The model is simple, until proven otherwise – how to cope in an ever changing world

A.C. Faul*, G. Pilikos

acf22@cam.ac.uk

Abstract

There are several challenges with which data present us nowadays. For one there is the abundance of data and the necessity to extract the essential information from it. When tackling this task a balance has to be struck between putting aside irrelevant information and keeping the relevant one without getting lost in detail, known as over-fitting. The law of parsimony, also known as Occam's razor should be a guiding principle, keeping models simple while explaining the data.

The next challenge is the fact that the data samples are not static. New samples arrive constantly through the pipeline. Therefore, there is a need for models which update themselves as the new sample becomes available. The models should be flexible enough to become more complex should this be necessary. In addition the models should inform us which samples need to be collected so that the collection process becomes most informative.

Another challenge are the conclusions we draw from the data. After all, as popularized by Mark Twain: "There are three kinds of lies: lies, damned lies, and statistics." An objective measure of confidence is needed to make generalized statements

The last challenge is the analysis. Can we build systems which inform us of the underlying structure and processes which gave rise to the data? Moreover, it is not enough to discover the structure and processes, we also need to add meaning to it. Here different disciplines need to work together.

Keywords— Bayesian inference; confidence measure; updating models.

1 Data and models

When analyzing data, we make the assumption that the data are a result of an underlying process which we do not know. Sometimes we know the principles of the process, but not the parameters which govern it. For example the physics of waves are well understood. However, they depend on the medium the wave travels in, the material and its properties. The medium or mixture of media are the unknown parameters of the process.

In data analysis we are given samples which are measurements $y_1, \ldots, y_N$, where each measurement depends on parameters we know $\mathbf{x}_1, \ldots, \mathbf{x}_N$. All these can be measured with more or less effort, but the effort is never prohibitive. Note that the notation $\mathbf{x}_n$ indicates $\mathbf{x}_n = (x_{n1}, \ldots, x_{np})^T$. That is each sample depends on p parameters. A real world application also depends on parameters which cannot be measured, or these measurements would be disproportionately difficult and costly.

If we had a solution to the underlying process, we could predict the measurement from a function $f(\mathbf{x})$ as

$$y_n = f(\mathbf{x}_n).$$

Here the argument to the function are the known parameters, while the unknown parameters are part of the function and depend on the process.

If we had a set of candidate functions $d_1(\mathbf{x}), \ldots, d_M(\mathbf{x})$, which all are solutions to the process for different unknown parameters, we could try which fits the measurements and thus infer the underlying structure. We say the functions $d_1(\mathbf{x}), \ldots, d_M(\mathbf{x})$ form a dictionary and assume

$$f(\mathbf{x}) = \sum_{m=1}^{M} c_m d_m(\mathbf{x}),$$

where $c_1, \ldots, c_M$ are coefficients and these need to be determined. These functions are also called basis functions and are the building blocks which build a model for the data and the model obeys the underlying process.

All analytic functions can be built from an *infinite* set of basis functions. However, computers remain to be finite machines and thus we need to restrict ourselves to a finite set, but we want to find the most suitable finite set of basis functions. Or in other words the smallest set which describes the data adequately.

The relationship to the measurements is

$$y_n = f(\mathbf{x}_n) + \varepsilon_n = \sum_{m=1}^{M} c_m d_m(\mathbf{x}_n) + \varepsilon_n,$$

where ε_n is noise intrinsic to the measurement process and assumed to be independent and identically, normally distributed, $\mathscr{N}(0,\sigma^2)$. Let D be the matrix with entries

$$D_{n,m} = d_m(\mathbf{x}_n)$$

and let $\mathbf{y}^T = (y_1,\dots,y_N)$, $\mathbf{c}^T = (c_1,\dots,c_M)$ and $\varepsilon^T = (\varepsilon_1,\dots,\varepsilon_N)$, then

$$\mathbf{y} = D\mathbf{c} + \varepsilon.$$

D is known as the design matrix. As it is written here D is an $N \times M$ matrix. However, N and M are not static. N varies with the number of samples, while M varies with the dictionary of basis functions. The rows of D, $\mathbf{d}_n = (d_1(\mathbf{x}_n),\dots,d_M(\mathbf{x}_n))^T$, $n = 1,\dots,N$, are defined by the data samples, while the columns of D, $\tilde{\mathbf{d}}_m = (d_m(\mathbf{x}_1),\dots,d_m(\mathbf{x}_N))$, $m = 1,\dots,M$, are defined by the model.

Since the noise is i.i.d. normal with mean 0 and variance σ^2, the log likelihood of observing $\mathbf{y}$ given the model specified by D, $\mathbf{c}$ and σ^2 is

$$\log\mathscr{L}(\mathbf{y}|D,\mathbf{c},\sigma^2) = -\frac{N}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-D\mathbf{c})^T(\mathbf{y}-D\mathbf{c}).$$

A large log likelihood means that the model explains the data well.

The challenge is to find the dictionary of basis functions and the coefficients. Once these are found, predictions for unseen parameters $\mathbf{x}_*$ can be made by

$$y_* = \sum_{m=1}^{M} c_m d_m(\mathbf{x}_*) = \mathbf{d}_*^T\mathbf{c},$$

where $\mathbf{d}_*^T = (d_1(\mathbf{x}_*),\dots,d_M(\mathbf{x}_*))$.

2 Sparse Bayesian Learning

In 2000/1 sparse Bayesian learning ([1], [2], [3]) was introduced. The central idea of this is that the coefficients $\mathbf{c}$ follow a distribution. We define a prior distribution $p(\mathbf{c})$ using all information apart from the samples themselves quantifying our belief about the coefficients. For example a simple assumption is that each coefficient c_m is a priori normally distributed with mean zero and variance α_m^{-1}. α_m is a hyper-parameter and known as the precision of the distribution. If α_m is very large the distribution becomes peaked at its mean and we have more confidence in the value of c_m than if α_m is small and the width of the distribution large. The multivariate prior distribution is given by

$$p(\mathbf{c}|\alpha) = (2\pi)^{-M/2}\sqrt{|A|}\exp\left(\mathbf{c}^T A\mathbf{c}\right),$$

where A is a diagonal matrix with entries $A_{mm} = \alpha_m$ and where $|\cdot|$ denotes the matrix determinant. The multivariate posterior distribution is also normal, since it is a convolution of Gaussians, with mean μ and variance Σ given by

$$\Sigma = \left(A + \sigma^{-2}D^T D\right)^{-1} \qquad \mu = \sigma^{-2}\Sigma D^T\mathbf{y}.$$

Given the posterior distribution of the coefficients, the probabilistic interpretation of the measurement y_n is that it is drawn from a univariate normal distribution with

$$\begin{array}{ll} \text{mean} & m_n = \mathbf{d}_n^T\mu, \\ \text{variance} & \sigma_n^2 = \sigma^2 + \mathbf{d}_n^T\Sigma\mathbf{d}_n. \end{array}$$

If the variance is small, it indicates that at this point the model explains the data well. If the variance is large, the model is not adequate at this point. This can indicate that the dictionary of basis functions is unsuitable for these data and needs to be amended.

The most suitable values for A are found by maximizing the logarithm of the marginal likelihood $\mathscr{L}(\mathbf{y}|\alpha,\sigma^2)$, which can be calculated analytically

$$\log\mathscr{L}(\mathbf{y}|\alpha,\sigma^2) = -\frac{1}{2}\left(N\log 2\pi + \log|C| + \mathbf{y}^T C^{-1}\mathbf{y}\right),$$

where $C = \sigma^2\mathbf{I} + DA^{-1}D^T$. In the process of maximization many of the hyper-parameters tend to infinity and the posterior distribution of those coefficients becomes infinitely peaked at zero. This means that these basis functions are not relevant for our model.

Faul and Tipping developed sparse Bayesian learning further, [4], [5], [6], by noticing that $\log\mathscr{L}(\mathbf{y}|\alpha,\sigma^2)$ can be maximized with respect to a single hyper-parameter. Instead of starting with all candidate basis functions from the dictionary in the model and then deleting those whose hyper-parameters tend to infinity, this version initializes the model with a single basis function and sets the hyper-parameters of the others notionally to infinity. Then the basis function d_m where setting its hyper-parameter α_m to its optimal value (given the current model) gives the largest increase in the marginal likelihood is found and the model updated accordingly. The algorithm converges if no significant increase in the marginal log likelihood can be achieved anymore. Note that the optimal value of α_m can be finite or infinite as shown in [5] and [6]. That means that if d_m is not in the model and the optimal α_m is finite, it gets added to the model. If d_m is in the model and the optimal α_m is infinite, it gets deleted from the model. The third option is that d_m is in the model and the optimal α_m is finite, in which case α_m is updated to this value. In all three cases, Σ and μ have to be updated, since A has changed. Fast update formulae are given in the appendix of [6].

It should be noted here that the dictionary of candidate basis functions does not need to be static. A new candidate basis function can be created, evaluated, and possibly added to the model at any point.

Thus sparse Bayesian learning addresses the first challenge. A probability distribution is associated with the coefficients and the posterior distribution gives probabilistic meaning to whether a basis function is relevant for the model to explain the given data or not.

3 New data

In this section we will address the challenge of new data arriving through the pipeline. Following the approach in [5] we calculate the change in the logarithm of the marginal likelihood for the current model, when a data sample $(y_*, \mathbf{x}_*)$ is added. This means adding a row to the design matrix D yielding

$$D_* = \left(\begin{array}{c} D \\ \hline \mathbf{d}_*^T \end{array} \right).$$

We then have

$$C_* = \sigma^2 \mathbf{I} + \left(\begin{array}{c} D \\ \hline \mathbf{d}_*^T \end{array} \right) A^{-1} \left(\begin{array}{c|c} D^T & \mathbf{d}_* \end{array} \right) = \left(\begin{array}{c|c} \mathbf{C} & \mathbf{v} \\ \hline \mathbf{v}^T & v \end{array} \right)$$

where $\mathbf{v} = DA^{-1}\mathbf{d}_*$ and $v = \mathbf{d}_*^T A^{-1} \mathbf{d}_* + \sigma^2$. Note that C_* is symmetric.

Using the formulae for block matrices we have

$$|C_*| = |C||v - \mathbf{v}^T C^{-1} \mathbf{v}|$$

and

$$C_*^{-1} = \left(\begin{array}{c|c} C^{-1} + \dfrac{C^{-1}\mathbf{v}\mathbf{v}^T C^{-1}}{v - \mathbf{v}^T C^{-1}\mathbf{v}} & -C^{-1}\mathbf{v}\dfrac{1}{v - \mathbf{v}^T C^{-1}\mathbf{v}} \\ \hline -\mathbf{v}^T C^{-1} \dfrac{1}{v - \mathbf{v}^T C^{-1}\mathbf{v}} & \dfrac{1}{v - \mathbf{v}^T C^{-1}\mathbf{v}} \end{array} \right).$$

Letting $\mathbf{y}_*^T = (y_1, \ldots, y_N, y_*)$, we can calculate

$$\mathbf{y}_*^T C_*^{-1} \mathbf{y}_* = \mathbf{y}^T C^{-1} \mathbf{y} + \frac{1}{v - \mathbf{v}^T C^{-1}\mathbf{v}} (\mathbf{y}^T C^{-1} \mathbf{v} - y_*)^2$$

Thus the logarithm of the marginal likelihood $\log \mathscr{L}(\mathbf{y}_*|\alpha, \sigma^2)$ is $\log \mathscr{L}(\mathbf{y}|\alpha, \sigma^2) + \Delta\mathscr{L}$, where

$$\begin{aligned} \Delta\mathscr{L} &= -\frac{1}{2} \left[\log 2\pi + \log |v - \mathbf{v}^T C^{-1} \mathbf{v}| + \right. \\ & \qquad \left. \frac{1}{v - \mathbf{v}^T C^{-1}\mathbf{v}} (\mathbf{v}^T C^{-1} \mathbf{y} - y_*)^2 \right]. \end{aligned}$$

This change can be interpreted probabilistically. To this end, note that the matrices Σ and C are related by the Woodbury matrix identity,

$$\Sigma = A^{-1} - A^{-1} D^T C^{-1} D \mathbf{A}^{-1}.$$

The predictive distribution for y_* has variance and mean

$$\begin{aligned} \sigma_*^2 &= \sigma^2 + \mathbf{d}_*^T \Sigma \mathbf{d}_* = v - \mathbf{v}^T C^{-1} \mathbf{v}, \\ m_* &= \mathbf{d}_*^T \mu = \sigma^{-2} \mathbf{d}_*^T \Sigma D^T \mathbf{y} = \mathbf{v}^T C^{-1} \mathbf{y}, \end{aligned} \tag{1}$$

where we used the fact that $DA^{-1}D^T = C - \sigma^2 I$. Thus $(\mathbf{v}^T C^{-1} \mathbf{y} - y_*)^2$ is the square of the difference of the sample measurement and its mean predicted by the current model.

Thus the change in the logarithm of the marginal likelihood is

$$\begin{aligned} \Delta\mathscr{L} &= -\frac{1}{2} \left[\log 2\pi + \log \sigma_*^2 + \left(\frac{m_* - y_*}{\sigma_*} \right)^2 \right] \\ &= \log \frac{1}{\sqrt{2\pi}\sigma_*} \exp \left(-\frac{(m_* - y_*)^2}{2\sigma_*^2} \right). \end{aligned}$$

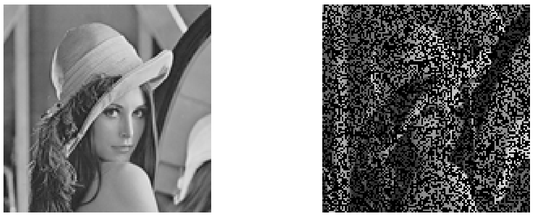
Hence the change is the logarithm of the likelihood of the new data value y_* at $\mathbf{x}_*$ given the predictive probability distribution $\mathscr{N}(m_*, \sigma_*^2)$.

Since $\sigma_* \geq \sigma$, the change lies between $-\infty$ and $\log \frac{1}{\sqrt{2\pi}\sigma}$. It can be positive. In this case the new sample affirms the model. If the likelihood of the data is small, the marginal likelihood is reduced, indicating that the model should be updated. To do so, all quantities need to be updated. Efficient update formulae are given in the Appendix.

To conclude, in this and the previous section we have developed an adaptive framework where new candidate basis functions and new data samples can be added. If this leads to a reduction in the marginal likelihood, the algorithm continues the process of updating the model by maximizing the marginal likelihood, by either adding, updating, and removing basis functions.

4 Uncertainty

In the previous section we have seen that sparse Bayesian learning infers a predictive distribution for y_* which is $\mathscr{N}(m_*, \sigma_*^2)$ with mean and variance as given in (1). This predictive distribution is heavily dependent on the model,



(a) (b)

Figure 1: Original (a) and decimated image (b)

since it depends on $\mathbf{d}_*$ which are the basis functions included in the model evaluated at $\mathbf{x}_*$. It is customary to choose basis functions for the dictionary which decay quickly when moving away from their centre, or basis functions with finite, compact support. Therefore the degenerate case is possible where $\mathbf{d}_*$ is close to, or even equal to zero, and thus the predictive probability distribution becomes $\mathscr{N}(0,\sigma^2)$ which is meaningless. This was noted in [7]. The solution proposed there is unsuitable, since it relies on the introduction of a basis function centred at $\mathbf{x}_*$, but the shape and width of this function can be varied or needs to be trained. However, the confidence we place in the predictions should only be informed by the data.

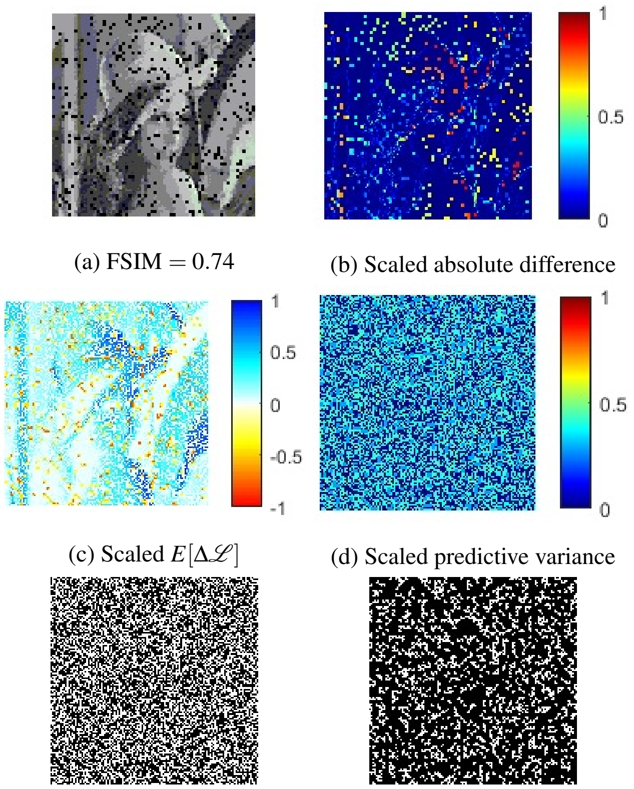
Let $\mathscr{S}$ be a subset of the samples. This could be all samples or a suitable set of neighbours of $\mathbf{x}_*$. We estimate the probability distribution of y_* to be normal with mean and variance

$$\begin{aligned}\bar{m} &= \operatorname*{mean}_{\mathbf{x}_i \in \mathscr{S}}\{y_i\},\\ \bar{\sigma} &= \operatorname*{var}_{\mathbf{x}_i \in \mathscr{S}}\{y_i\}.\end{aligned}$$

With this estimate the expected change when considering $\mathbf{x}_*$ in the logarithm of the marginal likelihood is

$$\begin{aligned}E[\Delta\mathscr{L}] &= \int_{-\infty}^{\infty}\left[\log\frac{1}{\sqrt{2\pi}\sigma_*} - \frac{(y_*-m_*)^2}{2\sigma_*^2}\right] * \\ &\quad \frac{1}{\sqrt{2\pi}\bar{\sigma}}\exp\left(-\frac{(y_*-\bar{m})^2}{2\bar{\sigma}^2}\right)dy_*\\ &= \log\frac{1}{\sqrt{2\pi}\sigma_*} - \frac{\bar{\sigma}^2+(\bar{m}-m_*)^2}{2\sigma_*^2}.\end{aligned}$$

The second term is the important one. If the predictive probability distribution does not match well the probability distribution estimated from the data in the neighbourhood the expected change in the logarithm of the marginal likelihood is negative. This expected change creates an uncertainty map with the largest negative values being the most uncertain regions. The uncertainty map can guide the data gathering, informing us where additional samples are necessary.



(a) FSIM $= 0.74$ (b) Scaled absolute difference

(c) Scaled $E[\Delta\mathscr{L}]$ (d) Scaled predictive variance

(e) Predictive variance ≥ 0.35 (f) Predictive variance ≥ 0.4

Figure 2: Reconstruction with Haar wavelets of scale 1.

5 Experiments

To illustrate the algorithm, data in the form of images were chosen. The reasoning being that images display structure such as edges, but are also very varied with different textures. From the image 55% of pixels were removed randomly and the remaining pixels were used to infer the values at the missing pixels. Figure 1 is an example.

Different basis functions were employed, Haar wavelets of scale 1 which have a support of 2×2 pixels, Haar wavelets of scale 2 with a support of 4×4 pixels and the Gaussian radial basis functions centered at each pixel with a radius of 8 pixels. These dictionaries were chosen to illustrate different aspects of the algorithm.

Figure 2 illustrates the degenerate case where $\mathbf{d}_*$ is zero. In the reconstructed image 2a this is visible as black areas, since zero is interpreted as black. The feature similarity index measure (FSIM) [8] between the original and the reconstruction is 0.74, where the closer the value to 1, the better the reconstruction. Figure 2b shows the absolute difference between the original and the reconstruction scaled to lie between 0 and 1, while 2c displays the expected change in the logarithm of the marginal likelihood scaled to lie between -1 and 1. The predictions for pixels with negative values in

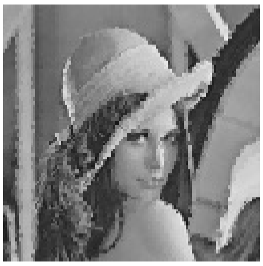


Figure 3: 5% more samples as informed by $E[\Delta\mathscr{L}]$, FSIM $= 0.93$

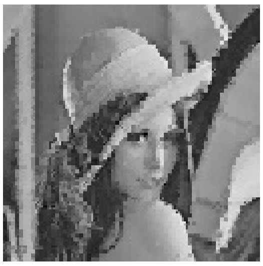
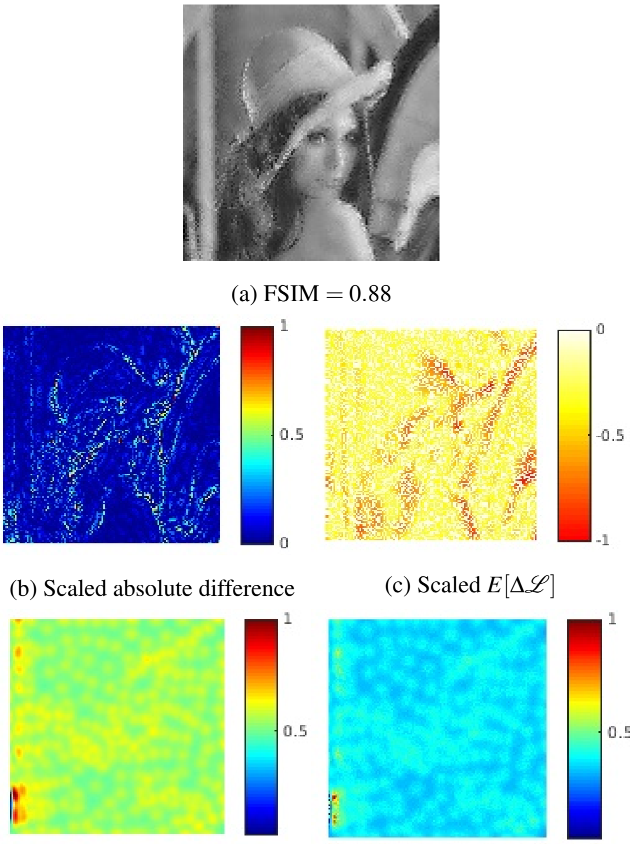


Figure 4: Improvements with Haar wavelets of scale 2, FSIM $= 0.91$

this change are not regarded as trustworthy, while the ones with positive change are accepted. These correspond well with pixels of large absolute difference between the original and the reconstruction. Contrast this with Figure 2d showing the predicted variance scaled to lie between 0 and 1. It is hard to set a threshold for accepting predictions as 2e and 2f illustrate. They show the pixels with a scaled predicted variance of 0.35 or more and 0.4 or more respectively. The former would disregard a lot more predictions than the latter.

The confidence measure is giving a good indication where predictions are problematic. There are two possibilities to improve the results. The first one is to obtain more samples in the problem areas and insert these into the algorithm as described in Section 3. This results in 5% more data being gathered and the resulting reconstruction is shown in Figure 3. The FSIM has moved up to 0.93.

The other possibility, arises from the analysis why the reconstruction is poor. In the decimated image there are areas of 2×2 missing pixels. Since Haar wavelets of scale 1 also have a support of 2×2 pixels, no reconstruction is made in these areas, since no information is available to base the reconstruction on. Thus we expect improvements with a different choice of basis functions. Before this, however, we accept all predictions where there is a positive change in the logarithm of the marginal likelihood, arguing that we have confidence in our model there. We then reconstruct the remaining missing pixels with a different basis function. Figure 4 shows the resulting reconstruction with Haar wavelets of scale 2. The FSIM is now 0.91.



(a) FSIM $= 0.88$

(b) Scaled absolute difference

(c) Scaled $E[\Delta\mathscr{L}]$

(d) Scaled predictive variance

(e) Scaled augmented variance

Figure 5: Reconstruction with Gaussian radial basis functions.

To complete the results, Figure 5 illustrates the reconstruction with Gaussian radial basis functions. With this reconstruction the FSIM is 0.88. The effect of Gaussians is that edges are smoothed. This is illustrated in Figure 5b of the absolute difference between the original and the reconstruction scaled to lie between 0 and 1. It shows large differences especially along the edges. Because of the smoothing effect of Gaussians, $\mathscr{S}$ was chosen to be the set of all samples to calculate the expected change in the logarithm of the marginal likelihood. This is shown in Figure 5c scaled to lie between -1 and 1. In fact the change was a reduction for all predictions. The problem areas highlighted by the confidence map again corresponds well with the areas of large absolute difference. Figure 5d shows the predicted variance scaled to lie between 0 and 1. Again this is not informative. Neither is the augmented variance as proposed by [7] which is displayed in Figure 5e. Both these variances are dominated by the choice of basis functions while the confidence measure proposed here removes this dependency.

6 Conclusions

We have presented a mathematical framework based on Bayesian inference, where the model can be augmented with more building blocks, while the Bayesian approach keeps the model sparse. Early runs can shed some light on the nature of the building blocks necessary for a good model. A possible line of investigation is whether this can be utilized to learn basis functions. Different dictionaries of basis functions are suitable for different data. The framework can also incorporate new data samples arriving. The confidence measure in form of the change in the logarithm of the marginal likelihood can inform which predictions are trustworthy and where more samples are necessary to obtain a model in which we have more confidence. The framework gives probabilistic interpretations and thus enables the expert community to add meaning to the results.

References

[1] M.E. Tipping, “The Relevance Vector Machine”, *Advances in Neural Information Processing Systems 12*, pp. 652–658, 2000.

[2] M.E. Tipping, C.M. Bishop, “Variational relevance vector machines”, *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pp. 46–53, 2000.

[3] M.E. Tipping, “Sparse Bayesian learning and the relevance vector machine”, *Journal of Machine Learning Research 1*, pp. 211–244, 2001.

[4] A.C. Faul, M.E. Tipping, “A variational approach to robust regression”, *Proceedings of ICANN’01*, pp. 95–102, 2001.

[5] A.C. Faul, M.E. Tipping, “Analysis of sparse Bayesian learning”, *Advances in Neural Information Processing Systems 14*, pp. 383–389, 2002.

[6] M.E. Tipping, A.C. Faul, “Fast marginal likelihood maximization for sparse Bayesian models”, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[7] C.E. Rasmussen, J. Quinoñero Candela, “Healing the Relevance Vector Machine through Augmentation”, *Proceedings of the 22nd International Conference on Machine Learning*, pp. 689–696, 2005.

[8] L. Zhang, L. Zhang, X. Mou, D. Zhang, ”FSIM: A Feature Similarity Index for Image Quality Assessment”, *IEEE Tranaction on Image Processing, vol. 20, no. 8* pp. 2378–2386, 2011.

Appendix

Following the notation of [6], the sparsity factor is updated by

$$\tilde{S}_m = S_m + \frac{1}{\sigma_*^2}\left[\frac{1}{\sigma^2}\mathbf{d}_*^T \Sigma D^T \hat{\mathbf{d}}_m - d_m(\mathbf{x}_*)\right]^2.$$

Note that the quantity in square brackets is the error the current model makes when inferring the value of d_m at $\mathbf{x}_*$. The quality factor becomes

$$\tilde{Q}_m = Q_m + \frac{1}{\sigma_*^2}\left[\frac{1}{\sigma^2}\mathbf{d}_*^T \Sigma D^T \hat{\mathbf{d}}_m - d_m(\mathbf{x}_*)\right][m_* - y_*].$$

The covariance matrix is updated as follows:

$$\tilde{\Sigma} = \Sigma - \frac{1}{\sigma_*^2}\Sigma \mathbf{d}_* \mathbf{d}_*^T \Sigma,$$

while the mean becomes

$$\tilde{\mu} = \mu - \frac{m_* - y_*}{\sigma_*^2}\Sigma \mathbf{d}_*.$$