Kalman Filtering for Compressed Sensing

Dimitri Kanevsky[1], Avishy Carmi[2,3], Lior Horesh[1], Pini Gurfil[2],
Bhuvana Ramabhadran[1], Tara N Sainath[1]
[1] *IBM T. J. Watson, Yorktown, NY 10598, USA*
[2] *Technion - Israel Institute of Technology, Haifa 32000, Israel*
[3] *Ariel University Center, Ariel 40700, Israel*

Abstract − *Compressed sensing is a new emerging field dealing with the reconstruction of a sparse or, more precisely, a compressed representation of a signal from a relatively small number of observations, typically less than the signal dimension. In our previous work we have shown how the Kalman filter can be naturally applied for obtaining an approximate Bayesian solution for the compressed sensing problem. The resulting algorithm, which was termed CSKF, relies on a pseudo-measurement technique for enforcing the sparseness constraint. Our approach raises two concerns which are addressed in this paper. The first one refers to the validity of our approximation technique. In this regard, we provide a rigorous treatment of the CSKF algorithm which is concluded with an upper bound on the discrepancy between the exact (in the Bayesian sense) and the approximate solutions. The second concern refers to the computational overhead associated with the CSKF in large scale settings. This problem is alleviated here using an efficient measurement update scheme based on Krylov subspace method.*

Keywords: Compressed sensing, Kalman filter, Krylov subspace method

1 Introduction

Recent studies have shown that sparse signals can be recovered accurately using less observations than what is considered necessary by the Nyquist/Shannon sampling principle; the emergent theory that brought this insight into being is known as compressed sensing (CS) [1, 2]. The essence of the new theory builds upon a new data acquisition formalism, in which compression plays a fundamental role. From a signal processing standpoint, one can think about a procedure in which signal recovery and compression are carried out simultaneously, thereby reducing the amount of required observations. Sparse, and more generally, compressible signals arise naturally in many fields of science and engineering. A typical example is the reconstruction of images from under-sampled Fourier data as encountered in radiology, biomedical imaging and astronomy [3, 4]. Other applications consider model-reduction methods to enforce sparseness for preventing over-fitting and for reducing computational complexity and storage capacities. The reader is referred to the seminal work reported in [1] and [2] for an extensive overview of CS theory.

The recovery of sparse signals is in general NP-hard [1, 5]. State-of-the-art methods for addressing this optimization problem commonly utilize convex relaxations, non-convex local optimization and greedy search mechanisms. Convex relaxations are used in various methods such as LASSO [6], the Dantzig selector [7], basis pursuit and basis pursuit de-noising [8], and least angle regression [9]. Non-convex optimization approaches include Bayesian methodologies such as the relevance vector machine, otherwise known as sparse Bayesian learning [10], as well as stochastic search algorithms that are mainly based on Markov chain Monte Carlo techniques [11–14]. Notable greedy search algorithms are the matching pursuit (MP) [15], the orthogonal MP [16], and the orthogonal least squares [17].

CS theory has drawn much attention to the convex relaxation methods. It has been shown that the convex l_1 relaxation yields an exact solution to the recovery problem provided two conditions are met: 1) the signal is sufficiently sparse, and 2) the sensing matrix obeys the so-called restricted isometry property at a certain level. A complementary result guarantees high accuracy when dealing with noisy observations, yielding recovery 'with overwhelming probability'. To put it informally, it is likely for the convex l_1 relaxation to yield an exact solution provided that the involved quantities, the sparseness degree s, and the sensing matrix dimensions $m \times n$ maintain relation of the type $s = \mathcal{O}(m/\log(n/m))$.

Recently, a Bayesian CS approach has been introduced in [18]. As opposed to the conventional non-Bayesian methods, the Bayesian CS has the advantage of providing the complete statistics of the estimate in the form of a posterior probability density function

(pdf). Adopting this approach, however, suffers from the fact that rarely one can obtain a closed-form expression of the posterior and therefore approximation methods should be utilized.

In this paper we extend our previous work in [19] where we have presented the so-called CSKF which is a novel approximate Bayesian CS scheme based on the Kalman filter. There are two major issues which were left unanswered in [19] that are addressed inhere. The first one concerns the validity of the so-called pseudo-measurement approximation technique used by the CSKF to enforce the sparseness-promoting prior. The second issue is related to the excessive computational overhead associated with the Kalman filter update stage in large scale settings. Our endeavor in providing a proper answer to the first concern involves a unique type of sparseness-promoting prior, termed here semi-Gaussian owing to its Gaussian-like characteristics. We show how an approximate variant of this prior is, in fact, the probabilistic interpretation of the pseudo-measurement constraint. This in turn allows us to formulate an upper bound on the discrepancy between the exact and the approximate posterior pdfs based exclusively on the computed estimation error covariance. Finally, we provide a computational scheme that relies on Krylov subspace method for alleviating the workload associated with the CSKF measurement update stage which thereby renders it highly efficient in large scale CS settings.

2 Sparse Signal Recovery

This section, which is taken from [19], briefly overviews the main concepts underlying compressed sensing.

Consider an $\mathbb{R}^n$-valued random discrete-time process $\{x_k\}_{k=1}^{\infty}$ that is sparse in some known orthonormal sparsity basis $\psi \in \mathbb{R}^{n\times n}$, that is $z_k = \psi^T x_k$, $s := \#\{\text{supp}(z_k)\} < n$, where $\text{supp}(z_k)$ and $\#$ denote the support of z_k and the cardinality of a set, respectively (i.e., $\#\{\text{supp}(z_k)\}$ denotes the number of non-zero elements of z_k). Assume that z_k evolves according to

$$z_{k+1} = Az_k + w_k, \quad z_0 \sim \mathcal{N}(\mu_0, P_0) \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ is the state transition matrix and $\{w_k\}_{k=1}^{\infty}$ is a zero-mean white Gaussian sequence with covariance $Q_k \succeq 0$. Note that (1) does not necessarily imply a change in the support of the signal. For example, A can be a block-diagonal matrix decomposed of A^d and A^n corresponding to the statistically independent elements $z^d \notin \text{supp}(z_k)$ and $z^n \in \text{supp}(z_k)$ where the respective noise covariance sub-matrices satisfy $Q^d = 0$ and $Q^n \succeq 0$. The process x_k is measured by the $\mathbb{R}^m$-valued random process

$$y_k = Hx_k + \zeta_k = H'z_k + \zeta_k \tag{2}$$

where $\{\zeta_k\}_{k=1}^{\infty}$ is a zero-mean white Gaussian sequence with covariance $R_k \succ 0$, and $H := H'\psi^T \in \mathbb{R}^{m\times n}$.

Letting $\mathscr{Y}_k := [y_1, \ldots, y_k]$, our problem is defined as follows. We are interested in finding a $\mathscr{Y}_k$-measurable estimator, $\hat{x}_k$, that is optimal in some sense. Often, the sought after estimator is the one that minimizes the mean square error (MSE) $E\left[\| x_k - \hat{x}_k \|_2^2\right]$. It is well-known that if the linear system (1), (2) is observable then the solution to this problem can be obtained using the Kalman filter (KF). On the other hand, if the system is unobservable, then the regular KF algorithm is useless; if, for instance, $A = I_{n\times n}$, then it may seem hopeless to reconstruct x_k from an under-determined system in which $m < n$ and $\text{rank}(H) < n$. Surprisingly, this problem may be circumvented by taking into account the fact that z_k is sparse.

2.1 The Combinatorial Problem and Compressed Sensing

Refs. [1,5] have shown that in the deterministic case (i. e., when z is a parameter vector), one can accurately recover z (and therefore also x, i.e., $x = \psi z$) by solving the optimization problem

$$\min \| \hat{z} \|_0 \ \text{ s.t. } \sum_{i=1}^{k} \| y_i - H'\hat{z} \|_2^2 \leq \epsilon \tag{3}$$

for a sufficiently small ϵ, where $\| v \|_p = \left(\sum_{j=1}^{n} v_j^p\right)^{1/p}$ is the l_p-norm of v, and the zero-norm, $\| v \|_0$, is defined as[1] $\| v \|_0 := \#\{\text{supp}(v)\}$.

Following a similar rationale, in the stochastic case the sought-after optimal estimator satisfies [2]

$$\min \| \hat{z}_k \|_0 \ \text{ s.t. } E_{z_k|\mathscr{Y}_k}\left[\| z_k - \hat{z}_k \|_2^2\right] \leq \epsilon \tag{4}$$

Unfortunately, the above optimization problems are NP-hard and cannot be solved efficiently. Recently, it has been shown that if the sensing matrix H' obeys a so-called *restricted isometry property* (RIP) then the solution of the combinatorial problem (3) can almost always be obtained by solving the constrained convex relaxation [1,2]

$$\min \| \hat{z} \|_1 \ \text{ s.t. } \sum_{i=1}^{k} \| y_i - H'\hat{z} \|_2^2 \leq \epsilon \tag{5}$$

This is a fundamental result in CS theory [1,2]. The main idea is that the convex l_1 minimization problem can be efficiently solved using a myriad of existing methods, such as LASSO [6], LARS [9], Basis pursuit [8], orthogonal matching pursuit [16] and relevance vector machine [10]. Other insights provided by CS are related to the construction of sensitivity matrices that satisfy the RIP. These underlying matrices are random

[1] For $0 \leq p < 1$, $\| v \|_p$ is not a norm; the common terminology is *zero norm* for $p = 0$ and *quasi-norm* for $0 < p < 1$.

by nature, which sheds a new light on the way observations should be sampled. For an extensive review of CS the reader is referred to [1, 2].

3 The CS-Embedded KF

The CSKF algorithm is aimed at solving a stochastic CS problem of the form

$$\min_{\hat{z}_k} E_{z_k|\mathscr{Y}_k}\left[\| z_k - \hat{z}_k \|_2^2\right] \text{ s.t. } \| \hat{z}_k \|_1 \leq \epsilon' \tag{6}$$

It can be shown that for proper values of the tuning parameters ϵ' and ϵ, the solution of both (6) and the convex l_1 relaxation of (4) coincide. The nonlinear inequality constraint in (6) is readily treated within the conventional KF framework using a so-called pseudo-measurement (PM) technique [19]. In practice this is carried out by recasting this constraint as

$$0 = \bar{H}_k z_k - \epsilon', \quad \bar{H}_k := \text{sign}(z_k) \tag{7}$$

where $\text{sign}(z_k)$ denotes a row vector consisting of the signs of the entries in z_k. This formulation facilitates the implementation of the standard KF update equations which are provided in Algorithm 1 for the sake of completeness. Notice that this approach assumes that

Algorithm 1 CSKF PM update stage

1: Set $\hat{z}^1 = \hat{z}_{k|k}$ and $P^1 = P_{k|k}$ (the posterior mean and covariance at time k).
2: **for** $\tau = 1, 2, \ldots, N_\tau - 1$ iterations **do**
3:

$$\bar{H}_\tau = \text{sign}(\hat{z}^\tau) \tag{8a}$$

$$\hat{z}^{\tau+1} = \left(I - \frac{P^\tau \bar{H}_\tau^T \bar{H}_\tau}{\bar{H}_\tau P^\tau \bar{H}_\tau^T + \sigma^2}\right)\hat{z}^\tau \tag{8b}$$

$$P^{\tau+1} = \left(I - \frac{P^\tau \bar{H}_\tau^T \bar{H}_\tau}{\bar{H}_\tau P^\tau \bar{H}_\tau^T + \sigma^2}\right)P^\tau \tag{8c}$$

4: **end for**
5: Set $\hat{z}_{k|k} = \hat{z}^{N_\tau}$ and $P_{k|k} = P^{N_\tau}$.

ϵ' is a random quantity, specifically a zero-mean Gaussian random variable with variance σ^2.

3.1 Bayesian Interpretation

The PM constraint (7) provides a convenient way for incorporating the nonlinear l_1 inequality within the KF. Its tightness, and therefore the overall estimation accuracy, is controlled by both σ^2 and the total number of iterations N_τ per measurement update. In [19] we have mentioned that the tradeoff between the associated computational complexity and the obtained accuracy can be regulated by properly setting these two parameters. In this work we provide a supporting result to this argument. Specifically, we derive an upper bound on the estimation accuracy which alludes the relation with the PM variance σ^2. In order to meet this end we shall restrict ourselves inhere to the conventional static linear model used in CS theory (i.e., $A_k = I_{n\times n}$ and $Q_k = 0_{n\times n}$ in (1)).

Our derivations in this section, which are concluded with the formulation of an upper bound on the discrepancy between the exact and the estimated posterior pdfs, rely on the Bayesian formulation of the KF update stage. For that reason we take the time to revisit some well-established elementary results from estimation theory. Under the restrictions above, the posterior pdf of the random parameter vector z conditioned on $\mathscr{Y}_k$ can be sequentially computed via the well-known Bayesian recursion

$$p(z \mid \mathscr{Y}_k) = \frac{p(y_k \mid z)p(z \mid \mathscr{Y}_{k-1})}{\int p(y_k \mid z)p(z \mid \mathscr{Y}_{k-1})dz} \tag{9}$$

where the likelihood $p(y_k \mid z) = p_{\zeta_k}(y_k - H'z)$. Because a Gaussian pdf is self-conjugate it readily follows that whenever the pdfs in the right-hand side of (9) are Gaussian the posterior can be described exclusively by its first two moments. These are given by the KF update equations

$$\hat{z}_k = \hat{z}_{k-1} + P_{k-1}H'^T\left(H'P_{k-1}H'^T + R_k\right)^{-1}[y_k - H'\hat{z}_{k-1}] \tag{10a}$$

$$P_k = \left[I - P_{k-1}H'^T\left(H'P_{k-1}H'^T + R_k\right)^{-1}H'\right]P_{k-1} \tag{10b}$$

where $P_k = E\left[(z-\hat{z}_k)(z-\hat{z}_k)^T \mid \mathscr{Y}_k\right]$ and $\hat{z}_k = E[z \mid \mathscr{Y}_k]$. It can be easily shown that in this case $\hat{z}_k$ is not only the minimum MSE estimator but also the maximum *a posteriori* (MAP) estimator, that is

$$\begin{aligned}\hat{z}_k &= \arg\max_z \log p(z \mid \mathscr{Y}_k) \\ &= \arg\min_z \sum_{i=1}^{k} \| y_i - H'z \|_{R_i}^2 + \| z - \mu_0 \|_{P_0}^2\end{aligned} \tag{11}$$

where $\| a \|_R^2 := a^T R^{-1} a$.

3.1.1 Semi-Gaussian Priors

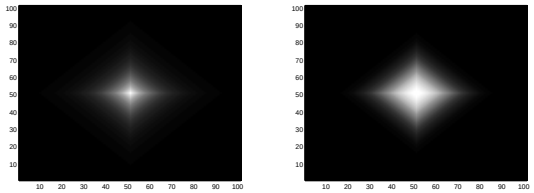
Compressed sensing was embedded in the framework of Bayesian estimation by utilizing sparseness-promoting priors such as Laplace and Cauchy [18]. As opposed to the conventional CS methods, which provide a point estimate, the Bayesian approach yields the statistics $p(z \mid \mathscr{Y}_k)$. We have seen that the posterior pdf is entirely described by (10) in the linear Gaussian case. When resorting to non-Gaussian (sparseness-promoting) priors, however, this recursion is strictly inadequate.

Let us now consider a new type of sparseness-promoting prior, which we will term *semi-Gaussian*

$$p(z) = c\exp\left(-\frac{1}{2}\frac{\| z \|_1^2}{\sigma^2}\right) \tag{12}$$

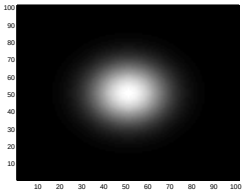
Compared with the Laplace distribution, the semi-Gaussian pdf concentrates greater portion of its mass in

the vicinity of the origin. This is illustrated in Fig. 1, in which the level maps are shown for Laplace, semi-Gaussian and Gaussian pdf's in the 2-dimensional case.



(a) Laplace

(b) Semi-Gaussian



(c) Gaussian

Figure 1. Laplace, semi-Gaussian and Gaussian pdf's in the 2-dimensional case.

Suppose for a moment that one embeds this prior within the Bayesian recursion (9) for promoting the sparseness of z. This in turn renders the closed form recursion (10) inadequate. This follows readily from the fact that the restrictions under which (10) are derived involve a purely-Gaussian prior and a likelihood pdf that is based on a deterministic sensing matrix H',

$$p(y_k \mid z) \propto \exp\left(-\frac{1}{2}(y_k - H'z)^T R_k^{-1}(y_k - H'z)\right) \tag{13}$$

Nevertheless, it turns out that the unique form (12) facilitates an approximate computation of the posterior $p(z \mid \mathscr{Y}_k)$ using the closed form recursion (10). As it is shown next, the approximate semi-Gaussian prior is closely related to the PM constraint (7) used by the CSKF.

3.1.2 Approximate Semi-Gaussian

The semi-Gaussian prior can be expressed based on the likelihood (13) using a sensing matrix similar to the one in (7). Thus, letting $\bar{H} := \mathrm{sign}(z)$ and setting $y = 0$ in (13) yields

$$p(z) = p(y = 0 \mid z) \propto \exp\left(-\frac{1}{2}\frac{(0 - \bar{H}z)^2}{\sigma^2}\right) \tag{14}$$

Although (14) resembles a Gaussian pdf its in fact a semi-Gaussian owing to the dependency of $\bar{H}$ on z. As the KF update stage (10) permits only deterministic, albeit varying, sensing matrices, (14) cannot be embedded straightforwardly. Nevertheless, its Gaussian approximation, in which the sensing matrix takes the form $\bar{H} := \mathrm{sign}(\hat{z}_k)$ (i.e., by substituting the estimator $\hat{z}_k$ rather than z), can be easily processed via (10). This modification renders $\bar{H}$ a $\mathscr{Y}_k$-measurable quantity, as it depends on $\hat{z}_k$, which is a function of the entire observation set.

The reader has probably realized by now that embedding the approximate semi-Gaussian prior within the Bayesian recursion (9) merely yields the PM stage described in Algorithm 1. Bearing in mind this, the unique formulation (14) allows us to assess the validity of the PM constraint (7) when enforcing the semi-Gaussian prior (12). Thus, the following theorem establish an upper bound on the discrepancy between the exact posterior, which relies on the semi-Gaussian prior (12), and the approximate posterior (i.e., the one that is computed by the CSKF) in terms of the CSKF estimation error covariance P_k.

Theorem 1. *Let $\hat{p}(z \mid \mathscr{Y}_k)$ be the Gaussian posterior pdf obtained by using the approximate semi-Gaussian prior. Let also $p(z \mid \mathscr{Y}_k)$ be the posterior pdf obtained by using the exact semi-Gaussian prior* (12)*. Then*

$$\mathrm{KL}\left(\hat{p}(z \mid \mathscr{Y}_k) \parallel p(z \mid \mathscr{Y}_k)\right) = \mathcal{O}\left(\sigma^{-2} \max\left\{\mathrm{Tr}(P_k), \mathrm{Tr}(P_k)^{1/2}\right\}\right) \tag{15}$$

where KL *and* Tr *denote the Kullback-Leibler divergence and the matrix trace operator, respectively.*

The proof is provided in the Appendix.

3.2 Discussion

The fundamental observation conveyed by Theorem 1 is that the approximation error of the PM stage in Algorithm 1 is affected by both the prior variance σ^2 and the estimation error covariance P_k. Consequently, regulating these factors is beneficial in getting close to the exact CS solution in the Bayesian sense. This can be attained by either increasing σ or having a sufficiently small P_k. The former approach, however, might bring upon an adverse effect as the sparseness constraint is less restrictive in such cases.

A prominent advantage of the bound (15) is that it involves quantities that are either known (σ^2) or computed (P_k). Whereas σ is a predetermined tuning parameter, the estimation error covariance P_k, which is computed at every step, decreases rapidly as $k \to \infty$ (in the sense that $P_k - P_{k+1} \succ 0$). This fact readily follows upon recognizing that (10b) and (8c) translate into $P_k^{-1} = P_{k-1}^{-1} + H'^T R_k^{-1} H'$ and $(P^{\tau+1})^{-1} = (P^{\tau})^{-1} + \sigma^{-2}\bar{H}^T\bar{H}$ by the matrix inversion lemma. Combining this observation with Theorem 1 explains why it is advantageous to perform N_τ consecutive updates of (8) using $\sigma = \sqrt{N_\tau}\sigma'$ rather than a single update with σ'.

It should be clarified that P_k does not represent the exact estimation error covariance (i.e., the one which is based on the semi-Gaussian prior (12)). Nevertheless, as stated by Theorem 1, the trace of this matrix indicates the proximity of the obtained solution to the exact Bayesian one. An upper bound on the exact estimation error covariance can be obtained based on the result in [2] (Theorem 4.1). Thus, assuming z is at most s-sparse and H' satisfies the RIP at the level $\delta_{3s}+\delta_{4s} < 2$, yields

$$\mathrm{Tr}\; E\left[(z-\hat{z}_k)(z-\hat{z}_k)^T \mid \mathscr{Y}_k\right] \leq \left(c_1\sqrt{\mathrm{Tr}(R_k)} + c_2 e\right)^2 \tag{16}$$

where e denotes the reconstruction error in the noiseless case. For reasonable values of δ_{4s}, the constants c_1 and c_2 in (16) are well behaved [2].

4 Computationally Efficient KF Update

When the state dimension n is large the computation of P_k becomes prohibitively expensive. In such cases we would like to avoid the explicit construction of large-scale dense matrices and their inverses. In addition, we recognize that exact computation of inverses is redundant in the course of defining a search direction within an optimization framework, especially when one is far from the solution.

Both objectives can be achieved by employing a Krylov subspace method. It is a family of iterative methods that are extremely memory efficient, as only storage of the previous search direction and the local gradient requires for forming a new search direction. The Krylov subspace solver requires only implicit access to the matrices via matrix vector products [20]. Unlike the common situation when such approach is practiced directly, here the difficulty is that matrix inverses are nested within each other, as shown by Equation (10). For that reason we first recast the initial estimate of z from (10) as follows. We define an auxiliary function that computes the product of $H'P_{k-1}H'^T + R_{k-1}$ by a vector v, which obeys the following property

$$\mathcal{H}v = H'P_{k-1}(H'^T v) + R_{k-1}v \tag{17}$$

Notice, that the bracketed product of H' by v, abstain the need for explicit formulation of the large-scale matrix $H'P_{k-1}H'^T$. The inversion of the matrix $H'P_{k-1}H'^T + R_{k-1}$ can now be obtained explicitly by solving the system

$$\mathcal{H}v = y_k - H'\hat{z}_{k-1} \tag{18}$$

In practice the above equation can be solved via the conjugate gradient method. Once (18) is solved it can be plugged back to yield the updated KF mean

$$\hat{z}_k = \hat{z}_{k-1} + P_{k-1}H'^T v \tag{19}$$

In a similar manner, the covariance update stage is carried out as follows

$$P_k = P_{k-1} - P_{k-1}H'^T M \tag{20}$$

where the columns m_i of the matrix $M \in \mathbb{R}^{m\times n}$ are obtained by solving

$$\mathcal{H}m_i = (H'P_{k-1})_i, \quad i = 1,\ldots,n \tag{21}$$

where $(H'P_{k-1})_i$ denotes the ith column of $H'P_{k-1}$.

5 Simulation Study

We demonstrate the previously described concepts using simple recovery examples taken from both [7] and [19]. In the first example we validate the relation conveyed by Theorem 1 which essentially implies that as the tuning parameter σ in (14) becomes larger the obtained approximate solution is closer to the exact one in the Bayesian sense. Thus, we would expect that increasing σ will result in smaller estimation errors.

An adequate measure of sparse reconstruction accuracy, typically known as the *ideal normalized estimation error*, is given in [7]. This measure, which is defined as

$$\parallel z - \hat{z}_k \parallel_2^2 \,/ \sum_{i=1}^{n} \min\left((z^i)^2, \mathrm{Tr}(R)/m\right) \tag{22}$$

where z^i denotes the ith element of z, quantifies the discrepancy between the attained estimation error and the ideal one which would have been obtained in the absence of measurement noise.

In this example we consider a sensing matrix $H \in \mathbb{R}^{72\times 256}$ of which the entries are samples from a zero-mean Gaussian distribution with variance $1/72$. This type of matrix is known to satisfy the RIP with high probability for sufficiently sparse signals [2]. The sparseness degree of the actual parameter vector z is set as $s = 10$.

The CSKF recovery errors in this case are shown for various values of σ in Fig. 2. This figure clearly demonstrates the idea underlying Theorem 1 as the attained estimation error (here after $N_\tau = 1000$ iterations of the PM stage) reduces as σ increases.

The two complementary panels in Fig. 3 illustrate the reconstruction performance of the CSKF when the parameter σ takes either of the two extremal values in Fig. 2, namely, $\sigma = 100$ and $\sigma = 800$. These rather qualitative figures visualize the estimation performance corresponding to the extremal values of the ideal normalized estimation error measure (as appears in Fig. 2).

The performance of the computationally efficient CSKF, which embeds the Krylov solver method, is assessed in two distinct settings. These consist of sensing matrices of the size 512×1024 and 3072×6144 which are composed similarly to the abovementioned

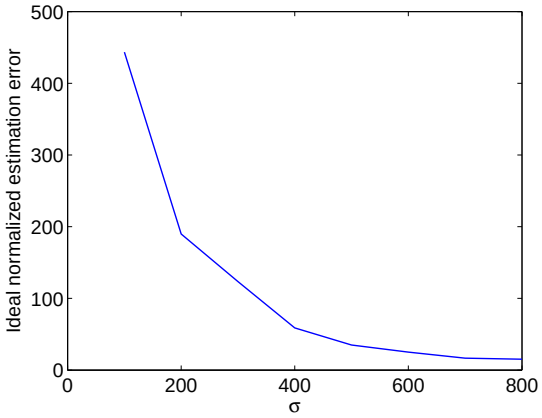


Figure 2. The recovery errors attained for various values of the tuning parameter σ. The problem dimension in this case is $H' \in \mathbb{R}^{72\times 256}$.

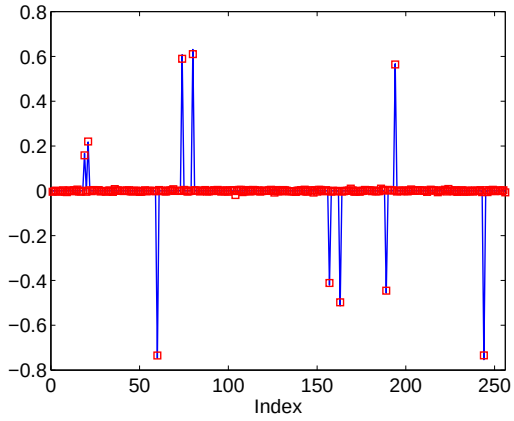
one. The accuracy of recovery is examined for various sparseness degrees s corresponding to the recovery index $s/m\log(n)$ which is derived from the relation $m \leq c(s\log n)$ [2] (for some $c > 0$). This index essentially refers to the probability of recovery assuming the sensing matrix obeys the RIP up to a certain sparseness degree. As it was already pointed out in [2], as this measure increases an exact recovery becomes less probable. In all runs the Krylov solver method is employed as described in Section 4 using Matlab's conjugate gradient (CG) function (assuming a tolerance of 10^{-3} and a maximal number of 200 iterations).

The results of this example are given in Fig. 4 where the estimation errors of both the CSKF and its computationally efficient variant (CSKF+CG) are depicted for various recovery indices. From this figure it can be clearly seen that the Krylov solver method does not significantly affect the reconstruction accuracy of the CSKF.

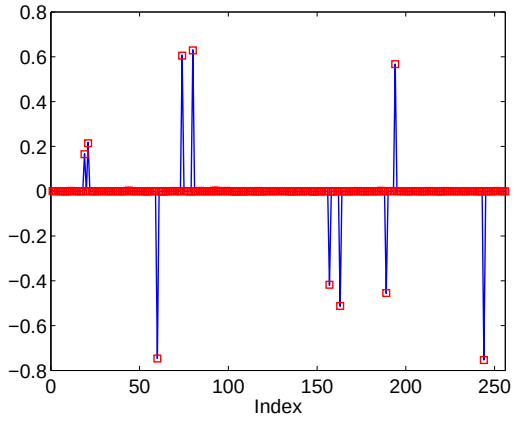
The prominent advantage of using the Krylov solver method is manifested in Fig. 5 where the timing data of both methods, the CSKF and its computational efficient variant, is shown for an increasing state dimension n. The sensing matrix used in this example is composed of samples from a zero-mean Gaussian distribution with variance $1/m$ with $m = n/2$ rows. In all runs, the sparseness degree s does not exceed $n/6$.

6 Conclusions

We have shown that the solution computed by the CSKF, which is a Kalman filtering-based approximate Bayesian compressed sensing algorithm, approaches the exact Bayesian solution as the PM tuning parameter σ increases. This result further justifies the iterative PM stage used by the CSKF (since as σ increases the sparseness constraint becomes less restrictive which thereby implies that more iterations are required).



(a) Ideal recovery error 443. $\sigma = 100$



(b) Ideal recovery error 15. $\sigma = 800$

Figure 3. The actual (blue line) and reconstructed (red squares) signals for two extremal values of the parameter σ.

Embedding the Krylov solver method within the CSKF/Kalman filter measurement update stage dramatically reduces the computational overhead, in particular, when the state dimension is prohibitively large.

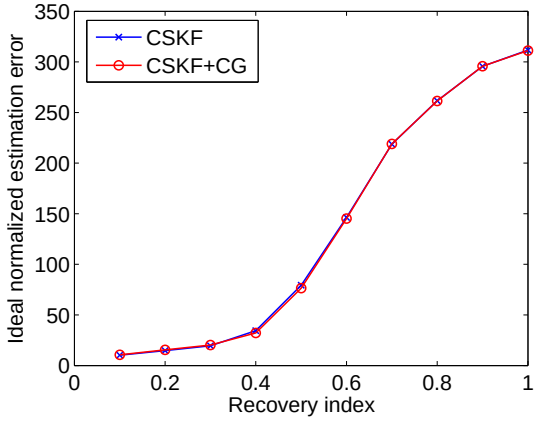
A Proof of Theorem 1

The exact and the approximate posterior pdfs are given by

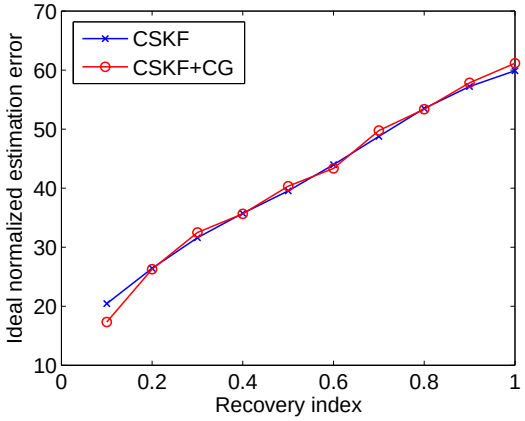
$$p(z \mid \mathscr{Y}_k) = c\exp\left(-\frac{1}{2}\frac{\| z \|_1^2}{\sigma^2}\right) p'(z \mid \mathscr{Y}_k) \qquad (23a)$$

$$\hat{p}(z \mid \mathscr{Y}_k) = \hat{c}\exp\left(-\frac{1}{2}\frac{(\bar{H}z)^2}{\sigma^2}\right) p'(z \mid \mathscr{Y}_k) \qquad (23b)$$

where c and $\hat{c}$ denote the appropriate normalization constants, and $p'(z \mid \mathscr{Y}_k)$ stands for the Gaussian posterior excluding the prior. Now, explicitly writing the



(a) 512×1024



(b) 3072×6144

Figure 4. The estimation errors of the CSKF with and without the Krylov solver method for various recovery indices. Showing the performance for two problem dimensions, 512×1024 **and** 3072×6144**.**

KL divergence between these two pdfs yields

$$\begin{aligned}
&\mathrm{KL}\left(\hat{p}(z \mid \mathscr{Y}_k) \parallel p(z \mid \mathscr{Y}_k)\right) \\
&\qquad = \int \hat{p}(z \mid \mathscr{Y}_k) \log \frac{\hat{p}(z \mid \mathscr{Y}_k)}{p(z \mid \mathscr{Y}_k)} dz \\
&= \log(\hat{c}/c) + \frac{1}{2\sigma^2} \int \hat{p}(z \mid \mathscr{Y}_k) \left[\parallel z \parallel_1^2 - (\bar{H}z)^2\right] dz \\
&= \log(\hat{c}/c) + \frac{1}{2\sigma^2} \hat{E}\left[\parallel z \parallel_1^2 \mid \mathscr{Y}_k\right] - \frac{1}{2\sigma^2} \hat{E}\left[(\bar{H}z)^2 \mid \mathscr{Y}_k\right]
\end{aligned} \tag{24}$$

where $\hat{E}[\cdot \mid \mathscr{Y}_k]$ denotes the expectation operator with respect to $\hat{p}(z \mid \mathscr{Y}_k)$. Applying the Jensen inequality while recalling that $\bar{H}\hat{z}_k = \parallel \hat{z}_k \parallel_1, \quad \hat{z}_k = \hat{E}[z \mid \mathscr{Y}_k]$,

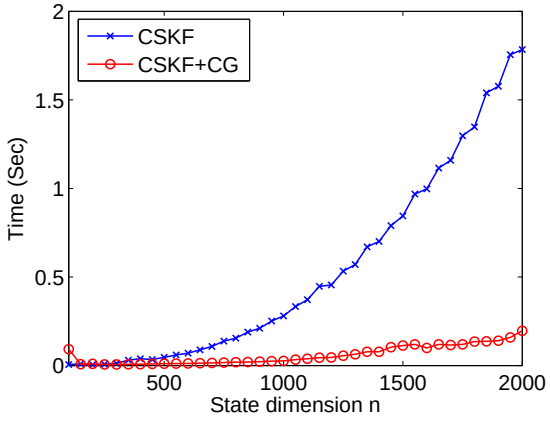


Figure 5. Computational load of the CSKF with and without the Krylov solver method.

gives

$$\begin{aligned}
&\mathrm{KL}\left(\hat{p}(z \mid \mathscr{Y}_k) \parallel p(z \mid \mathscr{Y}_k)\right) \\
&\leq \log(\hat{c}/c) + \frac{1}{2\sigma^2} \hat{E}\left[\parallel z \parallel_1^2 \mid \mathscr{Y}_k\right] - \frac{1}{2\sigma^2} \hat{E}\left[\bar{H}z \mid \mathscr{Y}_k\right]^2 = \\
&\qquad \log(\hat{c}/c) + \frac{1}{2\sigma^2} \hat{E}\left[\parallel z \parallel_1^2 \mid \mathscr{Y}_k\right] - \frac{1}{2\sigma^2} \parallel \hat{z}_k \parallel_1^2
\end{aligned} \tag{25}$$

Further letting $\delta z := z - \hat{z}_k$, Eq. (25) yields

$$\begin{aligned}
&\mathrm{KL}\left(\hat{p}(z \mid \mathscr{Y}_k) \parallel p(z \mid \mathscr{Y}_k)\right) \\
&\qquad \leq \log(\hat{c}/c) + \frac{1}{2\sigma^2} \hat{E}\left[\parallel z \parallel_1^2 \mid \mathscr{Y}_k\right] - \frac{1}{2\sigma^2} \parallel \hat{z}_k \parallel_1^2 \\
&\leq \log(\hat{c}/c) + \frac{1}{2\sigma^2} \hat{E}\left[\left(\parallel \hat{z}_k \parallel_1 + \parallel \delta z \parallel_1\right)^2 \mid \mathscr{Y}_k\right] - \frac{1}{2\sigma^2} \parallel \hat{z}_k \parallel_1^2 \\
&= \log(\hat{c}/c) + \frac{1}{\sigma^2} \hat{E}\left[\parallel \delta z \parallel_1 \mid \mathscr{Y}_k\right] \parallel \hat{z}_k \parallel_1 + \frac{1}{2\sigma^2} \hat{E}\left[\parallel \delta z \parallel_1^2 \mid \mathscr{Y}_k\right]
\end{aligned} \tag{26}$$

owing to the triangle inequality. Recalling that $\parallel \delta z \parallel_1 \leq \sqrt{n} \parallel \delta z \parallel_2$ we may then write

$$\begin{aligned}
&\mathrm{KL}\left(\hat{p}(z \mid \mathscr{Y}_k) \parallel p(z \mid \mathscr{Y}_k)\right) \\
&\leq \log(\hat{c}/c) + \frac{\sqrt{n}}{\sigma^2} \hat{E}\left[\parallel \delta z \parallel_2 \mid \mathscr{Y}_k\right] \parallel \hat{z}_k \parallel_1 + \frac{n}{2\sigma^2} \hat{E}\left[\parallel \delta z \parallel_2^2 \mid \mathscr{Y}_k\right] \\
&\qquad \leq \log(\hat{c}/c) + \frac{\sqrt{n}}{\sigma^2} \hat{E}\left[\parallel \delta z \parallel_2^2 \mid \mathscr{Y}_k\right]^{1/2} \parallel \hat{z}_k \parallel_1 \\
&\qquad\qquad + \frac{n}{2\sigma^2} \hat{E}\left[\parallel \delta z \parallel_2^2 \mid \mathscr{Y}_k\right]
\end{aligned} \tag{27}$$

where the 2nd line in (27) is due to the Jensen inequality. Recognizing that

$$\begin{aligned}
&\hat{E}\left[\parallel \delta z \parallel_2^2 \mid \mathscr{Y}_k\right] \\
&\qquad = \mathrm{Tr}\left(\hat{E}\left[(z - \hat{z}_k)(z - \hat{z}_k)^T \mid \mathscr{Y}_k\right]\right) = \mathrm{Tr}(P_k)
\end{aligned} \tag{28}$$

Eq. (27) can be written as

$$\begin{aligned}
&\mathrm{KL}\left(\hat{p}(z \mid \mathscr{Y}_k) \parallel p(z \mid \mathscr{Y}_k)\right) \\
&\leq \log(\hat{c}/c) + \mathcal{O}\left(\sigma^{-2} \max\left\{\mathrm{Tr}(P_k), \mathrm{Tr}(P_k)^{1/2}\right\}\right)
\end{aligned} \tag{29}$$

At this point the theorem immediately follows owing to the inequality $\log(\hat{c}/c) \leq 0$. In order to show this we first note that $\| z \|_1^2 \geq (\bar{H}z)^2$ owing to the fact that the left hand expression consists of summation of positive terms whereas the same terms on the right side of the equation have either positive or negative signs. This further implies

$$
\begin{aligned}
&- \| z \|_1^2 \leq -(\bar{H}z)^2 \Longrightarrow \\
&\exp\left(-\frac{1}{2}\frac{\| z \|_1^2}{\sigma^2}\right) \leq \exp\left(-\frac{1}{2}\frac{(\bar{H}z)^2}{\sigma^2}\right) \Longrightarrow \\
&\int \exp\left(-\frac{1}{2}\frac{\| z \|_1^2}{\sigma^2}\right) p'(z \mid \mathscr{Y}_k) dz \\
&\leq \int \exp\left(-\frac{1}{2}\frac{(\bar{H}z)^2}{\sigma^2}\right) p'(z \mid \mathscr{Y}_k) dz \Longrightarrow \\
&\left(\int \exp\left(-\frac{1}{2}\frac{\| z \|_1^2}{\sigma^2}\right) p'(z \mid \mathscr{Y}_k) dz\right)^{-1} \\
&\geq \left(\int \exp\left(-\frac{1}{2}\frac{(\bar{H}z)^2}{\sigma^2}\right) p'(z \mid \mathscr{Y}_k) dz\right)^{-1} \Longrightarrow c \geq \hat{c}
\end{aligned}
\tag{30}
$$

QED.

References

[1] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, pp. 489–509, 2006.

[2] E. J. Candes, "Compressive sampling," Madrid, Spain, 2006, European Mathematical Society, Proceedings of the International Congress of Mathematicians.

[3] M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: The application of compressed sensing for rapid MR imaging," *Magnetic Resonance in Medicine*, vol. 58, pp. 1182–1195, 2007.

[4] U. Gamper, P. Boesiger, and S. Kozerke, "Compressed sensing in dynamic MRI," *Magnetic Resonance in Medicine*, vol. 59, pp. 365–373, 2008.

[5] R. Chartrand, "Exact reconstruction of sparse signals via nonconvex minimization," *IEEE Signal Processing Letters*, vol. 14, pp. 707–710, 2007.

[6] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[7] E. Candes and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35, pp. 2313–2351, 2007.

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

[9] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407 – 499, 2004.

[10] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211 – 244, 2001.

[11] R. E. McCulloch and E. I. George, "Approaches for Bayesian variable selection," *Statistica Sinica*, vol. 7, pp. 339 – 374, 1997.

[12] J. Geweke, *Bayesian Statistics 5*, chapter Variable selection and model comparison in regression, Oxford University Press, 1996.

[13] B. A. Olshausen and K. Millman, "Learning sparse codes with a mixture-of-Gaussians prior," *Advances in Neural Information Processing Systems (NIPS)*, pp. 841 – 847, 2000.

[14] S. J. Godsil and P. j. Wolfe, "Bayesian modelling of time-frequency coefficients for audio signal enhancement," *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[15] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 4, pp. 3397 – 3415, 1993.

[16] Y. C. Pati, R. Rezifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," 27th Asilomar Conf. on Signals, Systems and Comput., 1993.

[17] S. Chen, S. A. Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of Control*, vol. 50, pp. 1873 – 1896, 1989.

[18] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, pp. 2346 – 2356, June 2008.

[19] A. Carmi, P. Gurfil, and D. Kanevsky, "Methods for sparse signal recovery using kalman filtering with embedded pseudo-measurement norms and quasi-norms," *IEEE Transactions on Signal Processing*, In print.

[20] L. Horesh, M. Schweiger, S.R. Arridge, and D.S. Holder, "Novel large-scale non-linear 3d reconstruction algorithms for electrical impedance tomography of the human head," *IFMBE*, vol. 14, pp. 3862–3865, 2007.