21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Multi-task Signal Recovery by Higher Level Hyper-parameter Sharing

Sakinah Ali Pitchay and Ata Kabán
School of Computer Science, University of Birmingham, Birmingham, B15 2TT, UK
{S.A.Pitchay, A.Kaban}@cs.bham.ac.uk, sakinah.ali@usim.edu.my

Abstract

Sharing of hyper-parameters is often useful for multi-task problems as a means of encoding some notion of task similarity. Here we present a multi-task approach for signal recovery by sharing higher-level hyper-parameters which do not relate directly to the actual content of the signals of interest but only to their statistical characteristics. Our approach leads to a very simple model and algorithm that can be used to simultaneously recover multiple natural images with unrelated content. We investigate the advantages of this approach in relation to state of the art multi-task compressed sensing and we discuss our findings.

1. Introduction

Multi-task signal recovery aims to perform several single-frame recovery tasks simultaneously by exploiting some form of similarity between the tasks. A recent paper tackles this complex problem by an approach termed as Multi-Task Bayesian Compressed Sensing (MT-BCS) [3]. In this approach the similarity of tasks is defined as a percentage of overlapping content — i.e. the positions of edges or smooth regions should have a non-negligible overlap. By its construction, MT-BCS is able to exploit this definition of similarity to recover multiple signals simultaneously in a single run more efficiently than multiple runs of a single-task recovery method would.

Here we propose and investigate a complementary approach in which we seek to exploit a much weaker notion of similarity that is unrelated to the actual content but only depends on the statistical characteristics of the signals to be recovered. We achieve this by building the model of MT-BCS to a further level and sharing higher level hyper-parameters in the resulting model. This turns out to yield a very simple model in terms of its model and experimental design. It has fewer hyper-parameters in which the edge-content related parameters are integrated out and the remaining shared higher-level hyper-parameters can be estimated automatically in a similar manner to what we have tackled previously [1, 2]. The next section describes our approach and its relation to MT-BCS, Section 3 presents comparative experiments and discussion, and the last section concludes the paper.

2. Multi-task Recovery Framework

Consider K different (though related) recovery tasks. We will denote by $\mathbf{z}^{(k)}$ the k-th high resolution signal (scene) of length N that we aim to recover. The observed low resolution (or compressed) signal $\mathbf{y}^{(k)}$ has length $M < N$ and is described by the following forward model:

$$\mathbf{y}^{(k)} = \mathbf{W}^{(k)}\mathbf{z}^{(k)} + \boldsymbol{\eta}^{(k)} \ \forall \ k\text{=}1,...,K \qquad (1)$$

where $\boldsymbol{\eta}$ is a mean-zero i.i.d. additive Gaussian noise with variance $\sigma^2 I$.

From eq. (1), we can write the likelihood as:

$$p(\mathbf{y}^{(k)}|\mathbf{z}^{(k)}, \mathbf{W}^{(k)}, \sigma^2) = \mathcal{N}(\mathbf{W}^{(k)}\mathbf{z}^{(k)}, \sigma^2) \qquad (2)$$

and in order to infer $\mathbf{z}^{(k)}, k = 1, ..., K$, we need to specify a model on these, which we do in the next subsection.

2.1 Prior for multiple signals

The gist of multi-task recovery is to exploit similarities between the multiple tasks in order to gain efficiency against performing the tasks individually. There are many ways to define similarity though, and this is a crucial aspect of designing a suitable prior. Before proceeding we define the notation $\boldsymbol{\theta}^{(k)} = \boldsymbol{D}\boldsymbol{z}^{(k)}$ where $\boldsymbol{D}$ could be e.g. a wavelet transform as in [3], or another linear transform that makes the representation of $\boldsymbol{z}^{(k)}$ sparse. In particular, we used a simple linear transform from pixel brightness values into neighbourhood-features by taking the difference between pixel brightness and the average of its four cardinal neighbours (see

978-4-9906441-1-6 ©2012 IAPR

e.g. [2]). With this latter choice of course the components of $\boldsymbol{\theta}^{(k)}$ are not completely statistically independent, however a pseudo-likelihood approximation (as in [2]) makes it possible to treat them as if they were. The transform $\boldsymbol{D}$ is invertible, so estimating $\boldsymbol{\theta}^{(k)}$ is equivalent to estimating $\boldsymbol{z}^{(k)}$, which allows us to simplify the exposition and make the link between the mult-task image prior of [3] and ours in the sequel.

2.1.1 Hyper-parameter sharing in [3]

Previous work in [3] posited the following Gaussian scale-mixture as a multi-task image prior:

$$p(\theta^{(k)}|\alpha) = \prod_{i=1}^{N} \mathcal{N}(\theta_i^{(k)}|0, \alpha_i^{-1}) \tag{3}$$

$$p(\alpha_i|c,d) = \mathcal{G}a(\alpha_i|c,d) \tag{4}$$

where α are hyper-parameters shared across the tasks.

The authors then propose to let $c = d \to 0$, which corresponds to a fat-tail uninformative improper prior. The estimates of α are then obtained by the so-called Type II Maximum Likelihood approach:

$$\alpha = \arg\max_{\alpha} \sum_{k=1}^{K} \log \int d\theta^{(k)} p(\mathbf{y}^{(k)}|\theta^{(k)}) p(\theta^{(k)}|\alpha) \tag{5}$$

Now, since the components of α are inverse variances of the (zero-mean) pixel neighbourhood features, a large entry in this hyper-parameter vector means a nearly zero variance i.e. a locally smooth region, whereas a small entry signifies a large departure from smoothness i.e. a spike or an edge. Sharing of this parameter vector across all the recovery tasks therefore defines a very strong and very specific kind of similarity: the positions of edges and smooth regions must have a considerable overlap. Hence, whenever we know a-priori that the high resolution images that we try to recover are similar to each other in this sense then we can expect that the method in [3] is best placed to exploit it. However, when the notion of similarity defined above is not satisfied, e.g. the images have independent content, then we conjecture that a weaker, higher level similarity of the natural image statistics could be exploited instead. This is what we investigate next.

2.1.2 Higher-level hyper-parameter sharing

We make two important changes to the model in [3]. First, we will not share the inverse-variances of $\boldsymbol{\theta}$ because we want to relax the definition that the extent of overlap in the positions of edges and smooth regions is what defines similarity. Secondly, we build the model further: Instead of letting hyper-parameters of the Gamma hyper-prior to zero, we will share these among the tasks and estimate them from all the data of the multiple recovery tasks. In addition, we make the model more flexible by introducing a width parameter λ. Summing up, our model is the following:

$$p(\theta^{(k)}|\alpha^{(k)}) = \prod_{i=1}^{N} \mathcal{N}(\theta_i^{(k)}|0, \lambda/\alpha_i^{(k)}) \tag{6}$$

$$p(\alpha_i^{(k)}|\nu) = \mathcal{G}a(\alpha_i^{(k)}|\nu/2, 1/2) \tag{7}$$

To estimate the remaining high-level hyper-parameters ν and λ we will use a type-II Maximum Likelihood (ML) on the prior term alone[1], and this will yield a simple and computationally convenient algorithm. That is, we take:

$$\{\nu, \lambda\} = \underset{\nu,\lambda}{\text{arg max}} \sum_{k=1}^{K} \log \int d\alpha^{(k)} p(\theta^{(k)}|\alpha^{(k)}, \lambda) p(\alpha^{(k)}|\nu) \tag{8}$$

The reason is, the integral in eq.(8) is analytically tractable and yields a product of Pearson type VII densities:

$$\int d\alpha^{(k)} p(\theta^{(k)}|\alpha^{(k)}, \lambda) p(\alpha^{(k)}|\nu) = ...$$

$$\prod_{i=1}^{N} \frac{1}{Z(\lambda,\nu)} [(\theta_i^{(k)})^2 + \lambda]^{-\frac{1+\nu}{2}} =: p(\boldsymbol{\theta}^{(k)}|\lambda,\nu) \tag{9}$$

where $Z(\nu, \lambda) = \frac{\Gamma(\frac{1+\nu}{2})\lambda^{\nu/2}}{\Gamma(\nu/2)\sqrt{\pi}}$

2.2 The joint model and parameter estimation

Putting everything together, our joint model for K recovery tasks are defined by:

$$p(\boldsymbol{y}^{(1)}, ..., \boldsymbol{y}^{(K)}, \boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(K)}|\boldsymbol{W}^{(1)}, ..., \boldsymbol{W}^{(K)}, \sigma^2, \lambda, \nu)$$

$$= \prod_{k=1}^{K} p(\boldsymbol{y}^{(k)}|\boldsymbol{\theta}^{(k)}, \boldsymbol{W}^{(k)}, \sigma^2) p(\boldsymbol{\theta}^{(k)}|\lambda,\nu) \tag{10}$$

The negative log of this joint probability will be our objective function that we minimise to get the MAP estimates of all $\boldsymbol{\theta}^{(k)}, k = 1, ..., K$ and ML estimates of

[1] Although a direct extension of the estimation approach in the previous section i.e. an evidence maximisation in the sense of a type-III ML would be interesting to investigate as well, our approach fits with the MAP estimation that we do for finding the most probable images $\boldsymbol{z}^{(k)}$, and we found it to work well in practice as we shall see in the experimental section.

λ, ν and σ^2. Note that we have now integrated out the full set of hyper-parameters $\boldsymbol{\alpha}$ (that appeared in [3]) and these do not need to be estimated at all in our approach. As we already mentioned, this, and our sharing of only ν and λ means a weaker and higher level notion of task similarity than that of [3]) — essentially we only assume similarity of the statistics of $\boldsymbol{\theta}^{(k)}$ and allow the content of the target signals to be different. We carried out the minimisation of the above objective using conjugate gradients in much the same way as described in our previous works [2].

3. Experiments

We investigate three research questions as follows: (i) To what extent our definition of relatedness can be exploited for multi-task recovery? (ii) How does the existing work in M-BCS [3] perform on data that only has our weaker notion of relatedness? (iii) What do we lose by exploiting only our weaker notion of similarity when the data really has the stronger one exactly as defined in MT-BCS[3]?

3.1 Results and Discussion

(i) To gain insight into our first question, we conduct experiments to compare the performance of multiple runs of a single-task recovery algorithm with the performance of one run of our multi-task recovery method. In both methods we use the Pearson type VII image model, however the single-task approach estimates the hyper-parameters ν, λ and the noise variance σ^2 separately for each task whereas the multi-task approach uses all the data to estimate these.

From our experiments we found that multiple runs of single-task recovery already performs very well in terms of means square error (MSE). Nevertheless, the multi-task approach works in a single run and from our experiments it performs no worse for a class of signals (e.g. natural images have similar statistics even when they have different content), and it may even yield a slight improvement in the quality of recovery since it has more data to estimate these hyper-parameters. Figure 1, shows the MSE results of three single-task recoveries versus one multi-task recovery of the same target images — natural images of size 80 × 80 pixels each, which have no overlapping content other than their naturally similar image statistics: '`woman face`', '`cameraman`', and '`castle`'. We varied the number of measurements (extent of compression), and we worked with $\boldsymbol{W}^{(k)}$ randomly generated matrices with i.i.d. standard Gaussian entries. We see the multi-task approach is able to get good recovery in a single run and it needs slightly less measurements for good recovery in this example.

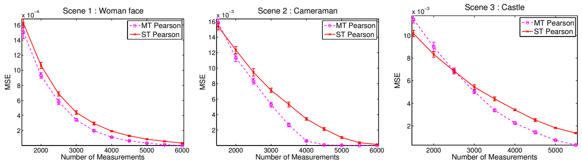

Figure 1: Comparing three separate runs of single-task (ST)-Pearson based recovery against one run of multi-task (MT) Pearson based recovery. The task is to recover three different high resolution images from only one randomly compressed and noisy frame of each. The noise standard deviation was $\sigma = 8 \times 10^{-5}$.

(ii) Next, we compare our multi-task approach presented in the earlier section against M-BCS [3] on data that has no overlapping content but exhibits only our weaker notion of similarity. To perform a systematic study, we first use synthetic 1D spikes signals modified from [3]. We try to recover two signals simultaneously, each having length 512, of which 20 entries are spikes (+1 or -1) and the rest of entries are zero. However, contrary to [3] the positions of these spikes are generated randomly for both signals, with no planned overlap in their positions. Figure 2 shows an example of the data, as well as the results of an extensive comparison when the number of measurements available is varied. Clearly, our MT-Pearson approach that only shares high level hyper-parameters performs significantly better in

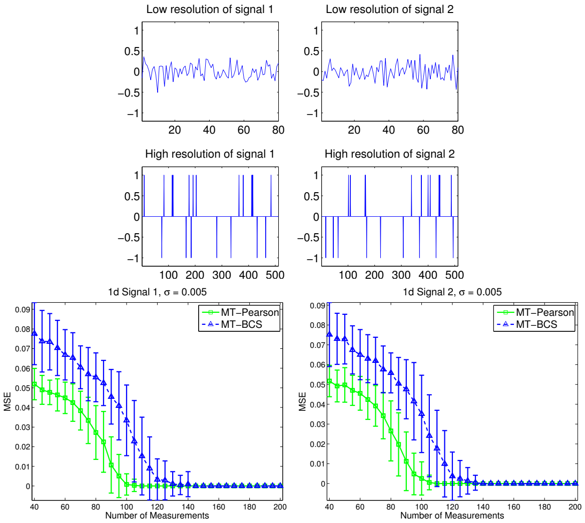

Figure 2: First 4 plots: Example input measurements and high resolution signals to be recovered. Last plot: Comparison of our MT-Pearson approach against MT-BCS on recovering two spike signals simultaneously.

this problem setting. It achieves lower MSE and needs less measurement to recover the high resolution signals. MT-BCS looses out because it expects a content-wise overlap, which is not present in the true signals in this setup.

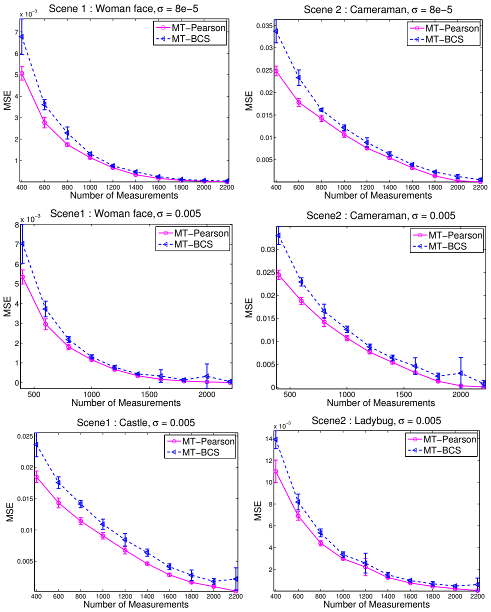

Figure 3: Three sets of experiments simultaneously recovering pairs of natural scenes of size 50 × 50.

To further validate this conclusion, Figure 3 shows multi-task comparison results on image recovery experiments where the task is to recover pairs of natural images simultaneously. Again, we see that our MT-Pearson approach outperforms MT-BCS, and this is because these images have similar statistics but no overlap in their content.

(iii) Finally, we test our approach in scenarios that do have content overlap of the kind that is hard-wired into MT-BCS. We use exactly the same 1D spike signals and use exactly the same experimental setup as [3], and also employ their experimental protocol: That is, the task is to recover two spike signals simultaneously when they have 25%, 50% or 75% of their spikes in the same positions, and the noise level is set to 0.005. By the design of MT-BCS, the larger the percentage of overlap the better MT-BCS will perform, whereas our MT-Pearson does not depend on any content-wise overlap but only on higher level statistical similarity.

The upper plot of Figure 4 shows the results of MT-BCS superimposed with our MT-Pearson. Interestingly, we see that our MT-Pearson is only outperformed by MT-BCS in 75% spike-overlap conditions. It comes out statistically equal to MT-BCS in the 50% overlap setting and it is significantly superior to MT-BCS in settings that have less content-wise overlap. The lower plots of Figure 4 detail all pairwise comparisons separately with error bars shown for completeness.

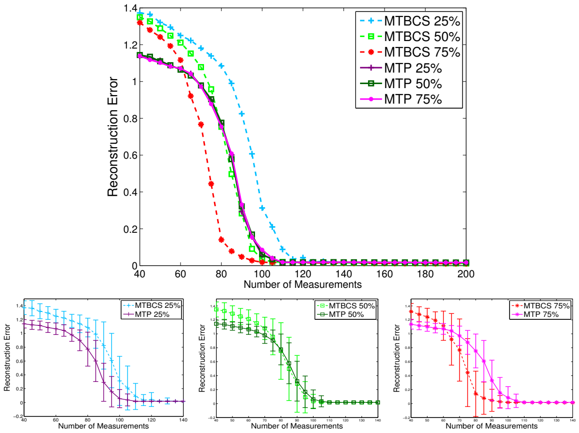

Figure 4: *Upper plot*: Reconstruction errors of MT-Pearson and MT-BCS [3], as a function of the number of compressive measurements. *Lower plots*: The variance of reconstruction errors for 25%, 50% and 75% similarity over 100 independent runs.

4. Conclusions

We presented a new approach to multi-task signal recovery where the target signals need not have any overlap in their content but only share their higher level statistical characteristics. This can be used for simultaneous recovery of sets of natural images in a single run. We compared our approach with multi-task BCS, which is the state of the art for multi-task signal recovery and we highlighted the settings in which our approach is advantageous.

References

[1] S.Ali Pitchay and A. Kabán. Single-frame Signal Recovery using a Similarity-Prior Based on Pearson Type VII MRF, 1^{st}International Conference on Pattern Recognition Applications and Methods (ICPRAM), Vilamoura, Algarve, Portugal, pp. 123-133, 2012.

[2] A. Kabán and S.Ali Pitchay. Single-frame Image Recovery using a Pearson type VII MRF, Special issue of Neurocomputing on Machine Learning for Signal Processing, MLSP2010 special issue. 80: 111-119, 2012.

[3] S. Ji, D. Dunson, and L. Carin. Multi-Task Compressive Sensing IEEE Trans. Signal Processing, vol. 57, no. 1, pp. 92-106, Jan. 2009.