Bayesian Experimental Design
for Compressed Sensing

Hannes Nickisch, Matthias Seeger

Abstract

Compressed sensing (CS) can be addressed as instance of Bayesian experimental design, which is based on similar models than common estimation (or energy minimisation) methods, but fundamentally differs from the latter, in that estimates of uncertainty and correlations of latent variables are maintained. The Bayesian computations are approximated by the expectation propagation fixed point algorithm, which is scaled up to the application of interest here through a novel scheduling mechanism, exploiting the fact that marginal uncertainty estimates are available at all times. An important application of CS is the optimisation of architectures for measuring natural images. In a large study, we compare various CS reconstruction methods utilising random measurement filters from different ensembles to a number of techniques which sequentially search for these filters, including our own, and Bayesian projection optimisation [1]. We find that a simple heuristic of measuring wavelet coefficients in a fixed, top-down ordering significantly outperforms CS methods using random measurements; the approach of [1] performs even worse. In contrast, our Bayesian design method learns filters that outperform the wavelet heuristic. Our results show that the property of incoherence of a measurement design, which plays a central role in the “unstructured except for random sparsity” theoretical CS setting, bears no significance for measuring real natural images. Our framework is not restricted to sparse signals, other notions of signal or noise structure can easily be accommodated. We give concrete ideas how our method can be scaled up to large signal representations.

Part of this work has been presented at a conference [2].

Index Terms

Experimental Design, Compressed Sensing, Bayesian Inference, Expectation Propagation

I. INTRODUCTION

Stable statistical structure in a signal, such as sparsity after some linear transform, can be used to compress its natural representation without perceptible loss. Surprisingly, such signals can be sampled significantly below the Nyquist/Shannon limit and still be reconstructed to satisfaction, if only regular undersampling designs are avoided. This important consequence of sparsity has led to compressed sensing (CS) [3], [4], theoretical underpinnings of which come from approximation theory [5]. In this paper, we address the CS problem within the general framework of statistical (Bayesian) experimental design.

Manuscript received March 27, 2008. Supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778.

Hannes Nickisch and Matthias Seeger are with the Max Planck Institute for Biological Cybernetics, Department of Empirical Inference, Tübingen, 72072 Germany (e-mail: {hn,seeger}@tuebingen.mpg.de).
The majority of current work on CS is based on theoretical considerations about the asymptotic minimax performance of certain penalised estimators. In general, signals are assumed to be unstructured except for random sparsity, mostly to render theoretical analyses feasible in the first place. However, in practice much more potentially useful structure is typically present, which may go unnoticed under overly restrictive assumptions. Moreover, the notion of “optimal measurements” certainly depends on whether we aim to hedge against the worst case, or whether we place ourselves in a more benevolent setting, where active reductions in uncertainty normally lead to better predictions. For example, natural images exhibit transform sparsity to some degree, yet random measurements favoured by CS theory can be suboptimal for them [6]. The reason is that there is more to low-level image statistics than sparsity. Much of this knowledge can be modelled tractably [7] and can therefore be incorporated into a Bayesian experimental design architecture.

We present experimental results shedding more light on how to make CS work for images. Similar to [6], we find that standard approaches to linear image measurement (wavelet coefficients) give significantly better reconstruction results than using random measurements, even if modern CS reconstruction algorithms are applied. Beyond that, we show that our efficient approximation to sequential Bayesian design can be used to learn measurement filters which indeed outperform measuring wavelet coefficients top-down. As we discuss below in more detail, our findings indicate that certainly for natural images, but also for other signals with non-Gaussian but structured statistics, measurement designs can be optimised in a data-driven way from little concrete prior knowledge, with outcomes that can be significantly superior to uniformed or even colored random designs. The main property driving the design optimisation in our case is the ability of the Bayesian reconstruction method to maintain valid uncertainty beliefs about its point estimates at all times.

Our setup is sequential, in that new filters are appended to the measurement design one at a time. Adaptive techniques, such as our Bayesian one, make use of all measurements obtained so far to decide upon the next, while non-adaptive methods neglect this sequential information. A simple non-adaptive approach is to sample the design matrix at random, using independent Gaussian or Bernoulli entries, or random rows of the discrete Fourier transform (DFT) matrix. Also, colored random projections have been proposed [8], to take into account second order structure of the signal besides sparsity. A different approach for a priori measurement design is given in [9], where the measurement matrix $X$ is optimised to make its rows maximally incoherent with the sparsifying transform. A similar argument lets [3] use the noiselet transform [10]: it is maximally incoherent to the Haar wavelet basis.

On the other hand, with adaptive techniques, the next filter (row of $X$) is chosen to maximise a criterion which depends on the measurements made so far. For example, the hierarchical nature of multi-scale wavelet coefficients motivates the adaptive heuristic proposed in [11]. An approximate Bayesian approach to compressed sensing (BCS) has been presented in [1], making use of sparse Bayesian learning [12]. Our method is based on a different, more general inference approximation, expectation propagation [13], and in our application to natural images outperforms theirs very significantly. As we argue below, their method seems to be over-aggressive in terms of sparsification, leading to avoidable mistakes on natural images, which are just not strictly sparse in general. Moreover, their uncertainty (posterior covariance) estimates seem to be adversely affected by their aggressiveness, which in
turn spoils their design adaptation. In addition, our framework is easily generalised to non-Gaussian observation likelihoods, skew prior terms, and generalised linear models [14], and our methodology and comparisons have a broader scope. Our method is an extension of the scheme in [15]. However, the application to images considered here is orders of magnitude larger than theirs, and several novel ideas are proposed here in order to increase computational efficiency substantially.

The structure of the paper is as follows. The experimental design approach to CS is detailed in Section II. Our framework for approximate inference is described in Section III, where we also show how to apply it to large problems, especially concerning sequential experimental design. Other approaches to the same problem are reviewed in Section IV. Our approach is validated through a series of experiments, comparing it to a range of others methods on artificial data (Section V-A), and on the problem of measuring natural images (Section V-B).

II. COMPRESSED SENSING AND EXPERIMENTAL DESIGN

Compressed sensing (CS) [3], [4], also known as compressive sampling, 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. 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 randomised designs are used, and if the signal reconstruction method specifically takes the compressibility into account.

It is important to distinguish between the CS problem, signal characteristics that make CS possible, reconstruction methods incorporating these properties, and theoretical results about the problem in principle, or about specific reconstruction methods. In the recent surge of activity on CS, such distinctions are not always precisely stated, which may lead to confusion. CS constitutes a problem, which in practice is amply motivated by cost reductions. Fewer measurements, or less precise sampling, can lead to similar quality in signal reconstruction, at the expense of having to design and run a more difficult reconstruction method, and also (in general) of having to modify “standard” measurement designs1. Not all types of signals are admissible to CS. For example, for band-limited random noise, the Nyquist theorem is tight. In general, CS is applicable to signals whose distribution has some structure that is known a priori, before any measurements are done. Since such knowledge can be used to compress samples, signals of that sort are also called compressible. A very important structure, which is characteristic to

1 An interesting point, not further discussed in this paper, is that in some applications, “standard” undersampling designs are particularly bad in terms of information content, once structural prior knowledge about the signal is used. The regularity of such designs (for example, sampling at equi-spaced locations) can mean that different measurements deliver essentially redundant information. In our opinion, this is probably the best explanation for the success of purely randomised designs in practice. However, Bayesian optimisation has no notion of regularity, and hardly ever produces designs which are very regular.
some extent for many signals, is sparsity: if the signal in its standard representation is transformed linearly, most coefficients are very close to zero, while a few can be large. We will discuss sparsity below in more detail. One can think about structural prior knowledge as a (partial) ordering on the representation space of the signal. In this ordering, a signal is “less complex” than another one, if it adheres better to prior knowledge.

Any solution to the CS problem has to master two related, but different tasks. First, for given measurements, an estimate of the signal has to be computed, taking into account the prior knowledge. This is called signal reconstruction. Second, the decision of how to measure in the first place, has to be taken. Bayesian experimental design offers a powerful way of addressing both points. The structural prior knowledge about a signal (its compressibility) is encoded in a prior distribution, under which signals of low complexity in general, or high (transform) sparsity in particular, have most mass. By the Nyquist theorem, all signals within some band are identifiable through the likelihood function of measurements spaced closely enough. A Bayesian (as well as a CS) reconstruction of the signal, however, is obtained by combining likelihood and prior: signals which are sufficiently likely under the prior, can often be reconstructed from a likelihood function of undersampled measurements, at lower cost than with a foolproof Nyquist-spaced sample.

The problem of optimising the measurement structure (or design), so that less measurements are needed to attain the same reconstruction quality, is harder in general. For this problem, Bayesian experimental design offers a powerful and general solution. The danger of misinterpretation of theoretical CS minimax results becomes apparent here: this problem is in essence “defined away” by requiring no more than incoherence from a design. In the context of natural images, maximally incoherent (random) designs perform rather poorly, while properly optimised designs can improve upon the engineering status quo. Remarkably, the same prior knowledge is available to both Bayesian design and CS reconstruction methods. While in our Bayesian setup, prior and observations are used in order to choose good subsequent measurements, this seems hard to do with CS point estimation techniques, and no principled approach based on such methods is known to us. The apparent divergence between CS theory and performance in practice is discussed in more detail below.

Bayesian experimental design is easily explained using the concrete model and task we are mainly interested in here. An image is represented as pixelised bitmap, which (for notational convenience only) is stacked into a vector $u \in \mathbb{R}^n$ (where $n$ is the number of pixels). In our example, $u_i$ are gray-scale values, but an extension to color images is straightforward. The task is to reconstruct $u$ (the latent variables) from noisy linear measurements

$$y = Xu + \varepsilon, \quad X \in \mathbb{R}^{m \times n}, \varepsilon \sim N(0, \sigma^2 I). \quad (1)$$

$X$ is called the design or measurement matrix, its rows are measurement filters. The filters are constrained to have unit norm$^3$. Note that $m < n$ in general, since measuring each pixel in turn is not considered an efficient design.

$^2$The Nyquist theorem states that there are always some signals that cannot be reconstructed properly from an undersampled likelihood, but a well-chosen design can ensure that most of these “bad signals” have very low prior probability.

$^3$When designing $X$, it is important to keep its rows of the same scale. Otherwise, a measurement can always be improved (at fixed noise level $\sigma^2$) simply by increasing its norm.
The reconstruction problem is therefore underdetermined, and \((X, y)\) constitutes an undersampling of \(u\). The task is to choose the filters in a sequential manner (one after the other), so as to obtain a satisfactory reconstruction of \(u\) with as small \(m\) as possible. Note that in real-world instances of this problem, for example in magnetic resonance imaging, additional constraints on the filters (beyond unit norm) may be present. Our solution presented here readily extends to constrained filter optimisation as well (see Section IV-A).

The prior distribution \(P(u)\) should encode properties which are characteristic of natural images, and this is where sparsity comes into play. While classical Bayesian analysis for the linear model (1) employs Gaussian priors for \(u\), and experimental design is well-developed in general for the Gaussian case [16], natural image statistics are distinctively non-Gaussian, and all we do in the following would work poorly with Gaussian \(P(u)\) (see [15] for a related problem). Our image prior here is composed of Laplace (or double exponential) potentials

\[
t_i(s_i) := \frac{\tau_i}{2} e^{-\tau_i |s_i|}, \quad s_i = [Bu]_i,\tag{2}
\]

whose coefficients \(s_i\) are linear functions of the image \(u\), collected in the transform matrix \(B\). In contrast to the Gaussian, the Laplacian is a sparsity-enforcing distribution: it concentrates more mass close to zero, but also has heavier tails. If \(P(s) = \prod_i t_i(s_i)\), then with Laplace potentials, the preference is for \(s\) to have most components very close to zero, allowing some components to be large, while with Gaussian potentials \(t_i\), no large \(s_i\) are tolerated, while there is also no pressure on the components to become very small. This notion is explained in more detail in [18], [12]. Our image prior employed here is the product of two potentials. The total variation potential is a product of Laplace terms on the image gradient (the vector of finite differences in the horizontal and vertical direction). If \(B_{tv} \in \mathbb{R}^{(n-\sqrt{2}) \times n}\) is the sparse structured matrix mapping \(u\) to its gradient, the potential is \(\exp(-\tau_{tv} \|B_{tv}u\|_1)\), where \(\|s\|_1 := \sum_j |s_j|\) denotes the \(L_1\) norm. The total variation potential encodes smoothness of images: neighbouring pixels tend to have similar gray-scale values, with occasional large differences due to edges, which agrees with the concentration at zero and the heavy tails of the Laplace density. The second transform sparsity potential encodes knowledge commonly employed in CS reconstruction methods, in that a wavelet transform of \(u\) should be sparse. If \(B_{sp} \in \mathbb{R}^{n \times n}\) is a multi-scale orthonormal wavelet transform, the potential is \(\exp(-\tau_{sp} \|B_{sp}u\|_1)\). In our experiments, we always use the Daubechies 4 wavelet [19].

\[B = (B_{tv}^T, B_{sp}^T)^T\], our setup becomes an instance of the model class of interest, where the Bayesian posterior

---

4 A potential (also called site here) is a non-negative function, and a prior distribution is commonly constructed as a product of one or more potentials. Distribution models of this sort are called undirected graphical models or Markov random fields [17].

5 \(P(u)\) is normalisable, because the transform sparsity potential is. Technically, the total variation potential is not normalisable on its own. However, it is still possible (and, in fact, works well) to use our method with \(\tau_{sp} = 0\), since in undirected graphical models, the “prior” \(P(u)\) need not be normalisable. In general, \(P(u)\) should not be understood as a sensible generator for natural images anyway, but rather as incorporating some important natural image characteristics.
distribution has the form
\[ P(u|y) \propto N(y|Xu, \sigma^2 I) \prod_{i=1}^{q} t_i(s_i), \quad s = Bu. \] (3)

For large numbers of image pixels \( n \), it is essential for computational efficiency that matrix-vector multiplications (MVMs) with \( B \) and \( B^T \) (less important: with \( X, X^T \)) can be done efficiently, and that \( B \) does not have to be stored explicitly. Our framework can readily be used with \( t_i(s_i) \) that are not Laplace. If the \( t_i \) are log-concave, as is the case here, our method can be shown to be numerically stable [18].

Many CS reconstruction methods (Section IV) can be understood as maximum a-posteriori (MAP) estimation
\[ \hat{u} = \arg\max_u \log P(u|y) = \arg\max_u \log P(y|u)P(u). \] (4)

Here, \( -\log P(y|u)P(u) \) is referred to as energy, and MAP estimation as energy minimisation. If \( -\log P(y|u) \) and \( -\log P(u) \) are convex in \( u \), as is the case for Gaussian and Laplace distributions, MAP estimation is a convex problem and can be solved efficiently. In this sense, the image prior constructed above is used in several CS estimation applications [20], which is the main reason for using it here as well. In contrast, the Bayesian estimate of \( u \) is given by the posterior mean
\[ E[u|y] = E_{P(u|y)}[u]. \]
Decision theory suggests that the posterior mean is a better estimate than the posterior mode, if the objective is to minimise the squared error [21, Chap. 4]. The mean is consistent under marginalisation (meaning that the Bayesian estimate of a part of the image is simply the corresponding part of the mean), while the mode is not. On the other hand, for the model considered here, no computationally tractable method for computing the exact mean is known (even though \( -\log P(u|y) \) is convex), and an approximation is harder to compute than solving for the mode (see Section III).

The problem of experimental design is to choose \( X \) among many candidates of the same cost, so that subsequent measurements allow the best reconstruction of \( u \). Importantly, the approach is at least partly “closed-loop”, in that it is not required to in fact do real measurements for most of the candidates. To understand this, keep in mind that (3) is only a model of the true measurement process, which however, combined with a growing number of real measurements, can successfully be used to predict the informativeness of new sampling not yet done. To do this, we need a quantitative statement about our uncertainty in \( u \) at the moment, which is the posterior \( P(u|y) \). An extension of our design means new rows in \( X \). Its informativeness is scored by imagining the new measurement being done with outcome \( y_\ast \), then measuring the decrease in uncertainty from \( P(u|y) \) to \( P(u|y, y_\ast) \). A common measure is the entropy difference \( \text{H}[P(u|y)] - \text{H}[P(u|y, y_\ast)] \). Since \( y_\ast \) is not known, it is integrated out using
\[ P(y_\ast|y) = \int P(y_\ast|u)P(u|y) du. \]
We now have information scores as criteria driving an optimisation of the design. It is clear that these are fundamentally based on a representation of uncertainty, the posterior in the Bayesian case, and that algorithms which merely estimate point solutions from given data, cannot be used directly in order to compute them. With such methods, either rules of thumb have to be followed to obtain a design (such as “do it at random”), or many measurements have to be taken in a trial-and-error fashion. The edge of Bayesian experimental design is that through a combination of the model and real measurements, a continuously refined uncertainty statement is obtained, based on which uninformative sampling can often be avoided. This way, often substantially fewer real measurements are required.
III. APPROSSIMATE INFERENCE

Bayesian inference is in general not analytically tractable for models of the form (3), and has to be approximated. Moreover, the application of interest here demands 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 [13], together with a robust and efficient representation for \(Q(u) \approx P(u|y)\). Our framework has previously been used in a different context [15], [18], 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 sped up by orders of magnitude.

Before we describe our approximation, we will give an intuitive view on what inference is about, and how algorithms to approximate it differ from estimation methods (see Section IV). Readers familiar with Bayesian methods may skip this paragraph. In many statistical problems — certainly the ones concerned with images — experience suggests that there are many potential constraints, which should to some degree be met by the underlying signal to be reconstructed. For example, observations imply constraints through likelihood terms, each of which may depend on all latent variables. Moreover, prior constraints for images are often local in nature, enforcing smoothness by constraining neighbouring pixels to have similar values, such as in the total variation potential described above. However, strictly enforcing all constraints is usually not possible, or leads to trivial solutions. Rather, the constraints have to be weighted against each other. In estimation methods, this constraint weighting is done in a rough way: either, some constraints have to be met (infinite weight), or the constraints are split in two groups (usually likelihood versus prior), with equal weighting within groups (see Section IV). In contrast, with Bayesian inference, all constraints are fundamentally probabilistic. An approximate inference method can be thought of as finding a proper weighting across all constraints, in an iterative process of negotiation between all model potentials. For example, in loopy belief propagation, applications of which have been revolutionary in coding and information theory [22], “messages” are exchanged between neighbouring sites (or potentials), until at convergence an equilibrium of mutual agreement is established. Importantly for our application here, these negotiation mechanisms are in terms of distributions (or beliefs), encoding uncertainties of sites about the state of neighbouring ones or about their own state. At convergence, these beliefs approximate posterior uncertainties, which in turn drive Bayesian experimental design. Moreover, we will see below how they can be used within the algorithm itself, in order to attain faster convergence. These additional information sources are not required, and therefore not present, in methods for pure MAP estimation.

In EP, the posterior \(P(u|y)\) is approximated by a Gaussian \(Q(u)\) with free (variational) parameters \(b, \pi\), which are formally introduced through replacing \(t_i(s_i)\) by \(\tilde{t}_i(s_i) := e^{b_i s_i - \pi_i s_i^2 / 2}\) in (3). Beyond \(b, \pi\), it is usually necessary to maintain a representation of \(Q\), so that marginal distributions \(Q(s_i)\) can be obtained rapidly. For an EP update at site \(i\), we compute the Gaussian moments of the tilted distributions

\[
\tilde{P}_i(u) \propto N(y| Xu, \sigma^2 I) \prod_{j \neq i} \tilde{t}_j(s_j) \tilde{t}_i(s_i)^{1-n} \tilde{t}_i(s_i)^{n},
\]
then update $Q(u)$ to match these moments. The latter can be done by modifying $(b_i, \pi_i)$ only. Here, $\eta \in (0, 1]$ is a fractional parameter\textsuperscript{6}. In each EP update, we merely need to compute mean and variance of the non-Gaussian marginal $\tilde{P}_i(s_i)$, and to update the $Q(u)$ representation to accommodate the novel $(b_i, \pi_i)$. As motivated above, the single updates form a process of negotiation between all sites $t_i(s_i)$, which is resolved at convergence, when the means and covariances of all $\tilde{P}_i$ are the same. Further details are given in the Appendix.

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

$$H[Q] - H[Q'] = \frac{1}{2} \log \left(1 + \sigma^{-2} x_*^T \text{Cov}_Q[u] x_*\right).$$

Since $\|x_*\| = 1$ by assumption, this score is maximised by choosing $x_*$ along the principal (leading) eigendirection of $\text{Cov}_Q[u]$, which can be calculated by the Lanczos method [23]. The same score is used in [1], yet the approximation of the posterior and its covariance is fundamentally different (see Section IV).

### A. Large Scale Applications

There are two major issues with trying to apply our method for large sizes $n$. First, the EP site updates are typically done in random sweeps over all sites, because it is not clear \textit{a priori} 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 $X, y$ (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 [18], which is used in the experiments here, requires $O(n^2)$ memory, and each update costs $O(n^2)$ (see Appendix). If $m \ll n$ at all times, a different representation of size $O(m^2)$ can be used. Beyond that, our method can also be run representation-free, requiring $O(n)$ storage only, if marginals are approximated by linear conjugate gradients and the Lanczos algorithm. Both of these extensions are described in [18]. In general, however, either of these modifications leads to a loss in numerical accuracy.

Our selective updating scheme for EP hinges on the fact that we can maintain all site marginals $h, \rho, Q(s_i) = N(h_i, \rho_i)$, up-to-date at all times. For a site $i$, we can quantify the change of $Q$ through an update there by $D[Q'_i(s_i) \| Q(s_i)]$ ($Q'_i$ the posterior after the update at $i$), which can be computed in $O(1)$. Here, $D[p(x) \| q(x)] = \int p(x) \log[p(x)/q(x)] dx$ is the relative entropy, or Kullback-Leibler divergence, measuring the gain in information from $q \rightarrow p$. Importantly, $D[Q'_i(u) \| Q(u)] = D[Q'_i(s_i) \| Q(s_i)]$, so the score precisely measures the global amount of change $Q \rightarrow Q'_i$. We maintain a list of candidate sites, which are scored before each EP update, and the update is done for the winner only. The list is then evolved by replacing the lower half of worst-scoring sites by others

\textsuperscript{6}$\eta = 1$ gives standard EP, but choosing $\eta < 1$ can increase the robustness of the algorithm on the sparse linear model significantly [18]. We use $\eta = 0.9$ in all our experiments.
randomly drawn from \( \{1, \ldots, q\} \). Importantly, the marginals \( h, \rho \) can be updated along with the representation of \( Q(u) \), at minor additional costs (see [18], Sect. 6.3).

Our sequential Bayesian design method is sketched in Algorithm 1. Here, \( d \) new rows are appended to \( X \) in each iteration (\( d = 3 \) in our experiments in Section V-B).

**Algorithm 1 Sequential Bayesian design method**

**Require:** Initial \( X, y, \tau_{sp}, \tau_{tv}, \sigma^2 \)

\[
\begin{aligned}
b &= 0, \quad \pi_{sp} = \frac{1}{2} \tau_{sp}^2 1, \quad \pi_{tv} = \frac{1}{2} \tau_{tv}^2 1.
\end{aligned}
\]

Compute initial \( Q \) representation, marginals \( h, \rho \)

repeat

\[
J = \{1, \ldots, q\} \quad \text{(for first update)}
\]

repeat

\[
\begin{aligned}
&\text{Compute } \Delta_i = D(\mathbb{Q}_i^\prime \| \mathbb{Q}) \text{ for all } i \in J, \text{ using } h, \rho. \\
&\text{EP update at site } \hat{i} = \arg\max_{i \in J} \Delta_i.
\end{aligned}
\]

Update of \( Q \) representation, marginals \( h, \rho \).

Evolve \( J \) (shrink to desired size after first iteration).

until \( \Delta_{\hat{i}} \) below threshold

Find \( X_* \in \mathbb{R}^{d \times n} : d \) leading unit norm eigendirections of \( \text{Cov}_Q[u] \) (Lanczos algorithm).

Measure image with \( X_* \rightarrow y_* \in \mathbb{R}^d \).

Append \((X_*, y_*)\) to \((X, y)\).

until \( X \) has desired size, or \( Q(u) \) has desired entropy

**IV. Related Work, Extensions**

In this section, we describe work related to ours, focusing on methods that we compare against in Section V-B. We also comment on constrained design optimisation within our framework.

Typically, CS reconstruction from incomplete measurements [3], [4] is done by minimizing a norm penalty under some observation constraints

\[
\hat{u} = \arg\min \{\|Bu\|_p \text{ s.t. } Xu = y\}, \quad p \in \{1, 2\}.
\]

Here, \( \|s\|_2 := \sqrt{s^T s} \) denotes the \( L_2 \) norm. Maximum sparsity in \( s = Bu \) is obtained for \( p = 0 \), yet this \( L_0 \) estimation problem is NP hard. If \( p = 1 \) is chosen instead, the corresponding solution can be found efficiently by solving a linear program. In highly sparse situations, this LP relaxation yields the exact solution to the \( L_0 \) problem [24]. In our experiments below, we consider several special cases. The simplest CS method (called \( L_1 \)) is obtained by choosing \( p = 1 \) and \( B = B_{sp} \) (the wavelet transform). It is also known as basis pursuit [25]. Classical least squares estimation (called \( L_2 \)) uses \( p = 2 \) and \( B = B_{sp} \). Since \( B \) is orthonormal, we have \( \|Bu\|_2 = \|u\|_2 \), and \( \hat{u} \) is given as solution of the normal equations: \( \hat{u} = X^T (XX^T)^{-1} y \).
We also consider a method with transform sparsity and total variation potential [20] (called $L_1 + TV$ here):

$$\hat{u} = \arg\min \{ \tau_{sp} \| B_{sp} u \|_1 + \tau_{tv} \| B_{tv} u \|_1 + (2\sigma^2)^{-1} \| y - Xu \|_2^2 \}.$$ 

Note that $L_1 + TV$ is the MAP estimator (4) for the same model we employ in our Bayesian method. It is also known as the Lasso [26]. $L_2$ and $L_1$ (5) can be seen as MAP estimators as well, if the noise variance $\sigma^2$ is set to zero, so that the likelihood constraints have infinite weight (see Section III).

The algorithm of [1] is called BCS. It comes with a transform sparsity potential only, so that $s = B_{sp} u$. BCS employs sparse Bayesian learning [12] in order to approximate Bayesian inference. This technique is specific to sparse linear models (all $t_i$ have to be Gaussian 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 [14]. We used the following code in our experiments:

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

A. Optimisation of Designs under Constraints

In our study on optimising image measurements, we assume that filters can be chosen anywhere on the unit sphere. In typical applications of this scenario, additional constraints have to be placed on the rows of $X$. For example, in magnetic resonance imaging, Fourier coefficients are measured along constrained paths in Fourier space. Or in digital photography, cameras may not be able to realise arbitrary filters $x_\ast$.

In many scenarios in practice, the number of candidates $x_\ast$ for the next measurement is finite and rather small [15]. In this case, it is easiest to score all candidates and pick the one maximising the information criterion. In one setup in Section V-B, we restrict our Bayesian experimental design technique to select among wavelet coefficient filters only. This case is very simple to deal with, since these coefficients feature in the transform sparsity prior potential. If $x_\ast = B_j^T$ is such a filter, then $x_\ast^T \text{Cov}[u] x_\ast$ is simply the variance of $Q(s_j)$, where $t_j(s_j)$ is the corresponding prior site. If selective site updating is used (see Section III-A), the variances for all these $s_j$ are maintained at all times, so the optimisation over all wavelet coefficient filters comes almost for free. Obviously, the marginals of any other set of linear projections of $u$ can be kept up-to-date alongside as well, independently of whether they feature in the sites of the model. Therefore, any extension of the setting considered here, based on a fixed candidate set, where the matrix containing all candidate filters as rows admits a fast matrix-vector product, can be implemented very efficiently.

However, in general the problem of maximising our information criterion, subject to further constraints, is not convex. The function $x_\ast^T \text{Cov}[u] x_\ast$ is convex in $x_\ast$, and the maximisation of a convex function, subject to convex constraints, can be hard. If the constraint set is a ball w.r.t. some Euclidean norm, centered at zero, the optimal $x_\ast$ is a (generalised) eigenvector, which is what we use in our setup here. In general, we recommend the simple
approach of keeping marginals up-to-date for a finite grid of candidate constraints, then to start some nonlinear
optimisation method from the maximiser $x^*$ on this grid.

V. Experiments

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

A. 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” [1], 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 that studies looking at the robustness of
CS theoretical claims are highly important. In this section, signals are sparse as such, so that $B = I$ and $u = s$ here.
We compare methods described in Section II and Section IV. It is important to stress that all methods compared
here (except for $L_2$) are based on exactly the same underlying model (3) with $B = I$, and differences arise only
in the nature of computations (approximate Bayesian inference versus maximum a-posteriori estimation), and in
whether $X$ is sequentially designed (EP, BCS) or chosen at random ($L_p$ reconstruction; we follow CS theory [3],
[4] and sample rows of $X$ uniformly of unit norm). Results are shown in Figure 1.

The “sparsity” (or sub-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 BCS (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. BCS 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 out-
performs all other methods, including designed BCS, with special benefits at rather small numbers of measurements.
BCS 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 BCS [1] seems to come from their desire
for “premature sparsification”. During their iterations, many $\pi_i$ 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 at present 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 $\pi_i$ 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.

B. Natural Images

In this section, we are concerned with finding linear filters which allow for good reconstruction of natural images from noisy measurements thereof. Natural images exhibit sparsity in a wavelet domain, fulfilling the basic requirement of CS. Theoretical results seem to suggest that measurement filters can be drawn at random, and there have been considerable efforts to develop hardware which can perform such random measurements cost-efficiently [27]. 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 coefficients, based on which natural image reconstruction should be more precise than for random measurements [6].

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 aims for a single wavelet coefficient, and the sequential ordering of the measurements is deterministic top-down, from coarse to fine scales\(^7\). This ordering is a pragmatic strategy: if mainly the coarse scale coefficients are far from zero, they should be measured first\(^8\). 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 II is used? Furthermore, how strong is the impact of the total variation potential? Note that no prior knowledge about typical ordering or dependence among wavelet coefficients is encoded in this model either.

Recall from Section II that every CS method has to address two problems: reconstruction of the signal $u$ from measurements $y$ for a fixed design $X$, and the choice of the design $X$. In our experiments, we pair five different reconstruction methods ($L_1$, $L_1 + TV$, $L_2$, BCS, and EP; see Section IV) with a number of non-adaptive (rand uni, rand colored, rand noiselet, heur wave) and adaptive (opt free, opt wave) measurement designs. The pairings we explored are summarised in Table I. For rand uni, entries are drawn uniformly at random: $X_{ij} \sim N(0, \frac{1}{n})$. For rand colored, filters are drawn respecting the second order structure of images. Inspired by [8], we applied a spectral

\(^7\)This 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.

\(^8\)Note 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.
Table I

<table>
<thead>
<tr>
<th>type of design X</th>
<th>reconstruction method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L₁</td>
</tr>
<tr>
<td>rand uni</td>
<td>–</td>
</tr>
<tr>
<td>rand colored</td>
<td>–</td>
</tr>
<tr>
<td>rand noiselet</td>
<td>–</td>
</tr>
<tr>
<td>heur wave</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>opt free</td>
<td>+</td>
</tr>
<tr>
<td>opt wave</td>
<td>+</td>
</tr>
</tbody>
</table>

low-pass filter to random Gaussian noise with a power spectrum decaying with \( f^{-2} \) [28]. For rand noiselet, we selected random rows of the noiselet transform [10], as was proposed for CS on images in [3]. We are grateful to Emmanuel Candès and Justin Romberg for providing us with their noiselet transform code. Finally, heur wave is the wavelet heuristic described above. While this heuristic is non-adaptive, in that the ordering is fixed in advance, we also considered the adaptive variant proposed in [11] (called heur Dekel below). We acknowledge Shai Dekel for sharing code and knowledge with us. The adaptive designs are both sequential, in that new rows \( x_\ast \) are added to \( X \) one at a time, based on all previous measurements. In opt free, the optimisation is done over all unit norm filters \( x_\ast \), while in opt wave, each filter has to correspond to a single wavelet coefficient. Note that opt wave is another adaptive alternative to the wavelet heuristic. The database for our study is a set of 75 natural gray-scale images frequently used in computer vision research (Figure 2), which were scaled to \( 64 \times 64 \) pixels. Results are given in the panels of Figure 3 (legend entries have the form “reconstruction method (type of design)”).

In the main panel a), we consider natural pairings: our Bayesian EP method, as well as BCS, with unconstrained experimental design (opt free), and current CS reconstruction methods (L₁, L₁ + TV) with randomly drawn measurement filters (rand uni). The wavelet heuristic is paired with least squares reconstruction (L₂). Note that EP(opt free) and L₂(heur wave) feature in all panels for reference. All methods in a) are started from ten initial filters drawn according to rand uni, except for BCS(opt free) which required 100 initial filters (rand uni) to attain a decent performance. The L₂ wavelet heuristic clearly outperforms all CS methods based on random designs. Among the latter, L₁ + TV does substantially better than L₁ or BCS, indicating the importance of the total variation prior potential\(^9\). Note that the BCS code supplied with [1] allows for a transform sparsity potential only. Moreover, our method EP(opt free) outperforms the wavelet heuristic, by selecting filters which are more informative than wavelet coefficients. Since EP(opt free) adjusts the design \( X \) specifically for each underlying image, it is natural to ask whether such designs are transferable to other images as well. In the setup EP(opt across), we reconstructed each image \( u \) using five measurement designs \( X \) adapted to different images (randomly chosen). The average

\(^9\)This is also witnessed in the scale parameters employed for the two potentials in EP: \( \tau_{sp} = 0.075, \tau_{tv} = 0.5 \). The total variation potential is much stronger. In fact, the results of EP with \( \tau_{sp} = 0, \tau_{tv} = 0.5 \) are only insignificantly worse.
reconstruction error is shown in a): as expected, it is slightly worse than for \( EP(\text{opt free}) \), yet still substantially better than the \( L_2 \) wavelet heuristic. Therefore, the filters found by \( EP(\text{opt free}) \) turn out to be transferable to other images, opening up the possibility to adapt such designs \textit{a priori}. Finally, the poor performance of BCS, compared to the simpler \( L_1 \) or \( L_1 + TV \), is remarkable.

In panel b), we consider other ensembles beyond \textit{rand uni}, which the designs \( X \) are drawn from. The random noiselet ensemble \textit{rand noiselet} proposed for CS in [3] has the theoretical advantage of being maximally incoherent with the Haar wavelet basis. Moreover, \( X \) does not have to be stored explicitly in this case, and MVMs with \( X \) or \( X^T \) can be computed very efficiently. There is no significant difference between \textit{rand uni} and \textit{rand noiselet} for \( L_1 + TV \). While the noiselet measurements lead to a more compact algorithm, they do not result in better reconstructions. The colored ensemble \textit{rand colored} results in filters which are more closely aligned with where signal energy is typically found. They lead to significant improvements over the uninformed ensembles, yet are again outperformed by the \( L_2 \) wavelet heuristic.

In panel c), we compare adaptive alternatives to the wavelet heuristic. The heuristic proposed in [11] does not improve upon \( L_2(\text{heur wave}) \) in our experiments. However, our EP method significantly outperforms the heuristic, even when constrained to measure wavelet coefficients only (see Section IV-A). The advantage may be due to EP choosing a better ordering of the coefficients, but also due to improved reconstruction (see also panel d). While \( EP(\text{opt free}) \) still outperforms the constrained variant \( EP(\text{opt wave}) \), we see that the design optimisation of our method is successful under structural constraints on the filters as well.

In panel d), we try to separate between reconstruction performance and the choice of measurement design. All methods shown there use the same wavelet heuristic design (except for \( EP(\text{opt free}) \), added for reference). First of all, \( L_2, L_1 \), and BCS provably give exactly the same reconstruction\(^{10} \) if \( X \) is a part of \( B_{sp} \). \( L_1 + TV \) and EP can do significantly better based on these measurements, while there is no significant difference between them. It is also interesting to compare \( EP(\text{heur wave}) \) here with \( EP(\text{opt wave}) \) in panel c). The latter does slightly better, although the major part of the improvement over \( L_2(\text{heur wave}) \) is due to EP being a better reconstruction method.

Intrigued by the fact that the wavelet heuristic with simple \( L_2 \) reconstruction outperformed all estimators based on random designs, we analysed their performance after giving them a warm-start, by supplying them with the first 100 and first 400 wavelet heuristic measurements. The results are shown in panel e) and f) respectively. In this setting, BCS with projection optimisation performed much worse than all other methods, the results are omitted to facilitate the comparison among the others. \( L_1 + TV \) profits from the warm-start to some extent, although its final performance (continuing with \textit{rand uni}) is worse than the \( L_2 \) wavelet heuristic. Both \( EP(\text{opt free}) \) and \( EP(\text{opt wave}) \) improve upon \( L_2(\text{heur wave}) \) from the moment they are allowed to choose filters by themselves, so the warm-start is in fact suboptimal for them. The deterioration of \( L_1 \) is rather striking, given that additional measurements provide novel information about the true \( u \). The failure is explained in the Appendix.

\(^{10}\)In this context, this is indeed a shortcoming of the measurements being exactly coherent to the transform sparsity prior potential (see also Appendix). However, note that BCS or \( L_1 \) do much worse based on all other designs tested here, also the maximally incoherent ones.
From these results we conclude, much as [6] argued on mostly theoretical grounds, that if natural images are to be measured successively by unit norm, but otherwise unconstrained linear filters, *drawing these filters at random leads to significantly worse reconstructions than standard wavelet coefficient filters top-down.* Moreover, the wavelet heuristic can be improved upon by adapting filters with our Bayesian experimental design technique. To put our findings into perspective, we note that the L2 wavelet heuristic is vastly faster to compute\(^\text{11}\) than all other methods considered here. Another finding is that the total variation potential seems to be more useful for natural images than the transform sparsity term. Our Bayesian design optimisation method, based on EP, can be used under structural constraints, and the designs can successfully be transferred to measure other images as well. CS theorems are mathematically intriguing, and there are certainly applications that benefit from these results, but linear image measurement is probably not among them.

Possible reasons for the failure of BCS on signals that are not truly sparse, were given in Section V-A. Premature sparsification, in light of not strictly sparse signals, leads to poor results even with random X. Their covariance estimates seem too poor to steer sequential design in a useful direction.

VI. DISCUSSION

We have shown how to address the compressed sensing problem with Bayesian experimental design, where designs are optimised to rapidly decrease uncertainty, rather than being chosen at random. In a large study 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 wavelet coefficients in a fixed ordering from coarse to fine scales. Our findings suggest that the impact of CS theoretical results 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 such as ours should be highly promising for CS applications in general.

Our best explanation for the marked differences between what theory seems to suggest, versus what is found in natural image applications in practice, is based on the explicit worst-case character of the theorems: while the signal is assumed to be sparse in some transform domain, no assumptions are made about where the non-zeros will lie. Moreover, the statements are usually of the minimax type, bounding the performance or success probability under the worst possible placing of the non-zero set. It is reassuring that random measurements and simple convex estimation methods are sufficient to give useful results within broad regimes of such a pessimistic setting. The impact in applications where high standards of security have to be met, or where adversarial signal constructions have to be detected, may be substantial. However, from practical statistics, it is well known that worst-case results are often not transferable to mathematically rather elusive “cases of practical interest”. While it is easy to see that experimental

\(^{11}\)EP sequential design is still very efficient. A typical run on one image took 53 min (on 64bit 2.53GHz AMD), for \(n = 4096\) and \(q = 12160\) sites: 16785 initial EP updates, then 308 increments of \(X\) by 3 rows each, with on average only 8.8 site updates needed to regain EP convergence (up to 85 updates after some increments).
design can fail badly in the worst case, a proper implementation often leads to significant cost reductions for non-adversarial tasks, whose properties can be modelled well. In minimax techniques, available prior knowledge can often be ignored, because the worst case may just as well be very unexpected. Moreover, making decisions about future sampling based on data observed so far, is usually not useful, because the “benign” assumptions underlying these techniques are violated in the worst case. It is therefore not reasonable to conclude from minimax results, or from results assuming the absence of any structure except for sparsity, that methods which perform close to optimal in these cases, set the standard in practice as well. In fact, while minimax CS theory requires $X$ and $B$ to be as “incoherent” w.r.t. each other as possible [3], and some methods strive for maximally incoherent designs [9], [3], on natural images, these methods are significantly outperformed by using wavelet coefficients in a certain ordering. The latter filters are rows of $B$, therefore maximally coherent with the sparsifying transform. If wavelet coefficients were sparse at random for the ensemble of natural images, incoherence would indeed be an important property of a measurement design. Since the sparsity of images is structured in a stable way, the completely coherent wavelet heuristic performs much better than worst-case optimal incoherent designs.

Our experience with the method of [1], which we compare against in our study, raises another interesting question. Several methods very frequently used in signal processing and 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, sparse Bayesian learning [12] is much more aggressive in this respect than our EP method here. In many cases, early sparsification seems to not hurt mean prediction performance much, and thus is generally embraced. 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 site updating method for EP is a step in that direction.

**APPENDIX**

$Q$ Representation, EP update

A numerically stable representation of $Q(u)$ maintains the $n \times n$ Cholesky factor $L$ and the $n$ vector $\gamma$, so that

$$LL^T = \sigma^{-2}X^TX + B^T\Pi B = \text{Cov}_Q[u]^{-1},$$

$$\gamma = L^{-1}(\sigma^{-2}X^Ty + B^Tb), \quad \Pi = \text{diag} \pi.$$ 

Note that $E_Q[u] = L^{-T}\gamma$. For an EP update at site $i$, we require $Q(s_i) = N(h_i, \rho_i)$, where $h_i = v^T\gamma$, $\rho_i = \|v\|^2$ with $v = L^{-1}B_i$. The back-substitution costs $O(n^2)$. The update constitutes in finding $b'_i, \pi'_i$, such that $\hat{P}_i(s_i)$ and $Q'(s_i)$ have the same mean and variance\(^{12}\). This moment matching is nontrivial to do in a numerically stable

\(^{12}\)The reader may wonder why an EP update, which is rather supposed to match joint mean and covariance of $\hat{P}_i(u)$ and $Q'(u)$, can be reduced to working on the marginals of $s_i$ only. This comes without loss of generality, see [18], Sect. 3.1.
manner, see [18], Appendix 1. Finally, $L$, $\gamma$ are updated, using numerical mathematics code for rank one Cholesky update/downdate, which costs $O(n^2)$ (see [18], Sect. 3.2).

For selective site updating, all marginals $h$, $\rho$ need to be present at all times (see Section III-A). This can be done using the Woodbury formula, as shown in [18], Sect. 6.3. The cost is two back-substitutions with $L$, rather than one only.

Failure of basis pursuit started with wavelet coefficients

In this section, we explain the failure of $L_1$ (basis pursuit) when started from coarse scale wavelet measurements. Initially, we have $X = B^T_{\gamma}, y = B^T_{\gamma}u + \varepsilon$ with $B$ the orthonormal wavelet transform and $I$ selecting coarse scale indices. Further, $\hat{s} = \arg\min\{\|s\|_1 | s_I = y\}$, thus $\hat{s}_I = y$ and $\hat{s}_{\backslash I} = 0$. Here, \(\backslash I\) is short for \(\{1, \ldots, n\} \backslash I\). If $x_\gamma$ is unit norm uniformly random, so is $v := Bx_\gamma$. A new measurement $(x_\gamma, y_\gamma)$ does not affect $\hat{s}_I = y$, but we need $\hat{s}_\backslash I = \arg\min\{\|s\|_1 | s_I = y\}$ such that $v^T_{\gamma} s_{\backslash I} = r := y_c - y^T_\gamma v_I$. If $i' = \arg\max\{|v_i| | i \notin I\}$, a solution is $\hat{s}_{i'} = r/v_{i'}$ and $\hat{s}_{\backslash I \cup \{i'\}} = 0$ (see below). But $i'$ is uniformly drawn from $\backslash I$, so the single new non-zero coefficient in $\hat{s}$ is most probably among the fine scales, since there are more of them. The initial choice of $\hat{s}_I = y$ cannot be revised. Novel observations are explained by wavelet coefficients randomly chosen from $\backslash I$ (at least initially), rather than preferring coarse scale ones, as the wavelet heuristic does.

We close with the derivation lacking above, using simplified notation. The minimum of $\|s\|_1$, such that $v^T s = r$, exists. Assume that $v \neq 0$, otherwise $s = 0$. Let $i' = \arg\max |v_i|$ (then, $v_{i'} \neq 0$). Suppose that $s_j \neq 0$ for $j \neq i'$. Now,

$$v_{i'} s_{i'} + v_j s_j = v_{i'} (s_{i'} + \frac{v_j}{v_{i'}} s_j) + v_j 0,$$

and $|s_{i'} + (v_j/v_{i'}) s_j| \leq |s_{i'}| + |v_j/v_{i'}| |s_j| \leq |s_{i'}| + |s_j|$, so that $\|s\|_1$ is not increased by setting $s_j = 0$ that way. Therefore, a minimiser is $s_{i'} = r/v_{i'}$, $s_{\backslash i'} = 0$ (unique if $i'$ is unique).

REFERENCES


Figure 1. Comparison on 6 random synthetic signals $\mathbf{u} \in \mathbb{R}^{512}$. Shown are $L_2$-reconstruction errors (mean±std.dev. over 100 runs). All methods start with same random initial $X (m = 40)$, then ("rand") add random rows, ("opt") optimise new rows sequentially. Noise variance $\sigma^2 = 0.005$, prior scale $\tau = 5$. BCS: [1], $L_p$: $L_p$ 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): $n_2$ of $u_i = 0, \frac{n}{4}$ exponential decay $1, \ldots, 0, \frac{n}{4}$ 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, $u_i \sim \frac{1}{n} + |t|, t \sim N(0, 1)$; as in [1]. Distributions in (d-f) normalised to unit variance.
Figure 2. We benchmarked the algorithms on 75 images frequently used in computer vision research. The bitmaps were obtained from http://decsai.ugr.es/cvg/dbimagenes/g512.php.
Figure 3. Experiments for measuring natural images of size $64 \times 64 = 4096$ pixels depicted in Figure 2. Shown is $L_2$-reconstruction error averaged over 75 gray-scale images ($\pm$std.dev./$\sqrt{75}$ for “*”). Noise level $\sigma^2 = 0.005$. BCS: $L_p$ reconstruction $p \in \{1, 2\}$, $L_1 + TV$: Lasso with TV/wavelet penalties, EP: our method. True $\sigma^2$ supplied, $\tau$ 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 $x_\ast$ of $X$ random unit norm (rand), actively designed (opt), according to wavelet heuristic (heur wave).

(a): Start from $m = 10$ with $X$ random uniform. (b): Comparison for $X$ drawn from different measurement ensembles. (c): Optimisation restricted to wavelet coefficients. (d) Different reconstruction methods based on same measurements (heur wave). (e,f): Start from $m = 100, 400$ with $X$ according to wavelet heuristic. See Table I for a complete list.

April 1, 2008