Summary The paper provides conditions under which dictionary learning is stable (in particular to noise), which apply in almost all cases. This, the authors claim, provides some support for the empirical observation that different dictionary learning methods lead to very similar dictionaries.

Comments The paper is quite well-written. I am not an expert in dictionary learning per se, but if these are truly the first perturbation bounds for dictionary learning, then they are worth publishing in TSP.

Other comments

1. "then each submatrix formed from $2k$ of its columns or less has a strictly positive lower bound." What is k here?

2. 3 lines below (5): why is $\cup\mathcal{H} = [m]$ required?

3. Are J and $\overline{J}$ in Theorem 1 of same size?

4. About Theorem 1. I am a little worried about C_1 depending on $\mathbf{A}$ and the $\mathbf{x}_i$'s. It could make some of parts of the statement void. Presumably, one sees these as fixed, and looks at the possible $\mathbf{B}$'s and possible $\overline{\mathbf{x}}_i$'s satisfying $\|\mathbf{A}\mathbf{x}_i - \mathbf{B}\overline{\mathbf{x}}_i\|_2 \leq \varepsilon$?

5. Also, in (7), using $\mathbf{x}_i$ to denote the restriction of $\mathbf{x}_i$ to J, and similarly for $\overline{\mathbf{x}}_i$, could lead to confusion, in particular if the reader glances at the result without reading the last sentence.

6. "A practical implication of Thm. 1 is the following: there is an effective procedure sufficient to affirm if a proposed solution to Prob. 1 is indeed unique (up to noise and inherent ambiguities). One need simply to check that the matrix and codes satisfy the (computable) assumptions of Thm. 1 on $\mathbf{A}$ and the $\mathbf{x}_i$." I question whether the assumptions of Theorem 1 are indeed computable. Could the authors elaborate on that?

7. I am not sure what "substitutions" means in Theorem 3. Does that mean "choices" of $\mathbf{A}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_N$ such that $\mathbf{A}\mathbf{x}_i = \mathbf{y}_i$? Isn't Theorem 3 a simple corollary of Theorem 1, together with the remark in the couple of paragraph above it? If so, I suggest making it a corollary. In the same vein, I am not sure that Proposition 1 is rigorous enough to make it a proposition. It could be stated as an informal claim instead.

8. The perturbation bound is used by the authors to provide some explanation for why, in practice, different methods lead to similar dictionaries. Can that also be used to explain the second (bold) part of their observation that "[these waveforms] appear in dictionaries learned by a variety of algorithms trained **with respect to different natural image datasets.**"? Just curious.