TRADING APPROXIMATION QUALITY VERSUS SPARSITY WITHIN INCREMENTAL AUTOMATIC RELEVANCE DETERMINATION FRAMEWORKS

Dmitriy Shutin

Institute for Communications and Navigation
German Aerospace Center (DLR)
Münchner Str. 20, 82234 Wessling Germany

*Thomas Buchgraber**

Signal Processing and Speech Comm. Lab
Graz University of Technology
Inffeldgasse 16c, 8010 Graz, Austria

ABSTRACT

In this paper a trade-off between sparsity and approximation quality of models learned with incremental automatic relevance determination (IARD) is addressed. An IARD algorithm is a class of sparse Bayesian learning (SBL) schemes. It permits an intuitive and simple adjustment of estimation expressions, with the adjustment having a simple interpretation in terms of signal-to-noise ratio (SNR). This adjustment allows for implementing a trade-off between sparsity of the estimated model versus its accuracy in terms of residual mean-square error (MSE). It is found that this adjustment has a different impact on the IARD performance, depending on whether the measurement model coincides with the used estimation model or not. Specifically, in the former case the value of the adjustment parameter set to the true SNR leads to an optimum performance of the IARD with the smallest MSE and estimated signal sparsity; moreover, the estimated sparsity then coincides with the true signal sparsity. In contrast, when there is a model mismatch, the lower MSE can be achieved only at the expense of less sparser models. In this case the adjustment parameter simply trades the estimated signal sparsity versus the accuracy of the model.

1. INTRODUCTION

The problem of recovering a sparse representations from a noisy measurement has received considerable attention in recent years [1–4]. Typically, the goal is to find an L-dimensional vector $\boldsymbol{w} = [w_1, \ldots, w_L]^T$, which is assumed to contain only K nonzero elements, $K \ll L$, from a noise-perturbed measurement vector $\boldsymbol{t}$ generated from the following linear measurement model:

$$\boldsymbol{t} = \boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{\xi}. \tag{1}$$

Here $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_L]$ is a dictionary matrix with L columns corresponding to some fixed signal waveforms $\boldsymbol{\phi}_l \in \mathbb{R}^N$, $l = 1, \ldots, L$. The additive perturbation $\boldsymbol{\xi}$ is typically assumed to be a white Gaussian random vector with zero mean and covariance matrix $\tau^{-1}\boldsymbol{I}$, where τ is a noise precision parameter. Despite the linearity of (1), the additional sparsity constraints imposed on the weight vector $\boldsymbol{w}$ make the estimation problem nontrivial. Specifically, learning the sparse vector $\boldsymbol{w}$ requires imposing formal constraints on the objective function for the model parameters $\boldsymbol{w}$ to ensure the resulting solution is sparse [2–6].

Sparse Bayesian learning (SBL) [7, 8] is a class of such learning schemes where these constraints are introduced through the use of a hierarchical prior. Specifically, the prior on $\boldsymbol{w}$ is introduced as $p(\boldsymbol{w}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})$, where $p(\boldsymbol{w}|\boldsymbol{\alpha})$ is a product of Gaussian probability density functions (pdfs) with zero mean and precision parameters $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_L]^T$ – also called sparsity parameters – that regulate the width of these pdfs.[1] The weights $\boldsymbol{w}$ and the sparsity parameters $\boldsymbol{\alpha}$ are then found from the mode of the corresponding posterior pdfs.

Recently, it has been shown that SBL can be realized very efficiently when the hyperprior $p(\boldsymbol{\alpha})$ is assumed to be flat [10,11] or non-informative [12][2]. These special cases of SBL, commonly referred to as automatic relevance determination (ARD), have several important theoretical as well as practical advantages. First, the ARD formulation of the hyperprior $p(\boldsymbol{\alpha})$ is related to weighted versions of minimum ℓ_1-norm regression and basis pursuit denoising (see [2,6,11,14]) – more traditional "non-Bayesian" methods for learning sparse representations. Second, the ARD formulation can be exploited to construct very fast and efficient inference schemes [10–12] – incremental ARD (IARD).

Typically, the weights $\boldsymbol{w}$ are determined from the mode of the posterior pdf $p(\boldsymbol{w}|\boldsymbol{t}, \boldsymbol{\alpha}, \tau)$ that can be computed analytically; the sparsity parameters $\boldsymbol{\alpha}$ and the noise precision τ are then estimated by maximizing the posterior pdf $p(\boldsymbol{\alpha}, \tau|\boldsymbol{t})$. The latter optimization is generally analytically intractable,

*This work was supported in part by the Austrian Science Fund (FWF) under Award S10610-N13 within the national research network SISE

[1] It is also possible to extend the SBL prior formulation to priors involving three layers of hierarchy [9].

[2] Note that while these priors are improper, i.e., they do not integrate to one, in our case the resulting weight posterior under such hyperprior is, nonetheless, a proper pdfs (see [13] for more details).

yet using the IARD scheme it can be solved very efficiently. Specifically, it can be shown [10, 12] that the optimum of the corresponding objective function with respect to the sparsity parameter of the lth component can be computed in closed form provided the other sparsity parameters α_k, $k \neq l$, and the noise precision parameter τ are fixed. The posterior pdf $p(\boldsymbol{\alpha}, \tau|\boldsymbol{t})$ is then optimized incrementally with respect to $\boldsymbol{\alpha}$ by cycling through L sparsity parameters in a round-robin fashion [10, 12]. An important consequence of this algorithm is that the analytical analysis of the corresponding inference expressions allows for deriving exact conditions, termed pruning conditions, that determine if the sparsity parameter that optimizes the posterior $p(\boldsymbol{\alpha}, \tau|\boldsymbol{t})$ is finite or infinite[3].

The pruning conditions are the key features of the IARD scheme that differentiate it from the other sparse learning techniques. In contrast to traditional minimum ℓ_1-norm methods, where the recovery of a sparse solution is established by guaranteeing (with some high probability) an upper bound on the norm of the error between the estimated vector $\widehat{\boldsymbol{w}}$ and a true sparse vector $\boldsymbol{w}_0$ [2, 15, 16], the pruning conditions obtained with IARD are "non-asymptotic"; they determine which elements in the weight vector $\boldsymbol{w}$ must be set to zero. More importantly, the structure of these conditions can be used to "explain" why a specific weight has been set to zero. In particular, it has been shown [12] that the pruning conditions depend on an estimate of the per-component signal-to-noise ratio (SNR). This interpretation allows for an empirical tuning of the pruning conditions and, as a result, of the estimated signal sparsity.

The goal of this paper is to analyze the performance of IARD schemes as a function of this adjustment in two practically relevant scenarios: (i) when the measurement model behind $\boldsymbol{t}$ coincides with the estimation model (1), and thus the measurement $\boldsymbol{t}$ can be optimally represented with exactly K components of $\boldsymbol{\Phi}$, and (ii) when (1) is used as an approximation to the actual measurement model. In the latter case $\boldsymbol{t}$ is only "approximately" sparse in $\boldsymbol{\Phi}$; some elements of $\boldsymbol{w}$ thus have small, yet larger than zero magnitudes even in noise-free cases. Therefore, it makes sense to inquire if in the second scenario it is possible to trade the sparsity of the resulting model versus the approximation quality and how this trading is to be realized. In this work we demonstrate that the IARD scheme with adjusted pruning condition effectively realizes this trade-off. Specifically, we show that in the model mismatch case the tuning of the pruning conditions effectively trades the sparsity of the estimated signal versus the quality of the resulting models in terms of mean-square error (MSE). In contrast, in the first scenario, there is no trade-off between sparsity and residual MSE; on the contrary, an optimal performance is achieved by setting the adjustment level to the actual SNR. This effectively minimizes the residual MSE, while at the same time estimates the correct signal sparsity.

[3] An infinite value of the hyperparameter forces the posterior value of the component weight to zero.

The rest of the paper is organized as follows. In Section 2 we give an outline of the SBL signal model. In Section 3 we explain the incremental ARD solution to the SBL problem and discuss the adjustment of the pruning conditions. Simulation results that demonstrate the impact of the adjusted pruning conditions on the performance of IARD schemes are discussed in Section 4.

2. SBL SIGNAL MODEL

A standard solution to SBL with ARD is a two step procedure that alternates between estimating the weight vector $\boldsymbol{w}$ and estimating the corresponding sparsity parameters $\boldsymbol{\alpha}$ and noise precision τ. Given an estimate of the noise precision parameter $\widehat{\tau}$ and sparsity parameters $\widehat{\boldsymbol{\alpha}}$, an estimate of the weight vector $\widehat{\boldsymbol{w}}$ is obtained as a mode of the posterior pdf $p(\boldsymbol{w}|\boldsymbol{t}, \widehat{\boldsymbol{\alpha}}, \widehat{\tau}) = p(\boldsymbol{t}|\boldsymbol{w}, \widehat{\tau})p(\boldsymbol{w}|\widehat{\boldsymbol{\alpha}})$, which can be shown to be a Gaussian pdf with the mean $\widehat{\boldsymbol{w}}$ and covariance matrix $\widehat{\boldsymbol{S}}$ given by

$$\widehat{\boldsymbol{S}} = \left(\widehat{\tau}\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \widehat{\boldsymbol{A}}\right)^{-1} \quad \text{and} \quad \widehat{\boldsymbol{w}} = \widehat{\tau}\widehat{\boldsymbol{S}}\boldsymbol{\Phi}^T\boldsymbol{t}, \tag{2}$$

where $\widehat{\boldsymbol{A}} = \mathrm{diag}(\widehat{\boldsymbol{\alpha}})$ is a diagonal matrix with the elements of $\widehat{\boldsymbol{\alpha}}$ on the main diagonal. The estimates of $\boldsymbol{\alpha}$ and τ are computed by maximizing $p(\boldsymbol{\alpha}, \tau|\boldsymbol{t}) \propto (\boldsymbol{t}|\boldsymbol{\alpha}, \tau)p(\boldsymbol{\alpha})p(\tau)$, where

$$\begin{aligned} p(\boldsymbol{t}|\boldsymbol{\alpha}, \tau) &= \int p(\boldsymbol{t}|\boldsymbol{w}, \tau)p(\boldsymbol{w}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{w} \\ &= \mathrm{N}(\boldsymbol{t}|\boldsymbol{0}, \tau^{-1}\boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^T) \end{aligned} \tag{3}$$

is also known as a marginal likelihood function [7]. The maximizers $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\tau}$ of $p(\boldsymbol{\alpha}, \tau|\boldsymbol{t})$ can be obtained using the EM algorithm where the weights $\boldsymbol{w}$ are used as complete data (see [7] for more details). For the ARD case the corresponding estimation expressions are then given as

$$\widehat{\alpha}_l = (|\widehat{w}_l|^2 + \widehat{S}_{ll})^{-1} \tag{4}$$

$$\text{and} \quad \widehat{\tau} = \frac{N}{\|\boldsymbol{t} - \boldsymbol{\Phi}\widehat{\boldsymbol{w}}\|^2 + \mathrm{Trace}(\widehat{\boldsymbol{S}}\boldsymbol{\Phi}^T\boldsymbol{\Phi})}, \tag{5}$$

where $\widehat{w}_l$ is the lth element of the vector $\widehat{\boldsymbol{w}}$, and $\widehat{S}_{ll}$ is the lth element on the main diagonal of the matrix $\widehat{\boldsymbol{S}}$. The update expressions (2) and (4)–(5) are then repeatedly evaluated until convergence [7].

In cases when $\boldsymbol{t}$ allows for a sparse representation in terms of $\boldsymbol{\Phi}$, the sparsity parameters $\widehat{\boldsymbol{\alpha}}$ of some of the components will diverge, forcing the posterior value of the corresponding weights in (2) to converge toward zero and, thus, encouraging a sparse solution.

3. INCREMENTAL ARD SCHEME

Unfortunately, due to the EM-based maximization of (3) the rate at which sparsity parameters diverge is very low;

many iterations are needed for the hyperparameters to reach a threshold at which they can be treated as "numerically" infinite.[4] This has motivated the use of alternative, more efficient schemes to maximize (3) with respect to $\boldsymbol{\alpha}$ [10–12, 17].

A more efficient algorithm can be obtained for a flat hyperprior $p(\boldsymbol{\alpha})$.[5] Surprisingly, the optimum of $p(\boldsymbol{t}|\boldsymbol{\alpha}, \tau)$ with respect to a single sparsity parameter α_l, assuming that the other sparsity parameters $\widehat{\boldsymbol{\alpha}}_{\overline{l}} = [\widehat{\alpha}_1, \ldots, \widehat{\alpha}_{l-1}, \widehat{\alpha}_{l+1}, \ldots, \widehat{\alpha}_L]^T$ and the noise precision parameter $\widehat{\tau}$ are fixed, can be computed in closed form. Specifically, the logarithm of $p(\boldsymbol{t}|\alpha_l, \widehat{\boldsymbol{\alpha}}_{\overline{l}}, \widehat{\tau})$ can be decomposed as [10]

$$\log p(\boldsymbol{t}|\alpha_l, \widehat{\boldsymbol{\alpha}}_{\overline{l}}, \widehat{\tau}) = \mathcal{L}(\alpha_l; \widehat{\boldsymbol{\alpha}}_{\overline{l}}, \widehat{\tau}) = \mathcal{L}(\widehat{\boldsymbol{\alpha}}_{\overline{l}}, \widehat{\tau}) + \frac{1}{2}\left(\log(\alpha_l) - \log(\alpha_l - s_l) + \frac{q_l^2}{\alpha_l + s_l}\right), \quad (6)$$

where

$$s_l = \boldsymbol{\phi}_l^T \widehat{\boldsymbol{C}}_{\overline{l}}^{-1} \boldsymbol{\phi}_l, \quad q_l = \boldsymbol{\phi}_l^T \widehat{\boldsymbol{C}}_{\overline{l}}^{-1} \boldsymbol{t}, \quad (7)$$

$$\widehat{\boldsymbol{C}}_{\overline{l}} = \widehat{\tau}^{-1}\boldsymbol{I} + \sum_{k \neq l} \widehat{\alpha}_k^{-1} \boldsymbol{\phi}_k \boldsymbol{\phi}_k^T, \quad (8)$$

and $\mathcal{L}(\widehat{\boldsymbol{\alpha}}_{\overline{l}}, \widehat{\tau})$ is a part of the marginal log-likelihood that is independent of α_l. Then, the maximum of (6) with respect to α_l is obtained at [10, 12]

$$\widehat{\alpha}_l = \begin{cases} s_l^2(q_l^2 - s_l)^{-1}, & q_l^2 > s_l, \\ \infty, & q_l^2 \leq s_l. \end{cases} \quad (9)$$

Using (9) the marginal likelihood $p(\boldsymbol{t}|\alpha_l, \widehat{\boldsymbol{\alpha}}_{\overline{l}}, \widehat{\tau})$ can be maximized incrementally with respect to the sparsity parameter of one component at a time. Note that the result (9) not only tells if a particular element in $\widehat{\boldsymbol{w}}$ is zero, but also dramatically accelerates the rate of SBL convergence since the optimum of the marginal likelihood with respect to a single sparsity parameters can be computed analytically [10, 12, 17].

3.1. Pruning condition of IARD schemes

The key to sparsity inducing properties of the incremental ARD schemes lies in the pruning condition (9) that determines which elements in the vector $\boldsymbol{w}$ are set to zero.[6] Here we study these conditions in greater detail.

Let us define $\mu_l = q_l/s_l$ and $\varsigma_l = s_l^{-1}$. In [12] it has been demonstrated that μ_l and ς_l correspond, respectively, to the lth element of the vector $\widehat{\boldsymbol{w}}$ and the lth diagonal element of the matrix $\widehat{\boldsymbol{S}}$ in (2) computed when the sparsity parameter $\widehat{\alpha}_l$ is set to zero, i.e., when the basis vector $\boldsymbol{\phi}_l$ is not regularized[7].

[4] For instance, this threshold can be set to 10^{15} or 10^{16}.

[5] Using variational Bayesian analysis a similar scheme can be derived for non-informative hyperpriors $p(\alpha_l) \propto \alpha_l^{-1}$, $l = 1, \ldots, L$, [12].

[6] Strictly speaking, the order in which basis vectors are processed also plays a role, in particular when $\boldsymbol{\Phi}$ has a high mutual coherence.

[7] From (2) it follows that $\widehat{\boldsymbol{\alpha}}$ is a vector of regularization parameters for computing $\widehat{\boldsymbol{w}}$.

Naturally, for $\varsigma_l > 0$, the pruning condition $q_l^2 > s_l$ in (9) is equivalent to

$$\frac{\mu_l^2}{\varsigma_l} > 1, \quad (10)$$

where μ_l^2/ς_l can recognized as an estimate of the lth component SNR. In other words, a finite estimate of $\widehat{\alpha}_l$ is obtained provided an estimate of the lth component SNR, given by μ_l^2/ς_l, exceeds a 0dB threshold. This is a very simple and intuitive result: it demonstrates that the condition (10) determines if the quality of the lth component (as measured by its SNR) is above the level of the residual noise after the impact of the other $L-1$ components is taken into account. Quite naturally, the condition (10) can be adjusted (or tuned) by requiring that

$$\frac{\mu_l^2}{\varsigma_l} > \eta_l \quad (11)$$

where η_l is some predefined SNR level; in other words, only components with a certain SNR level $\eta_l \geq 1$ have to be retained in the model. Note that (11) is an empirical adjustment since for $\eta_l > 1$ the resulting estimate $\widehat{\boldsymbol{\alpha}}$ no longer maximizes $p(\boldsymbol{t}|\boldsymbol{\alpha}, \widehat{\tau})$. Nonetheless, this adjustment can be motivated from the perspective of statistical hypothesis testing (see [18] for more details).

The condition (11) is a key to understanding how sparsity of the resulting model can be traded for the model approximation quality. Further in the text we will term the IARD scheme that makes use of pruning condition (11) as η-IARD algorithm.

In what follows we investigate the impact of the adjustment parameter η_l on the performance of η-IARD algorithm.

4. EXPERIMENTAL ANALYSIS

In order to demonstrate the impact of the parameter η_l on the performance of η-IARD scheme, three experimental setups are considered. In the first setting we consider a standard compressive sampling toy problem, where the data $\boldsymbol{t}$ is generated using (1); in other words, there is no model mismatch. In the second setting we consider an application of η-IARD scheme to two non-parametric regression problems: (a) a sinc regressions problem as used in [7], and (b) a sparse regression of real data taken from UCI Machine Learning Repository. The last two examples are aimed to demonstrate the performance of the η-IARD scheme when there is a model mismatch and parameter η_l is used to trade model sparsity versus residual MSE.

4.1. Compressive sampling toy problem

In this experiment we study the impact of η_l using a compressive sampling toy problem. Here the basis functions $\boldsymbol{\phi}_l \in \mathbb{R}^N$, $l = 1, \ldots, L$, are generated by drawing $N = 100$ samples from a Gaussian distribution with zero mean and variance 1. A sparse vector $\boldsymbol{w}$ is constructed by setting $K = 10$

elements of $\boldsymbol{w}$ to 1 at random locations. Note that in this setting each component has the same SNR; thus, we assume that $\eta_1 = \ldots = \eta_L = \eta$.[8] The target vector $\boldsymbol{t}$ is then generated according to (1).

We begin by computing the normalized mean-square error (NMSE) of the estimation schemes versus the SNR. For comparison purposes we also include the simulation results for a Relevance Vector Machine (RVM) [7] and reformulated ARD (R-ARD) [11]. η-IARD schemes are simulated with $\eta = 1, 3.16$ and 10, which corresponds to 0dB, 5dB, and 10dB adjustment SNR', respectively. For simplicity we will refer to the values of adjustment parameter η_l in dB-scale. The corresponding simulation results are summarized in Fig. 1. Note

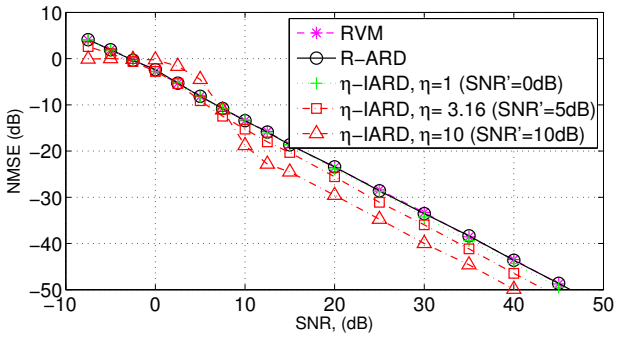

Fig. 1. Normalized mean-square error (NMSE) versus SNR.

that RVM, R-ARD and η-IARD with $\eta = 1$ perform comparably in terms of NMSE. Yet increasing the value of η improves the algorithm performance. The reason for that is an underestimated model sparsity due to the presence of noise. In fact, many of the estimated components have very small, yet non-zero weights (see Fig.2). The introduction of the adjustment η effectively prunes the noisy components, thus improving the sparsity estimation and, as the result, the NMSE.

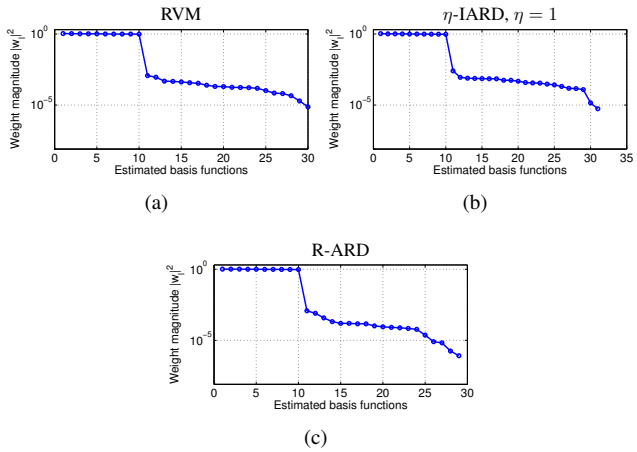

Fig. 2. The magnitude of the non-zero component weights for SNR fixed at 30 dB. a) RVM algorithm; b) η-IARD for $\eta = 1$ algorithm; c) R-ARD algorithm.

Let us also point to a distinctive threshold-like behavior of the η-IARD schemes: their performance in terms of NMSE is largely insensitive to the level of the noise when η is set above the actual SNR. However, when the actual SNR level approaches the selected value of the adjustment η, the NMSE improves rapidly and then continues to drop linearly[9] with the increasing SNR level.

Let, us now consider the performance of the η-IARD scheme in terms of NMSE value and estimated sparsity as a function of the adjustment level η. For that we fix the signal SNR at 10dB and consider the η-IARD algorithm with different settings of the adjustment parameter η. The corresponding results, averaged over 10000 Monte Carlo runs, are summarized in Fig. 3. Observe that for small η the number of

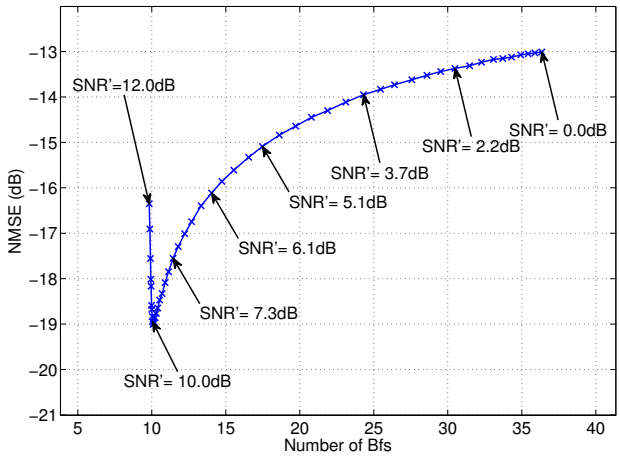

Fig. 3. NMSE and the estimated number of non-zero components as a function of the adjustment η for true SNR fixed at 10dB. SNR' corresponds to the adjustment η in dB-scale.

estimated components is high. As η grows, the NMSE value as well as the number of the estimated components drops. Note that the performance of the RVM and R-ARD schemes (not shown in Fig.3) is very similar to the performance of the η-IARD scheme with $\eta = 1$, i.e., the sparsity is significantly underestimated. In case of the η-IARD scheme, setting the value of η to the actual SNR of 10dB leads to the minimum of the NMSE; also, the estimated number of components coincides with the true signal sparsity. Obviously, there is a value of the parameter η that optimizes the performance of the η-IARD scheme. Specifically, in noisy scenarios the true sparsity can be recovered provided the value of η is selected to correspond to the actual signal SNR; at the same time the residual NMSE is minimized. This observation is also supported by our previous investigations [12, 17].

[8] In further experiments we will also make use of this assumption.

[9] Over a logarithmic scale.

4.2. Non-parameteric sinc regression problem

In this experiment we consider a reconstruction of a noise-perturbed $\mathrm{sinc}(x) = \sin(x)/x$ function based on $N = 100$ randomly-spaced samples in the interval $[-10, 10]$. The measurements $\boldsymbol{t}$ are generated by evaluating $\mathrm{sinc}(x)$ at sampled locations and perturbing them using a zero-mean normally distributed white noise with variance σ^2. The variance of the noise is chosen such as to ensure a desired SNR value. Reconstruction is implemented using a set of $L = 100$ Gaussian kernels $\boldsymbol{\phi}_l$, $l = 1, \ldots, L$ centered at measured samples and having a fixed variance of 2.3. Note that the measurement $\boldsymbol{t}$ generated with the sinc model is only "approximately" sparse in $\boldsymbol{\Phi}$. The noise precision τ has been automatically estimated for each run of the algorithm. This example demonstrates the performance of the η-IARD scheme in this case.

Due to space limitations we will not evaluate the NMSE performance as a function of SNR for the following examples. Instead, we go directly to the dependency of NMSE and estimated sparsity on the value of the adjustment parameter η for the SNR level fixed at 10dB. The corresponding results, averaged over 10000 Monte Carlo runs, are summarized in Fig. 4. Note that in contrast to the previous scenario,

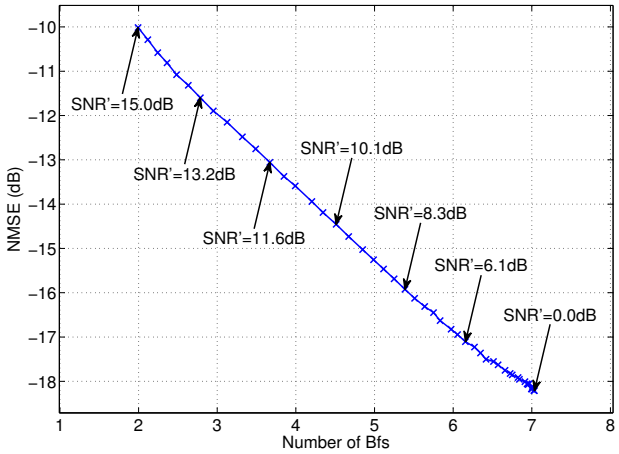

Fig. 4. NMSE and the estimated number of non-zero components as a function of the adjustment η for sinc regression problem.

the resulting curve does not have a distinct optimum. Instead, higher values of η increase the sparsity of the model at the expense of higher NMSE; in other words, the adjustment parameter η monotonically trades the approximation quality of the learned model versus estimated signal sparsity.

Let us stress that in practice the model (11) is often used as a mere approximation to the true measurement model. As a result the sparsity is not exact. The result (11) thus provides a mechanism to adjust the IARD scheme so as to account for the model mismatches. More specifically, the parameter η specifies the minimum quality of the component, as measured by its SNR, that is to be retained in the model. Quite naturally, a specific value of η is typically problem dependent.

4.3. Concrete Compressive Strength data

In this example we evaluate the algorithm performance on a real data set – the Concrete Compressive Strength (CCS) data from the UCI Machine Learning Repository. It is a multivariate regression data set with 8 attributes and 1030 instances. To evaluate the model performance a 20-fold cross-validation has been used. Both training and test data outcomes are normalized to have zero mean and unit variance. The dictionary matrix $\boldsymbol{\Phi}$ consists of Gaussian kernels $\boldsymbol{\phi}_l$ centered at the measurement samples. The variance of the kernels is fixed at 4.3. The noise precision τ has been automatically estimated for each run of the algorithm. In Fig. 5 we plot the dependency of NMSE and estimated signal sparsity on the value of η. Observe that the behavior of the NMSE-sparsity curve as a

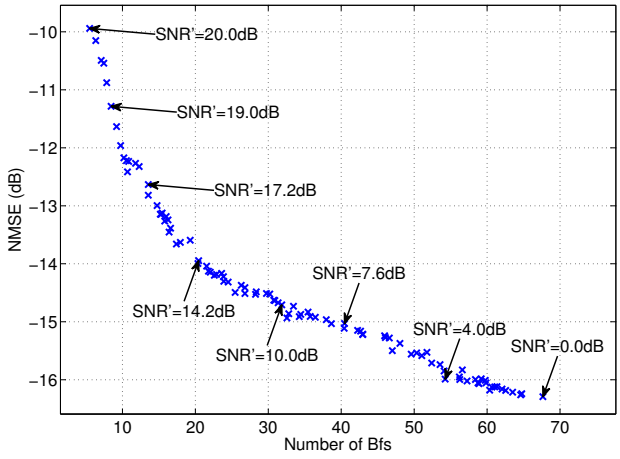

Fig. 5. NMSE and the estimated number of non-zero components as a function of the adjustment η for Concrete Compressive Strength data.

function of η is very similar to the sinc-regression case. Indeed, due to the model mismatch, the parameter η can be used to trade the sparsity of the resulting model versus the value of the residual NMSE.

5. CONCLUSION

In this paper the performance of the adjusted incremental automatic relevance determination (IARD) algorithm has been analyzed. The considered adjustment allows for a trade-off between the sparsity and the approximation quality of the learned models. The latter is realized by modifying the pruning condition of the IARD scheme. The modification has a simple interpretation in terms of the per-component signal-to-noise ratio (SNR).

The impact of the modification has been studied in two scenarios. In the first scenario the estimation model coincides with the measurement model; in the second one the estimation model is used as an approximation to the true model behind the measurement data. It has been determined that in the first case, the performance of the IARD scheme has a distinct optimum, which was achieved by setting the adjustment level to the true signal SNR. In this way the mean-square error (MSE) is minimized and the true signal sparsity is recovered. In the second case, the MSE is minimized only at the expense of less sparse models; fewer nonzero components inevitably leads to a higher MSE. In this case the adjustment of the pruning condition can be seen as a trade-off between signal sparsity and the achieved MSE.

6. REFERENCES

[1] R. Baraniuk, "Compressive sensing," *IEEE Signal Processing Magazine*, vol. 24, no. July, pp. 118–121, 4 2007.

[2] M. Duarte and Y. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Trans. on Sig. Proc.*, vol. 59, no. 9, pp. 4053 –4085, sept. 2011.

[3] D. Donoho, "For most large underdetermined systems of linear equations, the minimal ell-1 norm solution is also the sparsest solution," *Comm. on Pure and Appl. Math.*, vol. 59, no. 6, pp. 797–829, June 2006.

[4] M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[5] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[6] E. J. Cands, J. K. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communic. on Pure and Applied Math.*, vol. 59, no. 8, pp. 1207–1223, 2006.

[7] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Machine Learning Res.*, vol. 1, pp. 211–244, June 2001.

[8] M. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. on Pattern Analysis and Machine Intel.*, vol. 25, no. 9, pp. 1150–1159, 2003.

[9] N. L. Pedersen, D. Shutin, C. N. Manchon, and B. H. Fleury, "Sparse estimation using Bayesian hierarchical prior modeling for real and complex models," *submitted to IEEE Trans. on Sig. Proc.*, 2011. [Online]. Available: arXiv:1108.4324v1

[10] M. E. Tipping and A. C. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *Proc. 9th International Workshop on Artificial Intelligence and Statistics*, Key West, FL, January 2003.

[11] D. Wipf and S. Nagarajan, "A new view of automatic relevance determination," in *Proc. 21 Annual Conf. on Neural Inform. Process. Systems*. Vancouver, British Columbia, Canada: MIT Press, Dec. 2007.

[12] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor, "Fast variational sparse Bayesian learning with automatic relevance determination for superimposed signals," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6257 –6261, Dec. 2011.

[13] J. O. Berger, J. M. Bernardo, and D. Sun, "The formal definition of reference priors," *Ann. Statist.*, vol. 37, no. 2, pp. 905–938, 2009.

[14] E. J. Candes, M. B. Wakin, and S. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. of Fourier Analysis and Applic.*, vol. 14, no. 5, pp. 877–905, December 2008.

[15] E. Cands and J. Romberg, "Sparsity and incoherence in compressive sampling," *Inverse Prob.*, vol. 23, no. 3, pp. 969–985, 2007.

[16] M. J. Wainwright, "Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso)," *IEEE Trans. on Inform. Theory*, vol. 55, no. 5, pp. 2183–2202, May 2009.

[17] D. Shutin, S. R. Kulkarni, and H. V. Poor, "Incremental reformulated automatic relevance determination," *IEEE Trans. on Sig. Proc.*, 2012, to appear.

[18] ——, "A novel view of incremental sparse Bayesian learning with automatic relevance determination," submitted to *IEEE Transactions on Signal Processing*.