Tutorial: Sparse learning in linear regression

Rimple Sandhu
Department of Civil Engineering, Carleton University, Ottawa, Canada
Email: rimple_sandhu@outlook.com

November 19, 2018

Contents

1 Problem definition: Linear regression

An inverse problem is defined as the task of estimating unknown parameters of the proposed mathematical model using observations from the system. When the observations comprise of input-output samples the inverse problem is known as supervised learning. Supervised learning has two forms: 1) Regression, and 2) Classification. Regression deals with real-valued observations, while classification deals with discrete-valued observations. In this text we focus on the inverse problem of linear regression where the mathematical model is defined as

$$y^{(i)} = \sum_{j=0}^{N-1} \phi_j(\mathbf{x}^{(i)}) w_j + \epsilon^{(i)} \quad \Rightarrow \quad \mathbf{y} = \Phi \mathbf{w} + \boldsymbol{\epsilon} \tag{1}$$

where $\phi_j(.)$ is a proposed basis function, $\mathbf{x} \in \mathbb{R}^{d \times 1}$ is the model input (independent variable), $\mathbf{y} \in \mathbb{R}^{M \times 1}$ is the model output (dependent variable), $\mathbf{w} \in \mathbb{R}^{N \times 1}$ is the unknown model parameter, $\boldsymbol{\epsilon} \in \mathbb{R}^{M \times 1}$ is the model discrepancy error and $\Phi \in \mathbb{R}^{M \times N}$ is the design matrix. Model discrepancy error $\boldsymbol{\epsilon}$ quantifies the modelling inadequacy of the proposed model as well as the noise in the observations due to an imperfect data-acquisition process. In this text, model error $\boldsymbol{\epsilon}$ is assumed as Gaussian, meaning $\mathrm{p}(\epsilon_i) = \mathcal{N}(0, \rho^{-1})$ or $\mathrm{p}(\boldsymbol{\epsilon}) = \mathcal{N}(0, \Gamma)$, where ρ is the model error precision and $\Gamma = \rho^{-1}\mathbf{I}_M$.

The basis functions $\phi_j(\mathbf{x})$ in Eq. (1) are aimed at transforming the input space $\mathbf{x}$ into a new space of variables whose scale and variation aligns well with $\mathbf{y}$. This transformed space $\phi_j(\mathbf{x})$, also known as a *feature* of the model, ensures a good match between input and output observations. Most commonly used basis functions in practice are polynomial chaos expansion (PCE), Radial basis function (RBF), support vector machine (SVM), and many others. The model in Eq. (1) can be thought of as a decomposition of the underlying input-output relation as a weighted sum of unique features. These features or basis functions in Eq. (1) are proposed based on modeller's prior knowledge about the system and/or model usage requirements. In most practical scenarios the proposed models often contain more features than it is required to obtain a reasonably good fit to the observations (data-fit). This flexibility of the model degrades the predictive performance of the model for predicting unobserved data, despite providing a good fit to the available observations. Sparse learning algorithms are aimed at identifying the subset of proposed features that are sufficient enough for providing a reasonable data-fit and a good prediction capability. The removal of redundant features from the flexible model will also help expedite the learning and prediction phase due to the reduced model dimensionality (N). Sparse learning is also referred to as feature selection or variable selection in machine learning.

In this text, we will focus on following sparse learning techniques:

- Frequentist: Least absolute shrinkage and selection operator (LASSO) regression
- Bayesian: Sparse Bayesian learning (SBL)
- Bayesian: Bayesian compressive sensing (BCS)

We intend to provide a standalone mathematical exposition into the derivation of each of the above algorithms, followed by a rigorous numerical investigation. In Section 2 we detail

the frequentist techniques to parameter estimation including LASSO. In Section 3, we detail the Bayesian framework for analytical evaluation of $\mathbf{w}$ using the concept of conjugate priors, later used in deriving SBL and BCS. In section 4 we provide a detailed derivation of two popular forms of SBL algorithm. In section 5 we detail the derivation of BCS algorithm along with its relation to SBL. In Section 7 we report and discuss the numerical performance of the sparse learning techniques when applied to the case of polynomial regression.

2 Frequentist approach to sparse learning

Before proceeding to sparse learning, lets review the frequentist approaches to parameter estimation in the context of linear regression. The most popular frequentist approach of estimating $\mathbf{w}$ in linear regression (Eq. (1)) is the ordinary least-square (OLS) estimation. OLS is based on the assumptions of 1) linearity of the model ($\mathbf{y} = \Phi\mathbf{w} + \boldsymbol{\epsilon}$), 2) zero-mean model error ($\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$), 3) iid model errors ($\mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \mathbf{I}_M\rho^{-1}$), and 4) no multicolinearity within basis functions ($\mathrm{Rank}(\Phi) = N$). Following these assumptions, the OLS solution $\mathbf{w}_{\mathrm{ols}}$ is obtained as

$$\mathbf{w}_{\mathrm{ols}} = \arg\min_{\mathbf{w}}\{\mathrm{E}_D(\mathbf{w})\} = \arg\min_{\mathbf{w}}\left\{\frac{1}{2}||\mathbf{y} - \Phi\mathbf{w}||_2^2\right\} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y} \tag{2}$$

where $||.||_2$ denotes the L_2-norm and $\mathrm{E}_D(\mathbf{w})$ is the OLS cost function. The convex optimization task in Eq. (2) is solved by differentiating $\mathrm{E}_D(\mathbf{w})$ analytically with respect to $\mathbf{w}$ and then substituted to zero to obtain the OLS estimate $\mathbf{w}_{\mathrm{ols}}$=$(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$. The assumption of no multicolinearity among basis functions implies that no two columns of design matrix Φ are linearly dependent, thereby ensuring the matrix $\Phi^T\Phi$ in Eq. (2) is invertible. Also, the probability distribution of model error $\boldsymbol{\epsilon}$ is not needed for computing $\mathbf{w}_{\mathrm{ols}}$ using Eq. (2), and $\boldsymbol{\epsilon}$ only need to satisfy the properties $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \mathbf{I}_M\rho^{-1}$. Given $\mathbf{w}_{\mathrm{ols}}$, an unbiased estimate for error precision ρ can also be obtained analytically as $\rho_{\mathrm{ols}} = (M-N)/||\mathbf{y}-\Phi\mathbf{w}_{\mathrm{ols}}||_2^2$. The popularity of OLS estimation is attributed to the following special properties that $\mathbf{w}_{\mathrm{ols}}$ holds: 1) Linearity ($\mathbf{w}_{\mathrm{ols}}$ is a linear function of $\mathbf{y}$), 2) Unbiasedness ($\mathrm{E}[\mathbf{w}_{\mathrm{ols}}] = \mathbf{w}$), 3) Efficiency (Variance of $\mathbf{w}_{\mathrm{ols}}$ is minimum for all unbiased estimators of $\mathbf{w}$) [10]. As a result, $\mathbf{w}_{\mathrm{ols}}$ is also known as the best (efficient) linear unbiased estimator (BLUE) of $\mathbf{w}$.

Maximum likelihood estimation (MLE) is an another popular frequentist technique of estimating $\mathbf{w}$ by maximizing the log-likelihood function $\log \mathrm{p}(\mathbf{y}|\mathbf{w})$. Unlike OLS, MLE is applicable only when the probability distribution of model error $\boldsymbol{\epsilon}$ is known. Gaussian model errors are a common occurrence in practice since a Gaussian distribution contains maximum entropy (least information) for an unbounded, finite variance random variable [11]. Henceforth, assuming Gaussian model errors introduces little interference into estimation from modeller's perspective. In the case of Gaussian errors, i.e. $\mathrm{p}(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \mathbf{I}_M\rho^{-1})$, the likelihood function is obtained as $\mathrm{p}(\mathbf{y}|\mathbf{w}, \rho) = \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \mathbf{I}_M\rho^{-1})$. The MLE solution $\mathbf{w}_{\mathrm{mle}}$ is

obtained as

$$\mathbf{w}_{\text{mle}} = \arg\max_{\mathbf{w}} \{\log \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \mathbf{I}_M \rho^{-1})\} \tag{3}$$

$$= \arg\max_{\mathbf{w}} \left\{ -\frac{N}{2}\log 2\pi + \frac{1}{2}\log\rho - \rho \mathrm{E}_D(\mathbf{w}) \right\} \tag{4}$$

$$= \arg\min_{\mathbf{w}} \{\mathrm{E}_D(\mathbf{w})\} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y} = \mathbf{w}_{\text{ols}} \tag{5}$$

Hence, MLE provides the same solution as OLS for $\mathbf{w}$ when the model errors ϵ_i in Eq. (1) are Gaussian. The MLE solution for ρ is different from the OLS solution and is obtained analytically as $\rho_{\text{mle}} = M/||\mathbf{y} - \Phi\mathbf{w}_{\text{mle}}||_2^2$.

Ideally speaking, we should have multiple realizations of $\mathbf{y}$ where each realization leads to a unique point estimate $\mathbf{w}_{\text{ols}}$ or $\mathbf{w}_{\text{mle}}$. The histogram of multiple point estimates is known as the sampling distribution of $\mathbf{w}$, which is subsequently used in computing confidence interval on model predictions. However, in practice, we only have a limited set of observations and sampling distribution is not available, except in the following two scenarios: 1) Gaussian model errors, and 2) "Large" data (large M). In both the cases the sampling distribution of $\mathbf{w}$ can be obtained analytically using a single realization $\mathbf{y}$ as $\mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}}, (\rho\Phi^T\Phi)^{-1})$ for both OLS and MLE techniques.

When dealing with flexible models and noisy observations, the OLS/MLE technique suffer from the following disadvantages:

- **Under-determined system**: When the number of features exceed the number of observations ($N > M$), the OLS/MLE estimation is ill-posed as there exist infinite solutions for $\mathbf{w}$.

- **Multicollinearity**: Even for an over-determined system ($N < M$) the OLS/MLE estimation can become ill-posed if some features are linearly related to each other. This overlap among features is called multicolinearity, causing an increase in the condition number of matrix $\Phi^T\Phi$. The ill-conditioned matrix $\Phi^T\Phi$ causes the OLS/MLE solution $(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$ to be highly sensitive to the noise in observations $\mathbf{y}$, making the OLS/MLE estimation ill-posed. Note that matrix $\Phi^T\Phi$ is singular for the case of perfect collinearity among features since $\text{Rank}(\Phi^T\Phi) < N$.

- **Overfitting**: For a well-posed estimation involving a flexible model, the OLS/MLE solution is heavily dependent on a single realization of the observation. A flexible model tends to capture the noise in observations leading to overfitting and poor generalization to unseen observations.

Regularized least-square (RLS) is a family of methods designed to remedy these issues with OLS/MLE estimation by means of regularization. A general form of RLS estimation of $\mathbf{w}$ is defined as

$$\mathbf{w}_{\text{rls}} = \arg\min_{\mathbf{w}} \left\{ \frac{1}{2}||\mathbf{y} - \Phi\mathbf{w}||_2^2 + \frac{\lambda}{2}\sum_{i=1}^{N} |w_i|^p \right\} = \arg\min_{\mathbf{w}} \left\{ \mathrm{E}_D(\mathbf{w}) + \lambda \mathrm{E}_w(\mathbf{w}) \right\} \tag{6}$$

where $\mathrm{E}_w(\mathbf{w})$ is the regularization term, $\mathrm{E}_D(\mathbf{w})$ is the OLS cost function, and $\lambda > 0$ is the regularization coefficient. The mean term w_0 is excluded from the term $\mathrm{E}_w(\mathbf{w})$ to eliminate the dependence of the regularization on the origin (or scale) of the observations. The following two types of regularization is popular in practice [10]: 1) Least absolute shrinkage and selection operator (LASSO), 2) Ridge regression. LASSO uses L_1-norm regularization ($p = 1$ in Eq. (6)) while Ridge regression uses L_2-norm regularization ($p = 2$ in Eq. (6)), also known as Tikhonov regularization. Ridge regression cost function (p=2 in Eq. (6)) can be differentiated analytically and substituted to zero to obtain the estimate $\mathbf{w}_{\text{ridge}} = (\Phi^T\Phi + \lambda\mathbf{I}_M)^{-1}\Phi^T\mathbf{y}$. On the contrary, LASSO cost function (p=1 in Eq. (6)) cannot be differentiated analytically and therefore the LASSO solution $\mathbf{w}_{\text{lasso}}$ is computed through numerical optimization [10]. The least angle regression (LARS) algorithm provides an efficient alternative to the numerical optimization for providing the LASSO solution for a range of λ values though a stepwise regression approach [5]. A summary of the frequentist approach to linear regression in provided in Table 1.

Method	Cost function	$\mathbf{w}$ estimate	Precision of $\mathbf{w}$ estimate
OLS	$\frac{1}{2}\|\|\mathbf{y} - \Phi\mathbf{w}\|\|_2^2$	$(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$	$\rho\Phi^T\Phi$
LASSO	$\frac{1}{2}\|\|\mathbf{y} - \Phi\mathbf{w}\|\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{N}\|w_i\|$	Numerically (LARS*)	Numerically
Ridge	$\frac{1}{2}\|\|\mathbf{y} - \Phi\mathbf{w}\|\|_2^2 + \frac{\lambda}{2}\sum_{i=1}^{N}\|w_i\|^2$	$(\Phi^T\Phi + \lambda\mathbf{I})^{-1}\Phi^T\mathbf{y}$	$\rho(\Phi^T\Phi + \lambda\mathbf{I})(\Phi^T\Phi)^{-1}(\Phi^T\Phi + \lambda\mathbf{I})$

Table 1: Summary of frequentist estimation techniques for $\mathbf{w}$. (*LARS stands for Least-angle regression)

The regularization term $\mathrm{E}_w(\mathbf{w})$ with non-zero λ in Eq. (6) helps alleviate the aforementioned issues with the OLS/MLE techniques. For example, computing $\mathbf{w}_{\text{ridge}} = (\Phi^T\Phi + \lambda\mathbf{I}_M)^{-1}\Phi^T\mathbf{y}$ requires the inverse of matrix $\Phi^T\Phi + \lambda\mathbf{I}_M$, which is always invertible even when matrix $\Phi^T\Phi$ is singular either due to under-determined system ($N > M$) or due to multi-collinearity among features. The same conclusion goes for LASSO. Hence, LASSO and Ridge regression estimation is well-posed due to the regularization, even though the corresponding OLS estimation is ill-posed. In addition, the regularization term $\mathrm{E}_D(\mathbf{w})$ acts as a penalty on large values of $\mathbf{w}$ thereby limiting the flexibility of the model, leading to parsimonious models. The amount of penalty is dictated by the regularization coefficient λ. A large λ will over-penalize the model flexibility, thereby producing a model that fails to capture even the salient trends in the observations. On the contrary, a small λ value will be unable to limit the flexibility of the model leading to overfitting by allowing large $\mathbf{w}$ values. Therefore, there exist an optimal value of λ that provides the best balance between model parsimony and data-fit.

In practice, the optimal λ is chosen using the cross-validation (CV) technique of minimizing the model prediction error. The CV prediction error is computed for various values of λ and the optimal λ is chosen corresponding to the minimum CV prediction error. The calculation of CV prediction error operates by randomly grouping the M number of observations into K groups as $\{\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, \ldots, \mathbf{y}_{(K)}\}$ where each group contains $M' = M/K$ observations.

For each k^{th} group, LASSO or Ridge regression is used to estimate $\mathbf{w}$ using observations from all but k^{th} group, denoted as $\mathbf{y}_{(-k)}$. Subsequently, the CV prediction error for a given λ is computed as

$$\mathrm{E}_K(\lambda) = \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{M'}(y_{(-k)}^{(j)} - \hat{y}_{(-k)}^{(j)})^2 \tag{7}$$

where $\hat{y}_{(-k)}^{(j)}$ is the model prediction pertaining to j^{th} observation. The optimal λ is chosen corresponding to the minimum CV error $\mathrm{E}_K(\lambda)$. This form of cross-validation where the observations are randomly segmented into K groups is called K-fold cross-validation (K-fold CV). Multiple estimates of K-fold CV error $\mathrm{E}_K(\lambda)$ can be obtained for each λ through multiple instances of random grouping of observations. This repeated estimation of K-fold CV error provides the variation in CV error, thereby ensuring prudent selection of optimal λ. Studies have shown that choosing K between 5 to 10 ensures minimum variation of CV error for any given λ [10]. When $K = M$, the technique becomes leave-one-out cross-validation (LVOCV). The LVOCV error can be computed without performing the grouping task and just using the $\mathbf{w}$ estimate computed using all the observations $\mathbf{y}$ as [10]

$$\mathrm{E}_M(\lambda) = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{y^{(j)} - \hat{y}^{(j)}}{1 - h_{jj}}\right)^2 \tag{8}$$

where h_{jj} is the j^{th} diagonal term of matrix $\Phi(\Phi^T\Phi)^{-1}\Phi^T$. This estimation of λ by minimizing CV error ensures an optimal amount of penalty in the LASSO or Ridge regression cost function in Eq. (6).

The regularization has different effect on $\mathbf{w}$ values for LASSO and Ridge regression. In Ridge regression, increasing λ value will increasingly force many parameters towards zero, but never exactly zero [18]. On the contrary, increasing λ in LASSO forces increasingly number of parameters to be exactly zero, thereby providing a sparse solution to $\mathbf{w}$. This sparsity inducing property of LASSO is the reason behind its widespread usage as a sparse learning tool. We will revisit these differences between LASSO and Ridge regression and their applicability as a sparse learning tool during numerical investigations.

3 Bayesian linear regression

This section details the Bayesian perspective for estimating $\mathbf{w}$ in the linear regression setup presented in Eq. (1). This Bayesian estimation framework is termed as Bayesian linear regression (BLR) and it will eventually be used to lay the foundation of Bayesian sparse learning algorithms. In Bayesian framework, the unknown model parameters in $\mathbf{w}$ are treated as random variables whose posterior pdf $\mathrm{p}(\mathbf{w}|\mathbf{y}, \rho)$ is obtained using Bayes' theorem as

$$\mathrm{p}(\mathbf{w}|\mathbf{y}, \rho) = \frac{\mathrm{p}(\mathbf{y}|\mathbf{w}, \rho)\mathrm{p}(\mathbf{w})}{\mathrm{p}(\mathbf{y}|\rho)} \tag{9}$$

where $\mathrm{p}(\mathbf{y}|\mathbf{w}, \rho) = \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \Gamma)$ is the likelihood function, $\mathrm{p}(\mathbf{w})$ is the prior pdf and $\mathrm{p}(\mathbf{y}|\rho) = \int \mathrm{p}(\mathbf{y}|\mathbf{w}, \rho)\mathrm{p}(\mathbf{w})d\mathbf{w}$ is the model evidence or marginal likelihood. Although we have assumed

that ρ is known a priori, the joint posterior pdf $\mathrm{p}(\mathbf{w}, \rho|\mathbf{y})$ can be computed using a Normal-Gamma conjugate prior for $\mathbf{w}$ and ρ (not discussed here). Nevertheless, the Bayes' rule in Eq. (9) facilitates the inclusion of prior information through $\mathrm{p}(\mathbf{w})$ and the information contained in the observations through $\mathrm{p}(\mathbf{y}|\mathbf{w}, \rho)$ to provide the updated knowledge about $\mathbf{w}$ in the form of posterior pdf $\mathrm{p}(\mathbf{w}|\mathbf{y}, \rho)$. The model evidence $\mathrm{p}(\mathbf{y}|\rho)$ acts only as a normalizing constant in Eq. (9) and can be ignored for the sake of computing posterior pdf. Essentially, Eq. (9) can be rewritten as $\mathrm{p}(\mathbf{w}|\mathbf{y}, \rho) \propto \mathrm{p}(\mathbf{y}|\mathbf{w}, \rho)\mathrm{p}(\mathbf{w})$.

In most practical scenarios an analytical evaluation of posterior $\mathrm{p}(\mathbf{w}|\mathbf{y}, \rho)$ using Eq. (9) is not feasible and stochastic simulation techniques such as Markov chain Monte Carlo (MCMC) are used to generate stationary samples from the unnormalized posterior $\mathrm{p}(\mathbf{w}|\mathbf{y}, \rho) \propto \mathrm{p}(\mathbf{y}|\mathbf{w}, \rho)\mathrm{p}(\mathbf{w})$ [8]. The stochastic simulation techniques are often computationally expensive and thus create a bottleneck for an efficient application of Bayesian methods to inverse problems. However, an analytical evaluation of the posterior pdf is rendered feasible for the special case of linear regression where the prior pdf $\mathrm{p}(\mathbf{w})$ is chosen as a conjugate prior to the given likelihood function. A conjugate prior is defined as the prior that produces posterior in the same probability distribution family as the prior [7]. The popularity of conjugate priors in Bayesian linear regression is attributed to its ability to obtain closed-form solution to the posterior pdf for large parameter dimensionality with little computational effort.

For the Gaussian likelihood function $\mathrm{p}(\mathbf{y}|\mathbf{w}, \rho) = \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \Gamma)$ defined in Eq. (1), the conjugate prior for $\mathbf{w}$ is also Gaussian and is assigned as $\mathrm{p}(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{P}_0)$ where $\mathbf{m}_0 \in \mathbb{R}^{N\times 1}$ is the prior mean vector and $\mathbf{P}_0 \in \mathbb{R}^{N\times N}$ is the prior covariance matrix [7]. The analytical evaluation of posterior $\mathrm{p}(\mathbf{w}|\mathbf{y}, \rho)$ requires computation of following relations:

$$\mathrm{E}[\mathbf{y}] = \Phi\mathrm{E}[\mathbf{w}] + \mathrm{E}[\boldsymbol{\epsilon}] = \Phi\mathbf{m}_0 \tag{10a}$$

$$\mathrm{Var}(\mathbf{y}) = \Phi\mathrm{Var}(\mathbf{w})\Phi^T + \mathrm{Var}(\boldsymbol{\epsilon}) = \Phi\mathbf{P}_0\Phi^T + \Gamma$$

$$\mathrm{E}[(\mathbf{w} - \mathbf{m}_0)(\mathbf{y} - \Phi\mathbf{m}_0)^T] = \mathrm{E}[(\mathbf{w} - \mathbf{m}_0)(\Phi\mathbf{w} + \boldsymbol{\epsilon} - \Phi\mathbf{m}_0)^T] \tag{10b}$$

$$= \mathrm{E}[(\mathbf{w} - \mathbf{m}_0)(\mathbf{w} - \mathbf{m}_0)^T]\Phi^T + \cancelto{0}{\mathrm{E}[(\mathbf{w} - \mathbf{m}_0)\boldsymbol{\epsilon}^T]}$$

$$= \mathbf{P}_0\Phi^T \tag{10c}$$

Using the relations from Eq. (10), the joint pdf of $(\mathbf{w}, \mathbf{y})$ is computed as

$$\mathrm{p}(\mathbf{w}, \mathbf{y}|\rho) = \mathcal{N}\left(\begin{Bmatrix}\mathbf{w}\\ \mathbf{y}\end{Bmatrix} \middle| \begin{Bmatrix}\mathbf{m}_0\\ \Phi\mathbf{m}_0\end{Bmatrix}, \begin{bmatrix}\mathbf{P}_0 & \mathbf{P}_0\Phi^T\\ \Phi\mathbf{P}_0^T & \Phi\mathbf{P}_0\Phi^T + \Gamma\end{bmatrix}\right) \tag{11}$$

Using the Gaussian conditional distribution formula (listed in Appendix A.1) and Eq. (11), the posterior pdf of $\mathbf{w}$ is obtained as

$$\mathrm{p}(\mathbf{w}|\mathbf{y}, \rho) = \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P}) \tag{12a}$$

$$\mathbf{m} = \mathbf{m}_0 + \mathbf{K}(\mathbf{y} - \Phi\mathbf{m}_0) \tag{12b}$$

$$\mathbf{P} = (\mathbf{I} - \mathbf{K}\Phi)\mathbf{P}_0 \tag{12c}$$

$$\mathbf{K} = \mathbf{P}_0\Phi^T(\Phi\mathbf{P}_0\Phi^T + \Gamma)^{-1} \tag{12d}$$

where $\mathbf{m}$ is the posterior mean, $\mathbf{P}$ is the posterior covariance, and $\mathbf{K}$ is the gain matrix. Note that $\mathbf{m}$ is also the mode of posterior $\mathrm{p}(\mathbf{w}|\mathbf{y}, \rho)$ and is called as the *maximum a posteriori*

(MAP) estimate. Eq. (12) represents the Bayesian updating formula that also laid the foundation for many data assimilation and filtering techniques such as Kalman filter [9].

Model evidence $\mathrm{p}(\mathbf{y}|\rho)$ (denominator in Eq. (9)) can also be evaluated analytically using expressions for $\mathrm{E}[\mathbf{y}]$ and $\mathrm{Var}(\mathbf{y})$ in Eq. (10) to obtain $\mathrm{p}(\mathbf{y}|\rho) = \mathcal{N}(\mathbf{y}|\Phi\mathbf{m}_o, \Phi\mathbf{P}_0\Phi^T + \Gamma)$. Model evidence is a key entity in Bayesian model comparison due its automatic Ockham's razor property of providing optimal trade-off between data-fit and model complexity. This property of the evidence is realized when writing the logarithm of model evidence $\mathrm{p}(\mathbf{y}|\rho)$ as (using Eq. (9)) [15]

$$\begin{aligned}
\log \mathrm{p}(\mathbf{y}|\rho) &= \log \mathrm{p}(\mathbf{y}|\rho) \int \mathrm{p}(\mathbf{w}|\mathbf{y},\rho) d\mathbf{w} \\
&= \int \log\left(\frac{\mathrm{p}(\mathbf{y}|\mathbf{w},\rho)\mathrm{p}(\mathbf{w})}{\mathrm{p}(\mathbf{w}|\mathbf{y},\rho)}\right) \mathrm{p}(\mathbf{w}|\mathbf{y},\rho) d\mathbf{w} \\
&= \int \log \mathrm{p}(\mathbf{y}|\mathbf{w},\rho)\mathrm{p}(\mathbf{w}|\mathbf{y},\rho) d\mathbf{w} - \int \log\left(\frac{\mathrm{p}(\mathbf{w}|\mathbf{y},\rho)}{\mathrm{p}(\mathbf{w})}\right) \mathrm{p}(\mathbf{w}|\mathbf{y},\rho) d\mathbf{w} \\
&= \underbrace{\mathrm{E}[\log \mathrm{p}(\mathbf{y}|\mathbf{w},\rho)]}_{\text{Goodness-of-fit (GOF)}} - \underbrace{\mathrm{E}\left[\log\left(\frac{\mathrm{p}(\mathbf{w}|\mathbf{y},\rho)}{\mathrm{p}(\mathbf{w})}\right)\right]}_{\text{Expected Information Gain (EIG)}}
\end{aligned} \tag{13}$$

where the expectation $\mathrm{E}[.]$ is with respect to posterior $\mathrm{p}(\mathbf{w}|\mathbf{y},\rho)$. The goodness-of-fit (GOF) term quantifies the data-fit capability of the model as it relates to the likelihood function. The expected information gain (EIG) term captures the Kullback-Leibler (KL) divergence between the prior and posterior pdf, which in turn quantifies model complexity. For complex models the posterior is vastly different from the prior due to the overfitting caused by the flexibility of the model. In turn, the KL divergence or EIG is higher for complex models which then penalizes the model by reducing the log-evidence according to Eq. (13). This competing nature of GOF and EIG imparts model evidence the quantitative Ockham's razor property that promotes compromise between data-fit and model complexity.

For the current case of linear regression and Gaussian model errors $\boldsymbol{\epsilon}$, the GOF term in Eq. (13) can be calculated analytically using relations in Appendix A.2 and A.3 as

$$\begin{aligned}
\mathrm{GOF} &= \mathrm{E}[\log \mathrm{p}(\mathbf{y}|\mathbf{w},\rho)] \\
&= \mathrm{E}[\log \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho^{-1}\mathbf{I}_M)] \\
&= \mathrm{E}[\log(c\mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{mle}}, \mathbf{P}_{\mathrm{mle}}))] \\
&= \log c + \int \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{P}) \log \mathcal{N}(\mathbf{w}|\mathbf{w}_{\mathrm{mle}}, \mathbf{P}_{\mathrm{mle}}) d\mathbf{w} \\
&= \log c + \log \mathcal{N}(\mathbf{m}|\mathbf{w}_{\mathrm{mle}}, \mathbf{P}_{\mathrm{mle}}) - \frac{1}{2}\mathrm{Trace}(\mathbf{P}_{\mathrm{mle}}^{-1}\mathbf{P}) \\
&= \log \mathcal{N}(\mathbf{y}|\Phi\mathbf{m}, \rho^{-1}\mathbf{I}_M)) - \frac{1}{2}\mathrm{Trace}(\mathbf{P}_{\mathrm{mle}}^{-1}\mathbf{P})
\end{aligned} \tag{14}$$

The EIG term in Eq. (13) can also be evaluated analytically as

$$\begin{aligned}
\text{EIG} &= \text{E}\left[\log\left(\frac{\text{p}(\mathbf{w}|\mathbf{y},\rho)}{\text{p}(\mathbf{w})}\right)\right] \\
&= \text{E}[\log \text{p}(\mathbf{w}|\mathbf{y},\rho)] - \text{E}[\log \text{p}(\mathbf{w})] \\
&= \text{E}[\log \mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{P})] - \text{E}[\log \mathcal{N}(\mathbf{w}|\mathbf{m}_0,\mathbf{P}_0)] \\
&= \log \mathcal{N}(\mathbf{m}|\mathbf{m},\mathbf{P}) - \frac{1}{2}\text{Trace}(\mathbf{P}^{-1}\mathbf{P}) - \log \mathcal{N}(\mathbf{m}|\mathbf{m}_0,\mathbf{P}_0) + \frac{1}{2}\text{Trace}(\mathbf{P}_0^{-1}\mathbf{P}) \qquad (15) \\
&= \log\left(\frac{\mathcal{N}(\mathbf{m}|\mathbf{m},\mathbf{P})}{\mathcal{N}(\mathbf{m}|\mathbf{m}_0,\mathbf{P}_0)}\right) - \frac{N}{2} + \frac{1}{2}\text{Trace}(\mathbf{P}_0^{-1}\mathbf{P}) \qquad (16)
\end{aligned}$$

Note that when the posterior $\text{p}(\mathbf{w}|\mathbf{y},\rho)$ is entirely dictated by observations $\mathbf{y}$ the posterior precision $\mathbf{P}^{-1}$ will be same as the OLS/MLE precision $\mathbf{P}_{\text{mle}}^{-1}$ and hence $\text{Trace}(\mathbf{P}_{\text{mle}}^{-1}\mathbf{P}) \approx N$ in Eq. (14) and $\text{Trace}(\mathbf{P}_0^{-1}\mathbf{P}) \approx 0$ in Eq. (16).

Next, the posterior predictive distribution of unseen observations $\tilde{\mathbf{y}}$ can be computed using the new design matrix $\tilde{\Phi}$ (computed at new input $\tilde{\mathbf{x}}$ of our choice) as

$$\begin{aligned}
\text{p}(\tilde{\mathbf{y}}|\mathbf{y},\rho) &= \int \text{p}(\tilde{\mathbf{y}}|\mathbf{w},\rho)\text{p}(\mathbf{w}|\mathbf{y},\rho)d\mathbf{w} \\
&= \int \mathcal{N}(\tilde{\mathbf{y}}|\tilde{\Phi}\mathbf{w},\Gamma)\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{P})d\mathbf{w} \\
&= \mathcal{N}(\tilde{\mathbf{y}}|\tilde{\Phi}\mathbf{m},\tilde{\Phi}\mathbf{P}\tilde{\Phi}^T + \Gamma) \qquad (17)
\end{aligned}$$

where $\Gamma = \rho^{-1}\mathbf{I}_M$ is known. Table 2 provides the summary of BLR estimation of $\mathbf{w}$ using conjugate priors and a known model error precision.

Next, we provide an interesting contrast between the posterior pdf from BLR in Eq. (12) and the sampling distribution of $\mathbf{w}$ obtained using the frequentist approaches of OLS/MLE [1]. As demonstrated in Section 2 the OLS/MLE solution for the case of linear regression and Gaussian model errors $\boldsymbol{\epsilon}$ is identical and analytically tractable, wherein the point estimate $\mathbf{w}_{\text{ols}}$ and its sampling distribution has a closed-form solution of [1]

$$\mathbf{w}_{\text{mle}} = (\Phi^T\Gamma^{-1}\Phi)^{-1}\Phi^T\Gamma^{-1}\mathbf{y} \qquad (18a)$$

$$\text{Sampling distribution}(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}},(\Phi^T\Gamma^{-1}\Phi)^{-1}) \qquad (18b)$$

Rewriting the posterior covariance $\mathbf{P}$ from Eq. (11) using the Woodbury formula (Eq. 156 in [17]) leads to

$$\begin{aligned}
\mathbf{P} &= (\mathbf{I} - \mathbf{K}\Phi)\mathbf{P}_0 \\
&= \mathbf{P}_0 - \mathbf{P}_0\Phi^T(\Phi\mathbf{P}_0\Phi^T + \Gamma)^{-1}\Phi\mathbf{P}_0 \\
&= (\mathbf{P}_0^{-1} + \Phi^T\Gamma^{-1}\Phi)^{-1} \\
\Rightarrow \mathbf{P}^{-1} &= \mathbf{P}_0^{-1} + \Phi^T\Gamma^{-1}\Phi \qquad (19)
\end{aligned}$$

Hence, the posterior precision ($\mathbf{P}^{-1}$) of parameter $\mathbf{w}$ is the sum of its prior precision ($\mathbf{P}_0^{-1}$) and the precision ($\Phi^T\Gamma^{-1}\Phi$) of its OLS/MLE estimate (From Eq. (18)). Given that precision

is analogous to knowledge, this inference re-iterates the fact that the posterior knowledge is the sum total of prior knowledge and the knowledge provided by the observations. It is also realized that precision adds up when solving inverse problem, while variance adds up when solving a forward UQ problem.

Similarly, posterior mean in Eq. (11) can be re-written as

$$\begin{aligned}
\mathbf{m} &= \mathbf{m}_0 + \mathbf{K}(\mathbf{y} - \Phi\mathbf{m}_0) \\
&= (\mathbf{I} - \mathbf{K}\Phi)\mathbf{m}_0 + \mathbf{P}_0\Phi^T(\Phi\mathbf{P}_0\Phi^T + \Gamma)^{-1}\mathbf{y} \\
&= \mathbf{P}\mathbf{P}_0^{-1}\mathbf{m}_0 + (\mathbf{P}_0^{-1} + \Phi^T\Gamma^{-1}\Phi)^{-1}\Phi^T\Gamma^{-1}\mathbf{y} \\
&= \mathbf{P}\mathbf{P}_0^{-1}\mathbf{m}_0 + \mathbf{P}(\Phi^T\Gamma^{-1}\Phi)(\Phi^T\Gamma^{-1}\Phi)^{-1}\Phi^T\Gamma^{-1}\mathbf{y} \\
&= \left(\frac{\mathbf{P}_0^{-1}}{\mathbf{P}^{-1}}\right)\mathbf{m}_0 + \left(\frac{\Phi^T\Gamma^{-1}\Phi}{\mathbf{P}^{-1}}\right)\mathbf{w}_{\text{ols}}
\end{aligned} \tag{20}$$

Hence, posterior mean $\mathbf{m}$ is the weighted sum of prior mean $\mathbf{m}_0$ and the OLS/MLE estimate $\mathbf{w}_{\text{ols}}$, where the weights are equal to the ratio of its precision with the posterior precision $\mathbf{P}^{-1}$. Several observations can be made by analyzing Eq. (19) and Eq. (20). For example, in case of non-informative (flat) prior ($\mathbf{P}_0^{-1} \approx 0$) the posterior pdf is same as the sampling distribution $\mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{ols}}, (\Phi^T\Gamma^{-1}\Phi)^{-1})$ of the OLS/MLE estimate, meaning the posterior pdf is entirely dictated by observations with no influence from prior knowledge. Hence, BLR with flat priors provides the same solution as the sampling distribution from OLS/MLE. In the case of high observational noise ($\Gamma^{-1} = \rho \approx 0$) or large prior knowledge ($\mathbf{P}_0^{-1} >> \Phi^T\Gamma^{-1}\Phi$) the posterior pdf is same as the prior pdf $\mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{P}_0)$, meaning no information is gained from the data due to the lack of information in observations or abundance of knowledge in prior, respectively.

Model	$\mathbf{y} = \Phi\mathbf{w} + \boldsymbol{\epsilon}$
Observations	$\mathbf{y} = \{y_1, \ldots, y_M\}^T$
Model error	$\mathrm{p}(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}, \mathbf{0}, \mathbf{I}_M\rho^{-1}) = \mathcal{N}(\boldsymbol{\epsilon}\|\mathbf{0}, \Gamma)$
Prior pdf	$\mathrm{p}(\mathbf{w}) = \mathcal{N}(\mathbf{w}\|\mathbf{m}_0, \mathbf{P}_0)$
Likelihood function	$\mathrm{p}(\mathbf{y}\|\mathbf{w}, \rho) = \mathcal{N}(\mathbf{y}\|\Phi\mathbf{w}, \Gamma)$
Posterior pdf	$\mathrm{p}(\mathbf{w}\|\mathbf{y}, \rho) = \mathcal{N}(\mathbf{w}\|\mathbf{m}, \mathbf{P})$
	$\mathbf{m} = \mathbf{m}_0 + \mathbf{K}(\mathbf{y} - \Phi\mathbf{m}_0)$
	$\mathbf{P} = (\mathbf{I} - \mathbf{K}\Phi)\mathbf{P}_0$
	$\mathbf{K} = \mathbf{P}_0\Phi^T(\Phi\mathbf{P}_0\Phi^T + \Gamma)^{-1}$
Evidence	$\mathrm{p}(\mathbf{y}\|\rho) = \mathcal{N}(\mathbf{y}\|\Phi\mathbf{m}_o, \Phi\mathbf{P}_0\Phi^T + \Gamma)$
Predictive pdf	$\mathrm{p}(\tilde{\mathbf{y}}\|\mathbf{y}, \rho) = \mathcal{N}(\tilde{\mathbf{y}}\|\tilde{\Phi}\mathbf{m}, \tilde{\Phi}\mathbf{P}\tilde{\Phi}^T + \Gamma)$

Table 2: Summary: Bayesian linear regression using conjugate priors with known model error precision ρ

4 Sparse Bayesian learning

Sparse Bayesian learning (SBL) is a sparse learning algorithm first proposed by Tipping [19] to obtain sparse solution to linear-in-parameter models known as support vector machines (SVMs). SBL provided a Bayesian alternative to sparse SVM construction, and the resulting algorithm was termed as relevance vector machine (RVM). Nevertheless, SBL is applicable to any linear-in-parameter model of form Eq. (1) with additive and Gaussian model error $\boldsymbol{\epsilon}$. SBL employs the Bayesian framework to facilitate automatic pruning of redundant parameters from the model by means of prior regularization and hierarchical Bayesian modelling. Unlike the frequentist sparse learning method of LASSO, SBL allows for probabilistic viewpoint of uncertainty wherein posterior inferences are obtained in the form probability distributions. Furthermore, prior-posterior conjugacy relationships in Bayesian inference facilitate a semi-analytical evaluation of the sparse solution from SBL, thereby alleviating the need for numerical optimization. The mathematical exposition of SBL presented in this section is derived from [6, 19, 20].

Given the Gaussian likelihood function $\mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho^{-1}\mathbf{I}_M)$, SBL operates by assigning conjugate prior pdf to the unknown model parameter vector $\mathbf{w}$ and the model error precision ρ as

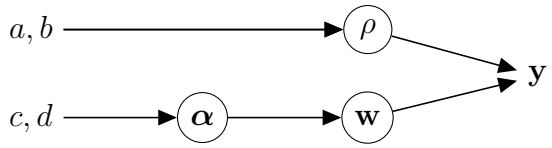
$$\mathrm{p}(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=0}^{N-1} \mathcal{N}(0, \alpha^{-1}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}) \tag{21}$$

$$\mathrm{p}(\boldsymbol{\alpha}) = \prod_{i=0}^{N-1} \mathcal{G}(\alpha_i|c, d)\ , \quad \mathrm{p}(\rho) = \mathcal{G}(\rho|a, b) \tag{22}$$

where $\boldsymbol{\alpha}$ is the unknown hyper-parameter vector, $\mathbf{A} = \mathrm{Diag}(\boldsymbol{\alpha})$ and $a, b, c, d > 0$ are some known parameters. The prior $\mathrm{p}(\boldsymbol{\alpha})$ is known as hyper-prior (prior for hyper-parameter). Note that $\mathcal{G}(x|r, s)$ denotes the gamma distribution with pdf

$$\mathcal{G}(x|r, s) = \frac{s^r}{\Gamma(r)} x^{r-1} e^{-sx}; \ \ \mathrm{E}[x] = \frac{s}{r}; \ \ \mathrm{Var}(x) = \frac{s}{r^2} \tag{23}$$

The zero-mean Gaussian prior pdf $\mathrm{p}(\mathbf{w}|\boldsymbol{\alpha})$ in Eq. (21) is the key behind the sparsity inducing property of SBL. The prior $\mathrm{p}(\mathbf{w}|\boldsymbol{\alpha})$ facilitates the addition/removal of each parameter w_i by varying the value of the corresponding hyper-parameter α_i. This behaviour of prior $\mathrm{p}(\mathbf{w}|\boldsymbol{\alpha})$ can be explained as following. When $\alpha_i = \infty$ the marginal prior pdf $\mathcal{N}(0, \alpha^{-1})$ becomes a Dirac-delta pdf centered at zero, or $\mathcal{N}(w_i|0, 0)$. According to Eq. (19) and Eq. (20), this prior pdf will force the posterior to be a Dirac-delta pdf, or $\mathcal{N}(w_i|0, 0)$. This posterior implies the basis $\phi_i(\mathbf{x})$ no longer contributes to the model since the parameter w_i is fixed at zero. On the contrary, when α_i has a finite value the corresponding prior pdf $\mathcal{N}(0, \alpha^{-1})$ is informative and the posterior of w_i parameter is a non Dirac-delta pdf, meaning the basis $\phi_i(\mathbf{x})$ has a non-zero contribution to the model. This property of the prior $\mathrm{p}(\mathbf{w}|\boldsymbol{\alpha})$ to control the contribution of each basis by varying the corresponding hyper-parameter is called automatic relevance determination (ARD) and the prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})$ in Eq. (21) is termed as ARD prior. The hierarchical structure of the ARD prior can be summarized as:



where the unknown parameters have been circled. Given the ARD prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})$, a sparse representation of $\mathbf{w}$ or the basis space can be obtained by estimating hyper-parameters $\boldsymbol{\alpha}$ using the observations. Given that $\mathbf{w}$ and ρ are also unknown, Bayes' rule can be used to compute the joint posterior distribution of unknowns $(\mathbf{w}, \boldsymbol{\alpha}, \rho)$ as

$$\begin{aligned}
\mathrm{p}(\mathbf{w}, \boldsymbol{\alpha}, \rho|\mathbf{y}) &= \frac{\mathrm{p}(\mathbf{y}|\mathbf{w}, \boldsymbol{\alpha}, \rho)\mathrm{p}(\mathbf{w}, \boldsymbol{\alpha}, \rho)}{\mathrm{p}(\mathbf{y})} \\
&\propto \mathrm{p}(\mathbf{y}|\mathbf{w}, \boldsymbol{\alpha}, \rho)\mathrm{p}(\mathbf{w}|\boldsymbol{\alpha})\mathrm{p}(\boldsymbol{\alpha})\mathrm{p}(\rho) \\
&\propto \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho^{-1}\mathbf{I}_M)\mathcal{N}(\mathbf{w}|0, \mathbf{A}^{-1}) \left\{ \prod_{i=1}^{N-1} \mathcal{G}(\alpha_i|c, d) \right\} \mathcal{G}(\rho|a, b)
\end{aligned}$$

Note that the solution to the joint posterior $\mathrm{p}(\mathbf{w}, \boldsymbol{\alpha}, \rho|\mathbf{y})$ is analytically intractable due to absence of prior-posterior conjugacy for the joint distribution of $\mathbf{w}, \boldsymbol{\alpha}, \rho$. As a result, the joint posterior $\mathrm{p}(\mathbf{w}, \boldsymbol{\alpha}, \rho|\mathbf{y})$ is decomposed as

$$\mathrm{p}(\mathbf{w}, \boldsymbol{\alpha}, \rho|\mathbf{y}) = \underbrace{\mathrm{p}(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \rho)}_{\text{Analytically tractable}} \mathrm{p}(\boldsymbol{\alpha}, \rho|\mathbf{y}) \tag{24}$$

As demonstrated in Section 3, the posterior pdf $\mathrm{p}(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \rho)$ is analytically tractable using prior-posterior conjugacy properties, and can be computed as

$$\begin{aligned}
\mathrm{p}(\mathbf{w}|\mathbf{y}, \boldsymbol{\alpha}, \rho) &= \frac{\mathrm{p}(\mathbf{y}|\mathbf{w}, \boldsymbol{\alpha}, \rho)\mathrm{p}(\mathbf{w}|\boldsymbol{\alpha})}{\mathrm{p}(\mathbf{y}|\boldsymbol{\alpha}, \rho)} && \text{(25a)} \\
&\propto \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho^{-1}\mathbf{I}_M)\mathcal{N}(\mathbf{w}|0, \mathbf{A}^{-1}) \\
&= \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P}) && \text{(25b)}
\end{aligned}$$

where $\mathbf{m}$ is the posterior mean and $\mathbf{P}$ is the posterior covariance of $\mathbf{w}$. The expression for $\mathbf{m}$ and $\mathbf{P}$ is obtained using Eq. (12) as

$$\mathbf{P} = (\mathbf{A} + \rho\Phi^T\Phi)^{-1} = (\mathbf{I} - \mathbf{K}\Phi)\mathbf{A}^{-1} \tag{26a}$$

$$\mathbf{m} = (\mathbf{A} + \rho\Phi^T\Phi)^{-1}(\rho\Phi^T\mathbf{y}) = \mathbf{K}\mathbf{y} \tag{26b}$$

$$\mathbf{K} = \mathbf{A}^{-1}\Phi^T(\Phi\mathbf{A}^{-1}\Phi^T + \mathbf{I}_M\rho^{-1})^{-1} = \mathbf{A}^{-1}\Phi^T\mathbf{B}^{-1} \tag{26c}$$

where $\mathbf{B} = \Phi\mathbf{A}^{-1}\Phi^T + \mathbf{I}_M\rho^{-1}$. Consequently, model evidence $\mathrm{p}(\mathbf{y}|\boldsymbol{\alpha}, \rho)$ (denominator in Eq. (25a)) can also be obtained analytically as (using relation in Eq. (10))

$$\mathrm{p}(\mathbf{y}|\boldsymbol{\alpha}, \rho) = \int \mathrm{p}(\mathbf{y}|\mathbf{w}, \rho)\mathrm{p}(\mathbf{w}|\boldsymbol{\alpha})d\mathbf{w} = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{B}) \tag{27}$$

Using hierarchical Bayes approach, the posterior $\mathrm{p}(\boldsymbol{\alpha}, \rho|\mathbf{y})$ in Eq. (25a) can be obtained as

$$\begin{aligned}\mathrm{p}(\boldsymbol{\alpha}, \rho|\mathbf{y}) &= \frac{\mathrm{p}(\mathbf{y}|\boldsymbol{\alpha}, \rho)\mathrm{p}(\boldsymbol{\alpha}, \rho)}{\mathrm{p}(\mathbf{y})} \propto \mathrm{p}(\mathbf{y}|\boldsymbol{\alpha}, \rho)\mathrm{p}(\rho)\mathrm{p}(\boldsymbol{\alpha}) \\ &\propto \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{B})\mathcal{G}(\rho|a, b) \prod_{i=0}^{N-1} \mathcal{G}(\alpha_i|c, d) \end{aligned} \tag{28}$$

For the sake of seeking sparsity in $\mathbf{w}$, our interest lies in finding the mode or the *maximum a posteriori* (MAP) estimate $(\boldsymbol{\alpha}^*, \rho^*)$ obtained by maximizing the posterior $\mathrm{p}(\boldsymbol{\alpha}, \rho|\mathbf{y})$ in Eq. (28). The MAP estimate $\boldsymbol{\alpha}^*$ will provide a sparse construction of $\mathbf{w}$ as many α_i^* will approach infinity, thereby forcing both prior and posterior of parameter w_i to be a Dirac-delta pdf centered at zero. Therefore, the problem of finding a sparse solution to $\mathbf{w}$ is transformed into finding the optimal hyper-parameter α that maximizes the posterior pdf $\mathrm{p}(\boldsymbol{\alpha}, \rho|\mathbf{y})$. Note that when using flat hyper-priors for ρ and $\boldsymbol{\alpha}$, meaning $\mathcal{G}(\alpha_i|c, d) \propto 1$ and $\mathcal{G}(\rho|a, b) \propto 1$, the maximization of posterior $\mathrm{p}(\boldsymbol{\alpha}, \rho|\mathbf{y})$ equate to maximizing the model evidence $\mathrm{p}(\mathbf{y}|\boldsymbol{\alpha}, \rho)$. This approach of maximizing model evidence to estimate hyper-parameters is also known as *type-II maximum likelihood method* [3] or *empirical Bayes* in statistical literature.

From implementation perspective, the MAP estimate $(\boldsymbol{\alpha}^*, \rho^*)$ is computed by maximizing the logarithm of posterior $\mathrm{p}(\boldsymbol{\alpha}, \rho|\mathbf{y})$ in Eq. (28) (ignoring constant terms) as

$$(\boldsymbol{\alpha}^*, \rho^*) = \arg\max\{\log \mathrm{p}(\boldsymbol{\alpha}, \rho|\mathbf{y})\} = \arg\max\{\mathcal{L}(\boldsymbol{\alpha}, \rho)\} \tag{29}$$

where $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ is the objective function being maximized, obtained as

$$\mathcal{L}(\boldsymbol{\alpha}, \rho) = -\frac{1}{2}\left\{\log|\mathbf{B}| + \mathbf{y}^T\mathbf{B}^{-1}\mathbf{y}\right\} + (a-1)\log\rho - b\rho + \sum_{i=0}^{N-1}\left\{(c-1)\log\alpha_i - d\alpha_i\right\} \tag{30}$$

where $\mathbf{B} = \Phi\mathbf{A}^{-1}\Phi^T + \mathbf{I}_M\rho^{-1}$. The optimization problem posed in Eq. (29) is a convex optimization problem (proved later in this section), and can be solved analytically through differentiation of $\mathcal{L}(\boldsymbol{\alpha}, \rho)$. Based on the approach undertaken to differentiate $\mathcal{L}(\mathbf{w}, \rho)$, there exist two type of SBL methods: 1) Original SBL method by Tipping [19], and 2) Accelerated SBL method by Tipping [20] and Faul [6].

Once the SBL algorithm has converged, the resulting sparse representation of $\mathbf{w}$ is obtained in the form of sparse posterior mean $\mathbf{m}$ and sparse posterior covariance $\mathbf{P}$ matrix computed from Eq. (11) by substituting $\boldsymbol{\alpha} = \boldsymbol{\alpha}^*$ and $\rho = \rho^*$. Subsequently, the posterior

predictive distribution of unseen observations $\tilde{\mathbf{y}}$ is computed as

$$\begin{aligned}
\mathrm{p}(\tilde{\mathbf{y}}|\mathbf{y}) &= \int\int\int \mathrm{p}(\tilde{\mathbf{y}}|\mathbf{w},\boldsymbol{\alpha},\rho)\mathrm{p}(\mathbf{w},\boldsymbol{\alpha},\rho|\mathbf{y})d\mathbf{w}d\boldsymbol{\alpha}d\rho \\
&= \int\int\int \mathrm{p}(\tilde{\mathbf{y}}|\mathbf{w},\boldsymbol{\alpha},\rho)\mathrm{p}(\mathbf{w}|\mathbf{y},\boldsymbol{\alpha},\rho)\mathrm{p}(\boldsymbol{\alpha},\rho|\mathbf{y})d\mathbf{w}d\boldsymbol{\alpha}d\rho \\
&= \int\int\int \mathrm{p}(\tilde{\mathbf{y}}|\mathbf{w},\boldsymbol{\alpha},\rho)\mathrm{p}(\mathbf{w}|\mathbf{y},\boldsymbol{\alpha},\rho)\delta(\boldsymbol{\alpha}-\boldsymbol{\alpha}^*,\rho-\rho^*)d\mathbf{w}d\boldsymbol{\alpha}d\rho \\
&= \int \mathrm{p}(\tilde{\mathbf{y}}|\mathbf{w},\boldsymbol{\alpha}^*,\rho^*)\mathrm{p}(\mathbf{w}|\mathbf{y},\boldsymbol{\alpha}^*,\rho^*)d\mathbf{w} \\
&= \int \mathcal{N}(\tilde{\mathbf{y}}|\tilde{\Phi}\mathbf{w},\mathbf{I}_M(\rho^*)^{-1})\mathcal{N}(\mathbf{w}|\mathbf{m},\mathbf{P})d\mathbf{w} \\
&= \mathcal{N}(\tilde{\mathbf{y}}|\tilde{\Phi}\mathbf{m},\tilde{\Phi}\mathbf{P}\tilde{\Phi}^T+\mathbf{I}_M(\rho^*)^{-1}) \qquad (31)
\end{aligned}$$

Next, we discuss in detail the two SBL methods to compute $(\boldsymbol{\alpha}^*,\rho^*)$ through analytical differentiation of $\mathcal{L}(\boldsymbol{\alpha},\rho)$ in Eq. (30).

4.1 Original algorithm

The Original SBL method corresponds to the iterative re-estimation technique proposed by Tipping [19]. The idea behind this algorithm is to isolate the effect of hyperparameter α_i and error precision ρ by expanding terms $\log|\mathbf{B}|$ and $\mathbf{y}^T\mathbf{B}^{-1}\mathbf{y}$ present in the objective function $\mathcal{L}(\boldsymbol{\alpha},\rho)$ in Eq. (30). To begin with, the $\log|\mathbf{B}|$ term in Eq. (30) is re-written using the matrix determinant identity

$$|\mathbf{A}||\Phi\mathbf{A}^{-1}\Phi^T+\mathbf{I}_M\rho^{-1}| = |\mathbf{I}_M\rho^{-1}||\mathbf{A}+\rho\Phi^T\Phi| \qquad (32)$$

to obtain

$$\begin{aligned}
\log|\mathbf{B}| &= \log|\Phi\mathbf{A}^{-1}\Phi^T+\mathbf{I}_M\rho^{-1}| \\
&= \log|\mathbf{I}_M\rho^{-1}| + \log|\mathbf{A}+\rho\Phi^T\Phi| - \log|\mathbf{A}| \\
&= -M\log\rho + \log|\mathbf{P}^{-1}| - \sum_{i=1}^{N-1}\log\alpha_i \qquad (33)
\end{aligned}$$

Furthermore, the term $\mathbf{y}^T\mathbf{B}^{-1}\mathbf{y}$ in Eq. (30) is evaluated using the Woodbury identity (Eq. 156 in [17]) of

$$\begin{aligned}
\mathbf{B}^{-1} &= (\Phi\mathbf{A}^{-1}\Phi^T+\mathbf{I}_M\rho^{-1})^{-1} \\
&= \mathbf{I}_M\rho - \rho\Phi(\mathbf{A}+\rho\Phi^T\Phi)^{-1}\Phi^T\rho \\
&= \mathbf{I}_M\rho - \rho^2\Phi\mathbf{P}\Phi^T \qquad (34)
\end{aligned}$$

to obtain

$$\begin{aligned}
\mathbf{y}^T\mathbf{B}^{-1}\mathbf{y} &= \mathbf{y}^T(\mathbf{I}_M\rho - \rho^2\Phi\mathbf{P}\Phi^T)\mathbf{y} \\
&= \rho\mathbf{y}^T(\mathbf{y} - \Phi(\rho\mathbf{P}\Phi^T\mathbf{y})) \\
&= \rho(\mathbf{y} - \Phi\mathbf{m} + \Phi\mathbf{m})^T(\mathbf{y} - \Phi\mathbf{m}) \\
&= \rho(\mathbf{y} - \Phi\mathbf{m})^T(\mathbf{y} - \Phi\mathbf{m}) + \rho(\Phi\mathbf{m})^T(\mathbf{y} - \Phi\mathbf{m}) \\
&= \rho||\mathbf{y} - \Phi\mathbf{m}||_2^2 + \mathbf{m}^T\mathbf{P}^{-1}(\rho\mathbf{P}\Phi^T\mathbf{y}) - \rho\mathbf{m}^T\Phi^T\Phi\mathbf{m} \\
&= \rho||\mathbf{y} - \Phi\mathbf{m}||_2^2 + \mathbf{m}^T(\mathbf{P}^{-1} - \rho\Phi^T\Phi)\mathbf{m} \\
&= \rho||\mathbf{y} - \Phi\mathbf{m}||_2^2 + \mathbf{m}^T\mathbf{A}\mathbf{m} \\
&= \rho||\mathbf{y} - \Phi\mathbf{m}||_2^2 + \sum_{i=0}^{N-1}\alpha_i m_i^2
\end{aligned} \tag{35}$$

where m_i is the posterior mean of parameter w_i. Substituting $\log|\mathbf{B}|$ from Eq. (33) and $\mathbf{y}^T\mathbf{B}^{-1}\mathbf{y}$ from Eq. (35) in the objective function $\mathcal{L}(\mathbf{w},\rho)$ in Eq. (30),

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\alpha},\rho) = -\frac{1}{2}\left\{-M\log\rho + \log|\mathbf{P}^{-1}| + \rho||\mathbf{y} - \Phi\mathbf{m}||_2^2 + \sum_{i=0}^{N-1}(\alpha_i m_i^2 - \log\alpha_i)\right\} \\
+ (a-1)\log\rho - b\rho + \sum_{i=0}^{N-1}\{(c-1)\log\alpha_i - d\alpha_i\}
\end{aligned} \tag{36}$$

The exact differentiation of the revised form of $\mathcal{L}(\boldsymbol{\alpha},\rho)$ in Eq. (36) is still not feasible since the posterior mean $\mathbf{m}$ has an implicit dependence on $\boldsymbol{\alpha}$ and ρ. This implicit dependence on $\mathbf{m}$ is easy to isolate but its differentiation with respect to $\boldsymbol{\alpha}$ and ρ is not tractable due to the complexity of the relationship (atleast to this end). It is realized that differentiating $\mathcal{L}(\boldsymbol{\alpha},\rho)$ locally by assuming posterior mean $\mathbf{m}$ remains unchanged for a small change in $\boldsymbol{\alpha}$ and ρ leads to an iterative re-estimation procedure. Note that the dependence of $\boldsymbol{\alpha}$ and ρ on the posterior covariance $\mathbf{P}$ is not ignored as $\mathbf{P}$ is directly responsible for inducing sparsity and moreover its differentation with respect to $\boldsymbol{\alpha}$ and ρ is analytically tractable. Differentiating $\mathcal{L}(\mathbf{w},\rho)$ from Eq. (36) under this assumption leads to

$$\frac{\partial\mathcal{L}(\boldsymbol{\alpha},\rho)}{\partial\alpha_i} = -\frac{1}{2}\left\{\frac{\partial(\log|\mathbf{P}^{-1}|)}{\partial\alpha_i} - \frac{1}{\alpha_i} + m_i^2\right\} + \frac{(c-1)}{\alpha_i} - d \tag{37}$$

where the differentiation of $\log|\mathbf{P}^{-1}|$ can be computed using the relation $\mathbf{P}^{-1} = \mathbf{A} + \rho\Phi^T\Phi$ and an identity involving derivative of a matrix determinant (Eq. 46 in [17]) as

$$\begin{aligned}
\frac{\partial(\log|\mathbf{P}^{-1}|)}{\partial\alpha_i} &= \frac{1}{|\mathbf{P}^{-1}|}\frac{\partial(|\mathbf{P}^{-1}|)}{\partial\alpha_i}) \\
&= \frac{1}{\cancel{|\mathbf{P}^{-1}|}}\cancel{|\mathbf{P}^{-1}|}\text{Trace}\left(\mathbf{P}\frac{\partial(\mathbf{P}^{-1})}{\partial\alpha_i}\right) \\
&= \text{Trace}\left(\mathbf{P}\frac{\partial(\mathbf{A}+\rho\Phi^T\Phi)}{\partial\alpha_i}\right) \\
&= \text{Trace}\left(\mathbf{P}\frac{\partial(\text{Diag}(\boldsymbol{\alpha}))}{\partial\alpha_i}\right) \\
&= P_{ii} \qquad (38)
\end{aligned}$$

where P_{ii} is the $(i,i)^{th}$ element of the posterior covariance matrix $\mathbf{P}$. Equating Eq. (37) to zero we get

$$\begin{aligned}
\frac{\partial\mathcal{L}(\boldsymbol{\alpha},\rho)}{\partial\alpha_i} = 0 &= -\frac{1}{2}\left\{P_{ii} - \frac{1}{\alpha_i} + m_i^2\right\} + \frac{(c-1)}{\alpha_i} - d \\
\Rightarrow \frac{2(c-1)+1}{\alpha_i} &= P_{ii} + m_i^2 + 2d \\
\Rightarrow \alpha_i &= \frac{1+2(c-1)}{P_{ii}+m_i^2+2d} \qquad (39)
\end{aligned}$$

This update formula for α_i is similar to the one obtained using *expectation-maximization* method (More details in appendix A in [19]). Similarly, differentiating $\mathcal{L}(\mathbf{w},\rho)$ in Eq. (36) with respect to ρ and equating it to zero,

$$\begin{aligned}
\frac{\partial\mathcal{L}(\boldsymbol{\alpha},\rho)}{\partial\rho} &= -\frac{1}{2}\left\{-\frac{M}{\rho} + \frac{\partial(\log|\mathbf{P}^{-1}|)}{\partial\rho} + ||\mathbf{y}-\Phi\mathbf{m}||_2^2\right\} + \frac{a-1}{\rho} - b \\
0 &= -\frac{M}{\rho} + \frac{1}{|\mathbf{P}^{-1}|}\frac{\partial(|\mathbf{P}^{-1}|)}{\partial\rho}) + ||\mathbf{y}-\Phi\mathbf{m}||_2^2 - \frac{2(a-1)}{\rho} + 2b \\
\frac{M+2a-2}{\rho} &= \frac{1}{\cancel{|\mathbf{P}^{-1}|}}\cancel{|\mathbf{P}^{-1}|}\text{Trace}\left(\mathbf{P}\frac{\partial(\mathbf{P}^{-1})}{\partial\rho}\right) + ||\mathbf{y}-\Phi\mathbf{m}||_2^2 + 2b \\
\frac{M+2a-2}{\rho} &= \text{Trace}\left(\mathbf{P}\frac{\partial}{\partial\rho}(\mathbf{A}+\rho\Phi^T\Phi)\right) + ||\mathbf{y}-\Phi\mathbf{m}||_2^2 + 2b \\
\rho &= \frac{M+2(a-1)}{\text{Trace}\,(\mathbf{P}\Phi^T\Phi) + ||\mathbf{y}-\Phi\mathbf{m}||_2^2 + 2b} \qquad (40)
\end{aligned}$$

In summary, Eq. (39) and Eq. (40) provide the iterative re-estimation equations for computing the MAP estimates $\boldsymbol{\alpha}^*$ and ρ^*, wherein the posterior statistics ($\mathbf{m}$ and $\mathbf{P}$) is updated using Eq. (26) after each re-estimation iteration.

Following Mackay [14], Tipping [20] suggested the use of a new entity $r_i = 1 - \alpha_i P_{ii}$ when writing the iterative re-estimation equations. The true nature of γ_i is revealed when

expanded as (using Eq. (26))

$$r_i = 1 - \frac{\alpha_i}{P_{ii}^{-1}} = 1 - \frac{\alpha_i}{\alpha_i + \rho \Phi_i^T \Phi_i} \in [0, 1] \tag{41}$$

where $\rho \Phi_i^T \Phi_i$ is the OLS/MLE contribution to the posterior precision P_{ii}^{-1} of parameter w_i. For an irrelevant parameter w_i the prior pdf will approach a Dirac-delta pdf during SBL, meaning prior precision α_i will be large (low prior variance) and the posterior will be dictated by prior so that $P_{ii}^{-1} \approx \alpha_i$ or $r_i \approx 0$. Alternatively, for a relevant parameter w_i the posterior precision P_{ii}^{-1} will be dictated by observations, thereby forcing $P_{ii}^{-1} \approx \rho^{-1} \Phi_i^T \Phi_i$ or $r_i \approx 1$. Hence, r_i can be thought of as a quantitative indicator of relevance for parameter w_i or basis ϕ_i. We will hereby refer to r_i as *sparsity indicator*.

Following the definition of r_i in Eq. (41), the iterative re-estimation equation for α_i in Eq. (39) is re-written as

$$\begin{aligned} \alpha_i(P_{ii} + m_i^2 + 2d) &= 1 + 2(c-1) \\ \alpha_i(m_i^2 + 2d) &= 1 - \alpha_i P_{ii} + 2(c-1) \\ \alpha_i &= \frac{r_i + 2(c-1)}{m_i^2 + 2d} \end{aligned} \tag{42}$$

Similarly, Eq. (40) is re-written using the relation $\text{Trace}\left(\mathbf{P}\Phi^T\Phi\right) = \rho^{-1} \sum r_i$, to obtain [19]

$$\begin{aligned} \rho(\text{Trace}\left(\mathbf{P}\Phi^T\Phi\right) + ||\mathbf{y} - \Phi\mathbf{m}||_2^2 + 2b) &= M + 2(a-1) \\ \rho(\rho^{-1}\sum r_i + ||\mathbf{y} - \Phi\mathbf{m}||_2^2 + 2b) &= M + 2(a-1) \\ \sum r_i + \rho(||\mathbf{y} - \Phi\mathbf{m}||_2^2 + 2b) &= M + 2(a-1) \\ \rho &= \frac{M - \sum r_i + 2(a-1)}{||\mathbf{y} - \Phi\mathbf{m}||_2^2 + 2b} \end{aligned} \tag{43}$$

Eq. (42) and Eq. (43) completes the iterative re-estimation procedure to compute the MAP estimate $(\boldsymbol{\alpha}^*, \rho^*)$ pertaining to the sparse representation of $\mathbf{w}$. Algorithm 1 summarizes the steps needed for implementing the SBL algorithm. Some important implementation details of the original SBL algorihtm are discussed below:

- Note that Tipping [19] suggested the use of uniform hyperprior for $\log \alpha_i$ and $\log \rho$ using values of $a, b, c, d \approx 0$ in Eq. (22). Later Babacan *et al.* [2] pointed out that Tipping [19] actually assigned flat (non-informative) hyperprior for α_i and ρ (and not $\log \alpha_i$ and $\log \rho$) obtained using values of $a, c \approx 1$ and $b, d \approx 0$. In this study we will assume flat hyper-prior for ρ using values of $a = 1$ and $b = 0$ in Eq. (43). For the hyperprior of α_i, we assign $c \approx 0$ and the effect of $a > 1$ on the performance of SBL is explored during numerical investigations reported in Section 7.

- The convergence of the original SBL algorithm is determined by monitoring the changes in 1) sparsity indicator $r_i = 1 - \alpha_i P_{ii}$, and 2) the objective function $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ from Eq. (36). When the change in all r_i's and $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ is below some predefined tolerance for a certain number of iterations, the SBL algorithm is said to be converged. Note

that choosing to monitor α_i values for convergence can be erroneous and impractical since the scale of α_i values depends on the scale of the corresponding parameter (w_i) value. This property of α_i is the primary reason behind the introduction of r_i into the SBL formulation as it provide a consistent way of determining relevancy and its convergence. We will revisit this assertion during numerical investigations reported in Section 7.

- Another entity that can be considered for the monitoring the convergence is the Kullback-Liebler (KL) divergence between the posterior and prior pdfs (marginals). The KL divergence quantifies the gap in information between the marginal posterior and prior pdfs and is computed as

$$\begin{aligned} \mathrm{D}_i^{\mathrm{KL}} &= \int \mathrm{p}(w_i|\mathbf{y},\boldsymbol{\alpha},\rho) \log \frac{\mathrm{p}(w_i|\mathbf{y},\boldsymbol{\alpha},\rho)}{\mathrm{p}(w_i|\boldsymbol{\alpha})} dw_i \\ &= \int \mathcal{N}(w_i|m_i,P_{ii}) \log \frac{\mathcal{N}(w_i|m_i,P_{ii})}{\mathcal{N}(w_i|0,\alpha_i^{-1})} dw_i \\ &= -\frac{1}{2}\left\{\log(\alpha_i P_{ii}) + \alpha_i(P_{ii}+m_i^2) - 1\right\} \end{aligned} \tag{44}$$

  When a parameter w_i is irrelevant the prior and posterior pdf are a Dirac-delta pdf centered at zero, meaning $\alpha P_{ii} \approx= 1$ and $m_i = 0$ in Eq. (44). Therefore, KL divergence is zero for irrelevant parameter and it is a non-zero positive value for a relevant parameter.

Algorithm 1: Original SBL method by Tipping [19]

Propose the candidate basis functions and construct the design matrix Φ according to Eq. (1);
Choose a starting value of $\boldsymbol{\alpha}$ and ρ;
Choose hyper-prior parameter a,b,c,d for $\boldsymbol{\alpha}$ and ρ;
while *not converged* **do**
  Compute posterior statistics using current $\boldsymbol{\alpha}$ and ρ values as

$$\mathbf{A} = \mathrm{Diag}(\boldsymbol{\alpha})\ ;\ \ \mathbf{B} = \Phi\mathbf{A}^{-1}\Phi^T + \mathbf{I}_M/\rho\ ;\ \ \mathbf{K} = \mathbf{A}^{-1}\Phi^T\mathbf{B}^{-1};$$
$$\mathbf{m} = \mathbf{K}\mathbf{y}\ ;\ \ \mathbf{P} = \mathbf{A}^{-1} - \mathbf{K}\Phi\mathbf{A}^{-1};\ r_i = 1 - \alpha_i P_{ii};$$

  Check convergence by tracking r_i and $\mathcal{L}(\boldsymbol{\alpha},\rho)$ (using Eq. (36)) values;
  Choose a trial basis ϕ_i (randomly or deterministically) and update the corresponding α_i and ρ as

$$\alpha_i = \frac{r_i + 2(c-1)}{m_i^2 + 2d}\ \ ;\ \ \rho = \frac{M - \sum r_i + 2(a-1)}{||\mathbf{y} - \Phi\mathbf{m}||_2^2 + 2b}$$

end

4.2 Accelerated algorithm

The accelerated SBL algorithm by Faul and Tipping [6] maximizes model evidence through explicit addition and deletion of basis functions from the model in Eq. (1). The addition of a basis function is performed by assigning the corresponding hyperparameter a finite value, while a basis is deleted by setting the corresponding hyperparameter to infinity. This approach is contrary to the original SBL algorithm discussed in section 4.1 that always has a finite hyper-parameter value which in-turn controls the contribution of each basis function. The mathematical exposition presented here is derived from Faul and Tipping [6].

The accelerated SBL algorithm assumes flat hyperpriors for $\boldsymbol{\alpha}$ and ρ using $a = b = 1$ and $c = d = 0$ in Eq. (21). In that case, the SBL objective function in Eq. (30) is same as log-evidence and is obtained as

$$\mathcal{L}(\boldsymbol{\alpha}, \rho) = -\frac{1}{2}\left\{\log|\mathbf{B}| + \mathbf{y}^T\mathbf{B}^{-1}\mathbf{y}\right\} \tag{45}$$

where $\mathbf{B} = \Phi\mathbf{A}^{-1}\Phi^T + \rho^{-1}\mathbf{I}_M$. The underlying structure of matrix $\mathbf{B}$ is revealed when expanding it as

$$\begin{aligned}
\mathbf{B} &= \Phi\mathbf{A}^{-1}\Phi^T + \rho^{-1}\mathbf{I}_M \\
&= \{\Phi_0\ \Phi_1\ \ldots \Phi_{N-1}\}\begin{bmatrix} \frac{1}{\alpha_0} & & & \\ & \frac{1}{\alpha_1} & & \\ & & \ddots & \\ & & & \frac{1}{\alpha_{N-1}} \end{bmatrix}\begin{Bmatrix} \Phi_0^T \\ \Phi_1^T \\ \vdots \\ \Phi_{N-1}^T \end{Bmatrix} + \rho^{-1}\mathbf{I}_M \\
&= \rho^{-1}\mathbf{I}_M + \sum_{j=0}^{N-1}\frac{\Phi_j\Phi_j^T}{\alpha_j}
\end{aligned} \tag{46}$$

where Φ_j is the j^{th} column of the design matrix Φ. The dependence of matrix $\mathbf{B}$ on each hyper-parameter α_i can be isolated by rewriting Eq. (46) as

$$\mathbf{B} = \left\{\rho^{-1}\mathbf{I}_M + \sum_{\substack{j=0 \\ j\neq i}}^{N-1}\frac{\Phi_j\Phi_j^T}{\alpha_j}\right\} + \frac{\Phi_i\Phi_i^T}{\alpha_i} = \mathbf{B}_{-i} + \frac{\Phi_i\Phi_i^T}{\alpha_i} \tag{47}$$

where $\mathbf{B}_{-i}$ quantifies the effect of all hyper-parameters but α_i, while the term $\Phi_i\Phi_i^T/\alpha_i$ is a sole function of α_i. Using this decomposition of $\mathbf{B}$ and the matrix determinant lemma $|X + vv^T| = |X||1 + v^TXv|$, the term $\log|\mathbf{B}|$ in Eq. (45) is expanded as

$$\begin{aligned}
\log|\mathbf{B}| &= \log\left|\mathbf{B}_{-i} + \frac{\Phi_i\Phi_i^T}{\alpha_i}\right| \\
&= \log\left(|\mathbf{B}_{-i}|\left|1 + \frac{\Phi_i^T\mathbf{B}_{-i}\Phi_i}{\alpha_i}\right|\right) \\
&= \log|\mathbf{B}_{-i}| + \log\left|\frac{\alpha_i + \Phi_i^T\mathbf{B}_{-i}\Phi_i}{\alpha_i}\right| \\
&= \log|\mathbf{B}_{-i}| + \log(\alpha_i + \Phi_i^T\mathbf{B}_{-i}\Phi_i) - \log\alpha_i
\end{aligned} \tag{48}$$

Using Sherman-Morrison identity (Eq. 160 in [17]), the term $\mathbf{y}^T\mathbf{B}\mathbf{y}$ in Eq. (45) is expanded as

$$\begin{aligned}\mathbf{y}^T\mathbf{B}^{-1}\mathbf{y} &= \mathbf{y}^T\left(\mathbf{B}_{-i} + \frac{\Phi_i\Phi_i^T}{\alpha_i}\right)^{-1}\mathbf{y} \\ &= \mathbf{y}^T\left(\mathbf{B}_{-i}^{-1} - \frac{\mathbf{B}_{-i}^{-1}\Phi_i\Phi_i^T\mathbf{B}_{-i}^{-1}}{\alpha_i + \Phi_i^T\mathbf{B}_{-i}^{-1}\Phi_i}\right)\mathbf{y} \\ &= \mathbf{y}^T\mathbf{B}_{-i}^{-1}\mathbf{y} - \frac{\mathbf{y}^T\mathbf{B}_{-i}^{-1}\Phi_i\Phi_i^T\mathbf{B}_{-i}^{-1}\mathbf{y}}{\alpha_i + \Phi_i^T\mathbf{B}_{-i}^{-1}\Phi_i} \end{aligned} \tag{49}$$

Using Eq. (48) and Eq. (49), the SBL objective function $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ in Eq. (45) is obtained as

$$\mathcal{L}(\boldsymbol{\alpha}, \rho) = \mathcal{L}(\boldsymbol{\alpha}_{-i}, \rho) + l(\alpha_i) \tag{50a}$$

$$\mathcal{L}(\boldsymbol{\alpha}_{-i}, \rho) = -\frac{1}{2}\left\{\log|\mathbf{B}_{-i}| + \mathbf{y}^T\mathbf{B}_{-i}^{-1}\mathbf{y}\right\} \tag{50b}$$

$$l(\alpha_i) = \frac{1}{2}\left\{-\log(\alpha_i + s_i) + \log\alpha_i + \frac{q_i^2}{\alpha_i + s_i}\right\} \tag{50c}$$

where $\boldsymbol{\alpha}_{-i}$ contains all hyper-parameters but α_i, $q_i = \Phi_i^T\mathbf{B}_{-i}^{-1}\mathbf{y}$ is the quality factor and $s_i = \Phi_i^T\mathbf{B}_{-i}^{-1}\Phi_i$ is the sparsity factor. This decomposition of $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ in Eq. (50) decouples the effect of each α_i on log-evidence in the form of function $l(\alpha_i)$ while function $\mathcal{L}(\boldsymbol{\alpha}_{-i}, \rho)$ quantifies the effect of rest of the hyper-parameters ($\boldsymbol{\alpha}_{-i}$).

Note that both q_i and s_i are not explicitly related to α_i but can be interpreted as qualitative indicators of the relevance of basis ϕ_i. As Faul and Tipping [6] pointed out, the quality factor q_i quantifies the alignment of each basis ϕ_i with the error in model predictions when ϕ_i is absent. Hence, higher the q_i is, the addition of ϕ_i into the model will lead to higher decrease in the model error, and therefore parameter w_i is more inclined to be relevant. On the contrary, the sparsity factor s_i quantifies the alignment of each basis ϕ_i along the basis already present in the model. Hence, higher the s_i is, the basis ϕ_i is inclined to duplicate the role of basis already present in the model, and therefore parameter w_i is less inclined to be relevant. We will revisit these statements during numerical investigations reported in Sectio 7.

To compute the MAP estimate $\boldsymbol{\alpha}^*$ the objective function $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ in Eq. (50) is differentiated with respect to α_i to obtain

$$\begin{aligned}\frac{\partial\mathcal{L}(\boldsymbol{\alpha}, \rho)}{\partial\alpha_i} &= \cancelto{0}{\frac{\partial\mathcal{L}(\boldsymbol{\alpha}_{-i}, \rho)}{\partial\alpha_i}} + \frac{\partial l(\alpha_i)}{\partial\alpha_i} \\ &= \frac{1}{2}\frac{\partial}{\partial\alpha_i}\left\{-\log(\alpha_i + s_i) + \log\alpha_i + \frac{q_i^2}{\alpha_i + s_i}\right\} \\ &= \frac{1}{2}\left\{-\frac{1}{\alpha_i + s_i} + \frac{1}{\alpha_i} - \frac{q_i^2}{(\alpha_i + s_i)^2}\right\} \\ &= \frac{1}{2}\left\{\frac{\alpha_i^{-1}s_i^2 - (q_i^2 - s_i)}{(\alpha_i + s_i)^2}\right\} \end{aligned} \tag{51}$$

The solution for α_i^* is conditioned on the sign of term $q_i^2 - s_i$, and is obtained by equating the derivative of log-evidence in Eq. (51) to zero to obtain

$$\alpha_i = \begin{cases} \dfrac{s_i^2}{q_i^2 - s_i}, & q_i^2 > s_i \\ \infty & q_i^2 \leq s_i \end{cases} \tag{52}$$

In other words, if $q_i^2 \leq s_i$ the basis ϕ_i is deemed irrelevant (set $\alpha_i^* = \infty$). Else, if $q_i^2 > s_i$ the basis ϕ_i is deemed relevant and the corresponding α_i is updated according to Eq. (52). The variation of log-evidence with varying α_i is further explored by computing the curvature of $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ (or $l(\alpha_i)$) using Eq. (51) as

$$\begin{aligned} \frac{\partial^2 \mathcal{L}(\boldsymbol{\alpha}, \rho)}{\partial \alpha_i^2} = \frac{\partial^2 l(\alpha_i)}{\partial \alpha_i^2} &= \frac{-\alpha_i^{-2} s_i^2}{2(\alpha_i + s_i)^2} - \frac{(\alpha^{-1} s_i^2 - (q_i^2 - s_i))}{(\alpha_i + s_i)^3} \\ &= -\frac{1}{2} \left\{ \frac{s_i}{\alpha_i (\alpha_i + s_i)} \right\}^2 - \frac{2}{(\alpha_i + s_i)} \frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \rho)}{\partial \alpha_i} \end{aligned} \tag{53}$$

At the optimal $\boldsymbol{\alpha}$ the slope $\partial l(\alpha_i)/\partial \alpha_i = 0$ and therefore the curvature of log-evidence in Eq. (53) is always negative (concave down), implying the maximization of log-evidence. Furthermore, when $q_i^2 \leq s_i$, the slope of log-evidence in Eq. (51) is always positive and hence log-evidence has negative curvature for all values of α, except at $\alpha_i = \infty$ when it is zero. Note that $s_i = \mathbf{y}^T \mathbf{B}_{-i} \mathbf{y}^T \geq 0$ since $\mathbf{B}_{-i}$ is a covariance matrix (a positive semi-definite matrix). These observations about the behaviour of slope/curvature of log-evidence will be revisited during numerical investigation of the SBL algorithm. Nevertheless, the optimal model error precision is obtained by differentiating log-evidence in Eq. (45) and following the same procedure as in section 4.1 to obtain

$$\rho = \frac{M - \sum r_i}{||\mathbf{y} - \Phi \mathbf{m}||_2^2} \tag{54}$$

where $r_i = 1 - \alpha_i P_{ii}$ for a relavant parameter and $r_i = 0$ for an irrelevant parameter. Note that Eq. (52) and Eq. (54) completes the iterative re-estimation equation for computing optimal α_i and ρ that maximizes the log-evidence or $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ in Eq. (45).

The accelerated SBL algorithm operates by selecting a basis function among the candidate basis functions, and then determining whether the basis needs to added or removed from the model based on $q_i^2 - s_i$ values in Eq. (52). This selection of a basis function can be done in a random or deterministic fashion. This addition/removal of basis requires the computation of factors $q_i = \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y}$ and $s_i = \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i$ using the updated $\mathbf{B}_{-i}^{-1}$ from Eq. (46) following every addition/removal of a basis. For large number of basis (large N) or large data size (large M), calculating inverse of a $M \times M$ matrix $\mathbf{B}_{-i}$ for all N basis at every iteration degrades the computational efficiency of the algorithm. As a result, entities $\bar{s}_m = \Phi_m^T \mathbf{B}^{-1} \Phi_m$ and $\bar{q}_m = \Phi_m^T \mathbf{B}^{-1} \mathbf{y}$ are introduced that depend on matrix $\mathbf{B}$ which is fixed for all N basis.

The relation between $\bar{s}_i$ and s_i is obtained using Eq. (47) as

$$\begin{aligned}
\bar{s}_i &= \Phi_i^T \mathbf{B}^{-1} \Phi_i = \Phi_i^T \left(\mathbf{B}_{-i} + \frac{\Phi_i \Phi_i^T}{\alpha_i} \right)^{-1} \Phi_i \\
&= \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i - \frac{\Phi^T \mathbf{B}_{-i}^{-1} \Phi_i \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i}{\alpha_i + \Phi_i^T \mathbf{B}_{-i}^{-1} \Phi_i} \\
&= s_i - \frac{s_i^2}{\alpha_i + s_i} \\
\Rightarrow s_i &= \frac{\alpha_i \bar{s}_i}{\alpha_i - \bar{s}_i}
\end{aligned} \tag{55}$$

Note that $s_i = \bar{s}_i$ when $\alpha_i = \infty$, while $s_i > \bar{s}_i$ for finite α_i. Similarly, the relation between $\bar{q}_i$ and q_i is obtained using Eq. (47) as

$$\begin{aligned}
\bar{q}_i &= \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y} - \frac{\Phi^T \mathbf{B}_{-i}^{-1} \Phi_i \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y}}{\alpha_i + \Phi_i^T \mathbf{B}_{-i}^{-1} \mathbf{y}} \\
&= q_i - \frac{q_i^2}{\alpha_i + s_i} \\
\Rightarrow q_i &= \frac{\alpha_i \bar{q}_i}{\alpha_i - \bar{s}_i}
\end{aligned} \tag{56}$$

Algorithm 2 summarizes the steps needed to implement the accelerated SBL algorithm. Further details about the practical aspect of the algorithm are discussed in the numerical investigation section. The accelerated SBL algorithm of altering the design matrix Φ provides higher computationally efficiency and the matrix Φ is free of multicolinearity as it only contains the basis functions that are currently present in the model.

Algorithm 2: Accelerated SBL algorithm by Faul and Tipping [6]

Propose the candidate basis functions and construct the design matrix Φ according to Eq. (1);

Choose a starting value of $\boldsymbol{\alpha}$ and ρ;

while *not converged* **do**

Compute posterior statistics using current $\boldsymbol{\alpha}$ and ρ values as:

$$\mathbf{A}^{-1} = \text{Diag}(1/\boldsymbol{\alpha}) \; ; \;\; \mathbf{B} = \Phi \mathbf{A}^{-1} \Phi^T + \mathbf{I}_M/\rho \; ; \;\; \mathbf{K} = \mathbf{A}^{-1} \Phi^T \mathbf{B}^{-1}$$

$$\mathbf{m} = \mathbf{K}\mathbf{y} \; ; \;\; \mathbf{P} = \mathbf{A}^{-1} - \mathbf{K}\Phi\mathbf{A}^{-1};$$

$$\bar{s}_m = \Phi_m^T \mathbf{B}^{-1} \Phi_m \; ; \;\; \bar{q}_m = \Phi_m^T \mathbf{B}^{-1} \mathbf{y};$$

$$\mathcal{L}(\boldsymbol{\alpha}, \rho) = -\frac{1}{2} \left\{ \log |\mathbf{B}| + \mathbf{y}^T \mathbf{B}^{-1} \mathbf{y} \right\}$$

$$r_i = \begin{cases} 1 - \alpha_i P_{ii}, & \alpha_i \text{ is finite} \\ 0, & \alpha_i = \infty \end{cases}$$

Check convergence by tracking r_i and $\mathcal{L}(\boldsymbol{\alpha}, \rho)$ values;

Compute $q_i = \bar{q}_i/(1 - \bar{s}_i/\alpha_i)$ and $s_i = \bar{s}_i/(1 - \bar{s}_i/\alpha_i)$;

Choose a trial basis ϕ_i;

if $q_i^2 > s_i$ **then**

if ϕ_i is present in the model (α_i is finite), re-estimate $\alpha_i = s_i^2/(q_i^2 - s_i)$;

if ϕ_i is absent from the model ($\alpha_i = \infty$), add ϕ_i to the model and set $\alpha_i = s_i^2/(q_i^2 - s_i)$;

else

if ϕ_i is present in the model (α_i is finite), set $\alpha_i = \infty$;

if ϕ_i is absent from the model, no change;

end

Re-estimate ρ as

$$\rho = \frac{M - \sum r_i}{||\mathbf{y} - \Phi\mathbf{m}||_2^2}$$

end

5 Bayesian compressive sensing

Bayesian compressive sensing (BCS) is a sparse learning technique designed to identify the sparse representation of $\mathbf{w}$ in Eq. (1) through the use of a parameterized Laplace prior on $\mathbf{w}$. The advent of BCS is attributed to the frequentist sparse learning technique of LASSO [18]. As demonstrated in Section 2, LASSO provides a sparse solution of $\mathbf{w}$ by solving a L1-constrained least-square minimization problem wherein the regularization parameter is estimated using cross-validation [18]. As demonstrated later in Section 6, LASSO minimization is equivalent to finding the MAP estimate for $\mathbf{w}$ when selecting independent, zero-mean Laplace prior with same hyper-parameter for all w_i [18]. Despite this Bayesian interpretation of LASSO, a point estimate of $\mathbf{w}$ fails to exploit the probabilistic viewpoint of Bayesian framework by not considering the posterior pdf of $\mathbf{w}$. Following the popularity of Bayesian statistics in inverse modelling, Park and Casella [16] proposed an alternative to LASSO called *Bayesian LASSO* wherein the onus of finding sparsity is transferred to the hyperparameter space. In particular, the hyperparameter of the Laplace prior is estimated by maximizing model evidence to obtain sparse solution to $\mathbf{w}$. This transformation also enabled probabilistic interpretation of uncertainty in $\mathbf{w}$ resulting in robust model predictions.

The use of Laplace prior over Gaussian prior (as in SBL) was being promoted due to the sparsity inducing property of L1-regularization (LASSO) over L2-regularization (Ridge regression) in least-square estimation. Due to this property of Laplace prior, Ji *et al.* [12] proposed BCS as an alternative to SBL and a fully Bayesian alternative to LASSO for application in signal reconstruction, albeit analytical evaluation of the sparse solution (like SBL) was not explored. Babacan *et al.* [2] later provided semi-analytical solution to BCS and provided a detailed comparison between SBL and BCS techniques. The mathematical exposition presented in this section is derived from Babacan *et al.* [2].

BCS operates by assigning a parametrized Laplace prior to each model parameter w_i. However, a Laplace prior on $\mathbf{w}$ is not a conjugate prior for a Gaussian likelihood function. This issue is resolved by constructing the Laplace prior through a two-level hierarchical prior structure. The hierarchical prior restores the prior-posterior conjugacy relationship and enables an analytical scrutiny of the Bayesian sparse learning process. The two-level hierarchical prior pdf is defined as

$$\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma}) = \prod_{i=0}^{N-1} \mathcal{N}(w_i|0,\gamma_i) = \mathcal{N}(\mathbf{w}|\mathbf{0},\mathbf{A}^{-1}) \tag{57a}$$

$$\mathrm{p}(\gamma|\lambda) = \prod_{i=0}^{N-1} \mathcal{G}(\gamma_i|1,\lambda/2) = \prod_{i=0}^{N-1} \frac{\lambda}{2} \exp\left\{-\frac{\lambda\gamma_i}{2}\right\} \tag{57b}$$

$$\mathrm{p}(\lambda) = \mathcal{G}(\lambda|c,d) \tag{57c}$$

$$\mathrm{p}(\rho) = \mathcal{G}(\rho|a,b) \tag{57d}$$

where $\boldsymbol{\gamma}$ is the unknown hyperparameter vector (prior variance of $\mathbf{w}$), λ is the unknown parameter that dictates the prior pdf of $\boldsymbol{\gamma}$, and $a, b, c, d > 0$ are some known constants. The prior $\mathcal{G}(\gamma_i|1,\lambda/2)$ is also called exponential distribution with rate parameter as $\lambda/2$. A Laplace prior on w_i is then obtained by integrating out hyperparameter γ_i from the joint

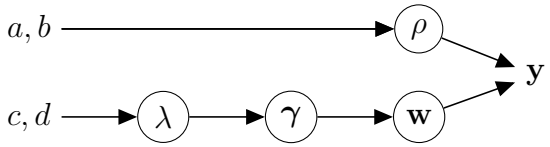
distribution of (w_i, γ_i) as [6]

$$\begin{aligned} \mathrm{p}(w_i|\lambda) &= \int \mathrm{p}(w_i, \gamma_i|\lambda) d\gamma_i \\ &= \int \mathrm{p}(w_i|\gamma_i)\mathrm{p}(\gamma_i|\lambda) d\gamma_i \\ &= \int \mathcal{N}(w_i|0, \gamma_i)\mathcal{G}(\gamma_i|1, \lambda/2) d\gamma_i \\ &= \mathcal{LP}(w_i|0, \lambda^{-0.5}) \end{aligned} \tag{58}$$

where $\mathcal{LP}(x|r, s)$ denotes a laplace distribution with properties:

$$\mathcal{LP}(x|r, s) = \frac{1}{2s} \exp\left\{-\frac{|x-r|}{s}\right\}; \;\; \mathrm{E}[x] = r; \;\; \mathrm{Var}(x) = 2s^2 \tag{59}$$

where s is the scale parameter. Therefore, the prior pdf of $\mathbf{w}$ is a laplace pdf conditioned on a single parameter λ, while it is a Gaussian pdf if conditioned on the hyperparameter vector $\boldsymbol{\gamma}$. The hierarchical structure of the prior can be summarized as



where the unknown parameters are circled.

Following the SBL formulation in Section 4, the joint posterior pdf $\mathrm{p}(\mathbf{w}, \boldsymbol{\gamma}, \lambda, \rho|\mathbf{y})$ of unknown parameters is decomposed as

$$\mathrm{p}(\mathbf{w}, \boldsymbol{\gamma}, \lambda, \rho|\mathbf{y}) = \underbrace{\mathrm{p}(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}, \lambda, \rho)}_{\text{Analytically tractable}} \; \mathrm{p}(\boldsymbol{\gamma}, \lambda, \rho|\mathbf{y}) \tag{60}$$

where $\mathrm{p}(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}, \lambda, \rho)$ is the posterior pdf of $\mathbf{w}$ given γ, λ, ρ are known. This scenario resembles the case of Bayesian linear regression (BLR) discussed in section 3, where the posterior is evaluated analytically as

$$\begin{aligned} \mathrm{p}(\mathbf{w}|\mathbf{y}, \boldsymbol{\gamma}, \lambda, \rho) &= \frac{\mathrm{p}(\mathbf{y}|\mathbf{w}, \boldsymbol{\gamma}, \lambda, \rho)\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma}, \lambda, \rho)}{\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}, \lambda, \rho)} \qquad (61) \\ &\propto \mathrm{p}(\mathbf{y}|\mathbf{w}, \rho)\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma}) \\ &\propto \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \mathbf{I}_M\rho^{-1})\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}) \\ &= \mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P}) \qquad (62) \end{aligned}$$

where the expressions for $\mathbf{m}$ and $\mathbf{P}$ are listed in Eq. (26) with $\mathbf{A}^{-1} = \mathrm{Diag}(\boldsymbol{\gamma})$. Consequently, model evidence $\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}, \lambda, \rho)$ (denominator in Eq. (61)) is also available analytically as

$$\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}, \lambda, \rho) = \int \mathrm{p}(\mathbf{y}|\mathbf{w}, \rho)\mathrm{p}(\mathbf{w}|\boldsymbol{\gamma}) d\mathbf{w} = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{B}) \tag{63}$$

where $\mathbf{B} = \Phi\mathbf{A}^{-1}\Phi^T + \mathbf{I}_M\rho^{-1}$. The remaining pdf $\mathrm{p}(\boldsymbol{\gamma}, \lambda, \rho|\mathbf{y})$ in Eq. (61) is evaluated using Bayes' theorem as

$$\mathrm{p}(\boldsymbol{\gamma}, \lambda, \rho|\mathbf{y}) = \frac{\mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}, \lambda, \rho)\mathrm{p}(\boldsymbol{\gamma}, \lambda, \rho)}{\mathrm{p}(\mathbf{y})} \tag{64a}$$

$$\propto \mathrm{p}(\mathbf{y}|\boldsymbol{\gamma}, \lambda, \rho)\mathrm{p}(\rho)\mathrm{p}(\lambda)\mathrm{p}(\boldsymbol{\gamma}|\lambda) \tag{64b}$$

$$\propto \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{B})\mathcal{G}(\rho|a, b)\mathcal{G}(\lambda|a, b)\prod_{i=0}^{N-1}\mathcal{G}(\gamma_i|1, \lambda/2) \tag{64c}$$

The sparse solution for $\mathbf{w}$ pertains to the MAP estimate $(\boldsymbol{\gamma}^*, \lambda^*, \rho^*)$ obtained by maximizing the posterior $\mathrm{p}(\boldsymbol{\gamma}, \lambda, \rho|\mathbf{y})$ in Eq. (64). As was discussed in Section 4, hyperparameters γ_i corresponding to redundant model parameters w_i are driven to zero during the maximization, resulting in a sparse posterior mean $\mathbf{m}$ and posterior covariance $\mathbf{P}$ matrix of $\mathbf{w}$. In practice, logarithm of the posterior in Eq. (64) is maximized, and therefore the objective function to be maximized is defined as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma}, \lambda, \rho) = &-\frac{1}{2}\left\{\log|\mathbf{B}| + \mathbf{y}^T\mathbf{B}^{-1}\mathbf{y}\right\} + N\log\frac{\lambda}{2} - \frac{\lambda}{2}\sum_{i=0}^{N-1}\gamma_i \\ &+ (a-1)\log\rho - b\rho + (c-1)\log\lambda - d\lambda\end{aligned} \tag{65}$$

Using the expansions of term $\log|\mathbf{B}|$ in Eq. (33) and $\mathbf{y}^T\mathbf{B}^{-1}\mathbf{y}$ in Eq. (35), the objective function $\mathcal{L}(.)$ in Eq. (65) is rewritten as

$$\begin{aligned}\mathcal{L}(\boldsymbol{\alpha}, \lambda, \rho) = &-\frac{1}{2}\left\{-M\log\rho + \log|\mathbf{P}^{-1}| + \rho||\mathbf{y} - \Phi\mathbf{m}||_2^2 + \sum_{i=0}^{N-1}\left(\frac{m_i^2}{\gamma_i} + \log\gamma_i\right)\right\} \\ &+ N\log\frac{\lambda}{2} - \frac{\lambda}{2}\sum_{i=0}^{N-1}\gamma_i + (a-1)\log\rho - b\rho + (c-1)\log\lambda - d\lambda\end{aligned} \tag{66}$$

This form of the objective function $\mathcal{L}(.)$ can be differentiated analyticaly leading to an iterative re-estimation technique to compute the MAP estimates for $\boldsymbol{\gamma}, \lambda, \rho$. However, BCS incorporates a fast sub-optimal approach similar to the accelerated SBL algorithm (Section 4.2) involving iterative addition/deletion of basis. Using the decomposition of matrix $\mathbf{B}$ in Eq. (47) and the subsequent expansion of $\log|\mathbf{B}|$ in Eq. (48) and $\mathbf{y}^T\mathbf{B}^{-1}\mathbf{y}$ in Eq. (49), the objective function is re-written as

$$\mathcal{L}(\boldsymbol{\gamma}, \lambda, \rho) = \mathcal{L}(\boldsymbol{\gamma}_{-i}, \lambda, \rho) + l(\gamma_i) \tag{67a}$$

$$\mathcal{L}(\boldsymbol{\gamma}_{-i}, \lambda, \rho) = -\frac{1}{2}\left\{\log|\mathbf{B}_{-i}| + \mathbf{y}^T\mathbf{B}_{-i}^{-1}\mathbf{y}\right\} + N\log\frac{\lambda}{2} - \frac{\lambda}{2}\sum_{\substack{j=0\\j\neq i}}^{N-1}\gamma_j \tag{67b}$$

$$+ (a-1)\log\rho - b\rho + (c-1)\log\lambda - d\lambda \tag{67c}$$

$$l(\gamma_i) = \frac{1}{2}\left\{\log\left(\frac{1}{1+\gamma_i s_i}\right) + \frac{\gamma_i q_i^2}{1+\gamma s_i} - \lambda\gamma_i\right\} \tag{67d}$$

where $q_i = \Phi_i^T \mathbf{B}_{-i}^{-1}\mathbf{y}$ is the quality factor and $s_i = \Phi_i^T \mathbf{B}_{-i}^{-1}\Phi_i$ is the sparsity factor. The calculation of $\bar{q}_i$ and $\bar{s}_i$ is performed by defining entities $\bar{s}_m = \Phi_m^T \mathbf{B}^{-1}\Phi_m$ and $\bar{q}_m = \Phi_m^T \mathbf{B}^{-1}\mathbf{y}$ and using relations $s_i = \bar{s}_i/(1-\gamma_i\bar{s}_i)$ and $q_i = \bar{q}_i/(1-\gamma_i\bar{s}_i)$ (from Eq. (55) and Eq. (56)). Subsequently, differentiating $\mathcal{L}(\boldsymbol{\gamma}, \lambda, \rho)$ with respect to γ_i leads to

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\gamma}, \lambda, \rho)}{\partial \gamma_i} &= \cancelto{0}{\frac{\partial \mathcal{L}(\boldsymbol{\gamma}_{-i}, \lambda, \rho)}{\partial \gamma_i}} + \frac{\partial l(\gamma_i)}{\partial \gamma_i} \\
&= \frac{1}{2}\left\{ \cancel{(1+\gamma_i s_i)} \frac{-1}{(1+\gamma_i s_i)^{\cancel{2}}} s_i + \frac{q_i^2}{1+\gamma_i s_i} - \frac{q_i^2 \gamma_i}{(1+\gamma_i s_i)^2} s_i - \lambda \right\} \\
&= \frac{1}{2}\left\{ \frac{-s_i - \gamma_i s_i^2 + q_i^2 + \cancel{q_i^2 \gamma_i s_i} - \cancel{q_i^2 \gamma_i s_i} - \lambda - \lambda\gamma_i^2 s_i^2 - 2\lambda\gamma_i s_i}{(1+\gamma_i s_i)^2} \right\} \\
&= -\frac{1}{2}\left\{ \frac{\gamma_i^2(\lambda s_i^2) + \gamma_i(s_i^2 + 2\lambda s_i) + (s_i - q_i^2 + \lambda)}{(1+\gamma_i s_i)^2} \right\}
\end{aligned} \tag{68}$$

Equating the derivative of the objective function in the above equation to zero leads to a quadratic equation

$$\gamma_i^2(\lambda s_i^2) + \gamma_i(s_i^2 + 2\lambda s_i) + (s_i - q_i^2 + \lambda) = 0 \tag{69}$$

whose solution is obtained as

$$\begin{aligned}
\gamma_i^* &= \frac{-s_i(s_i + 2\lambda) \pm \sqrt{(s_i^2 + 2\lambda s_i)^2 - 4\lambda s_i^2 (s_i - q_i^2 + \lambda)}}{2\lambda s_i^2} \\
&= \frac{-(s_i + 2\lambda) \pm \sqrt{(s_i + 2\lambda)^2 - 4\lambda(s_i - q_i^2 + \lambda)}}{2\lambda s_i}
\end{aligned} \tag{70}$$

Given that λ, γ_i, s_i are all greater than zero, following two scenarios exist regarding the MAP estimate γ_i^* obtained in Eq. (70):

1. $s_i - q_i^2 + \lambda \geq 0$ or $q_i^2 - s_i \leq \lambda$: For this case, both the roots of γ_i in Eq. (70) will be negative. Also, the slope of the objective function in Eq. (68) is negative for all $\gamma_i \geq 0$. Given that $\gamma_i \geq 0$, the only possible solution that can satisfy these conditions is $\gamma_i = 0$.

2. $s_i - q_i^2 + \lambda < 0$ or $q_i^2 - s_i > \lambda$: For this case, we have one positive and one negative solution for γ_i in Eq. (70). From Eq. (68) it can be deduced that the slope of the objective function $\mathcal{L}$ is positive at $\gamma_i = 0$ and is negative at $\gamma_i = \infty$. Given that $\gamma_i \geq 0$, the positive solution from Eq. (70) will maximize $\mathcal{L}$ and is taken to be the MAP estimate.

In summary, the MAP estimate of γ_i obtained by maximizing the posterior $\mathrm{p}(\boldsymbol{\gamma}, \lambda, \rho|\mathbf{y})$ in Eq. (61) or the objective function $\mathcal{L}(\boldsymbol{\gamma}, \lambda, \rho)$ in Eq. (67) is obtained as

$$\gamma_i^* = \begin{cases} \dfrac{-(s_i + 2\lambda) \pm \sqrt{(s_i + 2\lambda)^2 - 4\lambda(s_i - q_i^2 + \lambda)}}{2\lambda s_i}, & q_i^2 - s_i \geq \lambda \\ 0 & q_i^2 - s_i < \lambda \end{cases} \tag{71}$$

Further, differentiating $\mathcal{L}(\boldsymbol{\gamma}, \lambda, \rho)$ in Eq. (65) or Eq. (66) with respect to λ and equating it to zero leads to (matrices $\mathbf{B}$ and $\mathbf{P}$ is independent of λ)

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\gamma}, \lambda, \rho)}{\partial \lambda} &= \frac{\partial}{\partial \lambda}\left\{ N \log \frac{\lambda}{2} - \frac{\lambda}{2}\sum_{i=0}^{N-1} \gamma_i + (c-1)\log \lambda - d\lambda \right\} \\
&= \frac{N}{\lambda/2}.\frac{1}{2} - \frac{1}{2}\sum_{i=0}^{N-1} \gamma_i + \frac{c-1}{\lambda} - d\lambda = 0 \\
\implies \lambda^* &= \frac{N + (c-1)}{0.5\sum \gamma_i + d}
\end{aligned} \tag{72}$$

Similarly, differentiating $\mathcal{L}(\boldsymbol{\gamma}, \lambda, \rho)$ in Eq. (66) with respect to ρ (similar to Eq. (40)) and equating it to zero leads to

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\gamma}, \lambda, \rho)}{\partial \rho} &= \frac{\partial}{\partial \rho}\left\{ \frac{M \log \rho}{2} - \frac{1}{2}\log |\mathbf{P}^{-1}| - \frac{\rho ||\mathbf{y} - \Phi \mathbf{m}||_2^2}{2} + (a-1)\log \rho - b\rho \right\} \\
&= \frac{M}{2\rho} - \frac{\text{Trace}\left(\mathbf{P}\Phi^T\Phi\right)}{2} - \frac{||\mathbf{y} - \Phi \mathbf{m}||_2^2}{2} + \frac{(a-1)}{\rho} - b \\
&= \frac{M}{2\rho} - \frac{\sum r_i}{2\rho} - \frac{||\mathbf{y} - \Phi \mathbf{m}||_2^2}{2} + \frac{(a-1)}{\rho} - b = 0 \\
\implies \rho^* &= \frac{M - \sum r_i + 2(a-1)}{||\mathbf{y} - \Phi \mathbf{m}||_2^2 + 2b}
\end{aligned} \tag{73}$$

where $r_i = 1 - P_{ii}/\gamma_i$ is the sparsity indicator. Note that the expression for computing the MAP estimate ρ^* is the same for BCS and SBL. In summary, Eq. (71), Eq. (72), and Eq. (73) completes the derivation of iterative re-estimation equations for computing the MAP estimates $\gamma_i^*, \lambda^*, \rho^*$. Algorithm 3 summarizes the neccesary steps involved in the BCS algorithm.

Algorithm 3: BCS algorithm

Choose the basis functions for the model;
Assign initial values to hyper-parameter $\boldsymbol{\gamma}$, λ and error precision ρ;
while *not converged* **do**
Compute posterior statistics using current $\boldsymbol{\gamma}$ and ρ values as [$\mathbf{A}$ and Φ matrices only contain the basis currently in the model]

$$\mathbf{A} = \text{Diag}(1/\boldsymbol{\gamma}) \; ; \;\; \mathbf{B} = \Phi\mathbf{A}^{-1}\Phi^T + \mathbf{I}_M/\rho \; ; \;\; \mathbf{K} = \mathbf{A}^{-1}\Phi^T\mathbf{B}^{-1}$$
$$\mathbf{m} = \mathbf{K}\mathbf{y} \; ; \;\; \mathbf{P} = \mathbf{A}^{-1} - \mathbf{K}\Phi\mathbf{A}^{-1} \; ; \;\; \log \mathrm{p}(\mathbf{y}|\boldsymbol{\alpha}, \rho) = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{B})$$
$$\bar{s}_m = \Phi_m^T\mathbf{B}^{-1}\Phi_m \; ; \;\; \bar{q}_m = \Phi_m^T\mathbf{B}^{-1}\mathbf{y};$$
$$r_i = \begin{cases} 1 - P_{ii}/\gamma_i, & \gamma_i \text{ is non-zero} \\ 0, & \gamma_i = 0 \end{cases}$$

Choose a trial basis ϕ_i randomly and compute $q_i = \bar{q}_i/(1 - \gamma_i\bar{s}_i)$ and $s_i = \bar{s}_i/(1 - \gamma_i\bar{s}_i)$;
if $q_i^2 - s_i \geq \lambda$ **then**
if Φ_i is present in the model, re-estimate α_i using Eq. (71)
if Φ_i is absent from the model, add Φ_i to the model, modify $\mathbf{A}$ and Φ accordingly and set α_i value using Eq. (71);
else
if Φ_i is present in the model, remove Φ_i from the model (set $\alpha_i = \infty$), and modify $\mathbf{A}$ and Φ accordingly;
if Φ_i is absent from the model, no change;
end
Re-estimate λ (Eq. (72)) and ρ (Eq. (73)) as

$$\lambda^* = \frac{N + (c-1)}{0.5\sum \gamma_i + d}; \quad \rho^* = \frac{M - \sum r_i + 2(a-1)}{||\mathbf{y} - \Phi\mathbf{m}||_2^2 + 2b}$$

end

6 Relationship between LASSO, SBL and BCS

We first establish the relation between frequentist and Bayesian sparse learning techniques. The MAP estimate calculation is approached for the case of Gaussian likelihood function $\mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho^{-1}\mathbf{I}_M)$ and the following choices of prior pdf:

- We assign a zero-mean Gaussian prior pdf with precision $\rho\lambda$ to the model parameters, leading to the joint prior pdf

$$\mathrm{p}(\mathbf{w}) = \mathrm{p}(w_0) \prod_{i=1}^{N-1} \mathcal{N}(w_i|0, (\rho\lambda)^{-1}) = C \prod_{i=1}^{N-1} \mathcal{N}(w_i|0, (\rho\lambda)^{-1}) \tag{74}$$

where the mean parameter is assigned a flat prior. The MAP estimate calculation using Eq. (9) leads to

$$\begin{aligned}\mathbf{w}_{\text{map}} &= \arg\max_{\mathbf{w}}\{\text{p}(\mathbf{y}|\mathbf{w},\rho)\text{p}(\mathbf{w})\} = \arg\max_{\mathbf{w}}\{\log(\text{p}(\mathbf{y}|\mathbf{w},\rho)\text{p}(\mathbf{w}))\} \\ &= \arg\max_{\mathbf{w}}\left\{\log\mathcal{N}(\mathbf{y}|\Phi\mathbf{w},\rho^{-1}\mathbf{I}_M) + \log C + \log\left(\prod_{i=1}^{N-1}\mathcal{N}(w_i|0,(\rho\lambda)^{-1})\right)\right\} \\ &= \arg\max_{\mathbf{w}}\left\{-\frac{\rho}{2}||\mathbf{y}-\Phi\mathbf{w}||_2^2 - \frac{\rho\lambda}{2}\sum_{i=1}^{N-1}w_i^2\right\} \\ &= \arg\min_{\mathbf{w}}\left\{\frac{1}{2}||\mathbf{y}-\Phi\mathbf{w}||_2^2 + \frac{\lambda}{2}\sum_{i=1}^{N-1}w_i^2\right\} \end{aligned} \tag{75}$$

The cost function in the above equation is the same as Ridge regression cost function in Eq. (6) with $p = 2$. Therefore, the MAP estimate $\mathbf{w}_{\text{map}}$ for the case of Gaussian likelihood function and prior pdf $\mathcal{N}(w_i|0,(\rho\lambda)^{-1})$ for all but the mean parameter w_0 results in same solution as the Ridge regression estimate $\mathbf{w}_{\text{ridge}}$.

- Next, we assign a zero-mean Laplace prior pdf with the scale parameter $s = 2(\rho\lambda)^{-1}$ (see Eq. (59) for the definition of Laplace pdf) to all the model parameters except the mean parameter w_0. This assignment results in the joint prior pdf

$$\text{p}(\mathbf{w}) = \text{p}(w_0)\prod_{i=1}^{N-1}\mathcal{LP}(w_i|0,2(\rho\lambda)^{-1}) = C\prod_{i=1}^{N-1}\mathcal{LP}(w_i|0,2(\rho\lambda)^{-1}) \tag{76}$$

The MAP estimate calculation using Eq. (9) leads to

$$\begin{aligned}\mathbf{w}_{\text{map}} &= \arg\max_{\mathbf{w}}\{\text{p}(\mathbf{y}|\mathbf{w},\rho)\text{p}(\mathbf{w})\} = \arg\max_{\mathbf{w}}\{\log(\text{p}(\mathbf{y}|\mathbf{w},\rho)\text{p}(\mathbf{w}))\} \\ &= \arg\max_{\mathbf{w}}\left\{\log\mathcal{N}(\mathbf{y}|\Phi\mathbf{w},\rho^{-1}\mathbf{I}_M) + \log C + \log\left(\prod_{i=1}^{N-1}\mathcal{LP}(w_i|0,2(\rho\lambda)^{-1})\right)\right\} \\ &= \arg\max_{\mathbf{w}}\left\{-\frac{\rho}{2}||\mathbf{y}-\Phi\mathbf{w}||_2^2 - \frac{\rho\lambda}{2}\sum_{i=1}^{N-1}|w_i-0|\right\} \\ &= \arg\min_{\mathbf{w}}\left\{\frac{1}{2}||\mathbf{y}-\Phi\mathbf{w}||_2^2 + \frac{\lambda}{2}\sum_{i=1}^{N-1}|w_i|\right\} \end{aligned} \tag{77}$$

The cost function in the above equation is the same as LASSO cost function in Eq. (6) with $p = 1$. Therefore, the MAP estimate $\mathbf{w}_{\text{map}}$ for the case of Gaussian likelihood function and prior pdf $\mathcal{LP}(w_i|0,2(\rho\lambda)^{-1})$ for all but the mean parameter w_0 results in same solution as the LASSO regression estimate $\mathbf{w}_{\text{lasso}}$.

Next, we explore the relationship between the Bayesian sparse learning techniques of SBL and BCS. The most salient difference between SBL and BCS is the type of sparsity inducing prior pdf. The parametrized prior pdf assigned to the unknown parameter $\mathbf{w}$ in both SBL

and BCS is a zero-mean Gaussian prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1})$. The sparsity inducing property of this prior is realized when obtaining the marginal prior pdf of $\mathbf{w}$ by integrating out the hyperparameters ($\boldsymbol{\alpha}$ in SBL and $\boldsymbol{\gamma}$ in BCS), as was explored by Babacan *et al.* [2]. The marginal pdf of model parameters w_i for SBL and BCS is obtained as following:

- For SBL algorithm the marginal prior pdf for an unknown parameter w_i is obtained as

$$\begin{aligned} \mathrm{p}(w_i) &= \int \mathrm{p}(w_i|\alpha_i)\mathrm{p}(\alpha_i)d\alpha_i \\ &= \int \mathcal{N}(w_i|0, 1/\alpha_i)\mathcal{G}(\alpha_i|c, d) \\ &= \mathcal{T}(w_i|0, 2c, d/c) \end{aligned} \tag{78}$$

  where $\mathcal{T}(x|\mu, \nu, V)$ is a student's t-distribution with mean μ, dof ν and shape parameter V (section 7.7 in [17]). Note that the variance of the student's t-distribution is well-defined only for $2c > 2$ or $c > 1$. Also the solution to the integral in Eq. (78) is only possible if $c > 1$. This is one of the reason that parameter c is always chosen greater than one in the SBL algorithm.

- In the case of BCS, the marginal prior pdf $\mathrm{p}(w_i|\lambda)$ for BCS is obtained as

$$\begin{aligned} \mathrm{p}(w_i|\lambda) &= \int \mathrm{p}(w_i|\gamma_i)\mathrm{p}(\gamma_i|\lambda)d\gamma_i \\ &= \int \mathcal{N}(w_i|0, \gamma_i)\mathcal{G}(\gamma_i|1, \lambda/2)d\gamma_i \\ &= \mathcal{LP}(w_i|0, \lambda^{-0.5}) \end{aligned} \tag{79}$$

  where $\mathcal{LP}(x|m, b)$ denotes a laplace distribution whose properties are outlined in Eq. (59).

Hence, SBL uses a student's t-distribution as the marginal prior while BCS uses a Laplace distribution albeit through a two level hierarchical prior structure shown in Eq. (57). It is argued that this choice of Laplace prior results in a more sparser solution for BCS relative to SBL [2]. This fact is realized when posing the maximization problem by rewriting the objective function $l(.)$ for SBL (accelerated) using Eq. (45) and for BCS using Eq. (67) in terms of prior precision $\gamma_i = 1/\alpha_i$ as

$$\text{SBL: } \gamma_i^{SBL} = \arg\max\{l^{SBL}(\gamma_i)\} = \arg\max\left\{\frac{1}{2}\left(\log\left(\frac{1}{1+\gamma_i s_i}\right) + \frac{q_i^2\gamma_i}{1+\gamma_i s_i}\right)\right\} \tag{80}$$

$$\text{BCS: } \gamma_i^{BCS} = \arg\max\{l^{BCS}(\gamma_i)\} = \arg\max\left\{\frac{1}{2}\left(\log\left(\frac{1}{1+\gamma_i s_i}\right) + \frac{q_i^2\gamma_i}{1+\gamma_i s_i} - \lambda\gamma_i\right)\right\} \tag{81}$$

We can further deduce that

$$\left.\frac{\partial l^{BCS}(\gamma_i)}{\partial\gamma_i}\right|_{\gamma_i=\gamma_i^{SBL}} < 0 \;\;\&\;\; \left.\frac{\partial l^{BCS}(\gamma_i)}{\partial\gamma_i}\right|_{\gamma_i=0} > 0 \Rightarrow \gamma_i^{SBL} \geq \gamma_i^{BCS} \tag{82}$$

implying that the BCS solution is always sparser (lower γ_i values) than the SBL solution. Notice that the objective function for BCS has an additional term of $-\lambda\gamma_i$ in comparison to SBL, resulting in an additional penalty on large γ_i values. Also, SBL will provide the same solution as BCS when $\lambda = 0$. It is a fair statement to make that SBL is special case of BCS under these observations.

7 Application: Sparse learning for polynomial regression

In this section, we detail the performance of frequentist (LASSO) and Bayesian (SBL, BCS) sparse learning algorithms by taking an example of polynomial regression. A realization of noisy observations is generated using the polynomial $y_i = 1 + x_i^2 + \epsilon_i$, where $\epsilon \sim \mathcal{N}(0, 1/100)$ is the additive noise with the precision of 100. The noisy observations (number of data points $M = 25$) to be used for the application of sparse learning algorithms are shown in Figure 1. An inverse problem is posed wherein a flexible model is proposed in the form of a $(N-1)^{th}$ order polynomial $y_j = \sum_{i=0}^{N-1} w_i x_j^i + \epsilon_j$ where $\epsilon_j \sim \mathcal{N}(0, \rho^{-1})$ and unknown parameter vector is $\mathbf{w} = \{w_0, w_1, \ldots, w_{N-1}\}$. Next, we seek a sparse representation of $\mathbf{w}$ using the sparse learning algorithms.

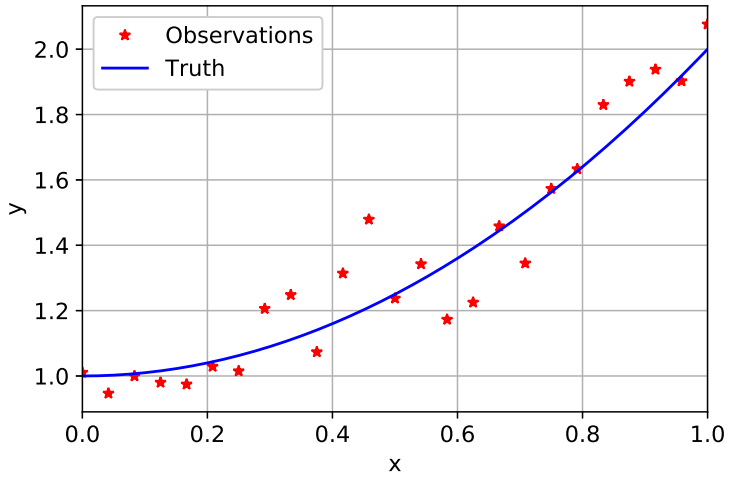


Figure 1: Noisy observations generated using $y_i = 1 + x_i^2 + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, 1/100)$

Frequentist approach:

We pose an inverse problem by considering a 5^{th} order polynomial to model the noisy observations shown in Figure 1. Table 3 lists the point estimates for the unknown model parameter $\mathbf{w} = \{w_0, w_1, w_2, w_3, w_4, w_5\}$ using the frequentist estimation techniques of OLS, LASSO and Ridge regression, as outlined in Section 2. For the case of LASSO and Ridge regression, Figure 2 shows the variation of cross-validation prediction error with varying regularization coefficient λ for the case of 5-fold cross-validation and leave-one-out cross-validation for both LASSO and Ridge regression. For the case of 5-fold cross-validation, each of the five groups of data contain five randomly chosen observation points (since $M = 25$). The cross-validation error shown in Figure 2 is the average of 20 cross-validation calculations wherein each calculation pertain to a different instance of random grouping of observations into five folds.

The optimal λ value is selected pertaining to the minimum cross-validation error, which is $\lambda = 0.15$ for LASSO and $\lambda = 0.4$ for Ridge regression. The point estimates listed in Table 3 are computed for three choices of λ: a low value of 0.001, the optimal value computed

using cross-validation, and a high value of 0.5. Note that the OLS solution is same as LASSO or Ridge regression solution with $\lambda = 0$. The mean parameter w_0 is excluded from the regularization to eliminate the dependence of penalty on the scale of observations. Figure 3 contrasts the true response with the predicted response computed using the $\mathbf{w}$ estimates shown in Table 3 for all three choices of λ. Following key inferences can be made by examining Table 3, Figure 2 and Figure 3:

1. The OLS estimates for $\mathbf{w}$ shown in Table 3 indicate large values for higher-order coefficients. This behaviour is the result of unrestricted flexibility of the model leading to overfitting and poor generalization of the learned model to unseen observations, as demonstrated by model predictions shown in Figure 3. This issue of overfitting hinders the use of OLS/MLE techniques when dealing with flexible models and noisy observations.

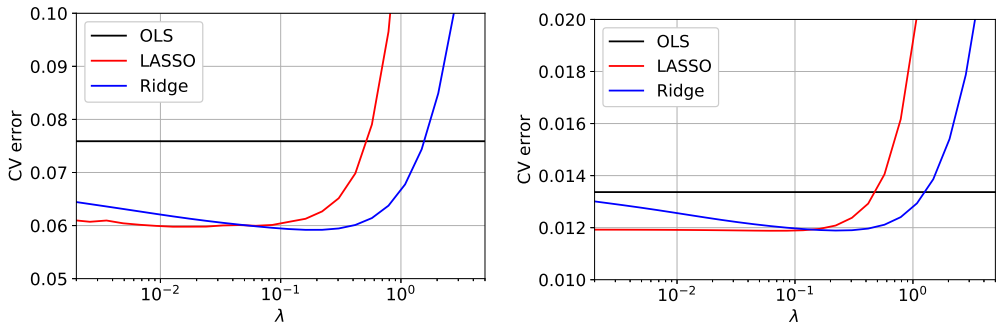
2. The $\mathbf{w}$ estimates in Table 3 for LASSO and Ridge regression have lower magnitude for the case of non-zero regularization coefficient. The regularization term in LASSO and Ridge regression helps restrict the flexibility of the model by imposing penalty on large values of $\mathbf{w}$ through the regularization. For low λ the data-fit dictates $\mathbf{w}$ values and hence the issue of overfitting is not fully resolved even though the CV prediction error is lower than it is for OLS (from Figure 2). Increasing λ increases the amount of penalty imposed by the regularization. For large λ the learned model fail to capture even the salient trends in the observations due to severe restrictions imposed on $\mathbf{w}$ values by the regularization. This over-penalty leads to a poor data-fit (case of $\lambda = 5$ in Figure 3) and higher cross-validation error (Figure 2). The optimal value of λ that ensures the best balance between data-fit and model generalization is chosen pertaining to minimum CV prediction error (Figure 2).

3. Unlike OLS/MLE, Ridge regression and LASSO estimation is well-posed even for the case of multicolinearity. For example, the condition number of matrix $\Phi^T\Phi$ for the case of a 10th-order polynomial model is of the order 10^{14} while that for matrix $\Phi^T\Phi + 0.001\mathbf{I}_M$ is of order 10^4. This is because the input x varies between 0 and 1 and therefore the basis, say, x^9 and x^{10} are almost colinear. OLS/MLE solution in this case is not possible since Rank($\Phi^T\Phi$)=9. However, Ridge regression and LASSO estimation is well-posed due to the regularization.

4. The penalty imposed by the regularization have different affects on $\mathbf{w}$ values for LASSO and Ridge regression. In case of LASSO, the penalty forces the redundant parameters to exactly zero, where the number of parameters driven to zero depends on the value of λ (Table 3). This property of LASSO is used in variable selection to produce sparse representation for $\mathbf{w}$. On the contrary, Ridge regression forces the parameter values towards zero but never exactly zero (Table 3). Even though the CV prediction error pertaining to optimal λ is comparable for LASSO and Ridge regression, LASSO is preferred in practice due to its sparsity inducing property. It can be seen from Table 3 that LASSO solution pertaining to optimal $\lambda = 0.15$ produces the same model as the data-generating model where the estimates for $\mathbf{w}$ are close to the true values.

w_i	True	OLS/MLE	LASSO			Ridge Regression		
		$\lambda = 0$	0.001	0.15*	5.0	0.001	0.4*	5.0
w_0	1.0	1.041	0.948	1.017	1.353	0.947	1.003	1.123
w_1	0.0	-3.350	0.588	0	0	0.616	0.297	0.163
w_2	1.0	29.78	-0.080	0.950	0	0.175	0.249	0.167
w_3	0.0	-81.73	0	0	0	0.067	0.207	0.156
w_4	0.0	93.26	0.735	0	0	0.805	0.179	0.144
w_5	0.0	-36.98	0	0	0	-0.153	0.155	0.133

Table 3: Frequentist estimation results for $\mathbf{w}$. (*optimal λ value computed using 5-fold cross-validation)



(a) 5-fold cross-validation

(b) Leave-one-out cross-validation

Figure 2: Cross-validation prediction error for varying regularization coefficient λ.

Bayesian approach: Bayesian linear regression

In this study, we investigate the applicability of BLR (detailed in Section 3) for the case of flexible polynomial model and noisy observations (Figure 1). We consider a fifth-order polynomial model where the joint prior pdf of $\mathbf{w}$ is chosen as a zero-mean Gaussian pdf $\mathrm{p}(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$, and the model error precision is assumed to be known as $\rho = 100$ (true value). Figure 4 shows the marginal posterior pdf of $\mathbf{w} = \{w_0, w_1, w_2, w_3, w_4, w_5\}$ obtained using BLR. Figure 5 shows the probabilistic predictions computed using this posterior pdf, along with the point predictions made using the OLS/MLE estimate and the true parameter values. Table 4 shows the precision estimate and the correlation coefficient matrix for $\mathbf{w}$ of the fifth-order polynomial model. Next we discuss some interesting observations made from the BLR parameter estimation results reported in Figure 4, Figure 5 and Table 4:

1. It is realized from Table 4 that the precision of model parameters is solely dictated by observations as the OLS/MLE precision (using Table 1) is same as the posterior precision for all $\mathbf{w}$. It also means that the prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$ has no influence on the posterior due to its large variance of 10^3. This sole dependence of model parameters on observations results in overfitting, as demonstrated in Figure 5 by a good data-fit for $0 \leq x \leq 1$ and a large bias and a large variance in predicting unseen observations for $x \geq 1$ and $x \leq 0$. The large variance in the prediction of unseen observations is attributed to the large variance in the posterior parameter pdf, as shown in Figure 4. Hence, BLR suffers from the issue of overfitting when dealing with flexible models, noisy observations and non-informative priors, thereby making the learned model in-

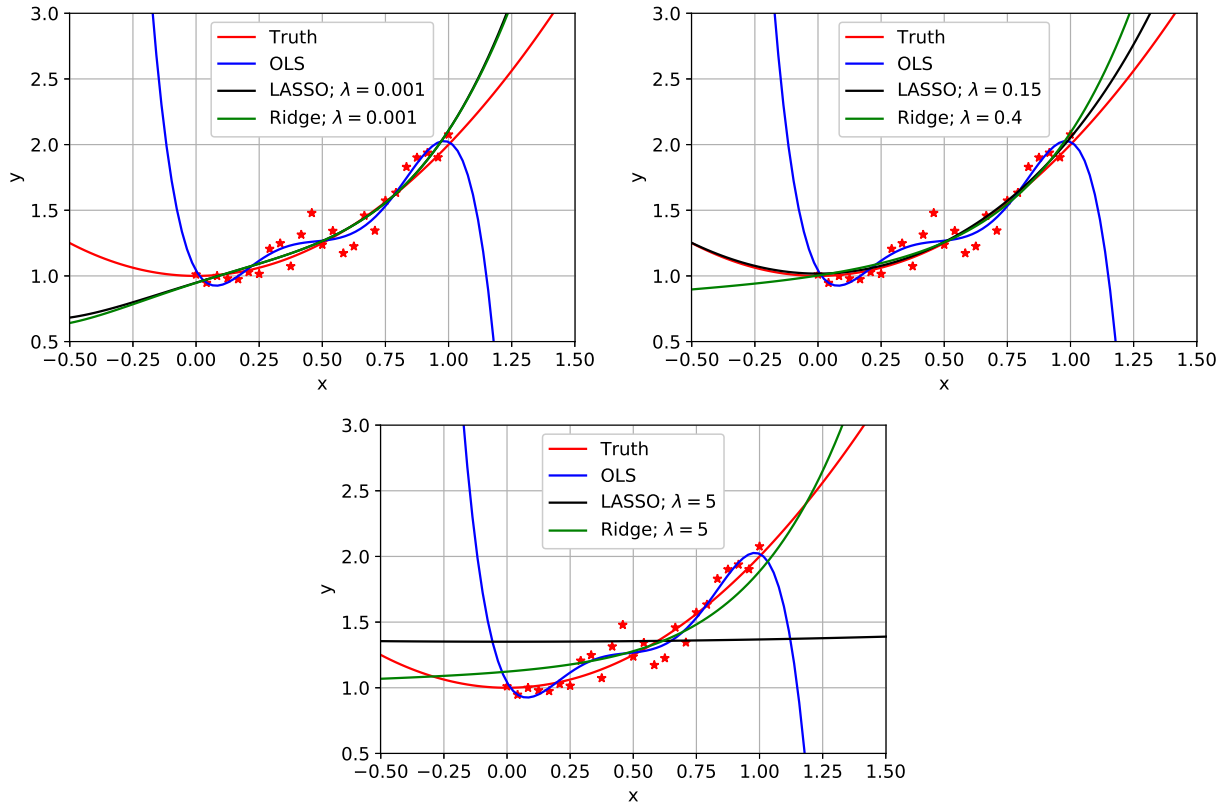


Figure 3: Model predictions using estimated $\mathbf{w}$ values from OLS, LASSO and Ridge regression, reported in table 3.

competent of making meaningful predictions.

2. The large, negative correlation among adjacent parameters shown in Table 4 allude to redundancy in the model parameter (or basis) space. For example, parameter w_4 and w_5 have a correlation of -0.98, which basically means that w_4 captures the same behaviour as w_5 and removing w_5 will not affect the data-fit capability of the model as w_4 will compensate for the removed basis. In fact, the parameters w_1, w_2, w_3, w_4, w_5 are all correlated to each other with large, negative correlation coefficients. This behaviour is expected since $0 \leq x \leq 1$ and the basis x, x^2, x^3, x^4, x^5 will exhibit some colinearity among each other. This colinearity implies that eliminating some of these basis from the model will have little effect on the data-fit capability of the model.

We realize from the BLR results that the proposed fifth-order polynomial model can be reduced to a simpler model in order to achieve good predictive capability for unseen data. As a result, we consider several polynomial models with increasing order and investigate the applicability of Bayesian model selection methodology in ranking these models. Table 5 reports log-evidence and its decomposition (according to Eq. (13)) in terms of goodness-of-fit (GOF) and expected information-gain (EIG) for varying order of the polynomial model. Note that GOF quantifies data-fit while EIG quantifies model complexity, and log-evidence equals to GOF minus EIG. Table 5 report results for three choices of prior pdf: 1) $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^0\mathbf{I}_N)$,

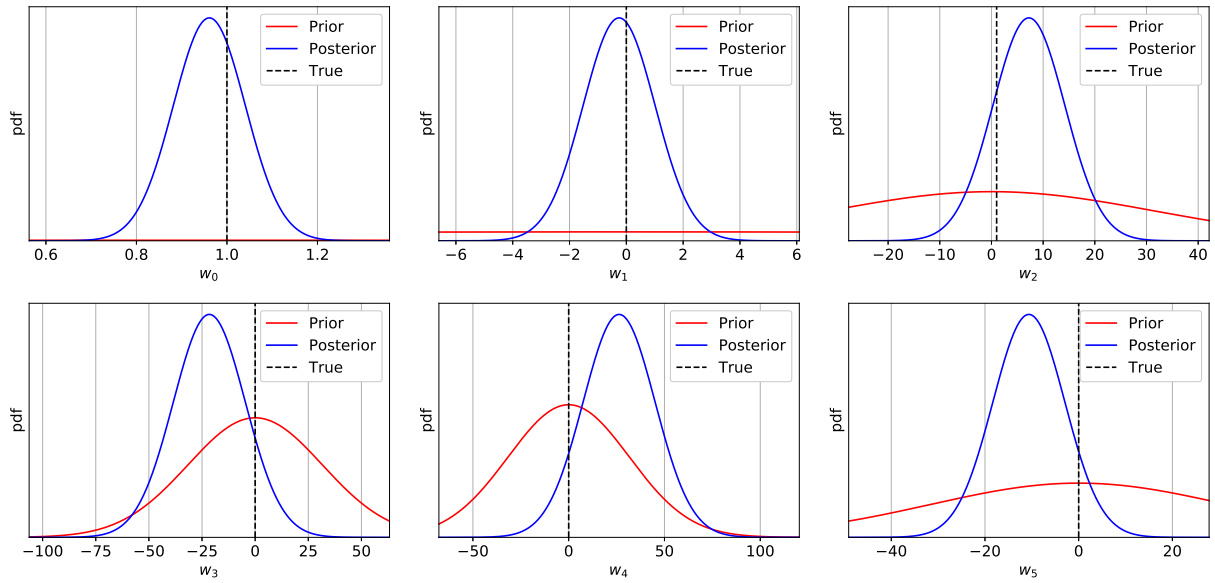


Figure 4: Marginal posterior pdf of $\mathbf{w}$ for a 5^{th} order polynomial with prior $\mathrm{p}(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$.

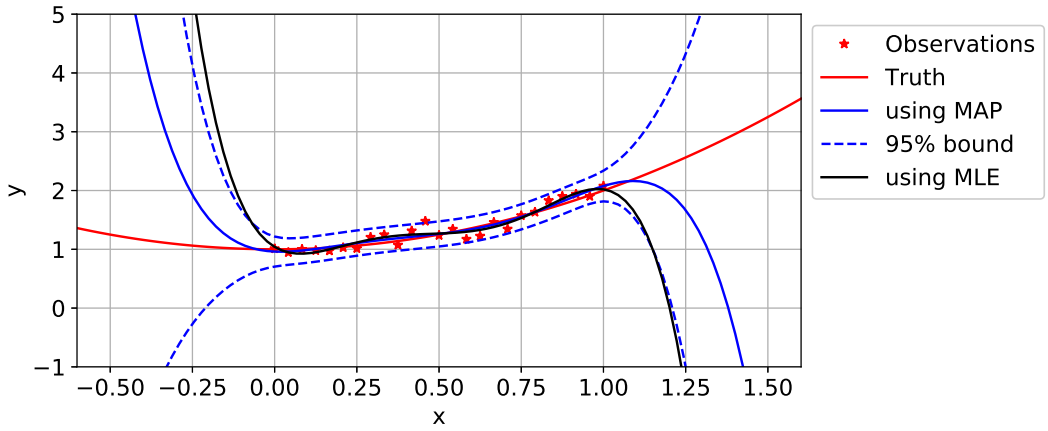


Figure 5: Comparison of model predictions using joint posterior pdf of $\mathbf{w}$ versus those from OLS/MLE estimation.

2) $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$ (base case), and 3) $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^6\mathbf{I}_N)$. Here are some observation made from the results reported in Table 5:

1. We first focus on the base case of prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$. Increasing complexity (order) of the model increases GOF only until the second-order polynomial, and the increase is marginal thereafter. On the other hand, EIG increases with increasing order of the polynomial for all models. This stagnation in GOF value while EIG increases with increasing order results in complex models being penalized heavily. This penalization of complex models leads to second-order model being the optimal model with highest log-evidence having a posterior model probability of 89%. This selection of optimal

w_i	Precision of w_i			Correlation coefficient					
	prior	MLE	posterior	w_0	w_1	w_2	w_3	w_4	w_5
w_0	1e-03	2.5e+03	2.5e+03	1.00					
w_1	1e-03	8.5e+02	8.5e+02	-0.76	1.00				
w_2	1e-03	5.3e+02	5.3e+02	0.52	-0.92	1.00			
w_3	1e-03	3.9e+02	3.9e+02	-0.34	0.76	-0.95	1.00		
w_4	1e-03	3.2e+02	3.2e+02	0.22	-0.60	0.85	-0.97	1.00	
w_5	1e-03	2.7e+02	2.7e+02	-0.13	0.47	-0.74	0.91	-0.98	1.00

Table 4: Bayesian linear regression results for precision and correlation coefficient for a 5^{th} order polynomial model.

model makes sense since the second-order model $y_i = w_0 + w_1 x_i + w_2 x_i^2 + \epsilon_i$ is the closest to the data-generating model of $y_i = 1 + x_i^2 + \epsilon_i$.

2. Next, we examine the trends in posterior model probability with increasing prior variance as shown in Table 5. It is realized that Bayesian model selection increasingly favors simpler models with increasing prior variance. This behaviour can be explained by noticing the corresponding trend in EIG for a given model. For example, for the second-order model the increase in prior variance has no influence on GOF (=20.0 for all prior choices) while EIG increases proportional to the increase in prior variance. This increase is higher for models with more parameters, which results in complex models being penalized more than simpler models. This influence of prior variance on log-evidence is well documented in literature, and has been a source of concern in practical Bayesian community as it leads to subjective conclusions that are conditioned on the choice of prior width [4, 13].

3. The issue of sensitivity of log-evidence to prior variance is further explored by considering the case of fifth-order polynomial with prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$. Figure 6 shows the variation in GOF, EIG and log-evidence with increase in prior variance for parameter w_0 and w_1. To start with, the prior variance of parameter w_0 is varied while the prior variance for all other parameters is held fixed at 10^3. The reduction of prior variance beyond a certain point results in a drastic decrease in GOF and hence the log-evidence. This reduction in GOF is due to posterior of w_0 being pushed towards zero by the zero-mean prior with a low variance. Since the data-generating model has a non-zero value of w_0, the removal of w_0 results in a drastic loss of data-fit. Hence, log-evidence is dictated by GOF for the case of a relevant parameter (like w_0).

4. Next, we vary the prior variance for w_1 while the prior variance for all other parameters is held fixed at 10^3. Figure 6 shows that the decrease in prior variance has no effect on GOF. This behaviour of GOF is expected since the data-generating model did not contain the $w_1 x$ term. Further, the removal of w_1 through reduction of prior variance decreases model complexity or EIG which then increases the log-evidence. This difference in the behaviour of log-evidence with reduction in prior variance of a relevant parameter (w_0) and an irrelevant parameter (w_1) laid the foundation for sparse learning algorithms. The notion that model complexity or basis space is controllable

through prior variance is exploited in sparse learning algorithms to obtain the optimal model nested under a flexible model.

Poly order	Prior: $\mathcal{N}(\mathbf{w}\|\mathbf{0}, 10^0\mathbf{I}_N)$			$\mathcal{N}(\mathbf{w}\|\mathbf{0}, 10^3\mathbf{I}_N)$			$\mathcal{N}(\mathbf{w}\|\mathbf{0}, 10^6\mathbf{I}_N)$		
	GOF	EIG	Log-evid	GOF	EIG	Log-evid	GOF	EIG	Log-evid
0	-112.5	4.3	-116.8	-112.5	6.9	-119.4	-112.5	10.3	-122.8
1	11.8	6.5	5.3	11.8	12.5	(2) -0.7	11.8	19.4	(40) -7.6
2	20.0	7.6	(26) 12.4	20.0	16.9	(89) 3.1	20.0	27.2	(59) -7.2
3	20.6	8.0	(32) 12.6	20.4	19.9	(7) 0.5	20.4	33.7	-13.3
4	20.7	8.4	(24) 12.3	20.0	21.6	-1.6	20.0	38.8	-18.8
5	20.6	8.7	(16) 11.9	21.4	23.5	-2.1	22.8	42.6	-19.8
True	20.5	6.7	13.8	20.5	12.6	7.9	20.5	19.5	1.0

Table 5: Bayesian model selection results for varying order of the polynomial model. The data in parenthesis is the posterior model probability in percentage (rounded), assuming all models were equally likely *a priori* and the true model $y_i = w_0 + w_2 x_i^2 + \epsilon_i$ was not included in the calculation. GOF stands for goodness-of-fit (quantifies data-fit) and EIG stands for expected information gain (quantifies model complexity), See Eq. (13) for details.

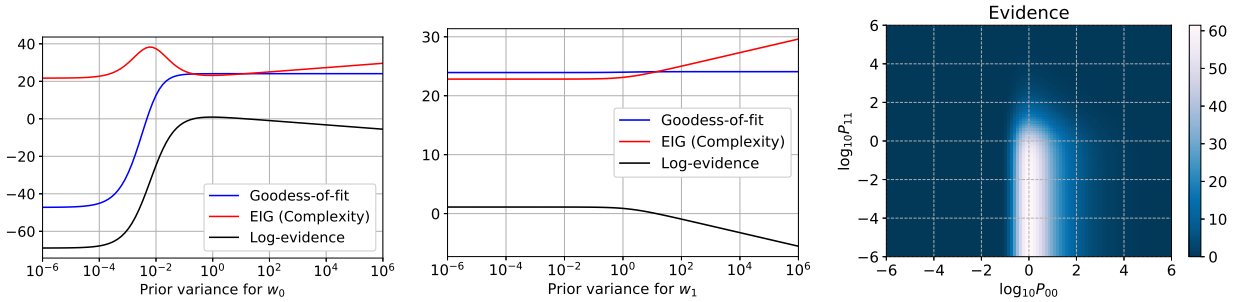


Figure 6: Variation in GOF, EIG and log-evidence with varying prior variance of parameter w_0 and w_1 for a fifth-order polynomial model where all other parameters have a fixed prior variance of 10^3.

Bayesian approach: Sparse Bayesian learning (SBL)

In this section we detail the performance of SBL (Section 4) in obtaining a sparse representation of $\mathbf{w}$ for a flexible polynomial model. The fifth-order polynomial model with prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$ is chosen as the input to SBL. Note that this model has been studied previously in this section using frequentist (OLS, LASSO, Ridge regression) and Bayesian (BLR) parameter estimation techniques. The BLR results for this choice of model and prior pdf has already been reported (and discussed therein) in Figure 4 and Table 4. Note that the prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$ equates to choosing $\mathbf{A} = 10^3\mathbf{I}_N$ or $\alpha_i = 10^3$ (for all w_i) in the ARD prior definition from Eq. (21). The starting value of ρ is chosen as 100.0 (the true value).

We first focus on the original SBL algorithm as detailed in Section 4.1 and summarized in Algorithm 1. The original SBL algorithm relies on an appropriate choice of parameters a, b, c, d of hyperpriors $\mathrm{p}(\rho) = \mathcal{G}(\rho|a, b)$ and $\mathrm{p}(\alpha_i) = \mathcal{G}(\alpha_i|c, d)$. The hyperprior for the

model error precision ρ is re-parametrized as $\mathcal{G}(\rho|a = \tilde{\rho}^2/h \geq 1, b = \tilde{\rho}/h \geq 0)$ where $\tilde{\rho}$ is the prior mean ($\tilde{\rho} = a/b$) and h is the prior variance ($a/b^2 = h$). A large h implies little prior knowledge about ρ while a small h implies an informative prior for ρ. On the other hand, the hyper-prior for α_i is always chosen to be non informative as we don't have any prior information regarding the relevance (or α_i) of model parameters, or else we won't be performing sparse learning in the first place. The non-informative prior for α_i can be obtained by choosing $c \approx 1$ and $d \approx 0$ in the Gamma distribution $\mathcal{G}(\alpha_i|c, d)$. In this study, we apply the original SBL algorithm to the fifth-order polynomial model with starting values of $\alpha_i = 10^3$ and $\rho = 10^2$ and hyperprior parameter values of $\tilde{\rho} = 10^2$, $h = 10^5$, $c = 1.0$ and $d = 10^{-6}$. Here is the discussion of results thus obtained:

1. The top-left plot in Figure 7 shows the variation in log-evidence, goodness-of-fit (GOF) and expected information gain (EIG) during SBL. The increase in log-evidence is primarily due to the decrease in EIG (quantifies model complexity) while GOF (quantifies data-fit) remains largely unchanged. This decrease in EIG is driven by the increase in α_i values (bottom-left plot in Figure 7) of irrelevant parameters during SBL, which eventually leads to their removal from the model. Hence, SBL automatically prunes redundant parameters from the flexible model, leading to decrease in EIG and increase in model evidence.

2. The sparsity indicator r_i plotted in the center-left plot in Figure 7 quantifies the relevance of each parameter w_i at each SBL iteration. A model parameter w_i is said to be contributing to the model if the sparsity indicator r_i is greater than 0.1. As SBL operates, the r_i value for parameter w_0 and w_2 approach the value of one, while r_i value for all other parameters approach the value of zero. Based on the threshold of $r_i = 0.1$, the optimal sparsity level for $\mathbf{w}$ is identified at around 35^{th} SBL iteration. Note that even though log-evidence, EIG and GOF does not vary significantly after 20^{th} iteration, sparsity indicator r_i is not converged until 60^{th} iteration. Based on our experience, the convergence of the original SBL algorithm must always include the convergence of r_i values for all parameters.

3. The variation of hyperparameter α_i values during SBL is shown in the bottom-left plot in Figure 7. The starting point for all parameters is $\alpha_i = 10^{-3}$, which is equivalent to the prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$. Note that this prior is a broad-ranged prior which allows all parameters to contribute to the model as the posterior is solely dictated by the observations (as outlined in Table 4). Beyond 30^{th} iteration all parameters except w_0 and w_2 have been eliminated from the model as their hyper-parameters are of order 10^2. This large value of hyper-parameter means both prior and posterior are a Dirac-delta pdf centered at zero. Note that even though $\mathbf{w}$ has reached its optimal sparsity level at around 35^{th} iteration, the hyper-parameter values keep on increasing till 80^{th} iteration. The hyperparameters pertaining to redundant parameters converge to a large value that is dictated by the value of parameter d of the hyperprior $\mathcal{G}(\alpha_i|c, d)$ (demonstrated later). Based on our experience, we recommend that the convergence of the original SBL algorithm should not be dictated by the convergence of α_i values as they keep increasing long after the optimal sparsity has been identified.

4. The center-right plot in Figure 7 shows the KL divergence between marginal prior and posterior pdf of each model parameter w_i during SBL. The KL divergence for w_0 and w_2 is non-zero while it is zero for all other parameters since their prior and posterior pdfs are identical (Dirac-delta pdf centered at zero). This behavior of KL divergence for redundant parameters fuels the decrease in EIG which then leads to increase in log-evidence (according to Eq. (13)).

5. Once the SBL algorithm has converged, the resulting sparse model is identified by the optimal hyper-parameter value $\boldsymbol{\alpha}^* = \{1.00, 2406,\ 0.915, 1102, 1676, 2016\}$ and $\rho^* = 87.8$. The prior and posterior pdf obtained using these optimal values is shown in Figure 8. As expected, both the prior and posterior pdf for redundant parameters w_1, w_3, w_4, w_5 is a Dirac-delta pdf centered at zero. On the other hand, the posterior pdf of relevant parameters w_0 and w_2 accurately captures the values used in the data-generating model. The MAP estimates for relevant parameters is computed as $m_0 = 1.00$ and $m_2 = 1.04$.

6. Figure 9 contrasts the model predictions computed using pre-SBL (using prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$) and post-SBL (using prior $\mathcal{N}(\mathbf{w}|\mathbf{0}, (\boldsymbol{\alpha}^*)^{-1}\mathbf{I}_N)$) posterior pdf. It is realized that the predictions from the sparse model (post-SBL) have lower bias and lower variance when predicting unseen data, albeit at the expense of higher bias than the flexible model when predicting current observations. This higher bias is however neglible in the comparison of the reduction in bias and variance when predicting unseen data using the sparse model. The sparse model thus balances model parsimony and data-fit as facilitated by the Bayesian model selection framework powered by model evidence.

7. The starting value of hyperparameters α_i can significantly affect the convergence characteristics of the original SBL algorithm. Figure 10 shows the variation in sparsity indicator r_i for three different choices of starting hyper-parameter value: 1) $\alpha_i = 10^0$, 2) $\alpha_i = 10^{-3}$, and 3) $\alpha_i = 10^{-6}$. Note that for all three choices the resulting sparse model is the same as the data-generating model. However, it is realized that the convergence is significantly hindered when choosing a large starting value of α_i. A large value of α_i pertains to a highly informative prior (prior variance $= 1/\alpha_i$) that will tend to dictate the posterior pdf of $\mathbf{w}$. As a result, SBL takes longer time to instill the effect of observations into the posterior pdf as SBL needs to see flexibility in the model for it to identify sparsity. If we assign informative priors, we are limiting the flexibility of the model, and hence hindering the ability of SBL to identify the sparse model. Based on our experience, we recommend the use of small (but not too small) starting value of α_i, i.e small enough to not alter the posterior space of $\mathbf{w}$.

8. In the case where the starting value of hyperparameter for SBL cannot be determined *a priori* as being appropriate (small enough), the issue of slow convergence can be remedied by supplying an appropriate value of parameter c of the hyperprior $\alpha_i \sim \mathcal{G}(\alpha_i|c, d)$. The effect of parameter c on SBL convergence can be examined by taking the case of SBL where the starting value of $\alpha_i = 10^1$ (the corresponding r_i variation already shown in Figure 10). Note that all the results reported so far pertained to $c = 1.0$. Figure 11 shows the variation in sparsity indicator r_i for three different

choices of c: 1.0, 1.5 and 10.0. Note that when $c = 1.5$ the convergence is expedited without affecting the resulting sparsity of $\mathbf{w}$. However, when $c = 10.0$, the sparse model contains only the mean term (w_0). Note that the objective function being maximized by SBL is the combination of hyperprior $\mathcal{G}(\alpha_i|c,d)$ and model evidence. Therefore, a large c allows the objective function to be controlled by the hyperprior rather than the model evidence, which then amplifies the sparsity identified by SBL. Based on our experience, we recommend a value of c in the range $[1.1, 2.0]$ which will allow the SBL maximization to be well-defined even in the case of poor starting value of hyperparameter $\boldsymbol{\alpha}$.

Next, we apply the accelerated SBL algorithm detailed in Section 4.2 to obtain the sparse solution of $\mathbf{w}$ for the fifth-order polynomial model with the initial values of $\alpha_i = 10^{-3}$ (prior pdf $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$) and $\rho = 100.0$. This model and prior combination was considered previously in this section for the application of original SBL algorithm. Note that we don't need to explicitly assign hyperprior $\mathcal{G}(\alpha_i|c,d)$ as the accelerated SBL algorithm is derived by assuming non-informative prior on α_i using $c \approx 1.00$ and $d \approx 0.00$. The hyper-prior for ρ is chosen as $\mathcal{G}(\rho|a = \tilde{\rho}^2/h, b = \tilde{\rho}/h)$ where $\tilde{\rho} = 10^2$ and $h = 10^5$. This choice of hyper-priors is alike to the hyperpriors used in the application of original SBL algorithm shown previously in this section. Figure 12 shows the variation of SBL entities for the accelerated SBL algorithm. Notice that both the SBL algorithms result in the same sparse solution for $\mathbf{w}$, wherein the accelerated algorithm converges faster than the original algorithm. The original SBL algorithm converges at 80^{th} iteration (from Figure 7) while accelerated SBL algotihm converges at 20^{th} iteration (from Figure 12). This increase in efficiency can be attributed to the fact that when a parameter w_i is deemed irrelevant the accelerated algorithm proceeds by removing the corresponding basis from the model by explicitly assigning $\alpha_i = \infty$ and $r_i = 0$. This explicit removal of basis expedites the convergence of SBL as α for redundant parameter is no longer being estimated (like original SBL) and is fixed at infinity. Since the resulting sparse solution is identical, the resulting posterior pdf of relevant parameters w_0 and w_2 and the model predictions are also identical for both the SBL algorithms, and have already been reported in Figure 8 and Figure 9, respectively.

Bayesian approach: Bayesian compressive sensing (BCS)

Next, we report the performance of BCS algorithm (as detailed in Section 5) in identifying the sparsity in $\mathbf{w}$ for the fifth-order polynomial model with the initial values of $\gamma_i = 1/\alpha_i = 10^3$ (prior pdf $\mathcal{N}(\mathbf{w}|\mathbf{0}, 10^3\mathbf{I}_N)$), $\lambda = 1.0$ and $\rho = 100.0$. The hyper-priors are chosen as $\mathcal{G}(\lambda|10, 10)$ and $\mathcal{G}(\rho|1, 10^5)$. Figure 13 shows the variation in BCS entities during BCS. The sparse solution obtained from BCS is identical to that obtained using SBL. Note that the efficiency of BCS is better than the original SBL algorithm and is comparable to the accelerated SBL algorithm.

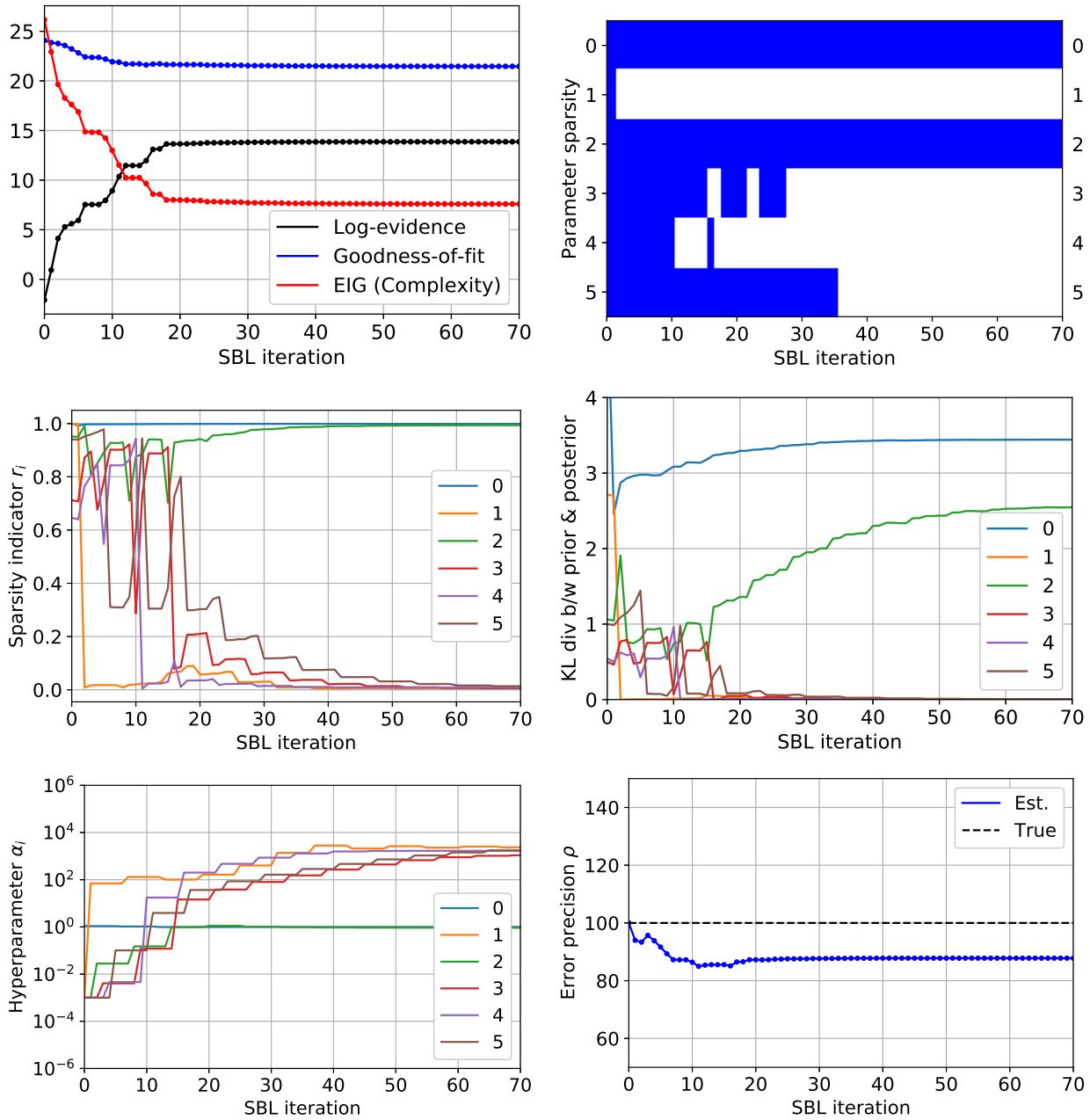


Figure 7: Original SBL algorithm results: Vairaiton of SBL entities for a fifth-order polynomial model with starting hyper-parameter values of $\alpha_i = 10^3$ (for all parameters) and $\rho = 100$. The hyperprior parameters were chosen as $a = 100 * 10^{-6}$, $b = 10^{-6}$, $c = 1.0$ and $d = 10^{-6}$.

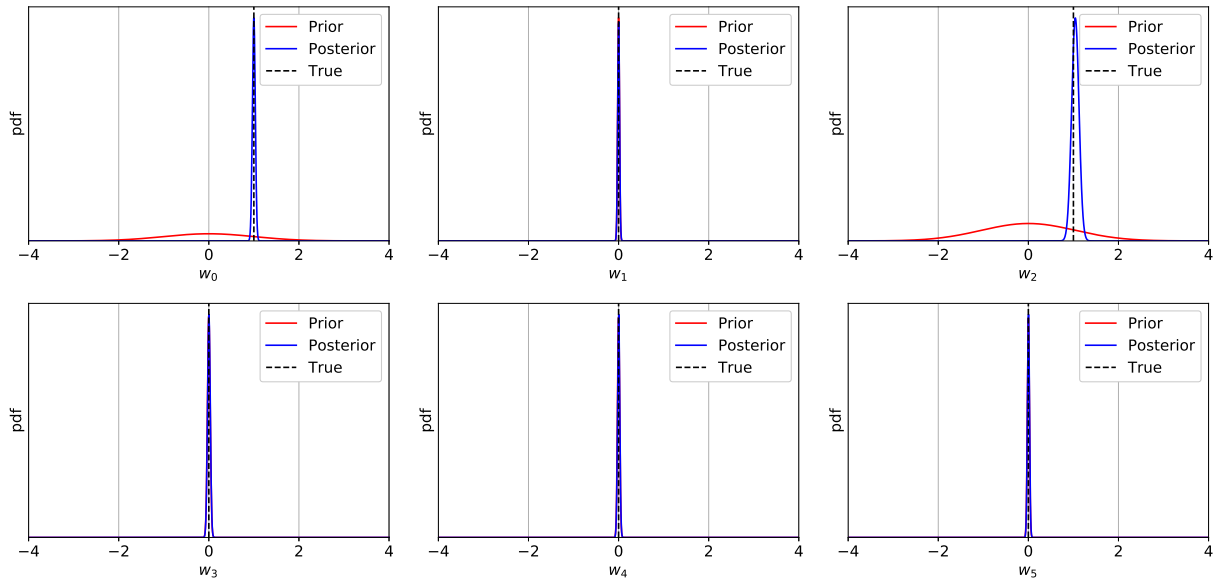


Figure 8: Original SBL algorithm results: Post-SBL marginal posterior pdf of $\mathbf{w}$ for a 5^{th} order polynomial.

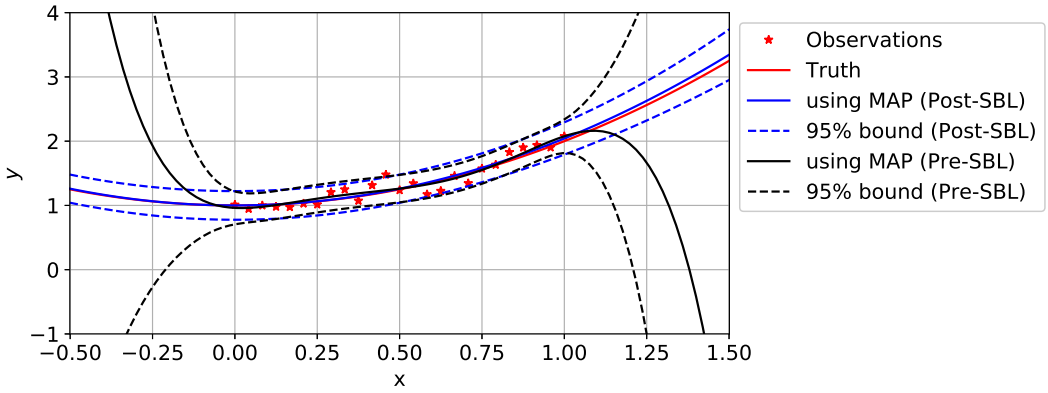


Figure 9: Original SBL algorithm results: Comparison of model predictions using joint posterior pdf of $\mathbf{w}$ before (pre-SBL) and after (post-SBL) running the original SBL algorithm.

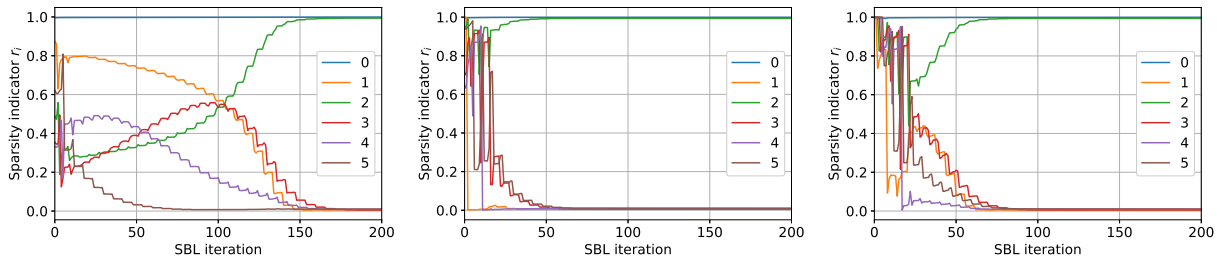


Figure 10: Original SBL algorithm results: Variation in sparsity indicator r_i for starting hyperparameter value of $\alpha_i = 10^1$ (left), $\alpha_i = 10^{-3}$ (center), and $\alpha_i = 10^{-6}$ (right).

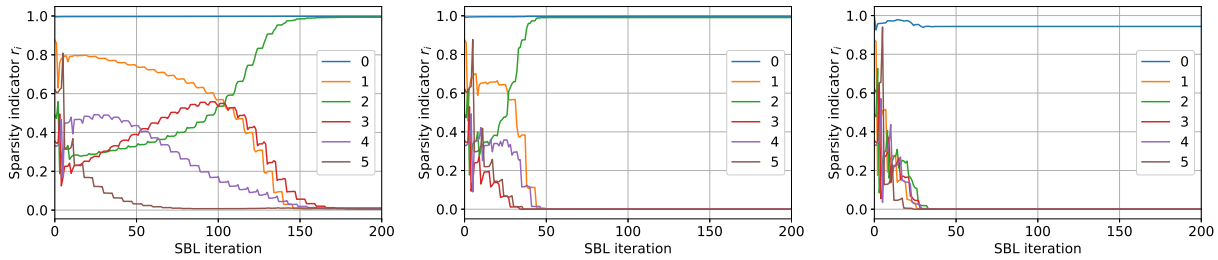


Figure 11: Original SBL algorithm results: Variation in sparsity indicator r_i for starting hyperparameter value of $\alpha_i = 10^1$ and $c = 1.0$ (left), $c = 1.5$ (center), and $c = 10.0$ (right).

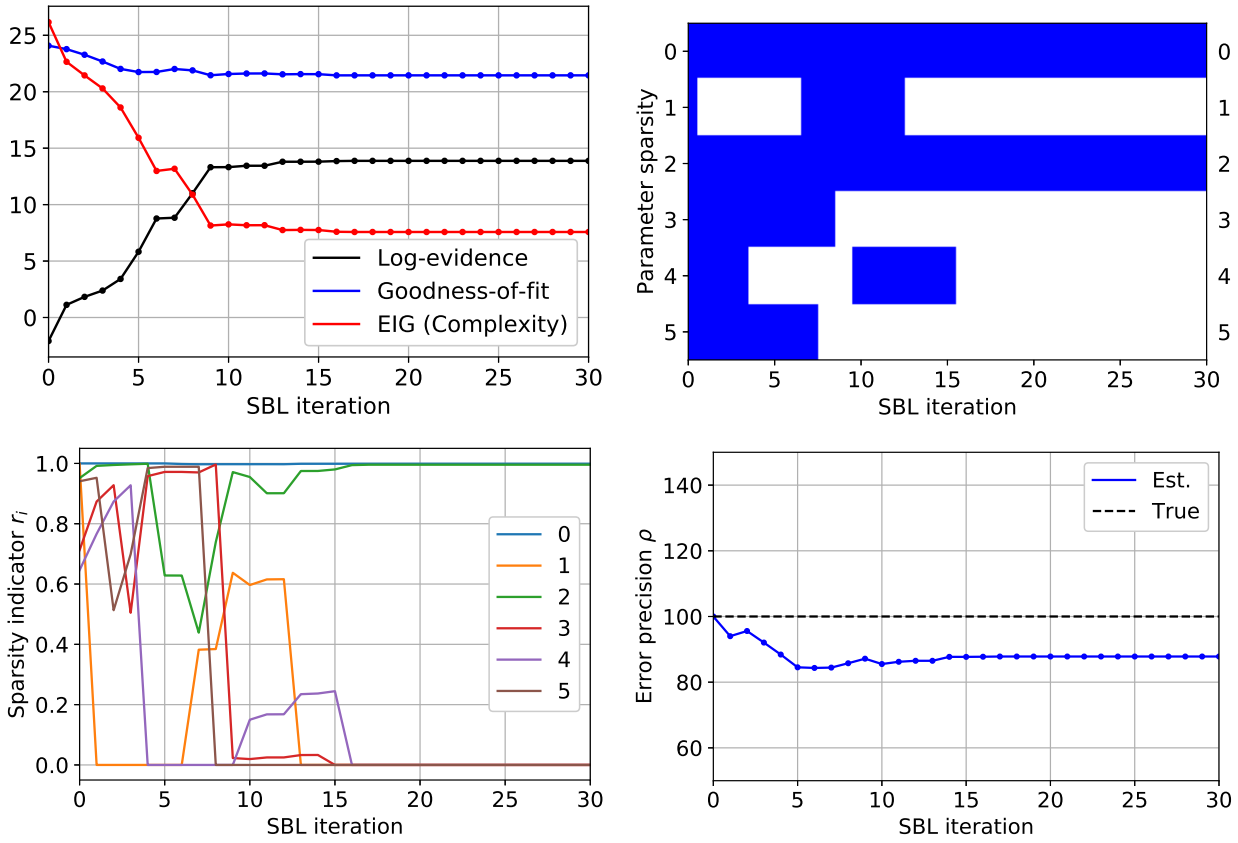


Figure 12: Accelerated SBL algorithm results: Variation of SBL entities for a fifth-order polynomial model with starting hyper-parameter values of $\alpha_i = 10^3$ (for all parameters) and $\rho = 100$. The hyperprior parameters were chosen as $a = 100 * 10^{-6}$, $b = 10^{-6}$, $c = 1.0$ and $d = 10^{-6}$.

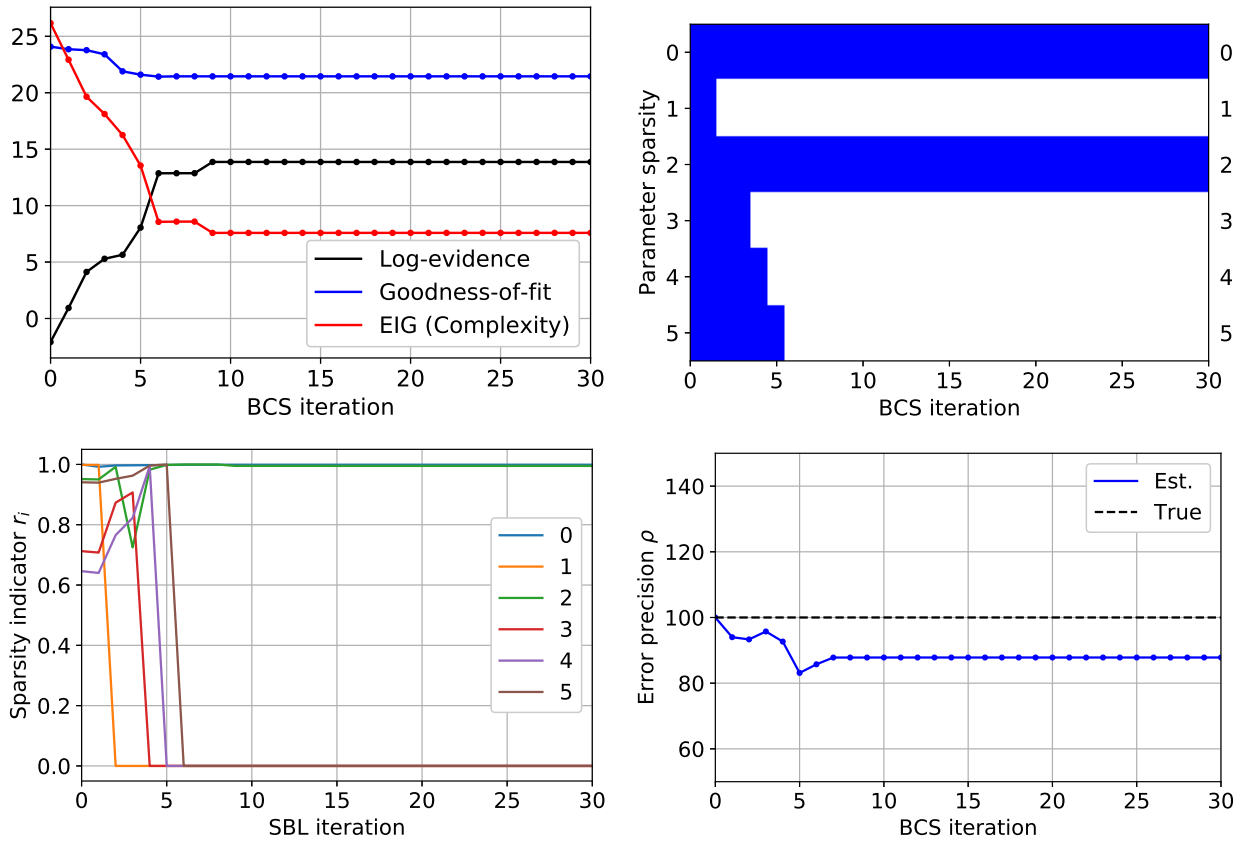


Figure 13: BCS results: Variation of BCS entities for a fifth-order polynomial model with starting hyper-parameter values of $\alpha_i = 10^3$ (for all parameters) and $\rho = 100$. The hyperprior parameters were chosen as $a = 100 * 10^{-6}$, $b = 10^{-6}$, $c = 1.0$ and $d = 10^{-6}$.

A Some important relations for Gaussian distributions

A.1 Gaussian conditional distribution formula

Given a jointly Gaussian random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$ with marginal distributions $\mathrm{p}(\mathbf{X}_1) = \mathcal{N}(\mathbf{X}_1|\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathrm{p}(\mathbf{X}_2) = \mathcal{N}(\mathbf{X}_2|\boldsymbol{\mu}_2, \Sigma_2)$ and the joint distribution as

$$\mathrm{p}(\mathbf{X}_1, \mathbf{X}_2) = \mathcal{N}\left(\begin{Bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{Bmatrix} \middle| \begin{Bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{Bmatrix}, \begin{bmatrix}\Sigma_1 & \Sigma_{12}\\ \Sigma_{12}^T & \Sigma_2\end{bmatrix}\right), \tag{83}$$

then the conditional distribution can be computed as

$$\mathrm{p}(\mathbf{X}_1|\mathbf{X}_2) = \mathcal{N}(\mathbf{X}_1|\mathbf{X}_2|\boldsymbol{\mu}_{1|2}, \Sigma_{1|2}) \tag{84a}$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_2^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) \tag{84b}$$

$$\Sigma_{1|2} = \Sigma_1 - \Sigma_{12}\Sigma_2^{-1}\Sigma_{12}^T \tag{84c}$$

where $\mathbf{K}_{1|2} = \Sigma_{12}\Sigma_2^{-1}$ is called the gain matrix and $\Sigma_{1|2}$ is also called the schur complement of matrix Σ_2.

A.2 Relation between likelihood function and sampling distribution of MLE estimate

The goal is to compute the constant of proportionality c in

$$\mathrm{p}(\mathbf{y}|\mathbf{w}, \rho) = c\mathcal{N}(\mathbf{w}|\mathbf{m}, \mathbf{P}) \tag{85}$$

for the case of linear regression with Gaussian model errors. For a linear regression setup presented in Eq. (1), the likelihood function is known as $\mathrm{p}(\mathbf{y}|\mathbf{w}, \rho) = \mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho\mathbf{I}_M)$. Also, the sampling distribution of the MLE estimate is also Gaussian and known as $\mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}}, \mathbf{P}_{\text{mle}})$ where $\mathbf{w}_{\text{mle}} = (\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}$ is the MLE estimate and $\mathbf{P}_{\text{mle}} = (\rho\Phi^T\Phi)^{-1}$ is the covariance in the MLE estimate. Note that this sampling distribution is same as the posterior distribution $\mathrm{p}(\mathbf{w}|\mathbf{y}, \rho)$ in Bayesian linear regression when using flat priors. Using Bayesian inference we can infer that likelihood function is proportional to this distribution, i.e. $\mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho\mathbf{I}_M) \propto \mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}}, \mathbf{P}_{\text{mle}})$. In this section we aim to find the constant of proportionality c that relates the likelihood function to the sampling distribution of MLE estimates as following:

$$\begin{aligned}
c &= \frac{\mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho^{-1}\mathbf{I}_M)}{\mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}}, \mathbf{P}_{\text{mle}})} = \frac{\mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho^{-1}\mathbf{I}_M)}{\mathcal{N}(\mathbf{w}|(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}, (\rho\Phi^T\Phi)^{-1})}\\
&= \frac{(2\pi)^{N/2}|(\rho\Phi^T\Phi)^{-1}|^{1/2}}{(2\pi)^{M/2}|\rho^{-1}\mathbf{I}_M|^{1/2}} \times \frac{\exp\left\{-\frac{1}{2}(\mathbf{y} - \Phi\mathbf{w})^T\rho\mathbf{I}_M(\mathbf{y} - \Phi\mathbf{w})\right\}}{\exp\left\{-\frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{mle}})^T\rho\Phi^T\Phi(\mathbf{w} - \mathbf{w}_{\text{mle}})\right\}}\\
&= \left(\frac{\rho}{2\pi}\right)^{\frac{M-N}{2}} \frac{1}{|\Phi^T\Phi|^{1/2}} \times \exp\left\{-\frac{\rho}{2}v\right\}
\end{aligned}$$

where the term v is evaluated as

$$\begin{aligned}
v &= (\mathbf{y}-\Phi\mathbf{w})^T(\mathbf{y}-\Phi\mathbf{w}) - (\mathbf{w}-\mathbf{w}_{\text{mle}})^T\Phi^T\Phi(\mathbf{w}-\mathbf{w}_{\text{mle}}) \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T\Phi^T\mathbf{y} + \cancel{\mathbf{w}^T\Phi^T\Phi\mathbf{w}} - \cancel{\mathbf{w}^T\Phi^T\Phi\mathbf{w}} + 2\mathbf{w}^T\Phi^T\Phi\mathbf{w}_{\text{mle}} - \mathbf{w}_{\text{mle}}^T\Phi^T\Phi\mathbf{w}_{\text{mle}} \\
&= \mathbf{y}^T\mathbf{y} - 2\mathbf{w}^T\Phi^T\mathbf{y} + 2\mathbf{w}^T\cancel{\Phi^T\Phi}\{\cancel{(\Phi^T\Phi)^{-1}}\Phi^T\mathbf{y}\} - \{(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y}\}^T\cancel{\Phi^T\Phi}\{\cancel{(\Phi^T\Phi)^{-1}}\Phi^T\mathbf{y}\} \\
&= \mathbf{y}^T\mathbf{y} - \cancel{2\mathbf{w}^T\Phi^T\mathbf{y}} + \cancel{2\mathbf{w}^T\Phi^T\mathbf{y}} - \mathbf{y}^T\Phi(\Phi^T\Phi)^{-1}\Phi^T\mathbf{y} \\
&= \mathbf{y}^T(\mathbf{I}_M - \Phi(\Phi^T\Phi)^{-1}\Phi^T)\mathbf{y}
\end{aligned}$$

Hence, the constant of proportionality c can be obtained as

$$c = \left(\frac{\rho}{2\pi}\right)^{\frac{M-N}{2}} \frac{1}{|\Phi^T\Phi|^{1/2}} \times \exp\left\{-\frac{\rho}{2}\mathbf{y}^T(\mathbf{I}_M - \Phi(\Phi^T\Phi)^{-1}\Phi^T)\mathbf{y}\right\} \tag{86}$$

$$\underbrace{\mathcal{N}(\mathbf{y}|\Phi\mathbf{w}, \rho\mathbf{I}_M)}_{\text{Likelihood function}} = c \qquad \underbrace{\mathcal{N}(\mathbf{w}|\mathbf{w}_{\text{mle}}, \mathbf{P}_{\text{mle}})}_{\text{Posterior pdf solely due to Likelihood}} \tag{87}$$

A.3 Cross-entropy between two multivariate Gaussian pdfs

Consider that we have a random vector $\mathbf{w}$ of size N that has two different probability distributions denoted as $r_1 = \mathcal{N}(\mathbf{w}|\mathbf{m}_1, \mathbf{P}_1)$ and $r_2 = \mathcal{N}(\mathbf{w}|\mathbf{m}_2, \mathbf{P}_2)$. The cross-entropy $H(r_1, r_2)$ between pdf r_1 and r_2 is defined as the expectation of logarithm of r_2 pdf where the expectation is taken with respect to pdf r_1. The cross-entropy $H(r_1, r_2)$ for Gaussian pdfs can be computed analytically as following:

$$\begin{aligned}
H(r_1, r_2) = \mathrm{E}[\log\mathcal{N}(\mathbf{w}|\mathbf{m}_2, \mathbf{P}_2)] &= \int \mathcal{N}(\mathbf{w}|\mathbf{m}_1, \mathbf{P}_1)\log\mathcal{N}(\mathbf{w}|\mathbf{m}_2, \mathbf{P}_2)d\mathbf{w} \\
&= \mathrm{E}\left[-\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{P}_2| - \frac{1}{2}(\mathbf{w}-\mathbf{m}_2)^T\mathbf{P}_2^{-1}(\mathbf{w}-\mathbf{m}_2)\right] \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{P}_2| - \frac{1}{2}\mathrm{E}\left[(\mathbf{w}-\mathbf{m}_2)^T\mathbf{P}_2^{-1}(\mathbf{w}-\mathbf{m}_2)\right]
\end{aligned}$$

Using the property $z^Tz = \text{Trace}(zz^T)$, we get

$$\begin{aligned}
H(r_1, r_2) &= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{P}_2| - \frac{1}{2}\mathrm{E}\left[\text{Trace}(\mathbf{P}_2^{-1}(\mathbf{w}-\mathbf{m}_2)(\mathbf{w}-\mathbf{m}_2)^T)\right] \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{P}_2| - \frac{1}{2}\text{Trace}(\mathrm{E}[\mathbf{P}_2^{-1}(\mathbf{w}-\mathbf{m}_2)(\mathbf{w}-\mathbf{m}_2)^T]) \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{P}_2| - \frac{1}{2}\text{Trace}(\mathbf{P}_2^{-1}\mathrm{E}[\mathbf{w}\mathbf{w}^T - 2\mathbf{m}_2\mathbf{w}^T + \mathbf{m}_2\mathbf{m}_2^T]) \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{P}_2| - \frac{1}{2}\text{Trace}(\mathbf{P}_2^{-1}(\mathrm{E}[\mathbf{w}\mathbf{w}^T] - 2\mathbf{m}_2\mathrm{E}[\mathbf{w}^T] + \mathbf{m}_2\mathbf{m}_2^T)) \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{P}_2| - \frac{1}{2}\text{Trace}(\mathbf{P}_2^{-1}(\mathbf{P}_1 + \mathbf{m}_1\mathbf{m}_1^T - 2\mathbf{m}_2\mathbf{m}_1^T + \mathbf{m}_2\mathbf{m}_2^T)) \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{P}_2| - \frac{1}{2}\text{Trace}(\mathbf{P}_2^{-1}(\mathbf{m}_1-\mathbf{m}_2)(\mathbf{m}_1-\mathbf{m}_2)^T) - \frac{1}{2}\text{Trace}(\mathbf{P}_2^{-1}\mathbf{P}_1) \\
&= -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{P}_2| - \frac{1}{2}(\mathbf{m}_1-\mathbf{m}_2)^T\mathbf{P}_2^{-1}(\mathbf{m}_1-\mathbf{m}_2) - \frac{1}{2}\text{Trace}(\mathbf{P}_2^{-1}\mathbf{P}_1) \\
&= \log\mathcal{N}(\mathbf{m}_1|\mathbf{m}_2, \mathbf{P}_2) - \frac{1}{2}\text{Trace}(\mathbf{P}_2^{-1}\mathbf{P}_1)
\end{aligned}$$

Note that if $\mathbf{P}_1 = \mathbf{P}_2$ then $\mathrm{E}[\log(\mathcal{N}(\mathbf{w}|\mathbf{m}_2, \mathbf{P}_2))] = \log(\mathcal{N}(\mathrm{E}[\mathbf{w}]|\mathbf{m}_2, \mathbf{P}_2))$ where the expectation is taken with respect to pdf $r_1 = \mathcal{N}(\mathbf{w}|\mathbf{m}_1, \mathbf{P}_1)$.

References

[1] T. W. Anderson. *An introduction to multivariate statistical analysis*. Hoboken, NJ John Wiley, 3rd ed edition, 2003.

[2] S. D. Babacan, R. Molina, and A. K. Katsaggelos. Bayesian compressive sensing using laplace priors. *IEEE Transactions on Image Processing*, 19(1):53–63, Jan 2010.

[3] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics, 1985.

[4] Kenneth P. Burnham and David R. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer, 2 edition, 2002.

[5] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[6] Anita C. Faul and Michael E. Tipping. Analysis of sparse bayesian learning. In *Advances in Neural Information Processing Systems 14*, pages 383–389. MIT Press, 2001.

[7] Daniel Fink. A compendium of conjugate priors, 1997.

[8] W. Gilks (Ed.), S. Richardson (Ed.), D. Spiegelhalter (Ed.), P. Van der Heijden, B. Morgan, and N. Keiding. *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1996.

[9] Mohinder S. Grewal. *Kalman Filtering*, pages 705–708. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[10] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The elements of statistical learning: data mining, inference, and prediction, 2nd Edition*. Springer series in statistics. Springer, 2009.

[11] E. T. Jaynes. Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630, May 1957.

[12] S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *IEEE Transactions on Signal Processing*, 56(6):2346–2356, June 2008.

[13] Charles C. Liu and Murray Aitkin. Bayes factors: Prior sensitivity and model generalizability. *Journal of Mathematical Psychology*, 52(6):362 – 375, 2008.

[14] David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[15] Matthew Muto and James L. Beck. Bayesian updating and model class selection for hysteretic structural models using stochastic simulation. *Journal of Vibration and Control*, 14(1-2):7–34, 2008.

[16] Trevor Park and George Casella. The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686, 2008.

[17] K. B. Petersen and M. S. Pedersen. The matrix cookbook, nov 2012. Version 20121115.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[19] Michael E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, September 2001.

[20] Michael E. Tipping and Anita C. Faul. Fast marginal likelihood maximization for sparse bayesian models. In *AISTATS*, 2003.