Linköping University Post Print

On the Optimal K-term Approximation of a Sparse Parameter Vector MMSE Estimate

Erik Axell, Erik G. Larsson and Jan-Åke Larsson

N.B.: When citing this work, cite the original article.

©2009 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Erik Axell, Erik G. Larsson and Jan-Åke Larsson, On the Optimal K-term Approximation of a Sparse Parameter Vector MMSE Estimate, 2009, Proceedings of the 2009 IEEE Workshop on Statistical Signal Processing (SSP'09), 245-248.
http://dx.doi.org/10.1109/SSP.2009.5278594

Postprint available at: Linköping University Electronic Press
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-25591

ON THE OPTIMAL K-TERM APPROXIMATION OF A SPARSE PARAMETER VECTOR MMSE ESTIMATE

Erik Axell, Erik G. Larsson and Jan-Åke Larsson

Department of Electrical Engineering (ISY), Linköping University, 581 83 Linköping, Sweden

ABSTRACT

This paper considers approximations of marginalization sums that arise in Bayesian inference problems. Optimal approximations of such marginalization sums, using a fixed number of terms, are analyzed for a simple model. The model under study is motivated by recent studies of linear regression problems with sparse parameter vectors, and of the problem of discriminating signal-plus-noise samples from noise-only samples. It is shown that for the model under study, if only one term is retained in the marginalization sum, then this term should be the one with the largest a posteriori probability. By contrast, if more than one (but not all) terms are to be retained, then these should generally *not* be the ones corresponding to the components with largest a posteriori probabilities.

***Index Terms*—** MMSE estimation, Bayesian inference, marginalization

1. INTRODUCTION

Bayesian mixture models for data observations are common in a variety of problems. One example of special interest for us is linear regression, where it is known a priori that a specific coefficient is constrained to zero with a given probability [1]. A similar model also occurs for the problem of discriminating samples that contain a signal embedded in noise from samples that contain only noise. The latter problem, for the case when the noise statistics are partially unknown, was dealt with in [2] and it has applications for example in spectrum sensing for cognitive radio [3, 4] and signal denoising [5].

Generally, optimal statistical inference in a Bayesian mixture model consists of marginalizing a probability density function. The marginalization sum typically has a huge number of terms, and approximations are often needed. There are various ways of approximating this marginalization sum. One approach is to find and retain only the dominant term. Another way is to keep only a specific subset of the terms. Then the question arises, how to choose this subset of terms. To understand the basic principles behind this question we formulate a simple toy example, which is essentially a degenerated case of the sparse linear regression problem of [1]. We show that if only one term in the marginalization sum is to be retained, then one should keep the one corresponding to the mixture component with the largest a posteriori probability. By contrast, we show that if more than one (but not all) terms are to be retained, then these are generally *not* the ones corresponding to the mixture components with the largest a posteriori probabilities. This observation is interesting, and it is different from what one might expect intuitively: if using K terms, the K terms that correspond to the most likely models (given the data) are not necessarily the ones that provide the best K-term approximation of the marginalization sum. Our findings are also verified numerically.

The research leading to these results has received funding from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement no. 216076. This work was also supported in part by the Swedish Research Council (VR) and the Swedish Foundation for Strategic Research (SSF). E. Larsson is a Royal Swedish Academy of Sciences (KVA) Research Fellow supported by a grant from the Knut and Alice Wallenberg Foundation.

2. MODEL

Throughout the paper we consider the model

$$\mathbf{y} = \mathbf{h} + \mathbf{e}, \tag{1}$$

where $\mathbf{y}$ is an observation vector, $\mathbf{h}$ is a parameter vector, and $\mathbf{e}$ is a vector of noise. All vectors are of length n. We assume that $\mathbf{h}$ has a sparse structure. Specifically, we assume that the elements h_m, $m = 1, 2, \ldots, n$, are independent and that

$$\begin{cases} h_m = 0, \text{ with probability } p, \\ h_m \sim N(0, \gamma^2), \text{ with probability } 1-p. \end{cases}$$

Furthermore, we assume that $\mathbf{e} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$. The task is to estimate $\mathbf{h}$ given that $\mathbf{y}$ is observed. This can be seen as a degenerated instance of linear regression with a sparse parameter vector [1] (with the identity matrix as regressor). It is also related to the problem of discriminating samples that contain a signal embedded in noise from samples that contain only noise. The latter problem was dealt with in [2] (for the case of unknown σ, γ).

We define the following 2^n hypotheses:

$$\begin{cases} H_0: & h_1, \cdots, h_n \text{ i.i.d. } N(0,\gamma^2), \\ H_1: & h_1 = 0; h_2, \cdots, h_n \text{ i.i.d. } N(0,\gamma^2), \\ H_2: & h_2 = 0; h_1, h_3, \cdots, h_n \text{ i.i.d. } N(0,\gamma^2), \\ \vdots & \\ H_n: & h_n = 0; h_1, \cdots, h_{n-1} \text{ i.i.d. } N(0,\gamma^2), \\ H_{n+1}: & h_1 = h_2 = 0; h_3, \cdots, h_n \text{ i.i.d. } N(0,\gamma^2), \\ \vdots & \\ H_{2^n-1}: & h_1 = \cdots = h_n = 0. \end{cases}$$

Since the elements of $\mathbf{h}$ are independent we obtain the following a priori probabilities:

$$\begin{cases} P(H_0) = (1-p)^n, \\ P(H_1) = P(H_2) = \cdots = P(H_n) = p(1-p)^{n-1}, \\ \vdots \\ P(H_{2^n-1}) = p^n. \end{cases}$$

Authorized licensed use limited to: Linkoping Universitetsbibliotek. Downloaded on February 10, 2010 at 05:18 from IEEE Xplore. Restrictions apply.

For each hypothesis, H_i, let S_i be the set of indices, m, for which $h_m \sim N(0, \gamma^2)$ under H_i:

$$\begin{cases} S_0 = \{1, 2, \cdots, n\},\ S_1 = \{2, 3, \cdots, n\},\ \cdots, \\ S_n = \{1, 2, \cdots, n-1\}, S_{n+1} = \{3, 4, \cdots, n\},\ \cdots, \\ S_{2^n-1} = \emptyset. \end{cases}$$

Let Ω_m denote the event that $h_m \sim N(0, \gamma^2)$. Then $\neg\Omega_m$ is the event that $h_m = 0$. The probability of Ω_m, given the observation $\mathbf{y}$, can be written

$$P(\Omega_m|\mathbf{y}) = \sum_{k:m\in S_k} P(H_k|\mathbf{y}).$$

In the model used here, the coefficients of $\mathbf{y}$ are independent. Thus, the probability of the mth component depends only on the mth component of $\mathbf{y}$, that is $P(\Omega_m|\mathbf{y}) = P(\Omega_m|y_m)$.

3. MMSE ESTIMATION AND APPROXIMATION

Assume that $\widehat{\mathbf{h}}$ is an estimate of $\mathbf{h}$, given the observation $\mathbf{y}$. It is well known that the minimum mean-square error (MMSE) estimate is given by the conditional mean:

$$\begin{aligned} \mathbf{h}_{MMSE} &= \underset{\widehat{\mathbf{h}}}{\operatorname{argmin}}\ E\left[\left\|\mathbf{h} - \widehat{\mathbf{h}}\right\|^2\right] \\ &= \underset{\widehat{\mathbf{h}}}{\operatorname{argmin}} \left\|E\left[\mathbf{h}|\mathbf{y}\right] - \widehat{\mathbf{h}}\right\|^2 = E\left[\mathbf{h}|\mathbf{y}\right]. \end{aligned}$$

If all the variables σ, γ and p are known, the MMSE estimate for each individual component h_m can be derived in closed form:

$$\begin{aligned} &E[h_m|\mathbf{y}] = E[h_m|y_m] \\ &= E[h_m|y_m, \Omega_m]P(\Omega_m|y_m) + E[h_m|y_m, \neg\Omega_m]P(\neg\Omega_m|y_m) \\ &= E[h_m|y_m, \Omega_m]\frac{p(y_m|\Omega_m)P(\Omega_m)}{p(y_m)} + \\ &\qquad E[h_m|y_m, \neg\Omega_m]\frac{p(y_m|\neg\Omega_m)P(\neg\Omega_m)}{p(y_m)} \\ &= E[h_m|y_m, \Omega_m]\frac{p(y_m|\Omega_m)P(\Omega_m)}{p(y_m|\Omega_m)P(\Omega_m) + p(y_m|\neg\Omega_m)P(\neg\Omega_m)}. \end{aligned}$$

The first equality follows when σ and γ are known because y_m are independent by assumption, and the last equality follows because $E[h_m|y_m, \neg\Omega_m] = 0$. The expectation and probabilities are given by:

$$\begin{cases} E[h_m|y_m, \Omega_m] = \frac{\gamma^2}{\gamma^2+\sigma^2} y_m \\ y_m|\Omega_m \sim N(0, \gamma^2 + \sigma^2),\ y_m|\neg\Omega_m \sim N(0, \sigma^2) \\ P(\Omega_m) = 1 - p,\ P(\neg\Omega_m) = p. \end{cases}$$

Inserting this and simplifying, we obtain

$$E[h_m|\mathbf{y}] = \frac{\frac{\gamma^2}{\gamma^2+\sigma^2} y_m}{1 + \frac{p}{1-p}\sqrt{\frac{\gamma^2+\sigma^2}{\sigma^2}} \exp\left(\frac{\gamma^2}{2\sigma^2\left(\gamma^2+\sigma^2\right)} y_m^2\right)}$$

In many (more realistic) versions of the problem the MMSE estimate cannot be obtained in closed form. For example, if σ and γ are unknown the equality $E[h_m|\mathbf{y}] = E[h_m|y_m]$ is no longer valid [2]. We must then compute $\mathbf{h}_{MMSE}$ via marginalization:

$$\mathbf{h}_{MMSE} = E[\mathbf{h}|\mathbf{y}] = \sum_{i=0}^{2^n-1} P(H_i|\mathbf{y})E[\mathbf{h}|\mathbf{y}, H_i]. \tag{2}$$

For our problem, the ingredients of (2) are straightforward to compute. For each hypothesis H_i, denote by $\mathbf{\Lambda}_i$ an $n\times n$ diagonal matrix where the mth diagonal element $\mathbf{\Lambda}_i(m,m) = 0$ if h_m is constrained to zero under hypothesis H_i and $\mathbf{\Lambda}_i(m,m) = 1$ if $h_m \sim N(0, \gamma^2)$ under H_i. That is:

$$\mathbf{\Lambda}_i(m,m) = \mathcal{I}_i(m) \triangleq \begin{cases} 1, \text{ if } h_m \sim N(0, \gamma^2) \text{ under } H_i, \\ 0, \text{ if } h_m = 0 \text{ under } H_i. \end{cases}$$

Then (see [1])

$$\mathbf{y}|H_i \sim N(\mathbf{0}, \gamma^2\mathbf{\Lambda}_i + \sigma^2\mathbf{I}),$$

and

$$E[\mathbf{h}|\mathbf{y}, H_i] = \frac{\gamma^2}{\gamma^2 + \sigma^2}\mathbf{\Lambda}_i\mathbf{y}.$$

The difficulty with (2) is that the sum has 2^n terms and for large n this computation will be very burdensome. Hence, one must generally approximate the sum, for example by including only a subset $\mathcal{H}$ of all possible hypotheses $H_0, \ldots, H_{2^n-1}$. The MMSE estimate is then approximated by

$$\widehat{\mathbf{h}}_{MMSE} \approx \frac{\sum_{i\in\mathcal{H}} p(\mathbf{y}|H_i)P(H_i)E[\mathbf{h}|\mathbf{y}, H_i]}{\sum_{j\in\mathcal{H}} p(\mathbf{y}|H_j)P(H_j)} = E[\mathbf{h}|\mathbf{y}, \vee_{\mathcal{H}} H_i]. \tag{3}$$

Our goal in what follows is to understand what subset $\mathcal{H}$ that minimizes the MSE of the resulting approximate MMSE estimate. Specifically, the task is to choose the subset $\mathcal{H}$ that minimizes the MSE, given that the sum is approximated by a fixed number of terms, that is, subject to $|\mathcal{H}| = K$:

$$\mathcal{H}_{MMSE} = \underset{\mathcal{H}:|\mathcal{H}|=K}{\operatorname{argmin}} \left\|E[\mathbf{h}|\mathbf{y}] - E[\mathbf{h}|\mathbf{y}, \vee_{\mathcal{H}} H_i]\right\|^2. \tag{4}$$

As a baseline, we describe an algorithm for choosing the subset $\mathcal{H}$ that was proposed in [1]. Intuitively, using hypotheses which are a posteriori most likely, should give a good approximation. The idea of the selection algorithm in [1] is to consider the hypotheses H_i, for which the a posteriori probability $P(H_i|\mathbf{y})$ is significant. The set $\mathcal{H}$ of hypotheses for which $P(H_i|\mathbf{y})$ is significant is chosen as follows:

1. Start with a set $\mathcal{B} = \{1, 2, \cdots, n\}$ and a hypothesis H_i (H_0 or H_{2^n-1} are natural choices).
2. Compute the contribution to (3), $P(H_i|\mathbf{y})$.
3. Evaluate $P(H_k|\mathbf{y})$ for all H_k obtained from H_i by changing the state of one parameter h_j, $j \in \mathcal{B}$. That is, if $h_j \sim N(0, \gamma^2)$ in H_i, then $h_j = 0$ in H_k and vice versa. Choose the j which yields the largest $P(H_k|\mathbf{y})$. Set $i := k$ and remove j from $\mathcal{B}$.
4. If $\mathcal{B} = \emptyset$ (this will happen after n iterations), compute the contribution of the last H_i to (3) and then terminate. Otherwise, go to Step 2.

This algorithm will change the state of each parameter once, and choose the largest term from each level. The set $\mathcal{H}$ will finally contain $n + 1$ hypotheses instead of 2^n.

As a comparison to this approximation, the numerical results will show a scheme that (brute force) finds the K hypotheses with

Authorized licensed use limited to: Linkoping Universitetsbibliotek. Downloaded on February 10, 2010 at 05:18 from IEEE Xplore. Restrictions apply.

maximum a posteriori probability, i.e., $\mathcal{H} = \mathcal{H}_K$ where

$$\mathcal{H}_k = \mathcal{H}_{k-1} \cup \left\{ \underset{H_i \notin \mathcal{H}_{k-1}}{\operatorname{argmax}} \; P(H_i|y) \right\}; \quad \mathcal{H}_0 = \emptyset. \tag{5}$$

4. OPTIMAL ONE-TERM APPROXIMATION, $K = 1$

If we take $K = 1$, then the set $\mathcal{H}$ contains only one element and the approximate estimate contains only one term. Hence, in this case the estimate (3) is just the conditional mean of one hypothesis H_i:

$$\widehat{\mathbf{h}}_{MMSE} = E[\mathbf{h}|\mathbf{y}, H_i]$$

Now the problem is to choose the hypothesis H_i that minimizes (4). Thus, the MMSE optimal estimate, from the MMSE perspective, is derived from the following minimization:

$$\begin{aligned} &\min_i \|E[\mathbf{h}|\mathbf{y}] - E[\mathbf{h}|\mathbf{y}, H_i]\|^2 \\ &= \min_i \left\| \sum_{k=0}^{2^n-1} P(H_k|\mathbf{y}) E[\mathbf{h}|\mathbf{y}, H_k] - E[\mathbf{h}|\mathbf{y}, H_i] \right\|^2 \\ &= \min_i \left\| \sum_{k=0}^{2^n-1} P(H_k|\mathbf{y}) \frac{\gamma^2}{\gamma^2+\sigma^2} \mathbf{\Lambda}_k \mathbf{y} - \frac{\gamma^2}{\gamma^2+\sigma^2} \mathbf{\Lambda}_i \mathbf{y} \right\|^2 . \end{aligned} \tag{6}$$

Let

$$\mathbf{\Lambda} \triangleq \sum_{k=0}^{2^n-1} P(H_k|\mathbf{y}) \mathbf{\Lambda}_k.$$

Then, $\mathbf{\Lambda}$ is also diagonal and the mth diagonal element $\mathbf{\Lambda}(m,m)$ is

$$\mathbf{\Lambda}(m,m) = \sum_{k: m \in S_k} P(H_k|\mathbf{y}) = P(\Omega_m|\mathbf{y}).$$

Hence, equation (6) can be written

$$\begin{aligned} &\min_i \left\| (\mathbf{\Lambda} - \mathbf{\Lambda}_i) \frac{\gamma^2}{\gamma^2+\sigma^2} \mathbf{y} \right\|^2 \\ &= \min_i \left(\frac{\gamma^2}{\gamma^2+\sigma^2} \right)^2 \sum_{m=1}^{n} \left| (P(\Omega_m|\mathbf{y}) - \mathcal{I}_i(m)) y_m \right|^2 . \end{aligned}$$

Since we minimize over all hypotheses, each element $\mathcal{I}_i(m)$ can be chosen to be zero or one independently of the other diagonal elements. As stated previously, the coefficients of $\mathbf{y}$ are independent by assumption and $P(\Omega_m|\mathbf{y}) = P(\Omega_m|y_m)$. Hence, each component of the sum is independent of all other components, and the minimization can be done componentwise. Thus, it is equivalent to solve

$$\min_i \left| (P(\Omega_m|y_m) - \mathcal{I}_i(m)) \right|^2 , \tag{7}$$

for all components $m = 1, 2, \ldots, n$. This is minimized when each coefficient $\mathcal{I}_i(m)$ is chosen as

$$\mathcal{I}_i(m) = \begin{cases} 0, \text{ if } P(\Omega_m|y_m) < \frac{1}{2} \\ 1, \text{ if } P(\Omega_m|y_m) \geq \frac{1}{2}, \end{cases}$$

or equivalently

$$\mathcal{I}_i(m) = \begin{cases} 0, \text{ if } P(\Omega_m|y_m) < P(\neg\Omega_m|y_m) \\ 1, \text{ if } P(\Omega_m|y_m) \geq P(\neg\Omega_m|y_m), \end{cases}$$

As a result of this, the optimal hypothesis H_i should be chosen to maximize the a posteriori probability for each coefficient. Since the coefficients of $\mathbf{h}$ are independent, the a posteriori probability of hypothesis H_i is the product of the marginal probabilities, i.e.

$$P(H_i|\mathbf{y}) = \prod_{m=1}^{n} P(H_i(m)|\mathbf{y})$$

Maximizing the a posteriori probability of each coefficient, also maximizes the a posteriori probability of the hypothesis. Thus, the optimal hypothesis should be chosen to maximize the a posteriori probability.

The same result is valid also if $\mathbf{y} = \mathbf{D}\mathbf{h} + \mathbf{e}$, where $\mathbf{D}$ is a diagonal matrix, $\mathbf{D} = \operatorname{diag}(d_1, d_2, \cdots, d_n)$. Then we can rewrite (1) as

$$\mathbf{y} = \mathbf{D}\mathbf{h} + \mathbf{e} \Leftrightarrow \mathbf{y} = \tilde{\mathbf{h}} + \mathbf{e},$$

where $\tilde{\mathbf{h}} \triangleq \mathbf{D}\mathbf{h}$. This problem is equivalent to the case where $\mathbf{y} = \mathbf{h} + \mathbf{e}$, except that the variances are not equal for all coefficients of $\tilde{\mathbf{h}}$. That is, the coefficients $\tilde{h}_m$ is either zero or $\tilde{h}_m \sim N(0, d_m^2\gamma^2)$. If we redefine Ω_m to be the event that $\tilde{h}_m \sim N(0, d_m^2\gamma^2)$, and use the same arguments as for $\mathbf{y} = \mathbf{h} + \mathbf{e}$, we will end up in (7). The result follows.

5. OPTIMAL APPROXIMATION WITH MULTIPLE TERMS, $K > 1$

When $K > 1$ the optimal approximation of (3) is

$$\begin{aligned} &\min_{\mathcal{H}} \|E[\mathbf{h}|\mathbf{y}] - E[\mathbf{h}|\mathbf{y}, \vee_{\mathcal{H}} H_i]\|^2 = \\ &\min_{\mathcal{H}} \left\| \frac{\gamma^2}{\gamma^2+\sigma^2} \mathbf{\Lambda}\mathbf{y} - \frac{\sum_{i \in \mathcal{H}} P(H_i|\mathbf{y}) \frac{\gamma^2}{\gamma^2+\sigma^2} \mathbf{\Lambda}_i \mathbf{y}}{\sum_{j \in \mathcal{H}} P(H_j|\mathbf{y})} \right\|^2 . \end{aligned} \tag{8}$$

Define

$$\mathbf{\Lambda}_{\mathcal{H}} \triangleq \frac{\sum_{i \in \mathcal{H}} P(H_i|\mathbf{y}) \mathbf{\Lambda}_i}{\sum_{j \in \mathcal{H}} P(H_j|\mathbf{y})}.$$

Then $\mathbf{\Lambda}_{\mathcal{H}}$ is a diagonal matrix and each diagonal element can be written

$$\mathbf{\Lambda}_{\mathcal{H}}(m,m) = \frac{\sum_{i \in \mathcal{H}: m \in S_i} P(H_i|\mathbf{y})}{\sum_{j \in \mathcal{H}} P(H_j|\mathbf{y})} = P(\Omega_m|\mathbf{y}, \vee_{\mathcal{H}} H_i). \tag{9}$$

Now (8) can be written as

$$\begin{aligned} &\min_{\mathcal{H}} \left\| (\mathbf{\Lambda} - \mathbf{\Lambda}_{\mathcal{H}}) \frac{\gamma^2}{\gamma^2+\sigma^2} \mathbf{y} \right\|^2 \\ &= \min_{\mathcal{H}} \left(\frac{\gamma^2}{\gamma^2+\sigma^2} \right)^2 \sum_{m=1}^{n} \left| (P(\Omega_m|\mathbf{y}) - P(\Omega_m|\mathbf{y}, \vee_{\mathcal{H}} H_i)) y_m \right|^2 . \end{aligned} \tag{10}$$

The expression for the probability $P(\Omega_m|\mathbf{y}, \vee_{\mathcal{H}} H_i)$ in (9) contains sums of probabilities over a *subset* $\mathcal{H}$ of the hypotheses, but not all. The denominator is simply the total probability of all hypotheses in the subset $\mathcal{H}$. Thus, the probability of each individual hypothesis affects the probability $P(\Omega_m|\mathbf{y}, \vee_{\mathcal{H}} H_i)$ for all m. This implies that the probability of the mth component depends also on other components of $\mathbf{y}$. Thus, the probabilities $P(\Omega_m|\mathbf{y}, \vee_{i \in \mathcal{H}} H_i)$ cannot be chosen independently for all m. Hence, unless $P(\Omega_m|\mathbf{y}, \vee_{\mathcal{H}} H_i)$ is equal for all m, the minimization *cannot* be done componentwise in this case. Furthermore, if the probabilities are not equal, the minimization also depends on $|y_m|^2$. Hence, it is *not* necessarily op-

Authorized licensed use limited to: Linkoping Universitetsbibliotek. Downloaded on February 10, 2010 at 05:18 from IEEE Xplore. Restrictions apply.

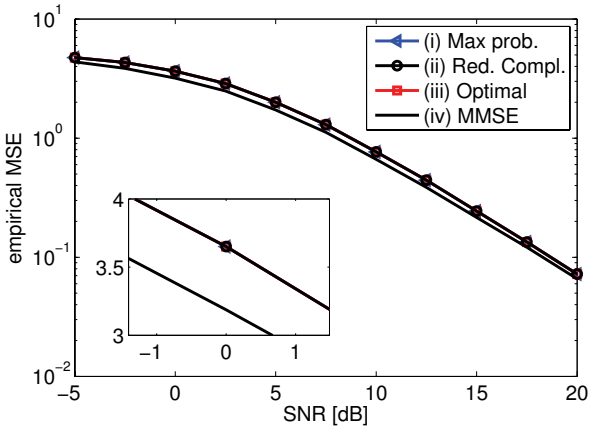

Fig. 1. Empirical MSE for different strategies. $K = 1$, $n = 10$.

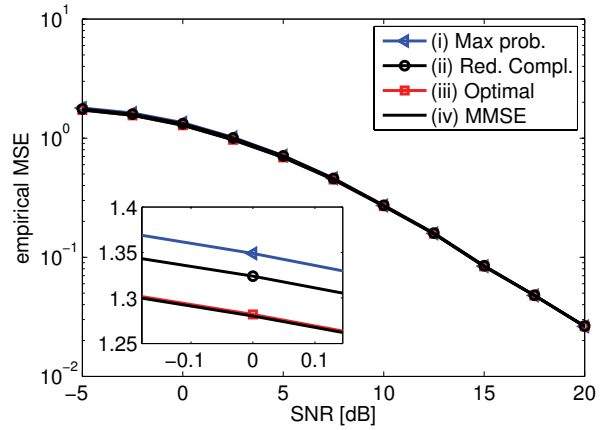

Fig. 2. Empirical MSE for different strategies. $K = 5$, $n = 4$.

timal to approximate the estimate using the K hypotheses with the largest a posteriori probability, not even for the toy example considered here.

6. NUMERICAL RESULTS

We verify our results by Monte-Carlo simulations. Performance is given as the empirical MSE as a function of SNR $\triangleq \gamma^2/\sigma^2$. In all simulations the true parameter values were $\gamma^2 = 1$ and $p = 0.5$. The noise variance σ^2 was varied from 5 dB down to -20 dB. We compare the following schemes:

(i) Maximum-probability approximation, see (5).

(ii) Reduced-complexity approximation of [1], see the end of Section 3.

(iii) Optimal K-term approximation, see (4).

(iv) Full MMSE, see (2).

Example 1: One-term approximation, $K = 1$ (Figure 1). In this example we verify the results from Section 4, for $n = 10$. That is, only one out of 2^{10} hypotheses is used in the approximations. Note that the reduced-complexity approximation algorithm originally selects $n + 1$ terms. However, in this case only the one term with the largest a posteriori probability is used in the approximation. As expected, the full MMSE scheme outperforms the approximate schemes. We also note that all three approximation schemes actually have the exact same performance. Especially, the optimal and the maximum probability schemes have equal performance, which verifies the results from Section 4. The reduced complexity approximation also performs the same, which shows that the selection algorithm always finds the term with globally maximum a posteriori probability. This is an effect of the independence of $\mathbf{y}$. Because of the independence, the hypothesis with maximum a posteriori probability can be found by maximizing the probability for one component at a time. This is exactly what the reduced complexity selection algorithm does.

Example 2: Multiple-terms approximation $K = 5$ (Figure 2). In this example, 5 hypotheses out of $2^4 = 16$ are used in the approximations. We note in this case that the optimal approximation outperforms the maximum-probability approximation, which verifies the results from Section 5. Notable is also that the reduced-complexity approximation outperforms the maximum-probability approximation. The idea of the selection algorithm is to find the hypotheses with large a posteriori probability, but it does not find the ones with largest probability. It is a sub-optimal algorithm to find the hypotheses with largest probability, but it seems like the chosen subset is actually closer to the optimal solution.

7. CONCLUDING REMARKS

We have dealt with the problem of approximating a marginalization sum, under the constraint that only K terms are retained. The approximation was exemplified in the context of MMSE estimation. One could argue that intuitively, the terms which correspond to the model components with the largest a posteriori probabilities should be used. We have shown that for a special case of the problem, if only one term is to be retained in the marginalization sum, then one should keep the one corresponding to the mixture component with the largest a posteriori probability. By contrast, if more than one (but not all) terms are to be retained, then these are generally *not* the ones corresponding to the components with the largest a posteriori probabilities. This holds even for the case when the parameters are assumed to be independent, and the variances of the noise and of the parameter coefficients are known. It is an open problem to what extent the observations that we have made can be extrapolated to other, more general marginalization problems.

8. REFERENCES

[1] E. G. Larsson and Y. Selen, "Linear regression with a sparse parameter vector," *IEEE Transactions on Signal Processing*, vol. 55, no. 2, pp. 451–460, Feb. 2007.

[2] E. Axell and E. G. Larsson, "A Bayesian approach to spectrum sensing, denoising and anomaly detection," *Proc. of IEEE ICASSP*, 19-24 Apr. 2009, To appear.

[3] A. Ghasemi and E.S. Sousa, "Spectrum sensing in cognitive radio networks: requirements, challenges and design trade-offs," *IEEE Communications Magazine*, vol. 46, no. 4, pp. 32–39, April 2008.

[4] R. Tandra and A. Sahai, "SNR walls for signal detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 1, pp. 4–17, Feb. 2008.

[5] E. Gudmundson and P. Stoica, "On denoising via penalized least-squares rules," *Proc. of IEEE ICASSP*, pp. 3705–3708, March 2008.

Authorized licensed use limited to: Linkoping Universitetsbibliotek. Downloaded on February 10, 2010 at 05:18 from IEEE Xplore. Restrictions apply.