathbf{m}_0^{(new)} \qquad \mathbf{S}_0 = \mathbf{L}^{(new)}\mathbf{L}^{(new)^T} \qquad \beta = \beta^{(new)} \tag{87}$$

Algorithm 1: EM Algorithm for estimating $\mathbf{m}_0$, $\mathbf{S}_0$ and β

Appendix D. Useful identities

In this appendix, a brief summary of the formulae and identities used in this report is provided.

D.1. **Linear Algebra.**

- Transposition. See [Spiegel and Liu, 1999]

$$(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T \tag{88}$$
$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \tag{89}$$
$$\mathbf{A} = \mathbf{A}^T \text{ if } \mathbf{A} \text{ is symmetric} \tag{90}$$

- Traces. See [Spiegel and Liu, 1999]

$$\mathrm{Tr}(\mathbf{A}+\mathbf{B}) = \mathrm{Tr}(\mathbf{A}) + \mathrm{Tr}(\mathbf{B}) \tag{91}$$
$$\mathrm{Tr}(\mathbf{ABC}) = \mathrm{Tr}(\mathbf{BCA}) = \mathrm{Tr}(\mathbf{CAB}) \tag{92}$$
$$\mathrm{Tr}(a) = a \ \forall a \in \mathbb{C} \tag{93}$$
$$\mathrm{Tr}(\mathbf{A}^T) = \mathrm{Tr}(\mathbf{A}) \tag{94}$$

- Determinant. See [Spiegel and Liu, 1999]

$$\det(\mathbf{AB}) = \det(\mathbf{A})\det(\mathbf{B}) \tag{95}$$
$$\det(\mathbf{A}^{-1}) = \frac{1}{\det(\mathbf{A})} \tag{96}$$
$$\det(\mathbf{A}^T) = \det(\mathbf{A}) \tag{97}$$

- The Frobenius norm. See [Petersen and Pedersen, 2012, Burden and Faires, 2010]

$$\|\mathbf{A}\|^2 = \mathrm{Tr}\left(\mathbf{AA}^T\right) \tag{98}$$

- The Woodbury identity. See [Petersen and Pedersen, 2012]

$$\left(\mathbf{A}+\mathbf{CBC}^T\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}\left(\mathbf{B}^{-1}+\mathbf{C}^T\mathbf{A}^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^T\mathbf{A}^{-1} \tag{99}$$

D.2. **Gaussian Distribution.** The probability density function (pdf) of a D-dimensional multivariate Gaussian distribution [Bishop, 2006] is given by

$$\mathcal{N}(\mathbf{x} \mid \mathbf{m}, \mathbf{S}) = \frac{1}{(2\pi)^D}\frac{1}{\sqrt{\det(\mathbf{S})}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T\mathbf{S}^{-1}(\mathbf{x}-\mathbf{m})\right\}, \tag{100}$$

where $\mathbf{x} \in \mathbb{R}^D$ is a random vector, $\mathbf{m} \in \mathbb{R}^D$ is the mean vector and $\mathbf{S} \in \mathbb{R}^{D\times D}$ is the covariance matrix.

Given a conditional distribution for $\mathbf{y}$ given $\mathbf{x}$ and a marginal distribution for $\mathbf{x}$ in the form

$$p(\mathbf{y} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}\right) \tag{101}$$

$$p(\mathbf{x}) = \mathcal{N}\left(\mathbf{x} \mid \mathbf{m}, \mathbf{\Lambda}^{-1}\right), \tag{102}$$

the conditional distribution of $\mathbf{x}$ given $\mathbf{y}$ and the marginal distribution of $\mathbf{y}$ are given by

$$p(\mathbf{x} \mid \mathbf{y}) = \mathcal{N}\left(\mathbf{x} \mid \mathbf{\Sigma}\left(\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda}\mathbf{m}\right), \mathbf{\Sigma}\right) \tag{103}$$

$$p(\mathbf{y}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{A}\mathbf{m} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\mathbf{\Lambda}^{-1}\mathbf{A}^T\right) \tag{104}$$

where

$$\mathbf{\Sigma} = \left(\mathbf{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A}\right)^{-1}. \tag{105}$$

For a detailed derivation of this result, see [Bishop, 2006, pp. 90-93].