
Compressed Sensing and Bayesian Experimental Design

Matthias W. Seeger

Hannes Nickisch

Max Planck Institute for Biological Cybernetics, Spemannstr. 38, Tübingen, Germany

SEEGER@TUEBINGEN.MPG.DE

HN@TUEBINGEN.MPG.DE

Abstract

We relate compressed sensing (CS) with Bayesian experimental design and provide a novel efficient approximate method for the latter, based on expectation propagation. In a large comparative study about linearly measuring natural images, we show that the simple standard heuristic of measuring wavelet coefficients top-down systematically outperforms CS methods using random measurements; the sequential projection optimisation approach of (Ji & Carin, 2007) performs even worse. We also show that our own approximate Bayesian method is able to learn measurement filters on full images efficiently which outperform the wavelet heuristic. To our knowledge, ours is the first successful attempt at “learning compressed sensing” for images of realistic size. In contrast to common CS methods, our framework is not restricted to sparse signals, but can readily be applied to other notions of signal complexity or noise models. We give concrete ideas how our method can be scaled up to large signal representations.

1. Introduction

There has been a lot of recent interest in the area of *compressed sensing* (CS) (Candès et al., 2006; Donoho, 2006), where it is argued that if signals can be expected to be compressible due to sparseness after some linear transform, then they can be reconstructed from a number of measurements significantly below the Nyquist/Shannon limit, if the measurement design is not too regular. In this paper, we relate CS to the more general notion of statistical (Bayesian) experimental design.

Appearing in *Proceedings of the 25th International Conference on Machine Learning*, Helsinki, Finland, 2008. Copyright 2008 by the author(s)/owner(s).

Through this view, characteristics of signals and algorithms, defined in an abstract mathematical way in the CS literature so far, become understandable and workable. The experimental design approach applies to signals of low complexity in general, not only to sparse ones. It has the potential to clearly outperform the randomised designs, favoured by theoretical CS arguments, in cases where signals are not well-described by common CS assumptions. For example, CS has been viewed with some scepticism so far by researchers in computer vision and image statistics (Weiss et al., 2007). While images exhibit transform sparsity to some degree, purely random measurement designs can be suboptimal for them. The reason is that there is more to low-level image statistics than sparsity. Much of this knowledge can be modeled tractably (Simoncelli, 1999) and could therefore be incorporated into a Bayesian experimental design architecture. To our knowledge, the current CS reconstruction schemes are purely estimation-based and lack proper representations of uncertainty (which is what fundamentally drives experimental design), and the theory deals exclusively with signals which are *unstructured except for random sparsity*. We present experimental results shedding more light on the relationship between CS and images. Similar to (Weiss et al., 2007), we find that standard approaches to linear image measurement (wavelet coefficients) give significantly better reconstruction results than using random measurements favoured by CS, even if modern CS reconstruction algorithms are applied. Yet, our experimental evidence is more substantial than theirs. Beyond that, we show that our efficient approximation to sequential Bayesian design can be used to learn measurements which indeed outperform measuring wavelet coefficients top-down. Our method provides a practically efficient solution to the problem posed in (Weiss et al., 2007), namely how to learn measurement filters automatically from data (using very little concrete knowledge about the signal class) which perform close to or even better than “standard” ones obtained through decades of research and experience. In contrast, the uncertain

components analysis algorithm suggested by them requires a large database of image patches to be run, and could hardly be scaled up to the realistic dimensions treated here¹.

An approximate Bayesian approach to compressed sensing has been presented in (Ji & Carin, 2007), making use of sparse Bayesian learning (SBL) (Tipping, 2001). Our method is based on a different, more general inference approximation, expectation propagation (Minka, 2001), and outperforms theirs very significantly, for prediction based on the same design and, even more so, for sequential design optimisation, as we show in comparative experiments below. Moreover, strongly underdetermined problems (many more variables than observations) are dealt with more efficiently in our framework. In addition, our framework is generalised easily to non-Gaussian observation likelihoods, skew prior terms, and generalised linear models (Gerwinn et al., 2008), and our methodology, our comparisons, as well as our discussion here have a broader scope. Our method is an extension of the scheme in (Seeger et al., 2007). However, the applications to images considered here are orders of magnitude larger than theirs, and several novel ideas are proposed here in order to increase computational efficiency substantially. While much work has been done in statistics on experimental design for the classical Gaussian-linear model, Gaussian priors are entirely inappropriate for images², and designs optimized for them are suboptimal (see also (Seeger et al., 2007)). We are not aware of existing methods for the model used here, which scale comparable to ours, with the exception of (Ji & Carin, 2007).

A different approach for optimising measurement design is given in (Elad, 2007), where \mathbf{X} is designed *a priori* with the aim of making its rows maximally de-coherent. In our setup, \mathbf{X} is designed sequentially, using Bayesian information criteria.

The structure of the paper is as follows: The experimental design view on CS is detailed in Section 2. Our framework for approximate inference is described in Section 3, where we also show how to apply it to large problems, especially in sequential experimental design. Our approach is validated through a series of experiments, comparing it to (Ji & Carin, 2007) and common CS methods on artificial data (Section 4.1),

¹Their experiments are on 4×4 image patches, while ours run efficiently on 64×64 images.

²Reconstruction under the Gaussian-linear model is simply the method of least squares, often referred to as “linear reconstruction”. Much of the improved performance through CS is due to the use of non-linear sparse reconstruction techniques.

and analysing the suitability of CS and Bayesian experimental design on natural images (Section 4.2).

2. Compressed Sensing and Experimental Design

Compressed sensing (CS) (Candès et al., 2006; Donoho, 2006) can be motivated as follows. Suppose a signal, such as an image or a sound waveform, is measured and then transferred over some channel or stored. Traditionally, the measurement obeys the Nyquist/Shannon theorem, allowing for an exact reconstruction of the (band-limited) signal if there is no measurement noise. However, what follows is usually some form of lossy compression, exploiting redundancies and non-perceptibility of losses. Given that, can the information needed for a satisfactory reconstruction not be measured below the Nyquist frequency (this is called undersampling)? In many key applications today, the measurement itself is the main bottleneck for cost reductions or higher temporal/spatial resolution. Recent theoretical results indicate that undersampling should work well if randomized designs are used, and if the signal reconstruction method specifically takes the “compressibility” into account.

Bayesian experimental design encompasses the CS problem. Here, the “compressibility” of signals is encoded in a *prior distribution*, under which signals of low complexity in general, or high (transform) sparsity in particular, have most mass. While an undersampling violates the Nyquist theorem, signals can often still be reconstructed if they are sufficiently likely under the prior. But not every way of undersampling will do. Experimental design is concerned with optimising the measurement structure (called design), so as to obtain the desired information at the lowest possible cost. This is easily explained by considering the model of interest here. Let $\mathbf{u} \in \mathbb{R}^n$ be latent variables (pixels of an image), and let $\mathbf{y} \in \mathbb{R}^m$ be noisy measurements thereof. The model class of interest is

$$P(\mathbf{u}|\mathbf{y}) \propto N(\mathbf{y}|\mathbf{X}\mathbf{u}, \sigma^2\mathbf{I}) \prod_{i=1}^q t_i(s_i), \quad \mathbf{s} = \mathbf{B}\mathbf{u}. \quad (1)$$

The likelihood $P(\mathbf{y}|\mathbf{u})$ is Gaussian and underdetermined ($n > m$). The prior³ is a product of univariate non-Gaussian potentials $t_i(s_i)$. It is computationally advantageous, yet not essential, that the $\log t_i$ be concave (Seeger, 2008), and in this paper we use Laplacian

³We do not require that the prior potential is actually a normalisable distribution over \mathbf{u} , the models of interest here are of the undirected Markov random field (or “energy-based”) type.

potentials

$$t_i(s_i) = \frac{\tau}{2} e^{-\tau|s_i|}, \quad (2)$$

which are of this sort. If number of image pixels n is large, it is important for computational efficiency that matrix-vector multiplications (MVMs) with \mathbf{B} and \mathbf{B}^T (less important: with \mathbf{X} , \mathbf{X}^T) can be done efficiently, and that \mathbf{B} does not have to be stored explicitly.

The unknown signal \mathbf{u} (an image for now) should be “compressible”, *i.e.* it should exhibit *transform sparsity*⁴: after some fixed linear mapping \mathbf{B} , such as a wavelet transform, $\mathbf{s} = \mathbf{B}\mathbf{u}$ has many coefficients s_i close to zero. An image coder would set these to exactly zero, thereby compressing the image. “Expected transform sparsity” is encoded in a sparsity prior, in our case the product of Laplacians (2). As opposed to a Gaussian, a Laplace distribution concentrates more mass close to zero, forcing coefficients to be very small. On the other hand, the Laplacian also has more mass in the tails, which allows for occasional large values. These points are explained further in (Seeger, 2008; Tipping, 2001).

Next, the design is \mathbf{X} , the measurement matrix. In our example, each row of \mathbf{X} is a linear filter specifying a single image measurement. In this paper, we assume that all rows of \mathbf{X} have unit norm⁵. The problem of experimental design is how to choose \mathbf{X} among many candidates of the same cost, so that subsequent measurements allow for the best reconstruction of \mathbf{u} . This decision has to be taken *without* doing real measurements for most candidates. In a Bayesian variant, the posterior distribution $P(\mathbf{u}|\mathbf{y})$ encodes all present knowledge. To score a candidate \mathbf{X}_* (new rows of \mathbf{X}), assume for the moment that the outcome \mathbf{y}_* is known. We can measure the decrease in uncertainty from $P(\mathbf{u}|\mathbf{y})$ to $P(\mathbf{u}|\mathbf{y}, \mathbf{y}_*)$ by the *entropy difference* $H[P(\mathbf{u}|\mathbf{y})] - H[P(\mathbf{u}|\mathbf{y}, \mathbf{y}_*)]$. Not knowing \mathbf{y}_* , we integrate it out using $P(\mathbf{y}_*|\mathbf{y}) = \int P(\mathbf{y}_*|\mathbf{u})P(\mathbf{u}|\mathbf{y})d\mathbf{u}$. This expected information score drives the optimisation of the design. It is clear that such scores are fundamentally based on the posterior as representation of uncertainty, so that algorithms which merely estimate good solutions from given data cannot be used directly in order to compute them⁶. With such methods, either

⁴In our experiments, we use an extended notion of sparsity, see Section 3.2.

⁵When designing \mathbf{X} , it is important to keep its rows of the same scale. Otherwise, a measurement can always be improved (at fixed noise level σ^2) simply by increasing its norm. Put differently, we place a prior on \mathbf{X} which is uniform over all matrices with rows of unit norm.

⁶It is one thing to learn to predict well, yet a different issue to estimate its own uncertainty well, and methods

rough rules of thumb have to be followed to obtain a design (“make it random” in CS), or many measurements have to be taken in a trial-and-error fashion. In Bayesian experimental design, a permanently refined uncertainty representation is used to avoid uninformative data sampling, so often many fewer real measurements are required.

3. Approximate Inference

Bayesian inference is in general not analytically tractable for models of the form (1), and has to be approximated. Moreover, the applications of interest here demand a high efficiency in many dimensions ($n = 4096$ in the natural image experiments here). Importantly, Bayesian experimental design does not only require inference just once, but many times in a sequential fashion. We make use of the *expectation propagation* (EP) method (Minka, 2001), together with a robust and efficient representation for $Q(\mathbf{u}) \approx P(\mathbf{u}|\mathbf{y})$. Our framework has previously been used in a different context (Seeger, 2008), where details can be found which are omitted here. As a novelty, we will show here how the framework can be run efficiently for large n , and how sequential design optimisation can be done orders of magnitude faster.

In EP, the posterior $P(\mathbf{u}|\mathbf{y})$ is approximated by a Gaussian $Q(\mathbf{u})$ with free (variational) parameters \mathbf{b} , $\boldsymbol{\pi}$, which are formally introduced by replacing $t_i(s_i)$ by $\tilde{t}_i(s_i) = e^{b_i s_i - \pi_i s_i^2/2}$ in (1). The distribution $Q(\mathbf{u})$ is represented by lower triangular \mathbf{L} and $\boldsymbol{\gamma}$,

$$\begin{aligned} \mathbf{L}\mathbf{L}^T &= \sigma^{-2}\mathbf{X}^T\mathbf{X} + \mathbf{B}^T\boldsymbol{\Pi}\mathbf{B} = \text{Cov}_Q[\mathbf{u}]^{-1}, \\ \boldsymbol{\gamma} &= \mathbf{L}^{-1}(\sigma^{-2}\mathbf{X}^T\mathbf{y} + \mathbf{B}^T\mathbf{b}), \quad \boldsymbol{\Pi} = \text{diag } \boldsymbol{\pi}, \end{aligned}$$

so that $\text{E}_Q[\mathbf{u}] = \mathbf{L}^{-T}\boldsymbol{\gamma}$. The (b_i, π_i) are then updated sequentially by matching the Gaussian moments of the *tilted distributions*

$$\hat{P}_i(\mathbf{u}) \propto N(\mathbf{y}|\mathbf{X}\mathbf{u}, \sigma^2\mathbf{I}) \prod_{j \neq i} \tilde{t}_j(s_j) \tilde{t}_i(s_i)^{1-\eta} t_i(s_i)^\eta$$

with the new $Q'(\mathbf{u})$. Here, $\eta \in (0, 1]$ is a fractional parameter⁷. In each local update, we need to compute the non-Gaussian moments of the marginal $\hat{P}_i(s_i)$, and to update the $Q(\mathbf{u})$ representation, which is done by an $O(n^2)$ Cholesky update of \mathbf{L} . Note that (Ji & Carin, 2007) employ the variational mean field approximation of (Tipping, 2001), which is specific to sparse linear models (more precisely, all t_i have to be Gaussian

employing “premature sparsification” often perform badly w.r.t. the latter (see Section 4.1).

⁷ $\eta = 1$ gives standard EP, but choosing $\eta < 1$ can increase the robustness of the algorithm on the sparse linear model significantly (Seeger, 2008). We use $\eta = 0.9$ in all our experiments.

scale mixtures, thus even functions), while EP can be applied with little modification to models with skew priors or non-Gaussian skew likelihoods as well (Gerrin et al., 2008).

In our applications of sequential design, we need to score the informativeness of new candidates \mathbf{x}_* (as row of \mathbf{X}), which we do by the entropy difference (see Section 2). If Q' is the approximate posterior after including \mathbf{x}_* , then $2\mathbb{H}[Q'] = \log |\text{Cov}_{Q'}[\mathbf{u}]| + C$, where Q' differs from Q in that $(\mathbf{X}')^T \mathbf{X}' = \mathbf{X}^T \mathbf{X} + \mathbf{x}_* \mathbf{x}_*^T$, and $\boldsymbol{\pi} \rightarrow \boldsymbol{\pi}'$. We approximate the entropy difference by assuming that $\boldsymbol{\pi}' = \boldsymbol{\pi}$, whence

$$\mathbb{H}[Q] - \mathbb{H}[Q'] = \frac{1}{2} \log (1 + \sigma^{-2} \mathbf{x}_*^T \text{Cov}_Q[\mathbf{u}] \mathbf{x}_*).$$

Since $\|\mathbf{x}_*\| = 1$ by assumption, this score is maximized by choosing \mathbf{x}_* along the principal (leading) eigendirection⁸ of $\text{Cov}_Q[\mathbf{u}]$. The same score is used by (Ji & Carin, 2007).

3.1. Large-Scale Applications

There are two major issues with trying to apply our method for large sizes n . First, the EP site updates are done in random sweeps over n sites, because it is not clear which particular site ordering leads to fastest convergence. This problem is severe in our sequential design application to natural images, since there are many small changes to \mathbf{X} , \mathbf{y} (individual new measurements), after each of which EP convergence has to be regained. We approach it by forward scoring many site candidates before each EP update, thereby always updating the one which gives the largest posterior change. This is detailed just below. Second, the robust Q representation of (Seeger, 2008) is of size $O(n^2)$, and each update costs $O(n^2)$. We sketch a different representation of size $O(m^2)$ below, which can be used to drive our framework as well. In contrast, (Ji & Carin, 2007) use a heuristic of setting many of the π_i to ∞ early in the iteration, which leads to much worse results than we obtain (see Section 4.1, Section 4.2).

Our selective updating scheme for EP hinges on the fact that we can maintain all site marginals \mathbf{h} , $\boldsymbol{\rho}$, $Q(s_i) = N(h_i, \rho_i)$, up to date at all times. For a site i , we can quantify the change of Q through an update there by $D[Q'(s_i) \| Q(s_i)]$ (Q' the posterior after the update at i), which can be computed in $O(1)$. Importantly, $D[Q'(\mathbf{u}) \| Q(\mathbf{u})] = D[Q'(s_i) \| Q(s_i)]$ (because $Q(\mathbf{u}|s_i) = Q'(\mathbf{u}|s_i)$), so the score precisely measures the global amount of change $Q \rightarrow Q'$. We maintain a list of candidate sites, which are scored before each EP update, and the update is done for the winner

⁸We compute \mathbf{x}_* by the Lanczos algorithm.

only. The list is then evolved by replacing the lower half of worst-scoring sites by others randomly drawn from $\{1, \dots, q\}$. Importantly, the marginals \mathbf{h} , $\boldsymbol{\rho}$ can be updated along with the representation, at the expense of only *one* additional \mathbf{L} backsubstitution and MVM with \mathbf{B} . Namely, if $\pi'_i = \pi_i + \Delta\pi_i$, $b'_i = b_i + \Delta b_i$, and $\mathbf{w} := \mathbf{B}\mathbf{L}^{-T}(\mathbf{L}^{-1}\mathbf{B}_{i,\cdot}^T)$, then

$$\boldsymbol{\rho}' = \boldsymbol{\rho} - \frac{\Delta\pi_i}{1 + \rho_i\Delta\pi_i} \mathbf{w} \circ \mathbf{w}, \quad \mathbf{h}' = \mathbf{h} + \frac{\Delta b_i - h_i\Delta\pi_i}{1 + \rho_i\Delta\pi_i} \mathbf{w}.$$

Here, $\mathbf{L}^{-1}\mathbf{B}_{i,\cdot}^T$ has to be computed for the \mathbf{L} update anyway. This idea is used in the experiments described in Section 4.2.

For large n , storing an $n \times n$ matrix in memory becomes prohibitive. In a less costly representation, we exploit $m \ll n$. We require⁹ that $\mathbf{B} = \mathbf{I}$. The Woodbury formula gives

$$\text{Cov}_Q[\mathbf{u}] = \boldsymbol{\Pi}^{-1} - \boldsymbol{\Pi}^{-1} \mathbf{X}^T \mathbf{L}^{-T} \mathbf{L}^{-1} \mathbf{X} \boldsymbol{\Pi}^{-1},$$

where $\mathbf{L}\mathbf{L}^T = \mathbf{I} + \mathbf{X}\boldsymbol{\Pi}^{-1}\mathbf{X}^T$, so \mathbf{L} (different from above) is of size m^2 only. An EP update requires $O(m^2)$ and two MVMs with \mathbf{X} , rather than $O(n^2)$ above. While this representation is exact, it is numerically less robust to update than the $O(n^2)$ one.

3.2. Image Model. Other Methods

In this section, we provide further details about the concrete model we use in our experiments with natural images. Our prior encourages two different notions of sparsity in an image. First, a multi-scale wavelet transform of \mathbf{u} should be sparse, modeling the observation that natural images can be compressed well in a wavelet domain. Second, the finite differences in the horizontal and vertical direction should exhibit sparsity, accounting for spatial smoothness often found in images¹⁰. A frequently used penalty term for the latter is the L_1 norm of the image gradient, also known as *total variation*.

Our model is an instance of (1), where all t_i are Laplacian (2). \mathbf{s} , and therefore \mathbf{B} , decompose into two different parts: $\mathbf{B}^T = (\mathbf{B}^{(sp)})^T \mathbf{B}^{(tv)T}$. Equivalently, the prior is the product of two potentials. The *transform sparsity* potential is a sparsity prior on the wavelet coefficients of \mathbf{u} . Note that the Laplace distribution is a sensible candidate to fit wavelet coefficient histograms from natural images (Simoncelli, 1999). Thus,

⁹More generally, $\mathbf{B}^T \boldsymbol{\Pi} \mathbf{B}$ must be easy to invert. If \mathbf{B} is invertible and \mathbf{B}^{-1} -MVM feasible, we represent $Q(\mathbf{s})$ rather than $Q(\mathbf{u})$.

¹⁰Recall what we mean by sparsity from Section 2: *most* coefficients are forced to be small, by allowing *some* to be large. Occasional large components in the gradient correspond to edges in the image.

$\mathbf{B}^{(sp)} \in \mathbb{R}^{n \times n}$ is a multi-scale orthonormal wavelet transform, and the potential is $\exp(-\tau_{sp} \|\mathbf{B}^{(sp)} \mathbf{u}\|_1)$. The *total variation* potential is a Laplace prior on the image gradient, *i.e.* the differences between horizontal and vertical pixel neighbours¹¹. $\mathbf{B}^{(tv)} \in \mathbb{R}^{2(n-\sqrt{n}) \times n}$ is a sparse structured matrix, mapping the image \mathbf{u} to its gradient. Here, we assume that $n = 2^{2k}$ for simplicity. The total variation potential is $\exp(-\tau_{tv} \|\mathbf{B}^{(tv)} \mathbf{u}\|_1)$.

Therefore, we have $q \approx 3n$ for the size of \mathbf{s} . Also, the potentials come with different scale parameters τ_{sp} , τ_{tv} . Importantly, neither of $\mathbf{B}^{(sp)}$, $\mathbf{B}^{(tv)}$ has to be stored in memory, and MVM with \mathbf{B} or \mathbf{B}^T can be done in $O(n)$.

We also briefly describe the methods we compare against. Most of them come with a transform sparsity potential only, so that $\mathbf{s} = \mathbf{B}^{(sp)} \mathbf{u}$. The method of (Ji & Carin, 2007) is called SBL here. In L_p reconstruction, $\hat{\mathbf{s}} = \arg\min\{\|\mathbf{s}\|_p \mid \mathbf{X}\mathbf{B}^{(sp)T} \mathbf{s} = \mathbf{y}\}$, $\hat{\mathbf{u}} = \mathbf{B}^{(sp)T} \hat{\mathbf{s}}$. For L_2 we just solve the normal equations, while for L_1 this is a linear program. Note that the latter is used in many CS publications (Candès et al., 2006; Donoho, 2006). A method with transform sparsity *and* total variation potential, called $L_1 + TV$ here, is given by the following quadratic program: $\hat{\mathbf{u}} = \arg\min \frac{1}{2} \|\mathbf{y} - \mathbf{X}\mathbf{u}\|_2^2 + \tau_{sp} \sigma^2 \|\mathbf{B}^{(sp)} \mathbf{u}\|_1 + \tau_{tv} \sigma^2 \|\mathbf{B}^{(tv)} \mathbf{u}\|_1$ (Candès & Romberg, 2004). We used the following code in our experiments:

SBL: www.ece.duke.edu/~shji/BCS.html
 L_1 : www.acm.caltech.edu/l1magic/
 $L_1 + TV$: www.stanford.edu/~mlustig/

4. Experiments

In this section, we provide experimental results for different instances of our framework, comparing to CS and approximate Bayesian methods on synthetic data (Section 4.1), and on the task of measuring natural images (Section 4.2).

4.1. Artificial Setups

It is customary in the CS literature to test methods on synthetic data, generated following the “truly sparse and otherwise unstructured” assumptions under which asymptotic CS theorems are proven. We do the same here, explicitly using the “(non-)uniform spikes” (Ji & Carin, 2007), but cover some other heavy-tailed distributions as well. It seems that not many signals of real-world interest are strictly and randomly sparse, so

¹¹This potential on its own is not normalisable as distribution over \mathbf{u} , being invariant against adding a constant to all pixels.

that studies looking at the robustness of CS theoretical claims are highly important. In this section, signals are sparse as such, so that $\mathbf{B} = \mathbf{I}$ and $\mathbf{u} = \mathbf{s}$ here. We compare methods described in Section 3.2. It is important to stress that all methods compared here (except for L_2) are based on exactly the same underlying model (1) with $\mathbf{B} = \mathbf{I}$, and differences arise only in the nature of computations (approximate Bayesian versus maximum a-posteriori optimisation) and in whether \mathbf{X} is sequentially designed (EP, SBL) or chosen at random (L_p reconstruction; we follow CS theory (Candès et al., 2006; Donoho, 2006) and sample rows of \mathbf{X} uniformly of unit norm). Results are shown in Figure 1.

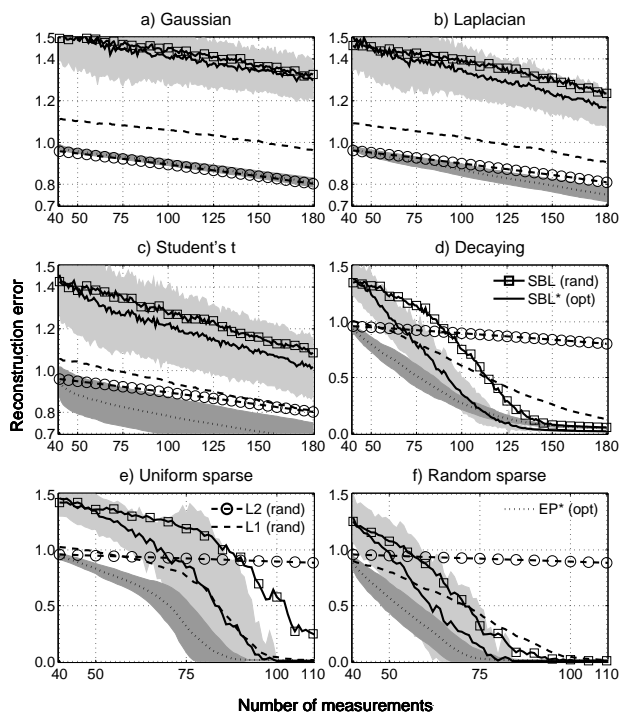


Figure 1. Comparison on 6 random synthetic signals $\mathbf{u} \in \mathbb{R}^{512}$. Shown are L_2 -reconstruction errors (mean \pm std.dev. over 100 runs). All methods start with same random initial \mathbf{X} ($m = 40$), then “(rand)” add random rows, “(opt)” optimise new rows sequentially. Noise variance $\sigma^2 = 0.005$, prior scale $\tau = 5$. SBL: (Ji & Carin, 2007), L_p : L_p reconstruction, EP: our method. (a-c): i.i.d. zero mean, unit variance Gaussian, Laplacian (Eq. 2), Student’s t (3 d.o.f.). (d): $\frac{n}{2}$ of $u_i = 0$, $\frac{n}{4}$ exponential decay $1, \dots, 0$, $\frac{n}{4}$ minus that, randomly permuted. (e-f): 20 $u_i \neq 0$ at random; (e) uniform spikes, $u_i \in \{\pm 1\}$; (f): non-uniform spikes, $u_i \sim \frac{1}{4} + |t|$, $t \sim N(0, 1)$; as in (Ji & Carin, 2007). Distributions in (d-f) normalised to unit variance.

The “sparsity” (or super-Gaussianity) of the signal distributions increases from (1a) to (1e-f). For Gaussian signals (1a), L_2 reconstruction based on random measurements is optimal. While all CS methods and SBL

(random and designed) lead to large errors, EP with design matches the L_2 results, thus shows robust behaviour. For Laplacian and Student’s t signals (1b-c), designed EP outperforms L_2 reconstruction significantly, while even the CS L_1 method still does worse than simple least squares. SBL performs poorly in all three cases with signals not truly sparse, thus is not robust against rather modest violations of the strict CS assumptions. Its non-robustness is also witnessed by large variations across trials.

On the other hand, L_2 performs badly on truly sparse signals. In all cases (1d-f), EP with design significantly outperforms all other methods, including designed SBL, with special benefits at rather small numbers of measurements. SBL does better now with truly sparse signals, and is able to outperform L_1 .

From the superior performance of EP with design on all signal classes, we conclude that experimental design can sequentially find measurements that are significantly better than random ones, even if signals are truly sparse. Moreover, the superior performance is robust against large deviations away from the underlying model, more so even than classical L_1 or L_2 estimation. The poor performance of SBL (Ji & Carin, 2007) seems to come from their desire for “premature sparsification”. During their iterations, many π_i are clamped to $+\infty$ early for efficiency reasons. This does not hurt mean predictions from current observations much, but affects their covariance approximation drastically: most directions not supported by the data right now are somewhat ruled out for further measurements, since posterior variance along them (which should be large!) is shrunk in their method. In contrast, in our EP method, none of the π_i become very large with modest m , and our covariance approximation seems good enough to successfully drive experimental design. Without premature sparsification, our scheme is still efficient, since the most relevant site updates are found actively, and the need to eliminate variables does not arise.

4.2. Natural Images

In this section, we are concerned about finding linear filters which allow for good reconstruction of natural images from noisy measurements thereof. Since natural images exhibit sparsity in wavelet or Fourier domains, CS theory seems to suggest that random measurements should be well-suited for this purpose, and there have been considerable efforts to develop hardware which can perform such random measurements cost-efficiently (Duarte et al., 2008). On the other hand, much is known about low level natural image

statistics, and powerful linear measurement transforms have emerged there, such as multi-scale wavelet transforms, based on which natural image reconstruction should be substantial better than for random measurements (Weiss et al., 2007).

The sparsity of images in a wavelet domain is highly structured, there is a clear ordering among the coefficients from coarse to fine scales: natural images typically have much more energy in the coarse-scale coefficients, and coefficients with very small values are predominantly found in the fine scales. In our experiments, we employ a simple heuristic for linearly measuring images, called *wavelet heuristic* in the sequel: every measurement computes a single wavelet coefficient, and the sequential ordering of the measurements is deterministic top-down, from coarse to fine scales¹². This ordering is a pragmatic strategy: if mainly the coarse-scale coefficients are far from zero, they should be measured first¹³. Do state-of-the-art CS reconstruction algorithms, based on random linear image measurements, perform better than simple L_2 reconstruction based on the wavelet heuristic? And how does Bayesian sequential design perform on this task, if the model described in Section 3.2 is used? Note that no prior knowledge about typical ordering or dependence among wavelet coefficients in encoded in this model either. Results of our study are given in Figure 2.

In fact, we started our exploration with what is shown in (2a), where 100 initial filters are drawn at random (except for L_2 (heur)). Intrigued by the fact that the wavelet heuristic method L_2 (heur) outperformed all CS variants significantly, we tried to give them a head-start, supplying $m = 100, 200, 400$ wavelet heuristic measurements initially (2b-d). However, the systematic under-performance of methods which have sparsity regularizers built in, yet do *random* rather than wavelet measurements, remains consistently present. From these results we conclude, much as (Weiss et al., 2007) argued on theoretical grounds, that if natural images are to be measured successively by unit norm, but otherwise unconstrained linear filters, then *drawing these filters at random leads to significantly worse*

¹²This ordering follows the recursive definition of such transforms: downsampling by factor two (coarse), horizontal differences, vertical differences, diagonal corrections at each stage. Our ordering is coarse \rightarrow horizontal \rightarrow vertical \rightarrow diagonal, descending just as the transform does.

¹³Note that another problem with common CS assumptions applied to images is that the typical scale of coefficients along a coarse-to-fine ordering follows a smooth power law, it does not exhibit the abrupt drop from “significantly above noise level” to “exactly zero” often required by CS theory.

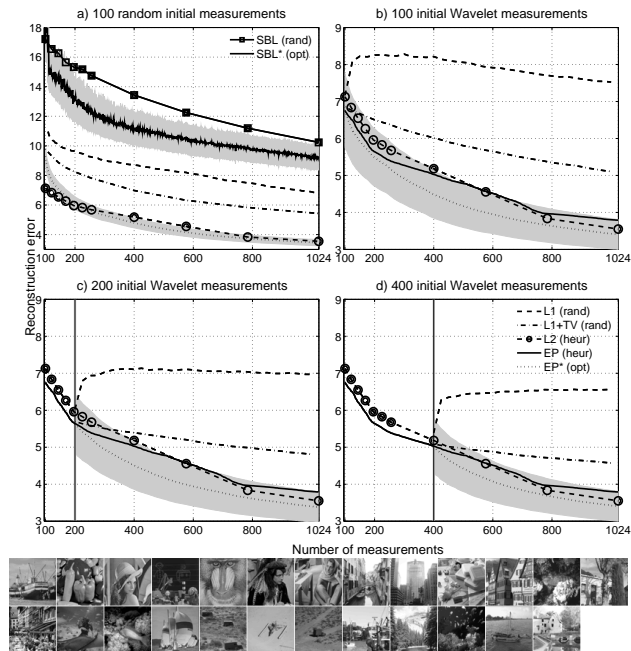


Figure 2. Experiments for measuring natural images ($64 \times 64 = 4096$ pixels). Shown are L_2 -reconstruction errors averaged over 25 grayscale images typically used in computer vision research (from decsai.ugr.es/cvg/dbimagenes/) ($\pm \frac{1}{4}$ std.dev. for “*”). Noise level $\sigma^2 = 0.005$. SBL: (Ji & Carin, 2007), L_p : L_p reconstruction, $L_1 + TV$: Lasso with TV/wavelet penalties, EP: our method. True σ^2 supplied, τ parameters chosen optimally for each method individually: $\tau_{sp} = \tau_{tv} = 0.075$ ($L_1 + TV$), $\tau_{sp} = 0.075$, $\tau_{tv} = 0.5$ (EP). New rows of \mathbf{X} random unit norm (rand), actively designed (opt), acc. to wavelet heuristic (heur).

(a): Start with $m = 100$, \mathbf{X} random unit norm. (b-d): Start with $m = 100, 200, 400$, \mathbf{X} acc. to wavelet heuristic.

reconstructions than using standard wavelet coefficient filters top-down. While CS theorems are mathematically intriguing, and while there certainly are important applications that benefit from these results¹⁴, linear image measurement is probably not among them.

On the other hand, the wavelet heuristic method is significantly outperformed by our EP method, where \mathbf{X} is designed sequentially. In (2a), EP quickly recovers from the suboptimal initial random \mathbf{X} . Moreover, even when started from the same point as the wavelet heuristic (2b-d), the designed measurements lead to improvements over the heuristic immediately.

¹⁴The theoretical CS setting is more extreme than what is really required here, in that there is no prior knowledge about where the non-zeros will lie. We speculate that more suitable applications could lie in steganography, spam or intrusion detection, where a signal has to be detected which has been hidden by an adversary.

EP(heur) is doing EP reconstruction, but based on the same measurements as L_2 (heur). While it slightly outperforms L_2 reconstruction, the significant difference is due to the choice of the measurements. Our method therefore provides an efficient solution to the problem posed in (Weiss et al., 2007), namely how to learn measurements automatically from data, starting from little concrete domain knowledge. On the particular problem of measuring images linearly, our findings should be put into perspective, by noting that the L_2 wavelet heuristic is vastly faster to compute¹⁵. Moreover, \mathbf{X} is optimised sequentially, particular to the image \mathbf{u} (but without knowing the underlying \mathbf{u}), while the wavelet heuristic filters are always the same. Finally, the final \mathbf{X} is dense and unstructured. However, our method can be used in the same way to address applications where strong structural constraints on allowable \mathbf{X} are present, and where wavelet (or purely random) measurements are not an option.

In this setting, SBL (Ji & Carin, 2007) performs much worse than all other methods tried, whether using random, wavelet or designed measurements. Results for SBL in cases (b-d) were even worse and are not included to facilitate comparison among the others. This is most probably an extreme instance of the problem noted in Section 4.1. Premature sparsification, in light of not strictly sparse signals, leads to poor results even with random \mathbf{X} . Their covariance estimates seem too bad to steer sequential design in a useful direction¹⁶.

Finally, the deterioration of L_1 , when adding random to initial wavelet measurements, is somewhat puzzling, especially since it does not happen for $L_1 + TV$. These additional measurements provide novel information about the true \mathbf{u} , so a valid inference method should rather improve.

5. Discussion

We have shown how to address the compressive sensing problem with Bayesian experimental design, where designs are optimised to rapidly decrease uncertainty and do not have to be chosen at random. In a large study

¹⁵EP sequential design is still very efficient. A typical run on one image took 53 min (on 64bit 2.33GHz AMD), for $n = 4096$ and $q = 12160$ sites: 16785 initial EP updates, then 308 increments of \mathbf{X} by 3 rows each, with on average only 8.8 site updates needed to regain EP convergence (up to 85 updates after *some* increments).

¹⁶In cases (b-d), top wavelet coefficients are measured initially, so their method confidently starts with a highly over-sparse solution and fails. Note that, as opposed to EP, we restarted the SBL code after each new measurement, so that poor current solutions are not inherited when new data is obtained.

about linearly measuring natural images, we show that CS reconstruction methods based on randomly drawn filters are outperformed significantly by standard least squares reconstruction measuring coarse-scale wavelet coefficients. Our findings suggest that the applicability of CS results (with their insistence on strict and unstructured signal sparsity) to natural image applications should be reconsidered. We also show that our Bayesian sequential design method, starting from a model with little domain knowledge built in, is able to find filters with significantly better reconstruction properties than top-down wavelet coefficients. Our findings indicate that efficient Bayesian experimental design techniques are highly promising for CS applications of different kinds just as well.

Why do random measurement filters enjoy good properties in CS theory, but are not useful in the case of natural images? We think that this seeming contradiction really comes from an erroneous “extrapolation” of what CS theorems really mean. Any structure apart from a randomly distributed sparsity pattern is ignored there. Also, they are *minimax* results, in that the reconstruction error for the worst sparsity pattern is bounded. But undersampled image reconstruction is not a worst-case problem, and much is known about the sparsity structure of natural images. It may be that L_1 or $L_1 + TV$ are minimax methods (for known \mathbf{B}), but that does not imply much about their *typical* performance. We suspect that our doubts about CS with random measurements extend beyond natural images to other signals of common interest in normal non-adversarial situations, since interest in a signal class implies that statistical knowledge about them beyond random sparsity has been obtained.

Our experience with the method of (Ji & Carin, 2007), which we compare against in our study, raises another more speculative, yet interesting point. Several methods very frequently used in machine learning today can loosely be summarised as trying to detect very sparse solutions early on, mainly with the aim of high computational efficiency. For example, SBL (Tipping, 2001) is much more aggressive in this respect than our EP method here. Early sparsification does not seem to hurt mean prediction performance much, and thus is embraced for efficiency. However, our experiences here indicate that it is the *covariance* (or uncertainty) estimates that can be badly hurt by such sparsity-by-elimination processes, and that in contexts such as experimental design, where covariances are more important than predictive means, their application should probably be avoided. The challenge is then to develop methods that run efficiently *without* eliminating many variables early on, and our selective site updat-

ing method for EP is a step in that direction.

References

- Candès, E., & Romberg, J. (2004). Practical signal recovery from random projections. *Proceedings of SPIE*.
- Candès, E., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theo.*, 52, 489–509.
- Donoho, D. (2006). Compressed sensing. *IEEE Trans. Inf. Theo.*, 52, 1289–1306.
- Duarte, M., Davenport, M., Takhar, D., Laska, J., Sun, T., Kelly, K., & Baraniuk, R. (2008). Single pixel imaging via compressive sampling. To appear in *IEEE Signal Processing Magazine*.
- Elad, M. (2007). Optimized projections for compressed sensing. *IEEE Transactions on Signal Processing*.
- Gerwinn, S., Macke, J., Seeger, M., & Bethge, M. (2008). Bayesian inference for spiking neuron models with a sparsity prior. *Advances in NIPS 20*.
- Ji, S., & Carin, L. (2007). Bayesian compressive sensing and projection optimization. *Proceedings of ICML 24*.
- Minka, T. (2001). Expectation propagation for approximate Bayesian inference. *Uncertainty in AI 17*.
- Seeger, M. (2008). Bayesian inference and optimal design in the sparse linear model. To appear in *Journal of Machine Learning Research*.
- Seeger, M., Steinke, F., & Tsuda, K. (2007). Bayesian inference and optimal design in the sparse linear model. *AI and Statistics 11*.
- Simoncelli, E. (1999). Modeling the joint statistics of images in the Wavelet domain. *Proceedings 44th SPIE* (pp. 188–195).
- Tipping, M. (2001). Sparse Bayesian learning and the relevance vector machine. *J. M. Learn. Res.*, 1, 211–244.
- Weiss, Y., Chang, H., & Freeman, W. (2007). Learning compressed sensing. Snowbird Learning Workshop, Allerton, CA.