

# KF-CS: Compressive Sensing on Kalman Filtered Residual

Namrata Vaswani

**Abstract**—We consider the problem of recursively reconstructing time sequences of sparse signals (with unknown and time-varying sparsity patterns) from a limited number of linear incoherent measurements with additive noise. The idea of our proposed solution, KF CS-residual (KF-CS) is to replace compressed sensing (CS) on the observation by CS on the Kalman filtered (KF) observation residual computed using the previous estimate of the support. KF-CS error stability over time is studied. Simulation comparisons with CS and LS-CS are shown.

## I. INTRODUCTION

Consider the problem of recursively and causally reconstructing time sequences of spatially sparse signals (with unknown and time-varying sparsity patterns) from a limited number of linear incoherent measurements with additive noise. The signals are sparse in some transform domain referred to as the sparsity basis. Important applications include dynamic MRI reconstruction for real-time applications such as MRI-guided surgery, single-pixel video imaging [3], or video compression. Due to strong temporal dependencies in the signal sequence, it is usually valid to assume that its *sparsity pattern (support of the sparsity transform vector) changes slowly over time*. This was verified in [4], [5].

The solution to the static version of the above problem is provided by compressed sensing (CS) [6], [7]. CS for noisy observations, e.g. Dantzig selector [8], Lasso [9], or Basis Pursuit Denoising (BPDN) [10], [11], have been shown to have small error as long as incoherence assumptions hold. Most existing solutions for the dynamic problem, e.g. [3], [12], are non-causal and batch solutions. Batch solutions process the entire time sequence in one go and thus have much higher reconstruction complexity. An alternative would be to apply CS at each time separately (simple CS), which is online and low-complexity, but since it does not use past observations, its reconstruction error is much larger when the number of available observations is small. *Our goal* is to develop a recursive solution that improves the accuracy of simple CS by using past observations, but keeps the reconstruction complexity similar to that of simple CS. By “*recursive*”, we mean a solution that uses only the previous signal estimate and the current observation vector at the current time.

In this work, we propose a solution called KF-CS-residual (KF-CS) which is motivated by reformulating the above problem as causal minimum mean squared error (MMSE) estimation with a slow time-varying set of dominant basis directions (or equivalently the support of the sparsity basis coefficients’ vector). If the support is known, and a linear Gaussian prior dynamic model is assumed for the nonzero coefficients, the causal MMSE solution is given by the Kalman filter (KF) [13] for this support. When the support is unknown and time-varying, the initial support can be estimated using CS. Whenever there is an addition to the support, it can be estimated by running CS on the KF residual, followed by thresholding. This new support estimate can be used to run the KF at the next time instant. If some coefficients become and remain nearly zero, they can be removed from the support set. Both the computational and storage complexity of KF-CS is similar to that of simple CS -  $O(m^3)$  at a given time where  $m$  is the signal length [14, Table 1] and  $O(Nm^3)$  for an  $N$  length sequence. This is significantly lower than  $O(N^3m^3)$  for

N. Vaswani is with the ECE dept at Iowa State University, Ames, IA (email: namrata@iastate.edu). This research was partially supported by NSF grants ECCS-0725849 and CCF-0917015. A part of this work appeared in [1], [2].

batch CS. Note that a full KF, that does not use the knowledge that the signal is sparse, is meaningless here, because the number of observations available is smaller than the signal dimension, and thus many elements of the state (sparsity basis coefficients vector) will be unobservable. Unless all unobservable modes are stable, the error will blow up [13], [1].

The most closely related work to KF-CS is our work on LS-CS [2], [4] which uses an LS residual instead of a KF residual. Thus it only uses the previous support estimate, not the previous signal estimates, to improve the current reconstruction. KF-CS uses both and hence it outperforms LS-CS when the available number of measurements is small, e.g. see Fig. 2. The work of [15] gives an approximate batch-CS approach for dynamic MRI. Bayesian approaches, but all for reconstructing a single sparse signal, include [16], [17], [14]. Related work, which appeared after [1], and in parallel with [2], includes [18] (addresses recursive sparse estimation but with time-invariant support), and our own later work on modified-CS [19].

*This paper is organized as follows.* The signal model and the algorithm are described in Sec. II. We analyze the CS-residual step of KF-CS in Sec. III. In Sec. IV, we prove KF-CS error stability and discuss why our result needs stronger assumptions than a similar result for LS-CS [4]. Simulation results comparing KF-CS with LS-CS and simple CS are given in Sec. V and conclusions in Sec. VI.

In this work, we do “CS”, whether in simple CS or in CS-residual, using the Dantzig selector (DS) [8]. This choice was initially motivated by the fact that its guarantees are stronger (depend only on signal support size, not support elements) than those for BPDN [11] and its results are simpler to apply and modify. In later work [5], we have also used BPDN. Between DS and Lasso [9], either can be used and everything will remain the same except for some constants.

### A. Notation and Problem Definition

The set operations  $\cup$ ,  $\cap$ , and  $\setminus$  have the usual meanings.  $T^c$  denotes the complement of  $T$  w.r.t.  $[1, m] := [1, 2, \dots, m]$ , i.e.  $T^c := [1, m] \setminus T$ .  $|T|$  denotes the size (cardinality) of  $T$ .

For a vector,  $v$ , and a set,  $T$ ,  $v_T$  denotes the  $|T|$  length sub-vector containing the elements of  $v$  corresponding to the indices in the set  $T$ .  $\|v\|_k$  denotes the  $\ell_k$  norm of a vector  $v$ . If just  $\|v\|$  is used, it refers to  $\|v\|_2$ . For a matrix  $M$ ,  $\|M\|_k$  denotes its induced  $k$ -norm, while just  $\|M\|$  refers to  $\|M\|_2$ .  $M'$  denotes the transpose of  $M$ . For a tall matrix,  $M$ ,  $M^\dagger := (M'M)^{-1}M'$ . For symmetric matrices,  $M_1 \leq M_2$  means that  $M_2 - M_1$  is positive semidefinite. For a fat matrix  $A$ ,  $A_T$  denotes the sub-matrix obtained by extracting the columns of  $A$  corresponding to the indices in  $T$ . The  $S$ -restricted isometry property (RIP) constant,  $\delta_S$ , and the  $S, S'$ -restricted orthogonality constant,  $\theta_{S, S'}$ , are as defined in equations 1.3 and 1.5 of [8] respectively.

For a square matrix,  $Q$ , we use  $(Q)_{T_1, T_2}$  to denote the sub-matrix of  $Q$  containing rows and columns corresponding to the entries in  $T_1$  and  $T_2$  respectively.  $I$  denotes an appropriate sized identity matrix. *The  $m \times m$  matrix  $I_T$  is defined as*

$$(I_T)_{T, T} = I, \quad (I_T)_{T^c, [1, m]} = 0, \quad (I_T)_{[1, m], T^c} = 0 \quad (1)$$

We use 0 to denote a vector or matrix of all zeros of appropriate size. The notation  $z \sim \mathcal{N}(\mu, \Sigma)$  means that  $z$  is Gaussian distributed with mean  $\mu$  and covariance  $\Sigma$ .

Let  $(z_t)_{m \times 1}$  denote the spatial signal at time  $t$  and  $(y_t)_{n \times 1}$ , with  $n < m$ , denote its noise-corrupted observation vector at  $t$ , i.e.  $y_t = Hz_t + w_t$ . The signal,  $z_t$ , is sparse in a given sparsity basis (e.g. wavelet) with orthonormal basis matrix,  $\Phi_{m \times m}$ , i.e.  $x_t \triangleq \Phi' z_t$  is a sparse vector. We denote its support by  $N_t$  and we use  $S_t := |N_t|$  to denote its size. Thus the observation model is

$$y_t = Ax_t + w_t, \quad A \triangleq H\Phi, \quad \mathbb{E}[w_t] = 0, \quad \mathbb{E}[w_t w_t'] = \sigma^2 I \quad (2)$$

where  $\mathbb{E}[\cdot]$  denotes expectation. We assume that  $A$  has unit norm columns. The observation noise,  $w_t$ , is independent identically distributed (i.i.d.) over  $t$  and is independent of  $x_t$ . Our goal is to recursively estimate  $x_t$  (or equivalently the signal,  $z_t = \Phi x_t$ ) using  $y_1, \dots, y_t$ . By *recursively*, we mean, use only  $y_t$  and the estimate from  $t-1$ ,  $\hat{x}_{t-1}$ , to compute the estimate at  $t$ .

*Definition 1 (Define  $S_*$ ,  $S_{**}$ ):* For  $A := H\Phi$ ,

- 1) let  $S_*$  denote the largest  $S$  for which  $\delta_S < 1/2$ ,
- 2) let  $S_{**}$  denote the largest  $S$  for which  $\delta_{2S} + \theta_{S,2S} < 1$ .

*Definition 2 (Define  $\hat{x}_t$ ,  $\hat{N}_t$ ):* We use  $\hat{x}_t$  to denote the final estimate of  $x_t$  at time  $t$  and  $\hat{N}_t$  to denote its support estimate.

*Definition 3 (Define  $T$ ,  $\Delta$ ,  $\Delta_e$ ):* We use  $T \equiv T_t := \hat{N}_{t-1}$  to denote the support estimate from the previous time. This serves as an initial estimate of the current support. We use  $\Delta \equiv \Delta_t := N_t \setminus T$  to denote the unknown part of the support at the current time. We use  $\Delta_e \equiv \Delta_{e,t} := T_t \setminus N_t$  to denote the “erroneous” part of  $T_t$ . To keep notation simple, we remove the subscript  $t$  in most places.

## II. KALMAN FILTERED CS RESIDUAL (KF-CS)

The LS-CS-residual (LS-CS) algorithm [4] only used the previous support estimate,  $T$ , to obtain the current reconstruction, but did not use the previous nonzero coefficient estimates,  $(\hat{x}_{t-1})_T$ . Because of temporal dependencies, these also change slowly and using this fact should improve reconstruction accuracy further. To do this we can replace LS by regularized LS. If training data is available to learn a linear prior model for signal coefficients’ change, this can be done by replacing the initial LS estimate of LS-CS by a Kalman filtered (KF) estimate. The KF will give the optimal (in terms of minimizing the Bayesian MSE) regularization parameters if the size of the unknown support,  $|\Delta| = 0$ . These will be close-to-optimal if  $|\Delta|$  is nonzero but small. We assume a simple linear model described below in Sec. II-A. We develop the KF-CS algorithm for it in Sec. II-B and discuss its pros and cons in Sec. II-C.

### A. Signal Model

We assume an i.i.d. Gaussian random walk model with support additions and removals occurring every  $d$  time instants. Additions occur at every  $t_j = 1 + jd$  and removals at every  $t_{j+1} - 1$  for all  $j \geq 0$ . The support sets,  $N_t$ , at all  $t$ , are deterministic unknowns, while the sequence of  $x_t$ ’s is a random process.

*Signal Model 1:* Assume the following model.

- 1) At  $t = 0$ ,  $x_0$  is  $S_0$  sparse with support  $N_0$  and  $(x_0)_{N_0} \sim \mathcal{N}(0, \sigma_{sys,0}^2 I)$ .
- 2) At every addition time,  $t_j = 1 + jd$ , for all  $j \geq 0$ , there are  $S_a$  new additions to the support. Denote the set of indices of the coefficients added at  $t_j$  by  $\mathcal{A}(j)$ .
- 3) At every removal time,  $t_{j+1} - 1 = (j+1)d$ , for all  $j \geq 0$ , there are  $S_r$  removals from the support.
- 4) The maximum support size is  $S_{\max}$ , i.e.  $|N_t| \leq S_{\max}$  at all  $t$ .
- 5) Every new coefficient that gets added to the support starts from zero and follows an independent Gaussian random walk model with zero drift and change variance  $\sigma_{sys}^2$ .
- 6) The value of every removed coefficient and the corresponding change variance both get set to zero.

The above model can be summarized as follows.

$$\begin{aligned} |N_t \setminus N_{t-1}| &= \begin{cases} S_a & \text{if } t = t_j \\ 0 & \text{otherwise} \end{cases} \\ |N_{t-1} \setminus N_t| &= \begin{cases} S_r & \text{if } t = t_{j+1} - 1 \\ 0 & \text{otherwise} \end{cases} \\ x_0 &\sim \mathcal{N}(0, Q_0), \quad \text{where } Q_0 = \sigma_{sys,0}^2 I_{N_0} \\ \nu_t &\sim \mathcal{N}(0, Q_t), \quad \text{where } Q_t = \sigma_{sys}^2 I_{N_t} \\ (x_t)_{N_t} &= (x_{t-1})_{N_t} + (\nu_t)_{N_t} \\ (x_t)_{N_t^c} &= (\nu_t)_{N_t^c} = 0 \end{aligned} \quad (3)$$

*Assumption 1:* We assume that

- 1) The support changes slowly over time, i.e.  $S_a \ll |N_t|$  and  $S_r \ll |N_t|$ . This is empirically verified in [4], [5].
- 2) The nonzero values also change slowly, i.e.  $\sigma_{sys}^2$  is small.

### B. KF CS-residual (KF-CS) algorithm

Recall that  $T := \hat{N}_{t-1}$  denotes the support estimate from  $t-1$ . KF CS-residual (KF-CS) runs a KF for the system in (2), (3) but with  $Q_t$  replaced by  $\hat{Q}_t = \sigma_{sys}^2 I_T$  and computes the KF residual, denoted  $\tilde{y}_{t,\text{res}}$ . The new additions, if any, to  $T$ , are detected by performing CS on  $\tilde{y}_{t,\text{res}}$  and thresholding the output. If the support set changes, an LS estimate is computed using the new support estimate. If it does not change, we just use the initial KF output as the estimate. We then use this estimate to compute deletions from the support by thresholding with a different (typically larger) threshold. Once again, if the support set changes, a final LS estimate is computed using the new support and if not, then we just use the initial KF output.

In this work, the CS-residual step in KF-CS uses the Dantzig selector [8] (but this can be easily changed to BPDN or Lasso or any greedy method such as OMP etc), i.e. it solves

$$\min_{\zeta} \|\zeta\|_1 \text{ s.t. } \|A'(\tilde{y}_t - A\zeta)\|_{\infty} < \lambda \quad (4)$$

with  $y$  replaced by the current KF residual,  $\tilde{y}_{t,\text{res}}$ .

Let  $P_{t|t-1}$ ,  $P_t$  and  $K_t$  denote the “assumed” prediction and updated error covariance matrices and the Kalman gain used by the KF in KF-CS. We say “assumed” since the KF does not always use the correct value of  $Q_t$  and so  $P_{t|t-1}$  or  $P_t$  are also not equal to the actual error covariances.

We summarize the complete KF-CS algorithm below.

*Initialization ( $t = 0$ ):* At  $t = 0$ , we run simple CS (Dantzig selector) with a large enough number of measurements,  $n_0 > n$ , i.e. we solve (4) with  $y = y_0$  and  $A = A_0$  ( $A_0$  will be an  $n_0 \times m$  matrix). This is followed by support estimation and then LS estimation as in the Gauss-Dantzig selector. We denote the final output by  $\hat{x}_0$  and its estimated support by  $\hat{N}_0$ . For  $t > 0$  do,

- 1) *Initial KF.* Let  $T = \hat{N}_{t-1}$ . Run Kalman prediction and update using  $\hat{Q}_t = \sigma_{sys}^2 I_T$  and compute the KF residual,  $\tilde{y}_{t,\text{res}}$ , using

$$\begin{aligned} P_{t|t-1} &= P_{t-1} + \hat{Q}_t, \quad \text{where } \hat{Q}_t := \sigma_{sys}^2 I_T \\ K_t &= P_{t|t-1} A' (A P_{t|t-1} A' + \sigma^2 I)^{-1} \\ P_t &= (I - K_t A) P_{t|t-1} \\ \hat{x}_{t,\text{init}} &= (I - K_t A) \hat{x}_{t-1} + K_t y_t \\ \tilde{y}_{t,\text{res}} &= y_t - A \hat{x}_{t,\text{init}} \end{aligned} \quad (5)$$

- 2) *CS-residual.* Do CS (Dantzig selector) on the KF residual, i.e. solve (4) with  $y = \tilde{y}_{t,\text{res}}$ . Denote its output by  $\hat{\beta}_t$ . Compute

$$\hat{x}_{t,\text{CSres}} = \hat{x}_{t,\text{init}} + \hat{\beta}_t \quad (6)$$

- 3) *Detection and LS.* Detect additions to  $T$  using

$$\tilde{T}_{\text{det}} = T \cup \{i \in T^c : |(\hat{x}_{t,\text{CSres}})_i| > \alpha\}$$

If  $\tilde{T}_{\text{det}}$  is equal to  $T$ , set  $\hat{x}_{t,\text{det}} = \hat{x}_{t,\text{init}}$ ,  
else,  
compute an LS estimate using  $\tilde{T}_{\text{det}}$ , i.e. compute

$$(\hat{x}_{t,\text{det}})_{\tilde{T}_{\text{det}}} = A_{\tilde{T}_{\text{det}}}^\dagger y_t, \quad (\hat{x}_{t,\text{det}})_{\tilde{T}_{\text{det}}^c} = 0 \quad (7)$$

4) *Deletion and Final LS.* Estimate deletions to  $\tilde{T}_{\text{det}}$  using

$$\hat{N}_t = \tilde{T}_{\text{det}} \setminus \{i \in \tilde{T}_{\text{det}} : |(\hat{x}_{t,\text{det}})_i| < \alpha_{\text{del}}\} \quad (8)$$

If  $\hat{N}_t$  is equal to  $T$ , set  $\hat{x}_t = \hat{x}_{t,\text{init}}$ ,  
else,  
compute an LS estimate using  $\hat{N}_t$  and update  $P_t$ , i.e.

$$\begin{aligned} (\hat{x}_t)_{\hat{N}_t} &= A_{\hat{N}_t}^\dagger y_t, \\ (\hat{x}_t)_{\hat{N}_t^c} &= 0 \\ (P_t)_{\hat{N}_t, \hat{N}_t} &= (A_{\hat{N}_t}^\dagger A_{\hat{N}_t})^{-1} \sigma^2, \\ (P_t)_{\hat{N}_t^c, [1,m]} &= 0, \quad (P_t)_{[1,m], \hat{N}_t^c} = 0 \end{aligned} \quad (9)$$

5) Output  $\hat{x}_t$  and  $\hat{z}_t = \Phi \hat{x}_t$ . Feedback  $\hat{x}_t$ ,  $P_t$ ,  $\hat{N}_t$ .

*Remark 1:* Notice that the final LS step re-initializes the KF whenever the estimated support changes. This ensures less dependence of the current error on the past, and makes the stability analysis easier.

*Remark 2:* For ease of notation, in (5), we write the KF equations for the entire  $x_t$ . But the algorithm actually runs a reduced order KF for only  $(x_t)_T$  at time  $t$ , i.e. we actually have  $(\hat{x}_t)_{T^c} = 0$ ,  $(K_t)_{T^c, [1:n]} = 0$ ,  $(P_{t|t-1})_{[1,m], T^c} = 0$ ,  $(P_{t-1})_{[1,m], T^c} = 0$ ,  $(P_{t|t-1})_{T^c, [1,m]} = 0$ , and  $(P_{t-1})_{T^c, [1,m]} = 0$ . For computational speedup, the reduced order KF should be explicitly implemented.

*Remark 3:* The KF in KF-CS does not always run with correct model parameters. Thus, even when  $\sigma_{\text{sys}}^2 / \sigma^2$  is small, it is not clear if KF-CS will always outperform LS-CS [4]. This will hold at times when the support is accurately estimated and the KF has stabilized [see Fig. 2(a)]. Also, this will hold when support changes occur slowly enough, and  $n$  is small so that LS-CS error becomes instable, but is just large enough to prevent KF-CS instability [see Fig. 2(c)].

### C. Discussion of the Signal Model

A more accurate model than Signal Model 1 would be random walk with nonzero and time-varying drift. If accurate knowledge of the time-varying drift is available, the KF estimation error can be reduced significantly. But, in practice, to estimate the time-varying drift values, one would need a large number of identically distributed training signal sequences, which is an impractical assumption in most cases. On the other hand, in the above model the parameters are time-invariant and their values can be estimated from a single training sequence. This is done in [5], [20].

Now, a random walk model at all times is not a realistic signal model since it implies that the signal power keeps increasing over time. The following is what is more realistic. A new sparse basis coefficient starts from zero and slowly increases to a certain roughly constant value, i.e. it follows a random walk model for sometime and then reaches steady state. Steady state can usually be accurately modeled by a (statistically) stationary model with nonzero mean. To design KF-CS for such a model one would either need to detect when a coefficient becomes stationary or one would need to know it ahead of time. The former will typically be very error prone while the latter is an impractical assumption. To avoid having to do this, we just assume a random walk model at all times.

In Sec. V, we show that the KF-CS algorithm of Sec. II-B works both for data generated from Signal Model 1 and for data generated from a more realistic bounded signal power model taken from [4], which is a deterministic version of what is discussed above. In [5], [20], we show that it works even for actual image sequences.

### III. ANALYZING (KF)CS-RESIDUAL STEP

The KF residual,  $\tilde{y}_{t,\text{res}}$ , can be rewritten as  $\tilde{y}_{t,\text{res}} = A\beta_t + w_t$  where

$$\begin{aligned} (\beta_t)_\Delta &= (x_t - \hat{x}_{t,\text{init}})_\Delta = (x_t)_\Delta \\ (\beta_t)_T &= (x_t - \hat{x}_{t,\text{init}})_T \\ &= [I - K_t A_T](x_t - \hat{x}_{t-1})_T - K_t A_\Delta(x_t)_\Delta - K_t w_t \\ (\beta_t)_{(T \cup \Delta)^c} &= 0 \end{aligned} \quad (10)$$

where  $T = \hat{N}_{t-1}$  and  $K_t \equiv (K_t)_{T, [1,n]}$ . Thus,  $\beta_t$  is  $|T \cup \Delta| = |\hat{N}_t \cup \Delta_e|$  sparse.

In Appendix A, we show that  $\|(\beta_t)_T\|$  is bounded as in (16). As we argue there, if (a) the support changes slowly enough, (b) the signal values change slowly enough, (c) the noise is small enough and (d) the previous reconstruction is accurate enough, this bound will be small, i.e.  $\beta_t$  will be compressible along  $T$ . In other words,  $\beta_t$  will be only  $|\Delta|$ -approximately-sparse. Because of (a) and (d),  $|\Delta|$  will be small compared to  $|\hat{N}_t|$ . Thus doing CS on  $\tilde{y}_{t,\text{res}}$  will incur much less error than doing CS on  $y_t$  (simple CS), which needs to reconstruct a  $|\hat{N}_t|$ -sparse signal,  $x_t$ . This statement can be quantified by using (16) to bound CS-residual error exactly like in [4, Theorem 1] and then doing the comparison with CS also as in [4].

The CS-residual error bound will be directly proportional to the bound on  $\|(\beta_t)_T\|$  given in (16). This can be used to argue why KF-CS outperforms LS-CS when  $n$  is smaller and support changes slowly enough. We do this in Appendix B.

### IV. KF-CS ERROR STABILITY

Analyzing the KF-CS algorithm of Sec. II-B, which includes the deletion step, is difficult using the approach that we outline below. Thus, in this section, we study KF-CS without the deletion step, i.e. we set  $\alpha_{\text{del}} = 0$ . KF-CS without deletion assumes that there are few and bounded number of removals and false detects. For simplicity, in this work, we just assume  $S_r = 0$  in Signal Model 1 and we will select  $\alpha$  so that there are zero false detects.  $S_r = 0$  along with the assumption that the maximum sparsity size is  $S_{\max}$  implies that there are only a finite number of addition times,  $K$ , i.e. for all  $t \geq t_{K-1}$ ,  $N_t = N_{t_{K-1}}$ . We summarize this in the following signal model.

*Signal Model 2:* Assume Signal Model 1 with  $S_r = 0$ . This implies that there are only a finite number of addition times,  $t_j$ , with  $j = 0, 1, \dots, (K-1)$  and  $K = \lceil \frac{S_{\max} - S_0}{S_a} \rceil$ . Let  $t_K := \infty$ .

Consider the genie-aided KF, i.e. a KF which knows the true support  $N_t$  at each  $t$ . It is the MMSE estimator of  $x_t$  from  $y_1, \dots, y_t$  if the support sets,  $N_t$ , are assumed known and the noise is Gaussian, and is the linear MMSE for any arbitrary noise. In this section, we find sufficient conditions under which, with high probability (w.h.p.), KF-CS for Signal Model 2 and observation model given by (2) gets to within a small error of the genie-KF for the same system, within a finite delay of the new addition time. Since the genie-KF error is itself stable w.h.p., as long as  $\delta_{S_{\max}} < 1$ , this also means that the KF-CS reconstruction error is stable w.h.p.

Our approach involves two steps. Consider  $t \in [t_j, t_{j+1})$ . First, we find the conditions under which w.h.p. all elements of the current support,  $N_t = N_{t_j}$  get detected before the next addition time,  $t_{j+1}$ . Denote the detection delay by  $\tau_{\text{det}}$ . If this happens, then during  $[t_j + \tau_{\text{det}}, t_{j+1})$ , both KF-CS and genie-KF run the same fixed dimensional and fixed parameter KF, but with different initial conditions. Next, we show that if this interval is large enough, then, w.h.p., KF-CS will stabilize to within a small error of the genie-KF within a finite delay after  $t_j + \tau_{\text{det}}$ . Combining these two results gives our stability result.

We are able to do the second step because, whenever  $\hat{N}_t \neq \hat{N}_{t-1}$ , the final LS step re-initializes the KF with  $P_t$ ,  $\hat{x}_t$  given by (9). This ensures that the KF-CS estimate,  $\hat{x}_t$ , and the Kalman gain,  $K_t$ , at  $t+1$  and future times depend on the past observations only through

$T := \hat{N}_t$ . Thus, conditioned on the event  $\{\hat{N}_t = N_t, \forall t \in [t_j + \tau_{\det}, t_{j+1}]\}$ , there will be no dependence of either  $\hat{x}_t$  or of  $K_t$  on observations before  $t_j + \tau_{\det}$ .

### A. The Stability Result

We begin by stating Lemma 1 which shows two things. First, if accurate initialization is assumed, the noise is bounded,  $S_{\max} \leq S_{**}$ ,  $\alpha_{\det} = 0$  and  $\alpha$  is high enough, there are no false detections. If the delay between addition times also satisfies  $d > \tau_{\det}(\epsilon, S_a)$ , where  $\tau_{\det}$  is what we call the “high probability detection delay”, then the following holds. If before  $t_j$ , the support was perfectly estimated, then w.p.  $\geq 1 - \epsilon$ , all the additions which occurred at  $t_j$  will get detected by  $t_j + \tau_{\det}(\epsilon, S_a) < t_{j+1}$ .

*Lemma 1:* Assume that  $x_t$  follows Signal Model 2. If

- 1) (*initialization* ( $t = 0$ )) all elements of  $x_0$  get correctly detected and there are no false detects, i.e.  $\hat{N}_0 = N_0$ ,
- 2) (*measurements*)  $S_{\max} \leq S_{**}$  and  $\|w\|_\infty \leq \lambda / \|A\|_1$ ,
- 3) (*algorithm*) we set  $\alpha_{\det} = 0$  and  $\alpha^2 = B_* := C_1(S_{\max})S_{\max}\lambda^2$ , where  $C_1(S)$  is defined in [8, Theorem 1.1],
- 4) (*signal model*) delay between addition times,  $d > \tau_{\det}(\epsilon, S_a)$ ,

$$\text{where } \tau_{\det}(\epsilon, S) := \left\lceil \frac{4B_*}{\sigma_{sys}^2[\mathcal{Q}^{-1}(\frac{(1-\epsilon)^{1/S}}{2})]^2} \right\rceil - 1, \quad (11)$$

[.] denotes the greatest integer function and  $\mathcal{Q}(z) := \int_z^\infty (1/\sqrt{2\pi})e^{-x^2/2}dx$  is the Gaussian Q-function,

then

- 1) at each  $t$ ,  $\hat{N}_t \subseteq N_t \subseteq N_{t+1}$  and so  $|\Delta_{e,t+1}| = 0$
- 2) at each  $t$ ,  $\|x_t - \hat{x}_{t,\text{CSres}}\|^2 \leq B_*$
- 3)  $\Pr(E_j | F_j) \geq 1 - \epsilon$  where  $F_j := \{\hat{N}_t = N_t \text{ for } t = t_j - 1\}$  and  $E_j := \{\hat{N}_t = N_t, \forall t \in [t_j + \tau_{\det}(\epsilon, S), t_{j+1} - 1]\}$ .

The proof is given in Appendix C. The initialization assumption is made only for simplicity. It can be easily satisfied by using  $n_0 > n$  to be large enough. Next we give Lemma 2 which states that if the true support set does not change after a certain time,  $t_{nc}$ , and if it gets correctly detected by a certain time,  $t_* \geq t_{nc}$ , then KF-CS converges to the genie-KF in mean-square and hence also in probability.

*Lemma 2:* Assume that  $x_t$  follows Signal Model 2;  $\delta_{S_{\max}} < 1$ ; and  $\alpha_{\det} = 0$ . Define the event  $D := \{\hat{N}_t = N_t = N_*, \forall t \geq t_*\}$ . Conditioned on  $D$ , the difference between the KF-CS estimate,  $\hat{x}_t$  and the genie-aided KF estimate,  $\hat{x}_{t,GAKF}$ ,  $\text{diff}_t := \hat{x}_t - \hat{x}_{t,GAKF}$ , converges to zero in mean square and hence also in probability. ■

The proof is similar to what we think should be a standard result for a KF with wrong initial conditions (here, KF-CS with  $t = t_*$  as the initial time) to converge to a KF with correct initial conditions (here, genie-KF) in mean square. A similar (actually stronger) result is proved for the continuous time KF in [22]. We could not find an appropriate citation for the discrete time KF and hence we just give our proof in Appendix E. After review, this can be significantly shortened. The proof involves two parts. First, we use the results from [13] and [21] to show that (a)  $P_{t|t-1}, P_t, K_t$  and  $J_t := I - K_t A_{N_*}$  converge to steady state values which are the same as those for the corresponding genie-KF; and (b) the steady state value of  $J_t$ , denoted  $J_*$ , has spectral radius less than 1 and because of this, there exists a matrix norm, denoted  $\|\cdot\|_\rho$ , s.t.  $\|J_*\|_\rho < 1$ . Second, we use (a) and (b) to show that the difference in the KF-CS and genie-KF estimates,  $\text{diff}_t$ , converges to zero in mean square, and hence also in probability (by Markov’s inequality).

A direct corollary of the above lemma is the following.

*Corollary 1:* Assume that  $x_t$  follows Signal Model 2;  $\delta_{S_{\max}} < 1$ ; and  $\alpha_{\det} = 0$ . Define the event  $D_f := \{\hat{N}_t = N_t = N_*, \forall t \in [t_*, t_{**}]\}$ . For a given  $\epsilon, \epsilon_{\text{err}}$ , there exists a  $\tau_{KF}(\epsilon, \epsilon_{\text{err}}, N_*)$  s.t. for

all  $t \in [t_* + \tau_{KF}, t_{**}]$ ,  $\Pr(\|\text{diff}_t\|^2 \leq \epsilon_{\text{err}} | D_f) > 1 - \epsilon$ . Clearly if  $t_{**} < t_* + \tau_{KF}$ , this is an empty interval.

The stability result then follows by applying Lemma 2 followed by Corollary 1 for each addition time,  $t_j$ .

*Theorem 1 (KF-CS Stability):* Assume that  $x_t$  follows Signal Model 2. Let  $\text{diff}_t := \hat{x}_t - \hat{x}_{t,GAKF}$  where  $\hat{x}_{t,GAKF}$  is the genie-aided KF estimate and  $\hat{x}_t$  is the KF-CS estimate. For a given  $\epsilon, \epsilon_{\text{err}}$ , if the conditions of Lemma 1 hold, and if the delay between addition times,  $d > \tau_{\det}(\epsilon, S_a) + \tau_{KF}(\epsilon, \epsilon_{\text{err}}, N_{t_j})$ , where  $\tau_{\det}(\cdot, \cdot)$  is defined in (11) in Lemma 1 and  $\tau_{KF}(\cdot, \cdot)$  in Corollary 1, then

- 1)  $\Pr(\|\text{diff}_t\|^2 \leq \epsilon_{\text{err}}) > (1 - \epsilon)^{j+2}$ , for all  $t \in [t_j + \tau_{\det}(\epsilon, S_a) + \tau_{KF}(\epsilon, \epsilon_{\text{err}}, N_{t_j}), t_{j+1} - 1]$ , for all  $j = 0, \dots, (K - 1)$ ,
- 2)  $\Pr(|\Delta| \leq S_a \text{ and } |\Delta_e| = 0, \forall t \geq (1 - \epsilon)^K)$
- 3)  $\Pr(|\Delta| = 0 \text{ and } |\Delta_e| = 0, \forall t \in [t_j + \tau_{\det}(\epsilon, S_a), t_{j+1} - 1], \forall j = 0, \dots, K - 1) \geq (1 - \epsilon)^K$ .

The proof is given in Appendix D. A direct corollary is that after  $t_{K-1}$  KF-CS will converge to the genie-KF in probability. This is because for  $t \geq t_{K-1}$ ,  $N_t$  remains constant ( $t_K = \infty$ ).

### B. Discussion

Consider a  $t \in [t_j, t_{j+1}]$ . Notice that  $\tau_{KF}$  depends on the current support,  $N_t = N_{t_j}$  while  $\tau_{\det}$  depends only on the number of additions at  $t_j$ ,  $S_a$ . Theorem 1 says that if  $n$  is large enough so that  $S_{\max} \leq S_{**}$ ;  $\alpha_{\det} = 0$  (ensures no deletions);  $\alpha = \sqrt{B_*}$  (ensures no false detects); and if the time needed for the current KF to stabilize,  $\tau_{KF}(\epsilon, \epsilon_{\text{err}}, N_{t_j})$ , plus the high probability detection delay,  $\tau_{\det}(\epsilon, S_a)$ , is smaller than  $d$ , then w.p.  $\geq (1 - \epsilon)^{j+2}$ , KF-CS will stabilize to within a small error,  $\epsilon_{\text{err}}$ , of the genie-KF before the next addition time,  $t_{j+1}$ . If the current  $\tau_{KF}$  is too large, this cannot be claimed. But as long as  $\tau_{\det}(\epsilon, S_a) < d$ , the unknown support size,  $|\Delta|$  remains bounded by  $S_a$ , w.p.  $\geq (1 - \epsilon)^K$ .

We give our result for the case of zero removals and zero false detects, but the same idea will extend even if  $|\Delta_e|$  is just bounded.

As explained in Sec. II-C, most signals do not follow a random walk model forever (such a model would imply unbounded signal power). In practice, a new coefficient may start with following a random walk model, but eventually reach steady state (stationary model). In this case, it should be possible to modify our result to claim that if, before reaching steady state, all coefficients become large enough to exceed the threshold plus upper bound on error, and if this happens before the next addition time, KF-CS remains stable.

Our result is weaker than that of LS-CS [4] - it needs  $S_{\max} \leq S_{**}$  (the LS-CS result only needs  $S_a \leq S_{**}$  and  $S_{\max} \leq S_*$ ); it uses a random walk model; it does not handle support removals; and the computed high-probability detection delay is quite loose<sup>1</sup>. This is due to two main reasons. One is that we assume a zero drift random walk model as the signal model both for defining KF-CS and for analyzing it, while LS-CS uses a model with nonzero drift for the analysis (the algorithm does not assume any signal model). The reason for our choosing this model is explained in Sec. II-C. The second and more important reason is that bounding KF error is more difficult than bounding LS error. This is because the KF error, and hence also the (KF)CS-residual error, depends on the previous reconstruction error. The (LS)CS-residual error only depends on  $|T|$ ,  $|\Delta|$  and if we can get a time-invariant bound on these, we can do the same for the error.

### V. SIMULATION RESULTS

We discuss two sets of simulation results. The first simulates data according to Signal Model 2 and verifies KF-CS stability. The second set of simulations compares KF-CS with LS-CS [4] and simple CS (Dantzig selector) [8]. This comparison uses the more realistic signal

<sup>1</sup>Our result may even go through if CS-residual was replaced by CS.

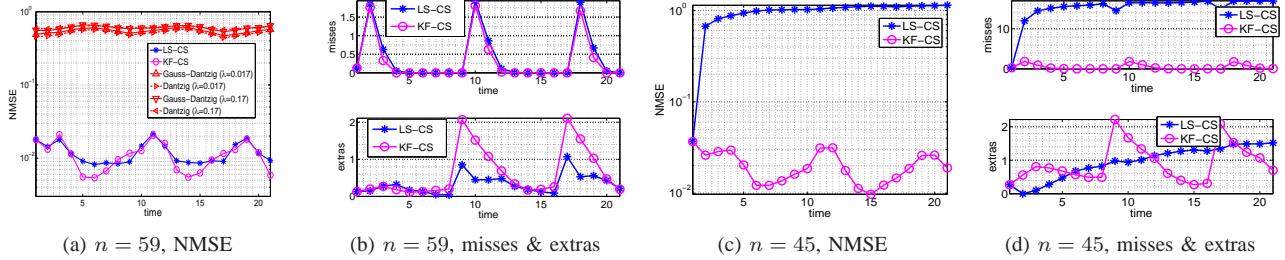


Fig. 2. Comparing KF-CS with CS and LS-CS. CS-residual in LS-CS or in KF-CS used  $\lambda = 0.17$ . Misses =  $\mathbb{E}[|N_t \setminus \hat{N}_t|]$ , Extras =  $\mathbb{E}[|\hat{N}_t \setminus N_t|]$ .

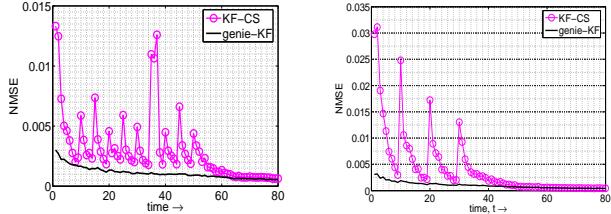


Fig. 1. Verifying KF-CS stability for Signal Model 2.

model assumed in [4], which has a roughly constant signal power and support size and allows regular additions and removals from support.

#### A. Signal Model 2: verify KF-CS Stability

We simulated Signal Model 2 with  $m = 256$ ,  $S_0 = 8$ ,  $S_a = 2$ ,  $d = 5$ ,  $S_{\max} = 26$  and  $\sigma_{sys} = 1$ . Thus additions occurred at  $t = 1, 6, 11, \dots, 46$ . The measurement model used  $n = n_0 = 72$  and Gaussian noise with  $\sigma = 0.16$ . The normalized MSE (NMSE) is plotted in Fig. 1(a). In a second simulation, we increased  $S_a$ , but we also increased  $d$ : we used  $S_a = 4$ ,  $d = 10$  and  $S_{\max} = 20$  and everything else was the same. We show the error plot in Fig. 1(b). Notice that in both cases, (i) KF-CS stabilizes to within a small error of the genie-KF within a short delay of a new addition time; and (ii) after the final set of new additions, KF-CS converges to the genie-KF. The difference between the two is that the peak errors at the new addition time are larger in the second case (since  $S_a$  is larger).

We implemented the KF-CS algorithm of Sec. II-B but without the deletion step, i.e. we set  $\alpha_{del} = 0$ . Since the observation noise is not truncated, occasionally the addition step can result in a very large number of false additions. To prevent this, we restricted the maximum number of allowed additions at a given time to  $\gamma n / \log_2 m$  ( $\gamma$  between 0.7 and 1.25) largest magnitude coefficients.

#### B. Bounded signal power model from [4]

For this comparison we used the signal model of [4]. This is a realistic signal model with roughly constant signal power and support size. We used  $m = 200$ ,  $S_0 = 20$ ,  $S_a = 2 = S_r$ ,  $a_i = 0.2$ ,  $M = 1$ ,  $d = 8$  and  $r = 3$ . Thus new additions occurred at  $t = 2, 10, 18$ . Coefficient decrease began at  $t = 7, 15$  and these got removed at  $t = 9, 17$  respectively. The measurement noise was  $uniform(-c, c)$ .

In the first simulation, we used  $n_0 = 150$ ,  $n = 59$  and  $c = 0.1266$ . LS-CS used  $\lambda = 0.176$ ,  $\alpha = c/2 = 0.06 = \alpha_{del}$ . Also, it restricted maximum number of additions at a time to  $S_a + 1$ . The KF-CS algorithm of Sec. II-B was implemented. It used the above parameters and it set  $\sigma^2 = c^2$  and  $\sigma_{sys}^2 = 0.01$ . For the signal model of [4], there are no correct choices of KF parameters. The average of  $(x_t - x_{t-1})_i^2$  over  $i$  and  $t$  was  $(0.04 * (5/8) * (2/20) + 0.11 * (3/8) * (2/20) + 0 * 1 * (16/20)) \approx 0.01$  and this motivated the choice of  $\sigma_{sys}^2$ . The noise variance is  $c^2/3$ , but we use a larger value to also model the effect of extra observation error due to the unknown support  $\Delta$ . The NMSE plot is shown in Fig. 2(a). The mean number of misses ( $\mathbb{E}[|N_t \setminus \hat{N}_t|]$ )

and of extras ( $\mathbb{E}[|\hat{N}_t \setminus N_t|]$ ) are plotted in Fig. 2(b). We averaged over 100 Monte Carlo runs. Notice that right after a new addition, both LS-CS and KF-CS have similar MSE, but in the stable state KF-CS stabilizes to a smaller value. The NMSEs for CS (Dantzig selector) and Gauss-Dantzig selector even with different choices of  $\lambda$  are much larger (40-60%).

In a second simulation, we used  $n_0 = 150$ ,  $n = 45$  and  $c = 0.15$  and everything else was the same as above. The error plots are shown in Fig. 2(c) and the number of extras and misses are plotted in Fig. 2(d). With such a small  $n$ , LS-CS error becomes instable. But  $n = 45$  (along with delay between addition times,  $d = 8$ ) is large enough to prevent KF-CS instability.

## VI. CONCLUSIONS AND FUTURE WORK

We proposed KF CS-residual (KF-CS) which replaces CS on the raw observation by CS on the KF residual, computed using the known part of the support. We proved KF-CS stability, but the assumptions used were somewhat strong (stronger than those used for LS-CS [4]). We demonstrated via simulations that KF-CS error is stable and small under much weaker assumptions. Also, it significantly outperformed LS-CS when the available number of measurements was very small.

A key direction of future work is to prove KF-CS stability under weaker assumptions. This will require assuming a signal model with nonzero drift (to get a tighter detection delay bound) and bounded signal power. It may also help to assume a statistical prior on support change, e.g. by using a model similar to [17]. A useful extension of KF-CS would be to replace CS-residual by modified-CS [19].

## APPENDIX

### A. Bounding $\|(\beta_t)_T\|$

Recall that  $T_t = \hat{N}_{t-1}$  and  $\Delta_t = N_t \setminus \hat{N}_{t-1}$ . Let  $\delta_t \triangleq \delta_{|T_t|}$  and  $\theta_t \triangleq \theta_{|T_t|, |\Delta_t|}$ . Also, let  $K_t \equiv (K_t)_{T,[1,n]}$ ,

$$M_t \triangleq A_{T'} A_T + (P_{t|t-1})_{T,T}^{-1} \sigma^2 \text{ and} \\ r \triangleq \sigma_{sys}^2 / \sigma^2 \quad (12)$$

We use the following simple facts in the discussion below [21]. For symmetric positive definite matrices,  $M$ ,  $\tilde{M}$ ,  $\|M\| = \lambda_{\max}(M) = 1/\lambda_{\min}(M^{-1})$ ,  $\lambda_{\min}(M + \tilde{M}) \geq \lambda_{\min}(M) + \lambda_{\min}(\tilde{M})$  while the inequality holds in the opposite direction for  $\lambda_{\max}$ . Here  $\lambda_{\max}(M)$ ,  $\lambda_{\min}(M)$  denote the maximum, minimum eigenvalue of  $M$ .

As is well known [13],  $K_t$ ,  $[I - K_t A_T]$ ,  $P_t$  can be rewritten as

$$K_t = M_t^{-1} A_{T'} \\ J_t := I - K_t A_T = M_t^{-1} (P_{t|t-1})_{T,T}^{-1} \sigma^2 \\ (P_{t|t-1})_{T,T} = (P_{t-1})_{T,T} + (\sigma_{sys}^2 I_T)_{T,T}, \text{ where} \\ (P_{t-1})_{T,T} = \begin{cases} M_{t-1}^{-1} \sigma^2 & \text{if } T_t = T_{t-1} \\ (A_{T_{t-1}}' A_{T_{t-1}})^{-1} \sigma^2 & \text{if } T_t \neq T_{t-1} \end{cases} \quad (13)$$

The third equation is repeated from (5). To bound  $\|(\beta_t)_T\|$ , defined in (10), we need to bound  $\|J_t\|$  and  $\|K_t A_{T'} A_{\Delta}\|$ , which in turn requires bounding  $\|M_t^{-1}\|$ ,  $\|(P_{t|t-1})_{T,T}^{-1}\|$  and  $\|A_{T'} A_{\Delta}\|$ . Using the definition of  $\theta_{S,S'}$  [8, eq 1.5], it is easy to see that  $\|A_{T'} A_{\Delta}\| \leq \theta_t$ .

Using (13),  $\|(P_{t|t-1})_{T,T}^{-1}\| \leq (\lambda_{\min}(M_{t-1}^{-1})\sigma^2 + \sigma_{sys}^2)^{-1}$  if  $T_t = T_{t-1}$  and  $\|(P_{t|t-1})_{T,T}^{-1}\| \leq ((1 + \delta_t)^{-1}\sigma^2 + \sigma_{sys}^2)^{-1}$  otherwise. Also,  $\|M_t^{-1}\| = \lambda_{\max}(M_t^{-1})$ . Thus bounding them requires upper bounding  $\lambda_{\max}(M_t^{-1})$  and lower bounding  $\lambda_{\min}(M_t^{-1})$ . Using the definition of the RIP constant [8, eq. 1.3],

$$\begin{aligned} \|M_t^{-1}\| &= \lambda_{\max}(M_t^{-1}) = \frac{1}{\lambda_{\min}(A_T' A_T + (P_{t|t-1})_{T,T}^{-1} \sigma^2)} \\ &\leq \frac{1}{1 - \delta_t + \frac{\sigma^2}{\lambda_{\max}((P_{t|t-1})_{T,T})}} \\ &\leq \begin{cases} \frac{1}{1 - \delta_t + \frac{1}{\|M_{t-1}^{-1}\| + r}} & \text{if } T_t = T_{t-1} \\ \frac{1}{1 - \delta_t + \frac{1}{(1 - \delta_t)^{-1} + r}} & \text{if } T_t \neq T_{t-1} \end{cases} \\ &\triangleq a_t \end{aligned} \quad (14)$$

Similarly, we can lower bound  $\lambda_{\min}(M_{t-1}^{-1})$  and use it to get

$$\|(P_{t|t-1})_{T,T}^{-1}\| \sigma^2 \leq \begin{cases} \frac{1}{1 + \delta_t + \frac{1}{\|M_{t-2}^{-1}\|^{-1} + r}} & \text{if } T_t = T_{t-1} = T_{t-2} \\ \frac{1}{1 + \delta_t + \frac{1}{(1 + \delta_{t-1})^{-1} + r}} & \text{if } T_t = T_{t-1} \neq T_{t-2} \\ \frac{1}{1 + \delta_t} & \text{if } T_t \neq T_{t-1} \end{cases} \triangleq \frac{1}{b_t} \quad (15)$$

From (10), (13),  $\|(\beta_t)_T\| \leq \|M_t^{-1}\| [\|(P_{t|t-1})_{T,T}^{-1}\| \sigma^2 \| (x_t - \hat{x}_{t-1})_T \| + \theta_t \| (x_t)_\Delta \| + \|A_T' w_t\| ]$ . Using this and the above bounds, we get

$$\begin{aligned} \|(\beta_t)_T\| &\leq a_t [T1 + \theta_t \| (x_t)_\Delta \| + \|A_T' w_t\|], \text{ where} \\ T1 &\triangleq \frac{\| (x_{t-1} - \hat{x}_{t-1})_{T \cap N_t} \| + \| (\hat{x}_{t-1})_{\Delta_e} \| + \sqrt{|T \cap N_t|} \|\nu_t\|_\infty}{b_t}, \end{aligned} \quad (16)$$

and  $a_t$  is defined in (14) and  $b_t$  in (15). Notice that  $a_t$  is an increasing function of  $\delta_t$  and  $r$ , and also of  $\|M_{t-1}^{-1}\| \leq a_{t-1}$  if  $T_t = T_{t-1}$ .

Now,  $\Delta \subseteq (N_{t-1} \setminus T) \cup (N_t \setminus N_{t-1})$  and  $\Delta_e \subseteq (T \setminus N_{t-1}) \cup (N_{t-1} \setminus N_t)$ . If the previous reconstruction is accurate enough, the previous support estimate will also be accurate enough. This combined with the slow support change assumption will imply that  $|\Delta|$  and  $|\Delta_e|$  are small enough.  $|\Delta_e|$  small enough will imply that  $|T|$  is small enough (since  $|T| \leq |N_t| + |\Delta_e|$ ) and hence  $\delta_t$  is small enough.  $\delta_t$  and  $\delta_{t-1}$  small ensure smaller  $a_t$  and larger  $b_t$ .  $|\Delta_e|$  and  $|\Delta|$  small enough will also imply that  $\theta_t$  is small enough. The noise being small along with  $|\Delta_e|$  small will imply that  $\|A_T' w_t\|$  is small.

Slow signal value change implies (i)  $r$  is small enough and (ii) at all  $t$ ,  $\|\nu_t\|_\infty$  is small enough w.h.p.. Small  $r$  implies that  $a_t$  is small, but it also implies that  $b_t$  is small. Small  $\|\nu_t\|_\infty$  at all  $t$ , along with small noise, also results in the previous reconstruction being accurate enough which, in turn means  $\| (x_{t-1} - \hat{x}_{t-1})_{T \cap N_t} \|$  is small. Using this and the fact that only small coefficients get falsely deleted or removed<sup>2</sup>,  $\| (\hat{x}_{t-1})_{\Delta_e} \|$  is also small. All this ensures that T1 in (16) is not very large even when  $b_t$  is small. This combined with the discussion of the previous paras ensures that the bound on  $\|(\beta_t)_T\|$  is small. Thus, if (a), (b), (c), (d) given in Sec. III hold,  $\|(\beta_t)_T\|$  is small, i.e.  $\beta_t$  is only  $|\Delta|$ -approximately-sparse; and  $|\Delta|$  is small.

### B. Comparing KF-CS and LS-CS using the bound on $\|(\beta_t)_T\|$

We will mention that we are only comparing upper bounds here.

Consider  $a_t$  defined in (14). Suppose  $r = 0.5$  and  $n$  is such that  $\delta_t = 0.8$  for all  $t$ . LS-CS can be interpreted as KF-CS with  $r = \infty$ . Thus for LS-CS  $a_t = 1/(1 - \delta_t) = 5$  always. For KF-CS, even if, at  $t$ ,  $T_t \neq T_{t-1}$ ,  $a_t = 1/(0.2 + (1/5.5)) = 2.62$  (almost half). If

<sup>2</sup>The fact that only small coefficients get removed from  $N_t$  is not modeled in Signal Model 1, but is true in practice. But it is modeled in our simulations.

$T_t$  does not change for just one time instant, then  $a_{t+1}$  reduces to  $1/(0.2 + 1/(a_t + 0.5)) = 1.92$ . If it does not change for two time instants, then  $a_{t+2}$  reduces to 1.63. If  $T_t$  does change and the change is a correct addition, the set  $\Delta_t$  becomes smaller and so the second term of (16),  $\theta_t \| (x_t)_\Delta \|$ , reduces. In either case, the bound reduces.

Of course for LS-CS,  $b_t = \infty$  and so the first term of (16), T1 = 0 while for KF-CS, T1  $\neq 0$ . But if  $\sigma_{sys}^2$  and  $\sigma^2$  are small and support changes slowly, T1 will also be small (argued earlier). When  $n$  is small, the net effect is that the KF-CS bound on  $\|(\beta_t)_T\|$ , and hence the bound on CS-residual error, is small compared to that for LS-CS. This is the main reason that, when  $n$  is very small, KF-CS error remains stable, while nothing can be said about LS-CS error. In simulations, we notice that it often becomes unstable.

### C. Proof of Lemma 1

With  $\|w\|_\infty \leq \lambda/\|A\|_1$ , all results of [8] hold w.p. 1 (because eq 3.1 of [8] holds w.p. 1). From Theorem 1.1. of [8], if a signal is  $S$ -sparse, and if  $S \leq S_{**}$ , then, the error after running the Dantzig selector is bounded by  $B_*$ .

The first two claims follow by induction. Consider the base case,  $t = 0$ . The first claim holds because condition 1 of the lemma holds and because  $S_r = 0$  in Signal Model 2. Since  $|N_0| \leq S_{\max}$  and condition 2 holds, [8, Theorem 1.1] applies. Thus the second claim holds at  $t = 0$ . For the induction step, assume that the first two claims hold for  $t - 1$ . Using the first claim for  $t - 1$ ,  $|\Delta_{e,t}| = 0$ . Thus,  $\beta_t$  is  $|N_t \cup \Delta_{e,t}| = |N_t|$  sparse. Since  $|N_t| \leq S_{\max}$  and condition 2 holds, we can apply [8, Theorem 1.1] to get  $\|\beta_t - \hat{\beta}_t\|^2 \leq B_*$ . But  $x_t - \hat{x}_{t,\text{CSres}} = \beta_t - \hat{\beta}_t$  and so the second claim follows for  $t$ . By setting  $\alpha = \sqrt{B_*}$  (condition 3), we ensure that for any  $i$  with  $x_i = 0$ ,  $(\hat{x}_{\text{CSres}})_i^2 = (x_i - (\hat{x}_{\text{CSres}})_i)^2 \leq \|x - \hat{x}_{\text{CSres}}\|^2 \leq B_* = \alpha^2$  (no false detections). Using this and  $S_r = 0$ , the first claim follows for  $t$ .

For the third claim, it is easy to see that for any  $i \in \Delta$ , if, at  $t$ ,  $(x_t)_i^2 > 2\alpha^2 + 2B_* = 4B_*$ , then  $i$  will definitely get detected at  $t$ . Consider a  $t \in [t_j, t_{j+1} - 1]$ . Since  $F_j$  holds, so at  $t = t_j$ ,  $\Delta = \mathcal{A}(j)$ . Also, since  $\alpha_{\text{del}} = 0$ , there cannot be false deletions and thus for any  $t \in [t_j, t_{j+1} - 1]$ ,  $|\Delta| \leq S_a$ . Consider the worst case: no coefficient has got detected until  $t$ , i.e.  $\Delta_t = \mathcal{A}(j)$  and so  $|\Delta_t| = S_a$ . All  $i \in \mathcal{A}(j)$  will definitely get detected at  $t$  if  $(x_t)_i^2 > 4B_*$  for all  $i \in \mathcal{A}(j)$ . From our model, the different coefficients are independent, and for any  $i \in \mathcal{A}(j)$ ,  $(x_t)_i^2 \sim \mathcal{N}(0, (t - t_j + 1)\sigma_{sys}^2)$ . Thus,

$$Pr((x_t)_i^2 > 4B_*, \forall i \in \mathcal{A}(j) \mid F_j) = \left( 2Q\left(\sqrt{\frac{4B_*}{(t - t_j + 1)\sigma_{sys}^2}}\right) \right)^{S_a}$$

Using the first claim,  $Pr(\hat{N}_t = N_t \mid F_j)$  is equal to this. Thus for  $t = t_j + \tau_{\text{det}}(\epsilon, S_a)$ ,  $Pr(\hat{N}_t = N_t \mid F_j) \geq 1 - \epsilon$ . Since there are no false detects; no deletions and no new additions until  $t_{j+1}$ ,  $\hat{N}_t = N_t$  for  $t = t_j + \tau_{\text{det}}$  implies that  $E_j$  occurs. This proves the third claim.

### D. Proof of Theorem 1

The events  $E_j$  and  $F_j$  are defined in Lemma 1. At the first addition time,  $t_0 = 1$ , using the initialization condition,  $\hat{N}_{t_0-1} = N_{t_0-1}$ , i.e.  $F_0$  holds. Thus, by Lemma 1,  $Pr(E_0) \geq 1 - \epsilon$ . Consider  $t_j$  for  $j > 0$ . Clearly<sup>3</sup>  $Pr(E_j \mid E_0, E_1, \dots, E_{j-1}) = Pr(E_j \mid E_{j-1}) = Pr(E_j \mid F_j)$ . By Lemma 1,  $Pr(E_j \mid F_j) \geq 1 - \epsilon$ . Combining this with  $Pr(E_0) \geq 1 - \epsilon$ , we get  $Pr(E_j) \geq (1 - \epsilon)^{j+1}$  for all  $j \geq 0$ .

Assume that  $E_j$  occurs and apply Corollary 1 with  $t_* = t_j + \tau_{\text{det}}(\epsilon, S_a)$  and  $t_{**} = t_{j+1} - 1$ . Combining the conclusion of Corollary 1 with  $Pr(E_j) \geq (1 - \epsilon)^{j+1}$ , the first claim follows.

The second and third claims follow directly from arguments in the proof of Lemma 1 and  $Pr(E_0 \cap E_1 \cap \dots \cap E_{K-1}) \geq (1 - \epsilon)^K$ .

<sup>3</sup>since  $E_j = \{(x_{t_j+\tau_{\text{det}}})_i^2 > 4B_*, \forall i \in \Delta_{t_j+\tau_{\text{det}}}\}$  and the sequence of  $x_t$ 's is a Markov process

### E. Proof of Lemma 2 and Corollary 1

Let  $\hat{x}_{t,GAKF}$  denote the genie-aided KF (GA-KF) estimate at  $t$ .

Assume that the event  $D$  occurs. Then, for  $t > t_*$ ,  $\hat{N}_t = N_t = N_*$ , i.e.  $\Delta_t := N_t \setminus \hat{N}_{t-1} = N_* \setminus N_* = \emptyset$  (empty set) and so  $\hat{x}_t = \hat{x}_{t,\text{init}}$ . Let  $e_t \triangleq x_t - \hat{x}_t$  and  $\tilde{e}_t \triangleq x_t - \hat{x}_{t,GAKF}$ .

For simplicity of notation we assume in this proof that all variables and parameters are only along  $N_*$ , i.e. we let  $\hat{x}_t \equiv (\hat{x}_t)_{N_*}$ ,  $e_t \equiv (e_t)_{N_*}$ ,  $\nu_t \equiv (\nu_t)_{N_*}$ ,  $P_{t|t-1} \equiv (P_{t|t-1})_{N_*,N_*}$ ,  $K_t \equiv (K_t)_{N_*,[1:n]}$ . Let  $J_t \triangleq I - K_t A_{N_*}$ . Similarly for  $\hat{x}_{t,GAKF}$ ,  $\tilde{e}_t$ ,  $\tilde{P}_{t|t-1}$ ,  $\tilde{K}_t$ ,  $\tilde{J}_t$ . Here  $\tilde{P}_{t|t-1}$ ,  $\tilde{K}_t$ ,  $\tilde{J}_t$  are the corresponding matrices for GA-KF.

Let  $\mathbb{E}[\cdot]$  denote expectation w.r.t. all random quantities conditioned on the event  $D$ , and let  $\mathbb{E}[\cdot | y_1, \dots, y_t]$  denote conditional expectation given  $y_1, \dots, y_t$  and the event  $D$ .

From (5), for  $t > t_*$ ,  $e_t$ ,  $\tilde{e}_t$  and  $\text{diff}_t = e_t - \tilde{e}_t$  satisfy

$$\begin{aligned} e_t &= J_t e_{t-1} + J_t \nu_t - K_t w_t \\ \tilde{e}_t &= \tilde{J}_t \tilde{e}_{t-1} + \tilde{J}_t \nu_t - \tilde{K}_t w_t \\ \text{diff}_t &= J_t \text{diff}_{t-1} + (J_t - \tilde{J}_t)(\tilde{e}_{t-1} + \nu_t) + (\tilde{K}_t - K_t)w_t \end{aligned} \quad (17)$$

For  $t > t_*$  both KF-CS and GA-KF run the same fixed dimensional and fixed parameter KF for  $(x_t)_{N_*}$  with parameters  $F \equiv I$ ,  $Q \equiv (\sigma_{sys}^2 I_{N_*})_{N_*,N_*}$ ,  $C \equiv A_{N_*}$ ,  $R \equiv \sigma^2 I$ , but with different initial conditions. KF-CS uses  $\hat{x}_{t_*}$ ,  $P_{t_*+1|t_*} \neq \mathbb{E}[e_{t_*+1} e'_{t_*+1} | y_1, \dots, y_{t_*}]$  while GA-KF uses the correct initial conditions,  $\hat{x}_{t_*,\text{GAKF}}$ ,  $\tilde{P}_{t_*+1|t_*} = \mathbb{E}[\tilde{e}_{t_*+1} \tilde{e}'_{t_*+1} | y_1, \dots, y_{t_*}] = \mathbb{E}[\tilde{e}_{t_*+1} \tilde{e}'_{t_*+1}]$ . Since  $|N_*| \leq S_{\max}$  and  $\delta_{S_{\max}} < 1$ ,  $C \equiv A_{N_*}$  is full rank. Thus  $(I, C)$  is observable. Also, since  $Q$  is full rank,  $(I, Q^{1/2})$  is controllable. Thus, starting from any initial condition,  $P_{t+1|t}$  will converge to a positive semi-definite,  $P_*$ , which is the unique solution of the discrete algebraic Riccati equation with parameters  $F, Q, C, R$  [13, Theorem 8.7.1]. Consequently  $K_t$  and  $J_t$  will also converge to  $K_* \triangleq P_* A_{N_*}' (A_{N_*} P_* A_{N_*}' + \sigma^2 I)^{-1}$  and  $J_* \triangleq I - K_* A_{N_*}$  respectively. For  $t > t_*$ , the GA-KF also runs the same KF. Thus,  $\tilde{P}_{t|t-1}$ ,  $\tilde{K}_t$ ,  $\tilde{J}_t$  will also converge to  $P_*$ ,  $K_*$ ,  $J_*$  respectively [13, Theorem 8.7.1]. Next, we use this fact to show that the estimation errors also converge in mean square.

Using [13, Theorem E.5.1],  $J_*$  is stable, i.e. its spectral radius  $\rho = \rho(J_*) < 1$ . Let  $\epsilon_0 = (1 - \rho)/2$ . By [21, Lemma 5.6.10], there exists a matrix norm, denoted  $\|\cdot\|_\rho$ , s.t.  $\|J_*\|_\rho \leq \rho + \epsilon_0 = (1 + \rho)/2 < 1$ .

Consider any  $\epsilon < (1 - \rho)/4$ . The above results imply that there exists a  $t_\epsilon > t_*$  s.t. for all  $t \geq t_\epsilon$ ,  $\|K_t - \tilde{K}_t\| < \epsilon$ ,  $\|J_t - \tilde{J}_t\| < \epsilon$  and  $\|J_t\|_\rho < \|J_*\|_\rho + \epsilon < (1 + \rho)/2 + (1 - \rho)/4 = (3 + \rho)/4 < 1$ . Now, the last set of undetected elements of  $N_*$  are detected at  $t_*$ . Thus at  $t_*$ , KF-CS computes a final LS estimate, i.e.  $\hat{x}_{t_*} = A_{N_*}^{-1} y_{t_*}$ ,  $P_{t_*|t_*-1} = \infty$ ,  $P_{t_*} = (A_{N_*}' A_{N_*})^{-1} \sigma^2$ ,  $K_{t_*} = (A_{N_*}' A_{N_*})^{-1} A_{N_*}'$  and  $J_{t_*} = 0$ . None of these depend on  $y_1, \dots, y_{t_*}$  and hence the future values of  $\hat{x}_t$  or of  $P_t, J_t, K_t$  etc also do not. Hence  $t_\epsilon$  also does not.

Since  $\tilde{P}_{t|t-1} \rightarrow P_*$ ,  $\tilde{P}_{t|t-1}$  is bounded. Since  $\tilde{P}_t \leq \tilde{P}_{t|t-1}$ ,  $\tilde{P}_t$  is also bounded, i.e. there exists a  $B < \infty$  s.t.  $\text{tr}(\tilde{P}_t) < B$ ,  $\forall t$ . Since  $\mathbb{E}[\tilde{e}_t \tilde{e}'_t | y_1, \dots, y_{t_*}] = \tilde{P}_t = \mathbb{E}[\tilde{e}_t \tilde{e}'_t]$ , thus  $\mathbb{E}[\|\tilde{e}_t\|^2] = \text{tr}(\tilde{P}_t) < B$ .

Thus, using (17), the following holds for all  $t \geq t_\epsilon$ ,

$$\begin{aligned} \mathbb{E}[\|\text{diff}_t\|^2]^{1/2} &\leq \\ \|M_{t,t_\epsilon}\| \mathbb{E}[\|\text{diff}_{t_\epsilon}\|^2]^{1/2} + \|L_{t,t_\epsilon}\| \sup_{t_\epsilon \leq \tau \leq t} \mathbb{E}[\|u_\tau\|^2]^{1/2}, \text{ where} \\ u_\tau &\triangleq (J_\tau - \tilde{J}_\tau)(\tilde{e}_{\tau-1} + \nu_\tau) + (K_\tau - \tilde{K}_\tau)w_\tau, \\ M_{t,t_\epsilon} &\triangleq \prod_{k=t_\epsilon+1}^t J_k, \quad L_{t,t_\epsilon} \triangleq I + J_t + J_t J_{t-1} + \dots \prod_{k=t_\epsilon+1}^t J_k \end{aligned} \quad (18)$$

Since neither  $t_\epsilon$ , nor the matrices  $J_t$  or  $K_t$  for  $t > t_*$ , depend on  $y_1, \dots, y_{t_*}$ , we do not need to condition the expectation on  $y_1, \dots, y_{t_*}$ .

Notice that

$$1) \sup_{t_\epsilon \leq \tau \leq t} \mathbb{E}[\|u_\tau\|^2]^{1/2} \leq \epsilon(\sqrt{B} + \sqrt{|N_*| \sigma_{sys}^2} + \sqrt{n \sigma^2}).$$

2)  $\|M_{t,t_\epsilon}\|_\rho \leq \prod_{\tau=t_\epsilon+1}^t \|J_\tau\|_\rho < a^{t-t_\epsilon}$  with  $a \triangleq (3 + \rho)/4 < 1$ . Thus  $\|M_{t,t_\epsilon}\| \leq c_{\rho,2} a^{t-t_\epsilon}$  where  $c_{\rho,2}$  is the smallest real number satisfying  $\|M\| \leq c_{\rho,2} \|M\|_\rho$ , for all size  $|N_*|$  square matrices  $M$  (holds because of equivalence of norms).

3)  $\|L_{t,t_\epsilon}\|_\rho \leq 1 + a + \dots a^{t-t_\epsilon} < \frac{1}{(1-a)}$ . Thus  $\|L_{t,t_\epsilon}\| \leq \frac{c_{\rho,2}}{(1-a)}$ .

Combining the above facts, for all  $t \geq t_\epsilon$ ,  $\mathbb{E}[\|\text{diff}_t\|^2]^{1/2} \leq c_{\rho,2} a^{t-t_\epsilon} \mathbb{E}[\|\text{diff}_{t_\epsilon}\|^2]^{1/2} + C\epsilon$  where  $a := (3 + \rho)/4$ ,  $C := \frac{c_{\rho,2}}{1-a}(\sqrt{B} + \sqrt{|N_*| \sigma_{sys}^2} + \sqrt{n \sigma^2})$ . Notice that  $a < 1$ . Consider an  $\tilde{\epsilon} < 2C(1-\rho)/4$  and set  $\epsilon = \tilde{\epsilon}/2C$ . It is easy to see that for all  $t \geq t_\epsilon/2C + \frac{\log(\mathbb{E}[\|\text{diff}_t\|^2]^{1/2}) + \log(2c_{\rho,2}) - \log \tilde{\epsilon}}{\log(1/a)}$ ,  $\mathbb{E}[\|\text{diff}_t\|^2]^{1/2} \leq \tilde{\epsilon}$ . Thus, conditioned on  $D$ ,  $\text{diff}_t$  converges to zero in mean square.

By Markov's inequality, this also implies convergence in probability, i.e. for a given  $\epsilon, \epsilon_{\text{err}}$ , there exists a  $\tau_{KF}(\epsilon, \epsilon_{\text{err}}, N_*) > 0$  s.t. for all  $t \geq t_* + \tau_{KF}(\epsilon, \epsilon_{\text{err}}, N_*)$ ,  $\Pr(\|\text{diff}_t\|^2 < \epsilon_{\text{err}} | D) \geq (1 - \epsilon)$ . The proof of Corollary 1 follows directly from this.

### REFERENCES

- [1] N. Vaswani, "Kalman filtered compressed sensing," in *ICIP*, 2008.
- [2] N. Vaswani, "Analyzing least squares and kalman filtered compressed sensing," in *ICASSP*, 2009.
- [3] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk, "Compressive imaging for video representation and coding," in *Proc. Picture Coding Symposium*, April 2006.
- [4] N. Vaswani, "LS-CS-residual (LS-CS): Compressive Sensing on Least Squares residual," *Accepted (with mandatory revisions) to IEEE Trans. Signal Processing, Arxiv preprint arXiv: 0911.5524v2*, 2010.
- [5] C. Qiu, W. Lu, and N. Vaswani, "Real-time dynamic mri reconstruction using kalman filtered cs," in *ICASSP*, 2009.
- [6] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Info. Th.*, vol. 52(2), pp. 489–509, February 2006.
- [7] D. Donoho, "Compressed sensing," *IEEE Trans. Info. Th.*, vol. 52(4), pp. 1289–1306, April 2006.
- [8] E. Candes and T. Tao, "The dantzig selector: statistical estimation when p is much larger than n," *Annals of Statistics*, vol. 35 (6), 2007.
- [9] E. Candes, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, pp. 589–592, 2008.
- [10] S. S. Chen, Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput*, vol. 20, pp. 33–61, 1998.
- [11] J. A. Tropp, "Just relax: Convex programming methods for identifying sparse signals," *IEEE Trans. Info. Th.*, pp. 1030–1051, March 2006.
- [12] U. Gamper, P. Boesiger, and S. Kozerke, "Compressed sensing in dynamic mri," *Magnetic Resonance in Medicine*, vol. 59(2), pp. 365–373, January 2008.
- [13] T. Kailath, A.H. Sayed, and B. Hassibi, *Linear Estimation*, Prentice Hall, 2000.
- [14] S. Sarvotham, D. Baron, and R. Baraniuk, "Compressed sensing reconstruction via belief propagation," in *Tech. rep. ECE-06-01, Dept. of ECE, Rice Univ.*, July 2006.
- [15] H. Jung, K. H. Sung, K. S. Nayak, E. Y. Kim, and J. C. Ye, "k-t foccus: a general compressed sensing framework for high resolution dynamic mri," *Magnetic Resonance in Medicine*, To appear.
- [16] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Sig. Proc.*, to appear.
- [17] P. Schniter, L. Potter, and J. Ziniel, "Fast bayesian matching pursuit: Model uncertainty and parameter estimation for sparse linear models," in *Information Theory and Applications (ITA)*, 2008.
- [18] D. Angelosante and G.B. Giannakis, "Rls-weighted lasso for adaptive estimation of sparse signals," in *ICASSP*, 2009.
- [19] N. Vaswani and W. Lu, "Modified-cs: Modifying compressive sensing for problems with partially known support," in *IEEE Int'l. Symp. Info. Th. (ISIT)*, 2009.
- [20] C. Qiu and N. Vaswani, "Compressive sensing on the least squares and kalman filtering residual for real-time dynamic mri and video reconstruction," *IEEE Trans. Image Proc.*, 2009, submitted.
- [21] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge Univ. Press, 1985.
- [22] D. Ocone and E. Pardoux, "Asymptotic stability of the optimal filter with respect to its initial condition," *SIAM Journal of Control and Optimization*, pp. 226–243, 1996.