Journal of Machine Learning Research 20 (2019) 1-42 Submitted 10/18; Revised 11/19; Published 11/19

Learning Overcomplete, Low Coherence Dictionaries with Linear Inference

Jesse A. Livezey JLIVEZEY@LBL.GOV
Biological Systems and Engineering Division
Lawrence Berkeley National Laboratory
Berkeley, California 94720, USA
Redwood Center for Theoretical Neuroscience
University of California, Berkeley
Berkeley, California 94720, USA

Alejandro F. Bujan AFBUJAN@GMAIL.COM
Friedrich T. Sommer FSOMMER@BERKELEY.EDU
Redwood Center for Theoretical Neuroscience
University of California, Berkeley
Berkeley, California 94720, USA

Editor: Aapo Hyvärinen

Abstract

Finding overcomplete latent representations of data has applications in data analysis, signal processing, machine learning, theoretical neuroscience and many other fields. In an overcomplete representation, the number of latent features exceeds the data dimensionality, which is useful when the data is undersampled by the measurements (compressed sensing or information bottlenecks in neural systems) or composed from multiple complete sets of linear features, each spanning the data space. Independent Components Analysis (ICA) is a linear technique for learning sparse latent representations, which typically has a lower computational cost than sparse coding, a linear generative model which requires an iterative, nonlinear inference step. While well suited for finding complete representations, we show that overcompleteness poses a challenge to existing ICA algorithms. Specifically, the coherence control used in existing ICA and other dictionary learning algorithms, necessary to prevent the formation of duplicate dictionary features, is ill-suited in the overcomplete case. We show that in the overcomplete case, several existing ICA algorithms have undesirable global minima that maximize coherence. We provide a theoretical explanation of these failures and, based on the theory, propose improved coherence control costs for overcomplete ICA algorithms. Further, by comparing ICA algorithms to the computationally more expensive sparse coding on synthetic data, we show that the limited applicability of overcomplete, linear inference can be extended with the proposed cost functions. Finally, when trained on natural images, we show that the coherence control biases the exploration of the data manifold, sometimes yielding suboptimal, coherent solutions. All told, this study contributes new insights into and methods for coherence control for linear ICA, some of which are applicable to many other nonlinear models.

Keywords: independent components analysis, dictionary learning, coherence

1. Introduction

Mining the statistical structure of data is a central topic of machine learning and is also a principle for computational models in neuroscience. A prominent class of such algorithms is dictionary learning, which reveal a set of structural primitives in the data, the dictionary, and a corresponding latent representation, often regularized by sparsity. In this work, we focus on overcomplete dictionary learning (Olshausen and Field, 1997; Hyvärinen, 2005; Le et al., 2011), the case when the dimension of the latent representation exceeds the dimension of the data and therefore the linear filters (dictionary) generating the data cannot all be mutually orthogonal.

Independent Components Analysis (ICA) (Comon, 1994; Bell and Sejnowski, 1997) is a technique for learning the underlying non-Gaussian and independent sources, S, in a dataset, X. ICA is commonly used in the complete, noiseless case, although methods which can be run on noisy data exist Hyvarinen (1999). When run as a complete, noiseless model, ICA is computationally light-weight because the learned mappings between data and sources are linear in both directions. ICA can be formulated as a noiseless linear generative model

$$X_i = \sum_{j=1}^{L} A_{ij} S_j, \tag{1}$$

where $A \in \mathbb{R}^{D \times L}$ is referred to as the *mixing matrix* wherein D is the dimensionality of the data, X, and L is the dimensionality of the sources, S. In the complete case ($D = L$), the goal of ICA is to find the *unmixing matrix* $W \in \mathbb{R}^{L \times D}$ such that the sources for all M data samples can be recovered, $S_j^{(i)} = \sum_k W_{jk} X_k$ with $W = A^{-1}$. The unmixing matrix W can then be obtained by minimizing the negative log-likelihood of the model

$$-\log P(X; W) = \sum_{i=1}^{M} \sum_{j=1}^{L} g(\sum_k W_{jk} X_k^{(i)}) - M \log(\det(W)) \tag{2}$$

where $g(\cdot)$ specifies the shape of the negative log-prior of the latent variables S and is usually a smooth version of the L_1 norm such as the $\log(\cosh(\cdot))$, which encourages the projections of X to be sparse, $X^{(i)}$ is the ith element of the dataset, X, which has M samples, and where the bases are constrained to have unit-norm. The log-determinant comes from the multivariate change of variables in the likelihood from X to S

$$P(X) = P(S) |\det \frac{dS}{dX}| = P(W \cdot X) |\det W|. \tag{3}$$

If the data has been whitened, the unconstrained optimization (Eq 2) can be replaced by a constrained optimization where the second term in the cost function is replaced with the constraint $WW^T = I$ (Hyvärinen and Oja, 1997).

In complete ICA, the log-determinant (or the identity constraint) will prevent multiple elements of the dictionary, W, from learning the same feature. In overcomplete ICA, the linear generative model (Eq 1) cannot be inverted, and therefore, overcomplete versions of Eqs 2 and 3 cannot be derived. One alternative to maximum likelihood learning is to create

an objective function by adding a new cost, $C(W)$, to the sparsity prior (Hyvärinen and Inki, 2002; Le et al., 2011). The new unconstrained objective function becomes

$$\text{Objective}(W) = \lambda \sum_{i=1}^{M} \sum_{j=1}^{L} g(\sum_{k} W_{jk} X_k^{(i)}) + C(W). \tag{4}$$

This form is also similar to the log-posterior of the sources, S, which appears in sparse coding models and is used in maximum a posteriori inference, although here, the cost is used to optimize the dictionary, W, not estimate sources. Overcomplete ICA models of this form are one case of analysis dictionary learning methods, where the projections of the data are assumed to have sparse structure rather than assuming a sparse linear generative (synthesis) model. In complete ICA methods, there is no distinction between the synthesis and the analysis view (Bell and Sejnowski, 1997; Hyvärinen et al., 2001; Elad et al., 2007; Teh et al., 2003; Ophir et al., 2011; Rubinstein et al., 2013; Chun and Fessler, 2018)

$$\begin{aligned} \text{Synthesis: } X = AS,\ S \text{ sparse} \\ \text{Analysis: } P = WX,\ P \text{ sparse.} \end{aligned} \tag{5}$$

In overcomplete ICA, these two formulations are no longer equivalent to each other.

The cost, $C(W)$, should be chosen to exert coherence control on the dictionary, that is, to prevent the co-alignment of the bases. The coherence of a dictionary is defined as the maximum absolute value of the off-diagonal elements of the Gram matrix of a unit-normalized dictionary (Davenport et al., 2011), W,

$$\text{coherence}(W) \equiv \max_{i \neq j} |\sum_{k} W_{ik} W_{jk}| = \max_{i \neq j} |\cos \theta_{ij}| \tag{6}$$

where $\sum_k W_{ik} W_{jk} = \cos \theta_{ij}$ is the cosine similarity between the unit normalized dictionary elements W_i and W_j. A dictionary with high coherence (near 1) will have duplicated or nearly duplicated bases.

A number of methods for coherence control in complete and overcomplete methods have been proposed including a quasi-orthogonality constraint (Hyvärinen et al., 1999), a reconstruction cost (Le et al., 2011) which is equivalent to the L_2 coherence cost in Eq 8 (Ramırez et al., 2009; Sigg et al., 2012; Bao et al., 2014; Chun and Fessler, 2018; Bansal et al., 2018), and a Random Prior cost (Hyvärinen and Inki, 2002) (see Section 3 for details). However, a systematic analysis of the properties of proposed coherence control methods and a comparison with methods that extend more naturally to overcomplete representations, for example, sparse coding, is still missing in the literature. In particular, the L_2 cost is often claimed to promote diversity or incoherence in overcomplete dictionaries elements, which we will show is not the case.

Our first theoretical result is that although the global minima of the L_2 cost have minimal coherence (coherence = 0) for a complete basis, in the overcomplete case, it has global minima with maximal coherence (coherence = 1). We introduce an analytic framework for evaluating different coherence control costs, and propose several new costs, which fix deficiencies in previous methods. Our first novel cost is the L_4 cost on the difference between the identity matrix and the Gram matrix of the bases. The second is a cost which we call

the *Coulomb* cost because it is derived from the potential energy of a collection of charged particles bound to the surface of an n-sphere. We also propose modifications to previously proposed methods of coherence control which we show allows them to learn less coherent dictionaries.

In addition to controlling coherence, we show empirically that these costs will influence the entire distribution of the learned bases in an overcomplete dictionary. We investigate the impact of different coherence control costs on recovering overcomplete synthesis and analysis models. Finally, we evaluate the coherence and diversity of bases learned on a dataset of natural image patches.

1.1. Related work

Studying methods for learning overcomplete dictionaries is motivated from applications in data analysis and theoretical neuroscience. In data analysis, overcomplete dictionaries become essential if data are either undersampled (Hillar and Sommer, 2015), or have a sparse structure with respect to a combination of orthobases (Donoho and Elad, 2003). In neuroscience, dictionary learning has not only been proposed for data analysis (Delorme et al., 2007; Agarwal et al., 2014; Hirayama et al., 2015), but also as a computational model for understanding the formation of sensory representations (Bell and Sejnowski, 1997; Olshausen and Field, 1996; Klein et al., 2003; Smith and Lewicki, 2006; Rehn and Sommer, 2007; Zylberberg et al., 2011; Carlson et al., 2012). It has been estimated from anatomical data that in primary sensory areas the number of neurons by far exceeds the number of afferent inputs (Barlow, 1981; Spoendlin and Schrott, 1989; Curcio and Allen, 1990; Leuba and Kraftsik, 1994; Northern and Downs, 2002; DeWeese et al., 2005). Further, it has been shown that dictionary learning forms more diverse sets of features when overcomplete, which more closely matches the diversity of receptive fields found in sensory cortex (Rehn and Sommer, 2007; Carlson et al., 2012; Olshausen, 2013).

Sparse coding is a linear generative model for dictionary learning, which unlike typical ICA models, also includes an additive noise term to the mixtures

$$X_i = \sum_{j=1}^{L} A_{ij} S_j + \epsilon_i, \tag{7}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma)$. Sparse coding requires an iterative, computationally complex maximum a posteriori estimation, posterior estimation step, or an approximation to these (Olshausen and Field, 1996; Lewicki and Sejnowski, 2000; Rehn and Sommer, 2007; Rozell et al., 2008; Gregor and LeCun, 2010; Hu et al., 2014). However, unlike ICA, sparse coding extends naturally to the overcomplete setting without modification. During inference, latent features in overcomplete sparse coding models (Lewicki and Olshausen, 1999) have an explaining-away effect on each other which discourages them from learning coherent solutions. Methods for incoherent overcomplete dictionary learning which add additional coherence costs, including the L_2 cost, with nonlinear inference have also been proposed (Ramırez et al., 2009; Sigg et al., 2012; Mailhé et al., 2012; Bao et al., 2014). Score matching (Hyvärinen, 2005) is another alternative to maximum likelihood learning which can be used for overcomplete ICA models.

In overcomplete dictionary learning, a distinction is made between synthesis and analysis methods. Synthesis methods posit that data is formed from a linear combination (dictionary) of sparse sources, and methods are designed to both recover the dictionary and the sources (Olshausen and Field, 1997; Chen et al., 2001; Davenport et al., 2011). Analysis dictionary learning assume that the atoms which combine linearly to makeup a signal have sparse projections from some analysis matrix (Ophir et al., 2011; Rubinstein et al., 2013; Chun and Fessler, 2018). In the overcomplete case, these two problems will deviate (Elad et al., 2007). Methods have been proposed for inference in analysis models (Elad et al., 2007) as well as analysis dictionary learning (Ophir et al., 2011; Rubinstein et al., 2013; Chun and Fessler, 2018).

2. Results

In this section we first prove that the L_2 cost has global minima with coherence = 1. We then propose new coherence control costs and evaluate them on synthetic datasets and natural images.

2.1. The L_2 cost has high coherence global minima

Dictionary or representation learning methods often augment their cost functions with additional terms aimed at learning less coherent features (Ramırez et al., 2009; Le et al., 2011; Sigg et al., 2012; Bao et al., 2014; Chun and Fessler, 2018; Bansal et al., 2018) or making learning through optimization more efficient (Howard et al., 2008). The L_2 cost, defined for a unmixing matrix, W, as

$$C_{L_2}(W) = \frac{1}{2}\sum_{ij}(\delta_{ij} - \sum_k W_{ik}W_{jk})^2 = \frac{1}{2}\sum_{ij}(\delta_{ij} - \cos\theta_{ij})^2, \tag{8}$$

has been used to augment dictionary learning methods motivated by the desire to learn more incoherent or diverse dictionaries (Strohmer and Heath Jr, 2003; Davenport et al., 2011). However, we show that minimizing the L_2 cost is a necessary but not sufficient condition for finding *equiangular tight frames* (see Section 3.1.1 for details and definitions), a certain class of minimum coherence solutions. Moreover, we prove that the L_2 cost has global minima with maximum coherence. This shows that the L_2 cost and its related costs are not providing coherence control in overcomplete dictionaries.

For the L_2 cost, it can be shown that for integer overcompleteness, there exists a set of global minima in which the angle between many pairs of bases is exactly zero and the coherence is 1, the maximum attainable value. We prove the following theorem:

Theorem 1 *Let $W_0 \in \mathbb{R}^{L\times D}$ be an overcomplete unmixing matrix with data dimension D and latent dimension $L = n \times D$, with $n > 1$, $\in \mathbb{Z}$ and unit-norm rows. There exist dictionaries, W_0, that are global minima of the L_2 cost with coherence = 1.*

This shows that the L_2 cost has global minima that have the exact property it was proposed to prevent (high coherence). The proof of this theorem also shows that, in the complete case ($n = 1$), an orthonormal basis is a global minimum of the L_2 cost. We also prove that there are operators which transform the pathological solution (coherence = 1) into non-pathological solutions (coherence < 1) to which the L_2 cost is invariant:

Theorem 2 *Let P be a projection operator from the L dimensional space of dictionary elements to a D dimensional subspace and P^C its complement projection. Φ is constructed as the sum of any rotation, $R \in O(L)$, projected within the D-dimensional subspace of the dictionary elements and an identity applied to the complement subspace, $\Phi = PR + P^C$. There exist non-trivial continuous transformations: Φ, on W_0 to which the L_2 cost is invariant. These transformed dictionaries, $W_0\Phi$, have coherence ≤ 1 for non-identity rotations and are global minima of the L_2 cost.*

Appendices B.1 and B.4 contain the proofs of these theorems.

These high coherence global minima are illustrated with a two dimensional, two times overcomplete example in Fig 1. It can be shown that there are pathological (high coherence) minima (Fig 1A) which can be continuously rotated into other low coherence minima (Fig 1B). These configurations are equivalent in terms of the value of the L_2 cost and lie on a connected global minimum. These families of configurations are minima if it can be shown that the gradient of the cost is zero, that is, they are critical points of the cost, and that the Hessian is positive definite in all directions but the one that rotates the configuration within the family of solutions. We will show these two things through an explicit derivation in the 2 dimensional case.

In order to understand these minima, we evaluate the L_2 cost in a two dimensional example analytically. The global rotational symmetry of the L_2 cost allows us to parameterize all solutions with respect to one fixed dictionary element: $(1, 0)$, without loss of generality. The four dictionary elements, shown in Fig 1, are

$$(1, 0), (\cos\theta_1, \sin\theta_1), (\cos\theta_2, \sin\theta_2), (\cos\theta_2 + \theta_3, \sin\theta_2 + \theta_3). \tag{9}$$

Setting θ_1 and θ_3 to $\pi/2$, that is, creating two sets of orthonormal bases, forms a ring of minima as θ_2 is varied. This can be shown by computing the gradient and the eigenvalues of the Hessian of the L_2 cost at these points. The cost function, gradient, and Hessian are tabulated in Appendix A and the eigenvalues are plotted individually in Fig 1.

The value of the L_2 cost is constant as a function of θ_2 (Fig 1C, purple, dashed line) even though the coherence is drastically changing as a function of θ_2. The Hessian of the L_2 cost along this path has one eigenvalue that is 0 as a function of θ_2 whose eigenvector corresponds to changing θ_2 with fixed θ_1 and θ_3 (Fig 1D, see Appendix A for the exact functional form). The other two eigenvalues are greater than zero and greater then zero for $\theta_2 \neq 0$ respectively which shows that the cost is a minimum almost everywhere along this path. At $\theta_2 = 0$, the second eigenvalue becomes 0. This eigenvalue has eigenvector $(-1, 0, 1)$. If we evaluate the cost along this direction centered at $\theta_1 = \theta_2 = \theta_3 = 0$, we find that although the second derivative is zero, the fourth derivative is positive showing that indeed, this point is a minimum (see Appendix A.2 for a derivation).

In many previous studies, the L_2 cost or variations of it were proposed in order to learn dictionaries with lower coherence (Ramırez et al., 2009; Le et al., 2011; Sigg et al., 2012; Bao et al., 2014; Chun and Fessler, 2018; Bansal et al., 2018). The results in this section show that the L_2 cost function does not provide coherence control in the overcomplete regime. In fact, dictionaries that should be maxima are part of a set of global minima. This indicates that there is a need for new forms of coherence control.

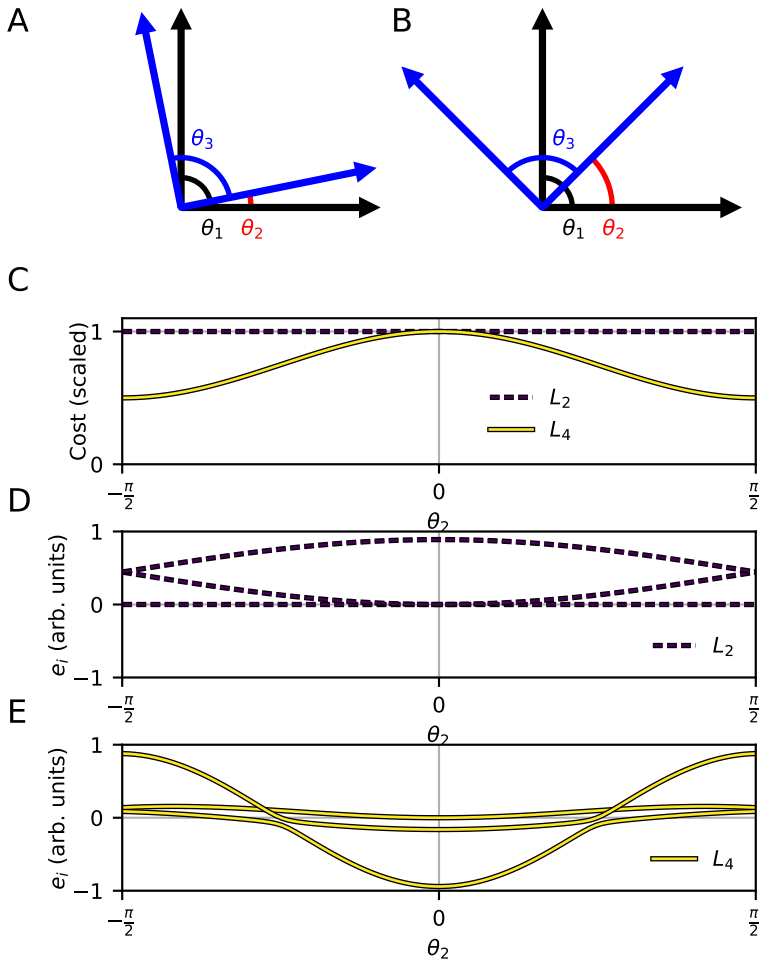

Figure 1: Structure of the pathological global minimum in the L_2 cost which the L_4 cost corrects. In **A** and **B**, each arrow represents a dictionary element in a 2-times overcomplete dictionary in a 2-dimensional space. **A** A dictionary with high coherence which has the same value of the cost as the dictionary in **B** for any θ_2 including the pathological solution $\theta_2 \to 0$. **B** A dictionary with low coherence. **C** The L_2 and L_4 costs are plotted at $\theta_1 = \theta_3 = \pi/2$ as a function of θ_2. The costs have been scaled so that their maximum value is 1. **D** The eigenvalues of the L_2 cost at $\theta_1 = \theta_3 = \pi/2$ as a function of θ_2 scaled between -1 and 1. **E** The eigenvalues of the L_4 cost at $\theta_1 = \theta_3 = \pi/2$ as a function of θ_2 scaled between -1 and 1.

2.2. Addressing high coherence solutions: L_4 and Coulomb costs

The rotational symmetry in the L_2 cost leads to its pathological (high coherence) global minima, and this insight motivates a simple modification which will not have high coherence minima. We propose a novel coherence control cost termed the L_4 cost, which removes the pathological minima of the L_2 cost. The motivation for this cost function is to more strongly penalize large inner products in the gram matrix. The L_4 cost function also acts on the gram matrix of W, but raises each off diagonal element to the fourth power which breaks the rotational symmetries which lead to the pathological minima

$$C_{L_4}(W) = \frac{1}{4}\sum_{ij}(\delta_{ij} - \sum_k W_{ik}W_{jk})^4 = \frac{1}{4}\sum_{ij}(\delta_{ij} - \cos\theta_{ij})^4. \tag{10}$$

Following the same analysis as in Section 2.1, we show that the pathological solutions are either reduced to saddle points at $\theta_2 = n\frac{\pi}{2}$ or local minima at $\theta_2 = (2n+1)\frac{\pi}{4}$, which correspond to incoherent solutions (Fig 1E). The L_4 cost as a function of θ_2 has a maximum at $\theta_2 = 0$ (coherent solutions) and minima at $\theta_2 = \frac{\pi}{2}$ (Fig 1C). The L_4 cost function, gradient, and Hessian are tabulated in Appendix A for this 2 dimensional example.

We also propose a second alternative cost, where the repulsion from high coherence is *Coulombic*: the Coulomb cost. Coherence control can then be related to the problem of characterizing the minimum potential energy states of L charged particles on an n-sphere, an open problem in electrostatics (Smale, 1998). The energy, E^{Coulomb}, of two charged point particles of the the same sign is proportional to the inverse of their distance, $\vec{r}_{ij}$

$$E_{ij}^{\text{Coulomb}} \propto \frac{1}{|\vec{r}_{ij}|}. \tag{11}$$

When constrained to the surface of the unit-radius n-sphere, the distance between two points can be written as a function of only the angle between the two points $|r_{ij}| = \sqrt{1-\cos^2(\theta_{ij}/2)}$. In the case of same-sign charged particles, the minimum energy is when the particles are at antipodal points. However, in ICA, there is no distinction between a dictionary element and its negative (the antipodal point). Instead, the minimum energy configuration should be when two elements are perpendicular. To map this problem onto ICA, the cost should be made symmetric around $\theta = \pi/2$ rather than $\theta = \pi$ since a dictionary element and its negative should have maximal pairwise energy, not minimal. This can be accomplished by replacing θ with 2θ, that is, $\sqrt{1-\cos^2(\theta_{ij}/2)} \to \sqrt{1-\cos^2\theta_{ij}}$. Therefore, the Coulomb cost can be formulated as

$$C_{\text{Coulomb}}(W) = \sum_{i\neq j}\frac{1}{\sqrt{1-\cos^2\theta_{ij}}} = \sum_{i\neq j}\frac{1}{\sqrt{1-(\sum_k W_{ik}W_{jk})^2}}. \tag{12}$$

In practice, we subtract the value of the cost for perpendicular bases, 1, for each pair $i \neq j$ to bring the cost into a better dynamic range. This cost diverges as coherence $\to 1$, which means it cannot have high coherence minima.

2.3. Numerical investigations of coherence control

The above analysis provides evidence of a failure of the L_2 cost to provide coherence control. The alternative coherence cost function can prevent high coherence solutions, but all costs

functions will act on the entire distribution of dictionary elements, not only the high coherence pairs. Deriving the distribution of pairwise angles in the minima of the cost functions is analytically difficult. However, understanding the influence of the coherence control cost function on the distribution of dictionary elements allows us to better understand their biases.

In order to understand the origin of the effects of the different coherence controls on the pairwise angle distributions, the coherence costs can be directly compared without the data dependent ICA sparsity prior. We use two different initializations of the bases and optimize the data-independent coherence costs. These initializations are: a noisy pathological initialization (as in Section 2.1) and a random uniform initialization on the surface of a n-sphere (Gaussian distributed entries normalized to unit-norm elements). We will numerically explore the minima of these cost function for a 2 times overcomplete dictionary in a 32 dimensional data space by minimizing the cost function with these two initializations.

The noisy pathological initialization tiles an orthonormal, complete basis two times and adds a relatively small ($\sigma = .01$) amount of zero-mean Gaussian noise to every basis element to create W. As shown by the red-dashed histogram in Fig 2A, most pairwise angles start close to either 0 or $\frac{\pi}{2}$ as shown in the two peaks in the initial distribution. Minimizing the L_2 cost (purple line) from this initialization gives a final solutions with high coherence, similar to the initial distribution. The other costs push the pairs of bases with initially small pairwise angles apart. This shows numerically that the L_2 does not provide coherence control for overcomplete dictionaries unlike other proposed methods. Appendix Fig C1 contains the same analysis for the full set of cost functions, and Appendix Fig C2 contains a comparison across powers from 1 to 6. Although the L_4 cost may have saddle points in the cost landscape (see Section 2.2), in practice they do not seem to be a problem for optimization (see Appendix Fig C3).

In the random uniform case, the elements of W are drawn independently from a uniform distribution on the unit n-sphere. The final distribution of pairwise angles for the L_2 cost peaks at $\frac{\pi}{2}$ but also has a longer tail towards small pairwise angles. The other costs have shorter tails and have varying amounts of density near $\frac{\pi}{2}$. Of all costs, the L_4 cost distributes the angles most evenly which is reflected by its distribution having the narrowest width and lowest coherence.

Together, these results show that the L_2 cost does not provide coherence control and is also sensitive to the initialization method. The proposed L_4 and Coulomb cost, as well as the previously proposed Random Prior (see Section 3), all provide coherence control. For these three costs, the distribution from which the dictionary was initialized does not have a large effect on the distributions at the numerical minima. These traits mean that they are better suited for providing coherence control in overcomplete dictionary learning methods.

2.4. Flattened costs

The previous analysis provides insight into why different cost function have different behavior for small angles (high coherence). However, the L_4, Coulomb, and Random Prior cost also show qualitatively different behavior in their distributions near $\frac{\pi}{2}$. Both the Coulomb and Random Prior have density near $\frac{\pi}{2}$ for the distribution of pairwise angles, meaning that a fraction of the bases are nearly orthogonal. The L_4 has much lower density near $\frac{\pi}{2}$,

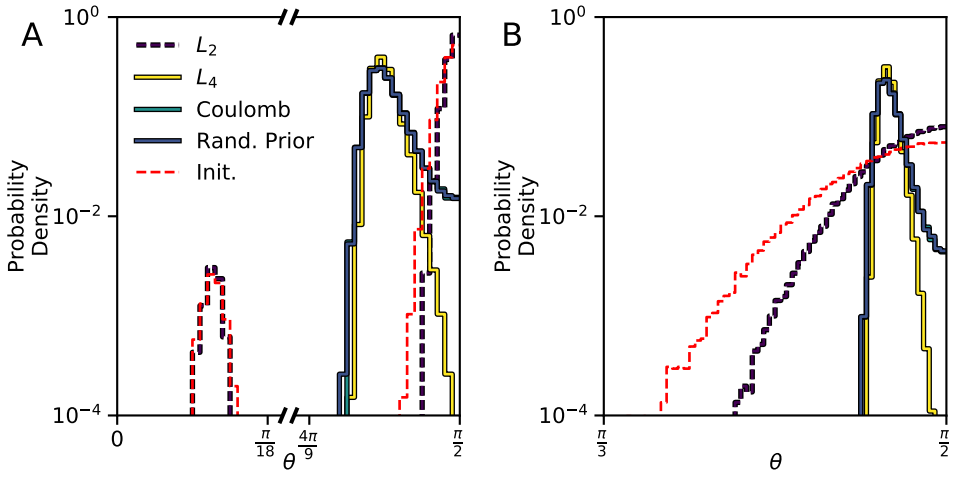

Figure 2: Coherence control costs have minima with varying coherence which can depend on initialization. Color legend is preserved across panels. For both panels a 2 times overcomplete dictionary with a data dimension of 32 was used and the distributions are averaged across 10 random initializations. **A** Distribution of pairwise angles (log scale) obtained by numerically minimizing a subset of the coherence cost functions for the pathological dictionary initialization. Red dotted line indicates the initial distribution of pairwise angles. Note that the horizontal axis is broken. **B** Angle distributions obtained through optimization from a uniform random dictionary initialization. Note that the horizontal axis only includes the range from $\frac{\pi}{3}$ to $\frac{\pi}{2}$. In both plots, the Coulomb and Random Prior lines are almost entirely overlapping.

and a correspondingly lower coherence (smallest pairwise angle). In order to achieve minimal coherence (equiangular tight frame) for an overcomplete dictionary, the distribution of pairwise angles should form a delta-function away from $\frac{\pi}{2}$. Therefore high density near $\frac{\pi}{2}$ may not be desirable for learning low coherence solutions.

In order to gain more insight into the causes of the qualitative differences in the distributions of angles, we analyze the behavior of the costs around $\theta = 0$ and $\theta = \frac{\pi}{2}$ (Fig 3A, B respectively). The gradient of the cost close to $|\cos\theta| = 1$ is proportional to the force the angles feel to stay away from zero which will influence the high coherence tail of the distribution. Taylor expanding all the costs near $\cos\theta = 0$ reveals that all cost functions have non-zero second order terms except for the L_4 cost which only has a fourth order term with linear and cubic terms in their gradients respectively as shown in Fig 3A. Cost with gradients that have lower-order Taylor expansions near $\cos\theta = 0$ encourage pairs of basis vectors to be more orthogonal at the expense of higher coherence. This may lead to distributions of pairwise angles which are less uniform over all pairs of elements of the dictionary. Since the L_4 cost only has higher order derivatives near $\cos\theta = 0$, there is no pressure to form exactly perpendicular pairs. This can also be seen in an L_1 cost which encourages many pair to be almost exactly perpendicular at the expense of many pairs with maximum coherence (see Appendix Fig C2).

We hypothesize that the quadratic terms are creating higher coherence minima with more pairwise angles close to $\frac{\pi}{2}$. This is additional motivation for the L_4 cost and leads us to propose modified versions of the Coulomb and Random Prior costs where the quadratic terms have been removed. The Random Prior cost (Hyvärinen and Inki, 2002) is derived from the distribution of angles expected between pairs of angles randomly drawn on the surface of an n-sphere and is described in Section 3. This can be done by subtracting the quadratic term in the Taylor series from the original cost function

$$C_{\text{Flat}}(\cos\theta_{ij}) = C(\cos\theta_{ij}) - \left.\frac{\partial^2 C(\cos\theta_{ij})}{\partial \cos\theta_{ij}^2}\right|_0 \cos^2\theta_{ij}. \tag{13}$$

This hypothesis can be validated numerically. We compared the distribution of pairwise angles when the Coulomb nad Random Prior costs were minimized with their flattened counterparts. Both the Flattened Coulomb and Random Prior costs (Fig 3C, dotted) show pairwise angle distributions which have lower coherence and fewer pairwise angles close to 90 degrees compared to the original costs (Fig 3C, solid). This shows that across costs, the quadratic terms dominate the behavior of the pairwise angle distributions near 90 degrees and can have a small effect on the coherence on the distributions.

These coherence control methods will also have different behaviors as a function of overcompleteness. To understand their behavior, we measured the coherence of their minima as a function of overcompleteness. Fig 3D shows the minimum pairwise angle (arccos of coherence, low coherence is high minimum pairwise angle) of these methods as a function of overcompleteness at fixed data dimensionality. The median over random initializations of the minimum pairwise angle between dictionary elements for numerically minimized coherence costs is shown. The cost functions evaluated here fall into three groups with quantitatively similar intra-group coherence as a function of overcompleteness. The L_2 cost has the highest coherence (smallest pairwise angle) for all overcompletenesses greater than

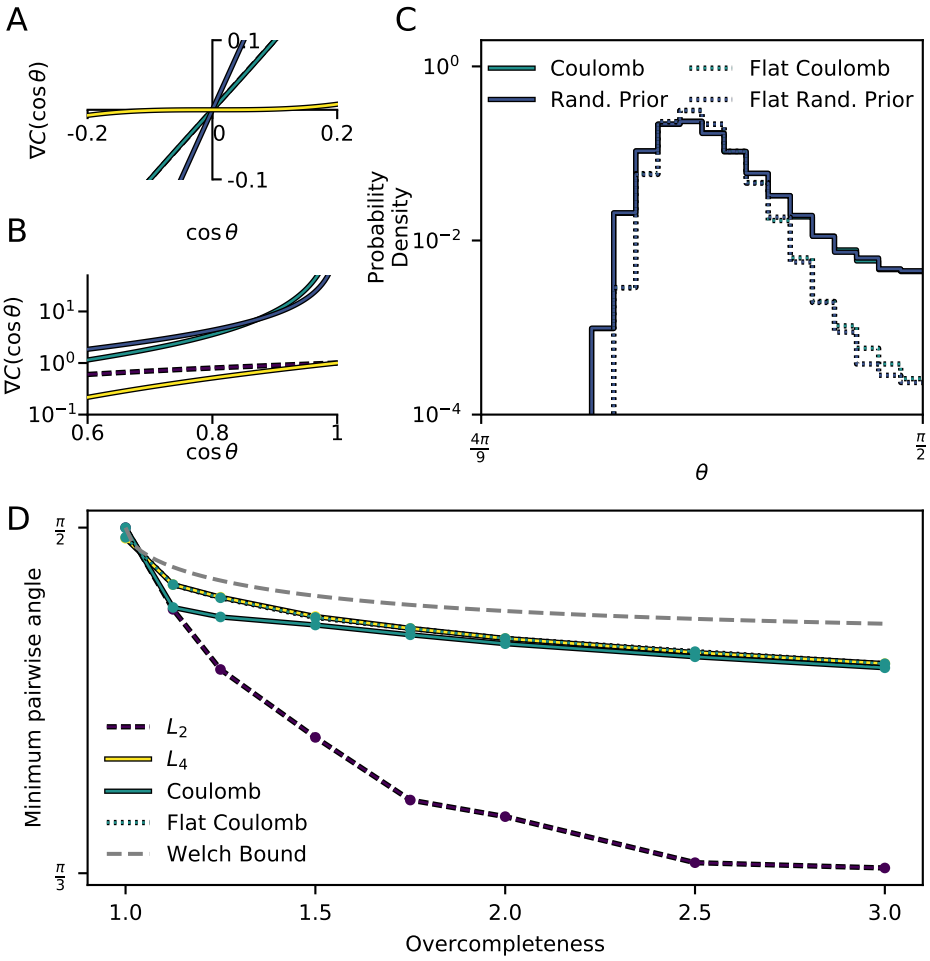

Figure 3: Quadratic terms dominate the minima of coherence control costs. **A** Gradient of the costs as a function of $\cos\theta$ near $\cos\theta = 0$. **B** Gradient of the costs as a function of $\cos\theta$ near $\cos\theta = 1$. **C** Distibution of pairwise angles for a 2 times overcomplete dictionary with a data dimension of 32 from 10 random uniform initializations. The Coulomb and Random Prior cost function distributions are shown (solid lines) along with their counterparts with quadratic terms removed ("flattened", dashed). The Coulomb and Random Prior lines are almost entirely overlapping. **D** The median minimum pairwise angle (arccosine of coherence) across 10 initializations is plotted as a function of overcompleteness for a dictionary with a data dimension of 32. The largest possible value (Welch Bound) is also shown as a function of overcompleteness. The L_4 and Flat Coulomb lines are almost entirely overlapping.

1. The L_4 cost and flattened versions of the Random Prior and Coulomb costs have the lowest coherence. The Random Prior and Coulomb costs behave similarly to the L_2 costs for low overcompleteness (less than 1.5) and then converge to be similar to the L_4 and flattened costs for high overcompletenesses (greater than 2). Fig C6 contains a detailed Coulomb and Random Prior comparison. The Welch Bound (Welch, 1974) is a lower bound for the smallest possible coherence (upper bound on the largest minimum pairwise angle) achievable (Fig 3D). The best coherence control cost functions approach, but do not saturate this bound. Note that constructing overcomplete dictionaries that saturate this bound for arbitrary overcompleteness is an open problem (Strohmer and Heath Jr, 2003; Fickus and Mixon, 2015). This shows that the quadratic terms in the cost function are dominating the coherence behavior of the cost functions and that removing the term as in the flattened costs or only including quartic terms as in the L_4 leads to lower coherence solutions.

These results show that proposed coherence control methods prevent high coherence to different degrees, and furthermore that the choice of coherence control, which is meant to affect the distribution of small pairwise angles, has an effect on the entire distribution of angles. Specifically, the L_2 cost does not provide coherence control and leads to solutions which are heavily biases by initialization unlike other proposed costs. These results also validate the relationship between second order terms in the cost function and the trade-off between coherence and orthogonality. Furthermore, since the costs were investigated without the data-dependent ICA prior, they should be useful for augmenting methods including sparse coding, deep learning, and anything that learns a overcomplete dictionary or weight matrix.

2.5. Recovery of the synthesis mixing matrix with overcomplete ICA

The previous analysis considered the data-independent coherence costs on their own. In ICA, the coherence costs will trade-off with the sparsity prior (Eq 4). Ideally, coherence costs would only prevent duplication of learned dictionary elements, but otherwise let the data shaping of the basis functions through the sparsity prior. In practice, we have shown that coherence control costs can have an effect on all dictionary elements, including those with large pairwise angles. It is not currently clear how these different costs will bias the learned dictionaries.

To investigate how the coherence control costs perform on data in overcomplete ICA, we compare different ICA cost functions and a sparse coding model on the task of recovering a known mixing matrix from k-sparse data with a Laplacian prior. We compare three classes of overcomplete dictionary recovery methods. The first is a sparse coding baseline (Olshausen and Field, 1997), the second are analysis ICA models described in Section 1 which combine the sparse prior from complete ICA and a coherence control cost, and the final is Score Matching (Hyvärinen, 2005), which is a non-maximum-likelihood method that can be used in overcomplete ICA. The data generated for this task comes from a noiseless overcomplete, sparse generative model. Sparse coding, as a dictionary recovery method, is designed to infer generative models. However, the overcomplete ICA models considered here are being fit assuming a sparse analysis model (Ophir et al., 2011; Rubinstein et al., 2013; Chun and Fessler, 2018). Therefore, these ICA models are mismatched to the underlying generative process of the data. Here, we evaluate in what regime overcomplete analysis ICA models

are able to recover a generative dictionary and what impact the coherence control cost has on recovery.

Overcomplete mixing matrices were generated from the Soft Coherence Cost (see Section 3) and used to generate a k-sparse dataset. The dictionary learning methods were then all trained on these datasets. Recovered unmixing matrices were compared to the ground-truth mixing matrix where the error for recovery is 0 for a perfect recovery ($W^T = A$) and 1 for a random recovery (see Section 3.5 for details). For a 32-dimensional data space, we vary the k-sparseness and overcompleteness of the data. For each of these datasets, where the number of dataset samples was 10-times the mixing matrix dimensionality, we fit all models to the data from 10 random initializations, for a range of sparsity weights: λ, if applicable, and then compare the recovery metric across models.

For a 12-sparse, 2-times overcomplete dataset, all methods can recover the mixing matrix well for some value of λ (Fig 4A). The L_2 and Score Matching costs perform slightly worse than the maximum-likelihood inspired ICA methods and sparse coding. All methods have a certain range of λ over which they recover the mixing matrix well and have differences in how they fail, for instance sparse coding has a very quick transition to poor recovery compared to ICA methods whose performance tends to decrease more slowly as λ moves outside of the optimal range.

At fixed k-sparsity ($k = 12$), we vary the overcompleteness and compare recovery costs (Fig 4B). As a function of overcompleteness, Score Matching recovers well in a smaller range of overcompleteness as compared to other ICA methods. Besides the L_2 cost, all other ICA methods have nearly identical recovery. The L_2 cost's performance breaks down at lower overcompleteness. All ICA methods fail to recover the mixing matrix once the overcompleteness becomes too large, while sparse coding continues to succeed in recovering the mixing matrix. Since the number of bases being recovered changes as the overcompleteness changes, it is not meaningful to compare the recovery metric between overcompletenesses, but it meaningful to compare different models at fixed overcompleteness.

At fixed overcompletenesss (OC=2), we vary the k-sparsity and compare recovery costs Fig (4C). Sparse coding performs well at all k-sparsenesses, but the ICA methods perform better with larger k-sparseness. The L_2 cost and Score Matching fails to recover well at a lower k-sparseness than other ICA methods. Since the number of bases being recovered is fixed as a function of the k-sparseness, the recovery metric can be compared across k-sparseness and models.

Fig 4D and E show the methods in a regime ($k = 6$ and 3-times overcomplete, respectively) where ICA methods do not recover the mixing matrix as well as sparse coding. Fig C4 contains the same analysis for the full set of cost functions.

In summary, we show that in general, ICA analysis methods have limited ability to recover generative dictionaries as a function of overcompleteness compared to sparse coding although the methods proposed here extend the range of applicability, which is consistent with Elad et al. (2007). Furthermore, we show that different ICA methods have different regimes of performance with Score Matching and the L_2 cost having the smallest ranges of applicability. Other ICA methods generally have similar performance. Score Matching did not always perform as well as other ICA methods as a function of overcompleteness or k-sparseness, although it is a hyperparameter-free cost (no λ hyperparameter). The more computationally costly sparse coding was able to recover the mixing matrix more

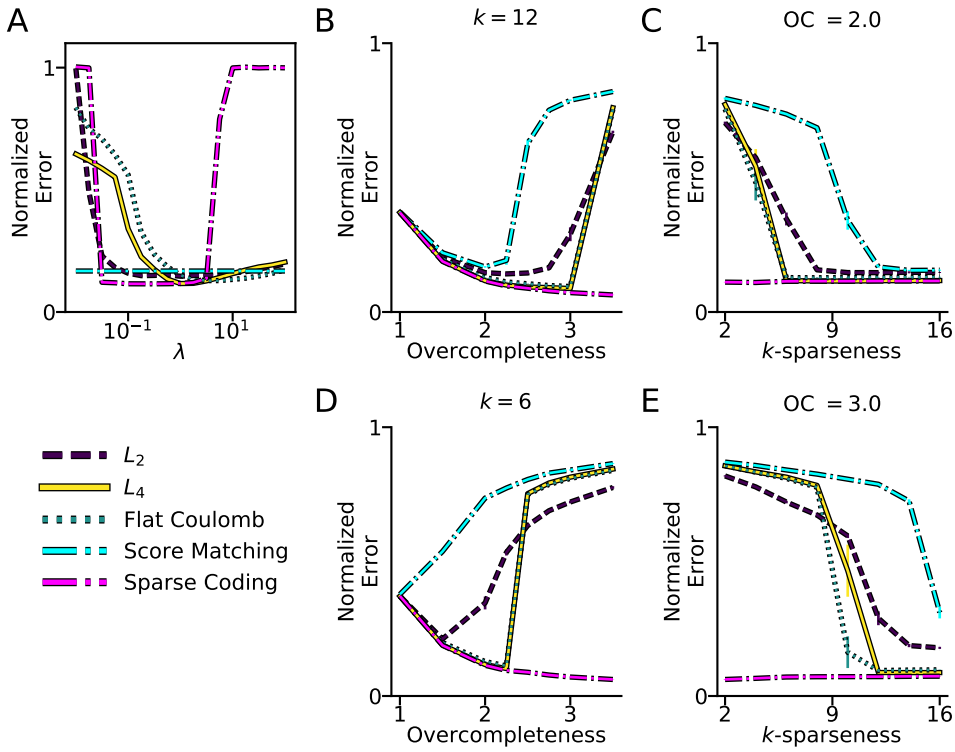

Figure 4: Coherence control costs do not all recover mixing matrices well. All ground truth mixing matrices were generated from the Soft Coherence cost and had a data dimension of 32. Color and line style legend are preserved across panels. **A** The normalized recovery error (see Section 3 for details) for a 2-times overcomplete mixing matrix and $k = 12$ as a function of the sparsity prior weight (λ). Since score matching does not have a λ parameter, it is plotted at a constant. **B** Recovery performance ($\pm$ s.e.m., $n = 10$) at the best value of λ as a function of overcompleteness at $k = 12$. **C** Recovery performance ($\pm$ s.e.m., $n = 10$) at the best value of λ as a function of k-sparseness at 2-times overcompleteness. **D, E** Same plots as **B** and **C** at a point where methods do not perform as well: $k = 6$ and 3-times overcomplete. In **B-E**, the L_4 and Flattened Coulomb lines are largely overlapping.

consistently than ICA models. This suggests that the linear inference in ICA models can only recover dictionaries for moderately overcomplete representations.

2.6. Recovery of the analysis matrix with overcomplete ICA

In analysis dictionary learning (Elad et al., 2007; Rubinstein et al., 2013; Ophir et al., 2011; Chun and Fessler, 2018), the goal is to find a set of dictionary elements such that each datapoint is orthogonal to many (or k) elements of the set, rather than to find a generative model for the data. Since overcomplete ICA models fit more naturally into an analysis framework rather than a synthesis framework, we perform a similar analysis as in Section 2.5, except here, we generate the data from an analysis model using the method described in Rubinstein et al. (2013) (see Section 3 for details).

Across overcompleteness (OC) and k, the methods are generally more similar than in Section 2.5. Note that in this analysis, k is the number of zeros in the projection per data point, which is different than k in the synthesis data, which is the number of elements included per data point. We find three main trends as a function of overcompleteness and k: the L_2 cost tends to perform worse or as well as the other costs, score matching can perform slightly better or slightly worse than other methods, and L_4, Coulomb, and Flattened Coulomb are not well separated. Although this analysis does not distinguish the costs proposed here, it does show that the L_2 cost is suboptimal for both the synthesis and analysis problems.

2.7. Experiments on natural images

When ICA is applied to real data, one typically does not know the exact sparse distribution of the data. For instance, for a natural images dataset, we no longer have a ground truth mixing matrix or known prior, and furthermore, it is not likely that natural image patches come from a simple generative model (Hyvärinen and Köster, 2007; Lücke et al., 2009). However, the effects of coherence control on the distribution of dictionary elements learned can be evaluated. Specifically, we can look at the coherence of learned dictionaries and whether different methods prevent duplicate features from being learned.

We train 2-times overcomplete ICA models on 8-by-8 whitened image patches from the Van Hateren database (van Hateren and van der Schaaf, 1998) at a fixed value of sparsity across costs found by binary search on λ. The score matching cost has no λ parameter to trade off sparsity versus coherence although it finds solutions of similar sparsity to the value chosen for the other costs. It is known that for natural images data sets, bases learned with ICA can be well-fit by Gabor filters (Bell and Sejnowski, 1997). Hence, we evaluate the distribution of the learned basis by inspecting the parameters obtained from fitting the bases to Gabor filters (see Section 3.6 for details).

The distributions of angles from the trained ICA models are in line with the theoretical results from Section 2.3. The L_2 cost has more pairwise angles close to zero compared to the other costs with the L_4 having the smallest coherence (Fig 6A). Although the probabilities for the L_2 cost in Fig 6A trend to 10^{-5} for small angles, there is a peak at zero which is closer to 10^{-3}. For the natural images analysis, each learned dictionary has $2 \times 8 \times 8 = 128$ elements which corresponds to 8128 pairwise angles. This means that for the L_2 cost there are approximately 8 pairs of elements that are nearly identical on average which means

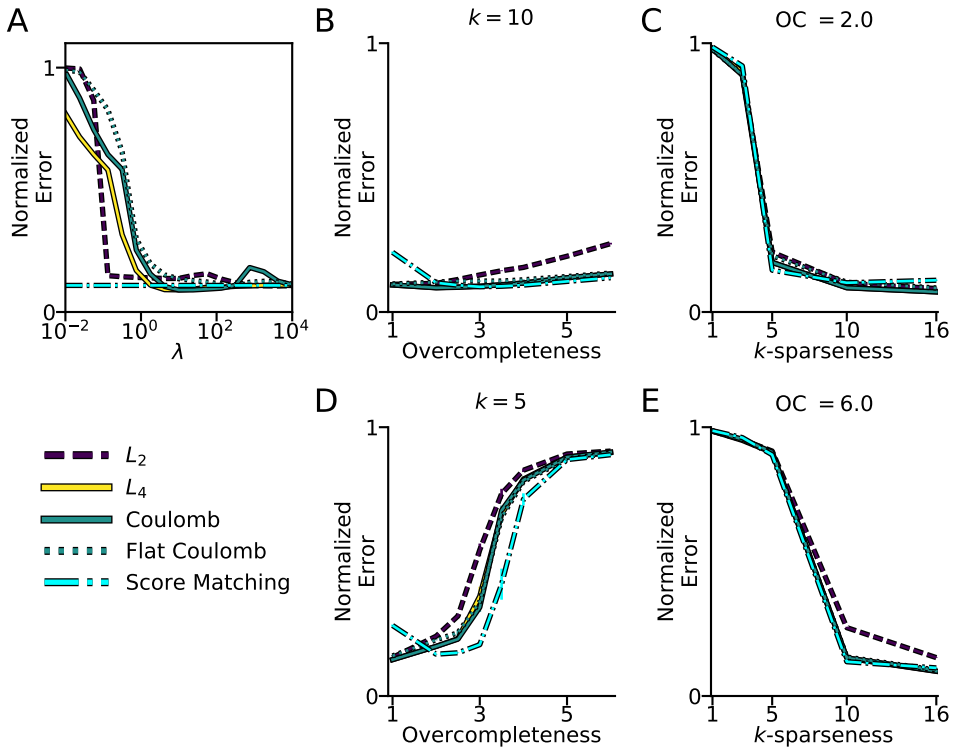

Figure 5: Coherence control costs do not all recover analysis matrices well. All ground truth analysis matrices were generated from the Soft Coherence cost and had a data dimension of 32. Color and line style legend are preserved across panels. **A** The normalized recovery error (see Section 3 for details) for a 2-times overcomplete analysis matrix and $k = 10$ as a function of the sparsity prior weight (λ). Since score matching does not have a λ parameter, it is plotted at a constant. **B** Recovery performance ($\pm$ s.e.m., $n = 10$) at the best value of λ as a function of overcompleteness at $k = 10$. **C** Recovery performance ($\pm$ s.e.m., $n = 10$) at the best value of λ as a function of k-sparseness at 2-times overcompleteness. **D, E** Same plots as **B** and **C** at a point where methods do not perform as well: $k = 5$ and 6-times overcomplete. In **B-E**, the L_4, Coulomb, and Flattened Coulomb lines are largely overlapping.

that about 5% of the elements are redundant. Similarly, as shown in Fig 6B, the Random Prior and Coulomb costs have lower coherence when the second order terms are removed and behave more similarly to the L_4 cost. These distributions also show that ICA models with the L_2 cost tend to learn duplicate bases from natural images.

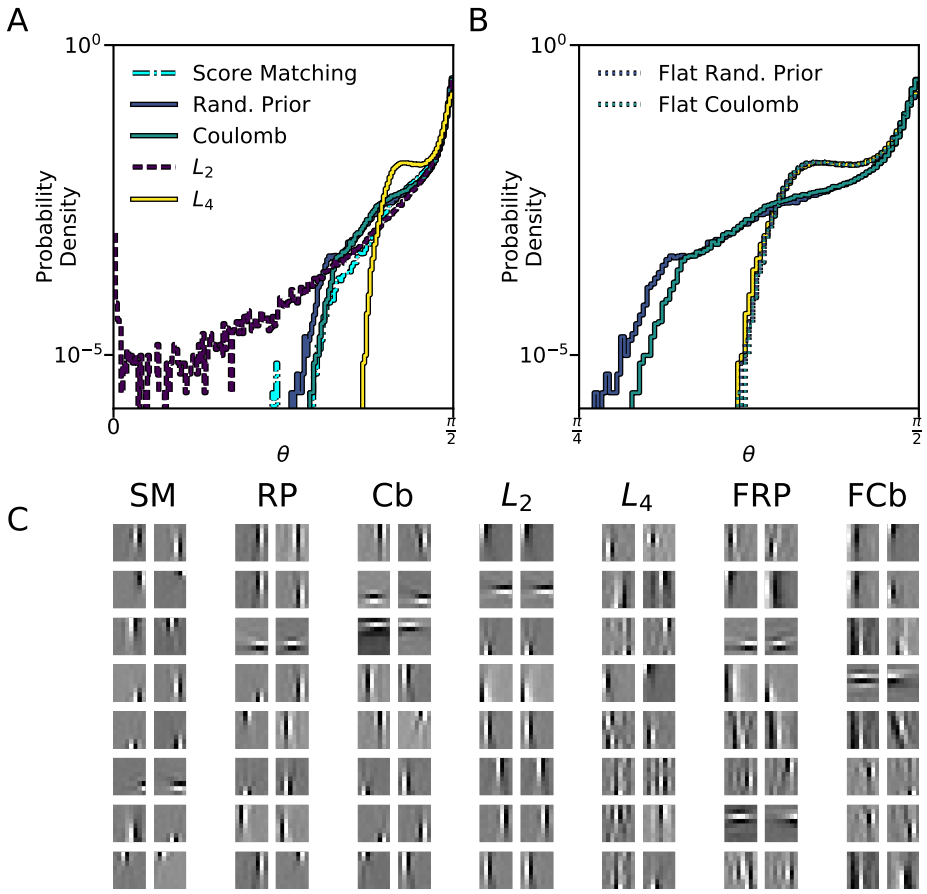

Figure 6: The coherence of an overcomplete dictionary learned from natural images depends on the coherence control cost. Results from fitting a 2-times overcomplete model on 8-by-8 natural image patches. **A, B** Pairwise angle distributions (log scale) across costs for the learned dictionaries for a fixed value of sparsity across costs. **B** Comparison between the Random Prior and Coulomb costs and their flattened versions. The L_4 distribution is also shown for comparison. Note that the horizontal axis covers 45 to 90 degrees and that the Flattened Random Prior and Flattened Coulomb lines largely overlap the L_4 line. **C** For each cost from **A** and **B**, the 8 pairs of bases with smallest pairwise angle are shown. Since the overall sign of a basis element is arbitrary, the bases have been inverted to have positive inner product, if needed, for visualization.

For the range of sparsities which were considered, the visual appearance of the individual bases is similar to results from previous ICA work and similar across costs (L_4 bases are shown in Fig 7A). The tiling properties of the learned dictionaries can also be visualized directly. The coordinates of the center of the fit Gabor filter, rotations, and scales tile the space for the L_2, L_4, and Flattened Coulomb costs (Fig 7B). The dimensions and rotation of the rectangle represent the envelope widths and planar rotation angle respectively. This is similarly true for the planar rotation angle against the oscillation wavelength of the Gabor (Fig 7C) and the envelope widths and wavelengths (Fig 7D). Although these distributions look qualitatively similar, the underlying dictionaries can have very different coherence.

These results demonstrate that the L_2 cost learns undesirable, high-coherence overcomplete dictionaries on real data. Visually inspecting the bases or even their tiling properties may not reveal the redundant set of basis functions. To reveal this type of redundancy one has to measure the coherence or the distribution of pairwise angles of a dictionary directly.

3. Methods

In this section we summarize previously proposed coherence control methods, our model implementations, and datasets used.

3.1. Previously proposed coherence control methods

3.1.1. Reconstruction cost and the L_2 cost

Le et al. (2011) propose adding a reconstruction cost to the ICA prior (RICA) as a form of coherence control, which they show is equivalent to a cost on the L_2 norm of the difference between the Gram matrix of the filters and an identity matrix for whitened data

$$\begin{aligned} C_{\text{RICA}} &= \frac{1}{N}\sum_{ij}(X_j^{(i)} - \sum_{kl} W_{kj}W_{kl}X_l^{(i)})^2 \\ \propto C_{L_2} &= \frac{1}{2}\sum_{ij}(\delta_{ij} - \sum_k W_{ik}W_{jk})^2 = \frac{1}{2}\sum_{ij}(\delta_{ij} - \cos\theta_{ij})^2, \end{aligned} \tag{14}$$

where W_{ij} is the component of the ith source for the jth mixture , $X_j^{(i)}$ is the jth element of the ith sample, θ_{ij} is the angle between pairs of basis, and δ_{ij} is the Kronecker delta.

The L_2 cost has also been proposed as a form of coherence control (Ramırez et al., 2009; Sigg et al., 2012; Bao et al., 2014; Chun and Fessler, 2018). Equiangular tight-frames (ETFs) are frames (overcomplete dictionaries) which have minimum coherence. The fact that an ETF has minimum coherence is used to motivate the L_2 cost as a form of coherence control. A matrix $W \in \mathbb{R}^{L\times D}$ is an ETF if

$$|\sum_k W_{ik}\cdot W_{jk}| = \cos\alpha,\ \forall i \neq j \tag{15}$$

for some angle, α, and

$$\sum_k W_{ki}W_{kj} = \frac{L}{D}\delta_{ij}. \tag{16}$$

The L_2 cost will encourage Eq 16 to be satisfied, but does not encourage Eq 15 to be satisfied as we show in Theorem 1.

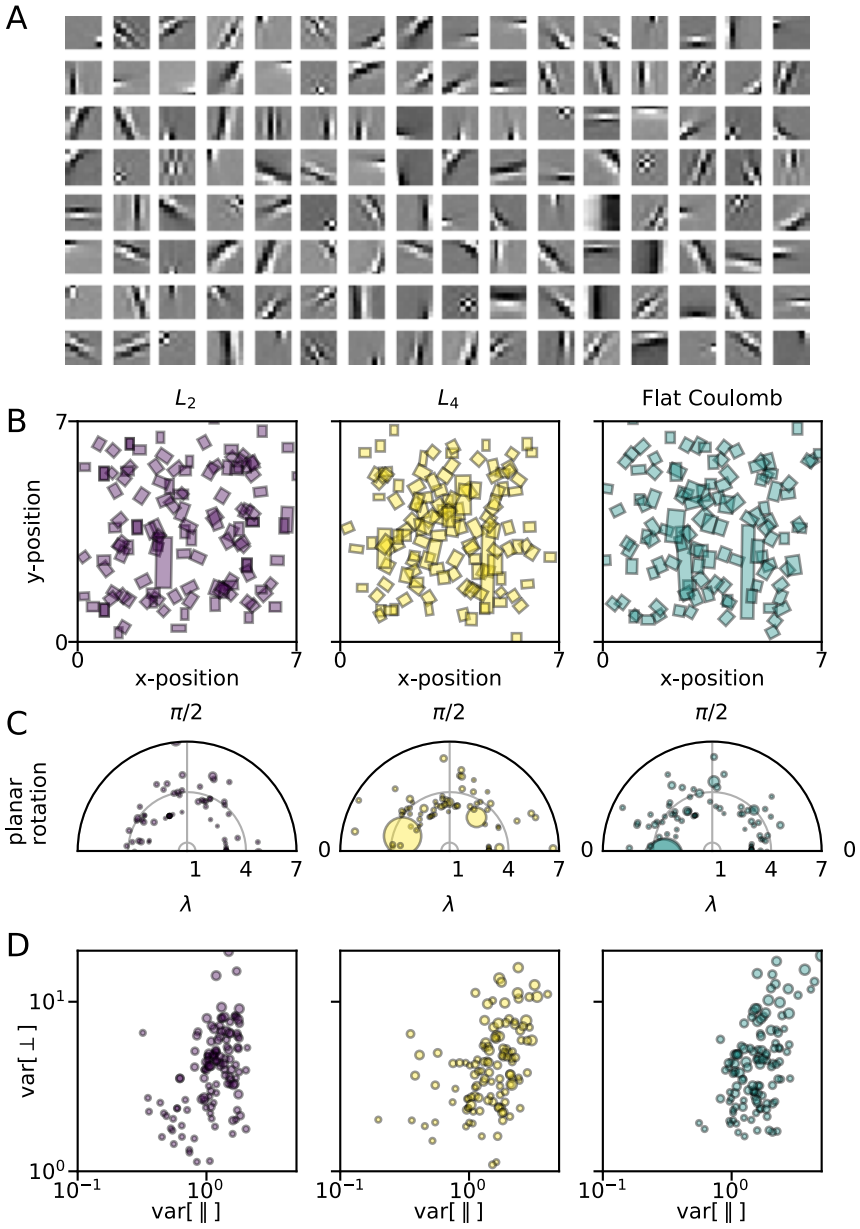

Figure 7: All coherence costs learn a dictionary that approximately tiles the space of Gabor Filters. **A** Dictionary learned using the L_4 cost on 8-by-8 natural image patches. **B** Distributions of locations, envelope scales, and rotations. Rectangle position: center of Gabor fit in pixel coordinates, rectangle rotation: planar-rotation of the Gabors, rectangle shape: envelope width parallel and perpendicular to the oscillation axis. **C** Distributions of rotations, log-wavelengths (λ), and envelope widths. Polar plots of planar-rotation angle and log-spatial wavelength of the Gabors. Marker size scales with geometric mean of envelope widths. **D** Distributions of envelope scales and log-wavelengths. Log-scale plot of envelope widths-squared parallel and perpendicular to the oscillation axis of the Gabors. Circle size scales with log-wavelength.

3.1.2. Quasi-orthogonality constraint

Hyvärinen et al. (1999) suggest a quasi-orthogonality update which approximates a symmetric Gram-Schmidt orthogonalization scheme for an overcomplete basis, W, which is formulated as

$$W \leftarrow \frac{3}{2}W - \frac{1}{2}WW^TW. \tag{17}$$

3.1.3. Random prior cost

A prior on the distribution of pairwise angles was proposed to encourage low coherence (Hyvärinen and Inki, 2002). The prior is the distribution of pairwise angles for two vectors drawn from a uniform distribution on the n-sphere[1]

$$C_{\text{Random prior}} = -\sum_{i \neq j} \log P(\cos\theta_{ij}) \propto -\sum_{i \neq j} \log(1 - \cos^2\theta_{ij}). \tag{18}$$

3.1.4. Score Matching

Score matching is a training objective function for non-normalized statistical models of continuous variables(Hyvärinen, 2005). It has been used to learn overcomplete ICA models. The score function is derivative of the log-likelihood of the model or data distribution with respect to the data

$$\psi(X;\Theta) = \nabla_X \log p(X;\Theta) \tag{19}$$

The score matching objective is the mean-squared error between the model score, $\psi(X;\Theta)$, and data score, $\psi_{\mathcal{D}}(X;\Theta)$ averaged over the data, $\mathcal{D}$

$$J(\Theta) = \frac{1}{2}\int_X p_{\mathcal{D}}(X)||\psi(X;\Theta) - \psi_{\mathcal{D}}(X;\Theta)||^2. \tag{20}$$

3.2. Coherence-based costs

The coherence of a dictionary is defined as the maximum absolute value of the off-diagonal elements of the Gram matrix (Davenport et al., 2011) as in Eq 6. Using the coherence as a cost is equivalent to using the L_∞ norm version of the L_p cost function. We find that this cost is difficult to numerically optimize with both second order methods and gradient descent since the derivative through the max operation will only act on one pair of bases at each optimization step, although it should find solution with local minima of coherence. An easier to optimize, but heuristic, version of this cost is the sum over all off-diagonal elements whose squares are larger than the mean squared value

$$C_{\text{Soft Coherence}} = \sum_{i \neq j \text{ s.t. } \cos^2\theta_{ij} > \cos\hat{\theta}^2} |\cos\theta_{ij}|, \text{ with } \cos\hat{\theta}^2 = \underset{i \neq j}{\text{mean}}(\cos^2\theta_{ij}). \tag{21}$$

We find that this cost does not optimize well for coherence control in ICA when fit with data, but it can be used to create low-coherence mixing matrices for generating data with known structure in Section 2.5.

1. For both the Random Prior and the Coulomb cost, we regularize the costs and their derivatives near $|\cos\theta| = 1$ by adding a small positive constant in the objective: $1 - \cos^2\theta_{ij} \rightarrow 1 + |\epsilon| - \cos^2\theta_{ij}$.

3.3. Model implementation

All models were implemented in Theano (Theano Development Team, 2016). ICA models were trained using the L-BFGS-B (Byrd et al., 1995) implementation in SciPy (Jones et al., 2001–2017). FISTA (Beck and Teboulle, 2009) was used for MAP inference in the sparse coding model and the weights were learned using L-BFGS-B. All weights were training with the norm-ball projection (Le et al., 2011) to keep the bases normalized. A repository with code to reproduce the results is available[2]. For ICA models with coherence costs, the coherence control cost with no sparsity penalty ($\lambda = 0$) was used as the objective for Figs 2 and 3.

3.4. Datasets

For all datasets and models, the number of samples in a dataset was equal to 10 times the number of model parameters, that is, $10 \times n_{\text{sources}} \times n_{\text{mixtures}}$. Datasets were mean-centered and whitened.

3.4.1. k-sparse and analysis datasets

For both datasets, dictionaries were generated by minimizing the Soft Coherence cost. The synthesis data was generated by keeping k random elements per data sample from draws of a diagonal multivariate Laplacian distribution, zeroing out the rest, and combining them with the mixing matrix, $X^{(i)}_j = \sum_{j=1}^{L} A_{jk} S^{(i)}_k$, with k-sparse S.

For the analysis dataset, we use the method proposed by Rubinstein et al. (2013). $X^{(i)}$ is initialized to i.i.d. Gaussian samples. Then, a subset of k elements, $W^{A_{\text{sub}}}$, of the analysis matrix, W^A, are chosen per data sample and are used to form a projection matrix which removes the subspace spanned by $W^{A_{\text{sub}}}$ from $X^{(i)}$, $X^{(i)}_j = \sum_{k=1}^{L} (I - \Omega^T \Omega)^{(i)}_{jk} X^{(i)}_k$, where Ω is a basis for the subspace spanned by $W^{A_{\text{sub}}}$. This ensures that at least k elements of $W^A X^{(i)}$ will be zero.

3.4.2. Natural images dataset

Images were taken from the Van Hateren database (van Hateren and van der Schaaf, 1998). We selected images where there was no evident motion blur and minimal saturated pixels. 8-by-8 patches were taken from these images and whitened using PCA.

3.5. Dictionary recovery error

If the mixing matrix A is recovered perfectly, W^T will be a permutation of A. To estimate the closeness to a permutation matrix, the matrix $P_{ij} = |A^T_i \cdot W_j|$ is created. For the analysis dictionaries W^A, $P_{ij} = |W^A_i \cdot W_j|$. The Hungarian method (Kuhn, 1955) is used to find the best assignment between A^T_i (or W^A_i) and W_j. Given this best assignment, the median angle between the elements is returned.

This error is normalized by calculating the same quantity for matrices, W^*, which were recovered from mixing matrices A^*, which were from the same distribution as A but

2. `https://github.com/JesseLivezey/oc_ica`

with different random initializations. After this normalization, perfect recovery gives a normalized error of 0 and a random recovery gives a normalized error of 1.

3.6. Fitting Gabor parameters

We fit the Gabor parameters (Ringach, 2002) to the learned bases using an iterative grid-search and optimization scheme which gave the best results on generated filters. The learned parameters were the center vector: $\{\mu_x, \mu_y\}$, planar-rotation angle: θ, phase: ϕ, oscillation wave-vector k, and envelope variances parallel and perpendicular to the oscillations: $\sigma_{\hat{x}}^2$ and $\sigma_{\hat{y}}^2$ respectively. Because they are constrained to be positive, the log of the parameters: $\sigma_{\hat{x}}^2$ and $\sigma_{\hat{y}}^2$ are optimized. To keep the wavelength of the Gabor larger than 2 pixels, instead of optimizing k directly we optimize ρ with $k = \frac{2\pi}{2\sqrt{2}+\exp(\rho)}$. Shorter wavelengths would be aliased by the pixel sampling.

$$\begin{aligned}
\hat{x} &= \cos(\theta)x + \sin(\theta)y \\
\hat{y} &= -\sin(\theta)x + \cos(\theta)y \\
\hat{\mu}_x &= \cos(\theta)\mu_x + \sin(\theta)\mu_y \\
\hat{\mu}_y &= -\sin(\theta)\mu_x + \cos(\theta)\mu_y \\
\text{Gabor}(x, y; \mu_x, \mu_y, \theta, \sigma_{\hat{x}}, k, \sigma_{\hat{y}}, \phi) &= \exp\left(-\frac{(\hat{x}-\hat{\mu}_x)^2}{2\sigma_{\hat{x}}^2} - \frac{(\hat{y}-\hat{\mu}_y)^2}{2\sigma_{\hat{y}}^2}\right)\sin(k\hat{x}+\phi)
\end{aligned} \tag{22}$$

A global, gradient-based optimization leads to many local minia where one lobe of the Gabor would be well fit, but the other would not. We found that a combination of an iterative approach with gradient-based optimization of subsets of the parameters worked well. The procedure for finding the best Gabor kernel parameters was to save the parameter set with best mean-squared error after the following iterations

1. for different initial envelope widths, fit the center location for the envelope to the blurred, absolute value of the basis element,
2. for different initial planar rotations and frequencies, numerically optimize the planar rotation, phase, and frequency of the Gabor
3. for the best fit from above, re-optimize the centers, widths, and phases,
4. re-optimize all parameters from best previous fit.

A repository with code to fit the Gabor kernels is posted online [3].

4. Discussion

Learning overcomplete sparse representations of data is often an extremely informative first stage in analyzing multivariate data. In the field of neuroscience, overcomplete dictionary learning serves as a theoretical model of how the brain analyzes sensory inputs (Olshausen

3. `https://github.com/JesseLivezey/gabor_fit`

and Field, 1996; Klein et al., 2003; Smith and Lewicki, 2006; Rehn and Sommer, 2007; Zylberberg et al., 2011; Carlson et al., 2012), motivating the study of methods suitable in the overcomplete regime. For all of these purposes, the heavy computational cost of the nonlinear inference step involved in common sparse coding approaches can be an obstacle for large datasets. For learning complete sparse representations, ICA with just a linear inference mechanism is a viable alternative with drastically reduced computational demand. Here, we investigated the limitations of linear inference in overcomplete dictionary learning.

Any multidimensional method for extracting signal components needs a form of coherence control to prevent components from becoming co-aligned and therefore redundant. We first compared different coherence costs' ability to prevent the learning of coherent dictionary elements in the overcomplete case. We show theoretically and by simulation, that the L_2 cost, which successfully achieves minimum coherence (orthogonality) in the complete case, exhibits pathological global minima with maximum coherence in the overcomplete case. Encouraging diverse, incoherent, or orthogonal solutions has been proposed as a desirable additional cost function for dictionary learning and deep network applications (Le et al., 2011; Sigg et al., 2012; Ramırez et al., 2009; Bao et al., 2014; Brock et al., 2017; Chun and Fessler, 2018; Bansal et al., 2018), typically by applying a power/norm to the difference between an identity matrix and the gram matrix. However, the impact of previously proposed costs (L_1 and L_2) on coherence has not been directly explored, and we have shown that they do not encourage lower coherence.

We propose novel cost functions which do not suffer from pathological minima in the overcomplete case. Specifically, we propose the L_4 cost and the flattened versions of the Coulomb and Random Prior costs, and show that they yield dictionaries with lower coherence than the cost functions that have been proposed earlier. At the same time, these new cost functions have smaller effects on incoherent basis pairs, thus leading to dictionaries that reflect the structure of the data rather than effects from the coherence term.

We show that the methods of coherence control proposed here can extent the regime of overcompleteness and sparseness, in which ICA methods can successfully learn recover synthetic dictionaries with linear inference. However, this expansion of the regime of applicability is still limited. Even the improved methods begin to fail when overcompleteness grows beyond two-fold (for 32 dimensional data) or if the data is k-sparse with small k. The problem to deal with extremely k-sparse data is counterintuitive at first, because nonlinear inference methods usually do better as k is decreased because the combinatorial search for the best sparse support in the inference becomes easier (Davenport et al., 2011). However, linear inference in ICA models cannot recover extremely sparse sources unlike sparse coding models, which do not fail in the small-k limit.

Based on the observations in Section 2.4 that the leading terms in the Taylor expansion will have the largest influence on the distribution of angles, we proposed a modification to the Coulomb and Random Prior costs to make them more similar to the L_4 costs which resulted in lower coherence solutions. We do not have a strong way of distinguishing the L_4 cost and the flattened Coulomb and Random Prior costs. However, it may be possible that the higher order terms become important in specific situations we have not considered here and may distinguish these costs.

In this work, coherence costs were investigated as augmented costs for overcomplete ICA. Although, the proposed methods only modestly extended the ability of linear inference to

perform overcomplete model recovery, the proposed coherence control costs lead to learned dictionaries with significantly higher coherence. In the case of dictionaries learned on natural image patches, we show that the L_4 cost prevents the model from learning duplicated bases, unlike the L_2 cost. For other application where low coherence might be desirable (e.g., sparse coding, deep learning, the proposed cost functions could provide improved coherence control. We note that variations of the ICA sparsity prior and mismatch with data sparsity structure have not been systematically explored here and are another potential topic of further investigation. All told, our study explores the power and limitations of linear inference for overcomplete dictionary learning.

Acknowledgments

We thank Yubei Chen, Alexander Anderson, and Kristofer Bouchard for helpful discussions. We also thank the editor and reviewers for their helpful comments which improved the quality of the manuscript. We thank the reviewer for pointing out a simplified proof of Theorem 1. JAL was supported by the Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231. JAL and AFB were supported by the Applied Mathematics Program within the Office of Science Advanced Scientific Computing Research of the U.S. Department of Energy under contract No. DE-AC02-05CH11231. FTS was supported by the National Science Foundation grants IIS1718991, IIS1516527, INTEL, and the Kavli Foundation. We acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

Appendix A. Minima analysis for the L_2 and L_4 costs for a 2-dimensional space.

We can write a 2 dimensional, 2 times overcomplete dictionary as

$$W = \begin{pmatrix} 1 & 0 \\ \cos\theta_1 & \sin\theta_1 \\ \cos\theta_2 & \sin\theta_2 \\ \cos(\theta_2+\theta_3) & \sin(\theta_2+\theta_3) \end{pmatrix} \tag{23}$$

If we restrict the dictionary to be composed of two orthogonal pairs, without loss of generality, we can write W as

$$W|_{\theta_1=\theta_3=\frac{\pi}{2}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \cos\theta_2 & \sin\theta_2 \\ -\sin\theta_2 & \cos\theta_2 \end{pmatrix} \tag{24}$$

For the general dictionary For the L_p cost for even p is

$$
\begin{aligned}
C_{L_p} =& \frac{1}{p}\sum_{ij}(\delta_{ij} - \sum_k W_{ik}W_{jk})^p \\
=& \frac{1}{p}\sum_{ij}\left(\begin{pmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix} - \begin{pmatrix}1 & \cos\theta_1 & \cdots \\ \cos\theta_1 & 1 & \\ \cos\theta_2 & \cos\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2 & \\ \cos(\theta_2+\theta_3) & \cos\theta_1\cos(\theta_2+\theta_3) + \sin\theta_1\sin(\theta_2+\theta_3) & \end{pmatrix}\right. \\
& \left.\begin{matrix} \cos\theta_2 & \cos(\theta_2+\theta_3) \\ \cos\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2 & \cos\theta_1\cos(\theta_2+\theta_3) + \sin\theta_1\sin(\theta_2+\theta_3) \\ 1 & \cos\theta_2\cos(\theta_2+\theta_3) + \sin\theta_2\sin(\theta_2+\theta_3) \\ \cos\theta_2\cos(\theta_2+\theta_3) + \sin\theta_2\sin(\theta_2+\theta_3) & 1 \end{matrix}\right)^p \\
& \cdots \\
=& \frac{1}{p}\left(2\cos^p\theta_1 + 2\cos^p\theta_2 + 2\cos^p(\theta_2+\theta_3)\right. \\
& + 2(\cos\theta_1\cos\theta_2 + \sin\theta_1\sin\theta_2)^p + 2(\cos\theta_1\cos(\theta_2+\theta_3) + \sin\theta_1\sin(\theta_2+\theta_3))^p \\
& \left. +2(\cos\theta_2\cos(\theta_2+\theta_3) + \sin\theta_2\sin(\theta_2+\theta_3))^p\right)
\end{aligned}
\tag{25}
$$

This can be simplified for the case of 2 orthogonal bases (Eq 24)

$$
\begin{aligned}
C_{L_p}|_{\theta_1=\theta_3=\frac{\pi}{2}} &= \frac{1}{p}\left(2\cos^p\theta_2 + 2\sin^p\theta_2 + 2\sin^p\theta_2 + 2\cos^p\theta_2 + 2(-\cos\theta_2\sin\theta_2 + \sin\theta_2\cos\theta_2)^p\right) \\
&= \frac{1}{p}\left(4\sin^p\theta_2 + 4\cos^p\theta_2\right).
\end{aligned}
\tag{26}
$$

For $p = 2$ and $p = 4$ this is

$$
\begin{aligned}
C_{L_2}|_{\theta_1=\theta_3=\frac{\pi}{2}} &= \frac{1}{2}(4\cos^2\theta_2 + 4\sin^2\theta_2) = 2 \\
C_{L_4}|_{\theta_1=\theta_3=\frac{\pi}{2}} &= \cos^4\theta_2 + \sin^4\theta_2.
\end{aligned}
\tag{27}
$$

Here we tabulate the full Hessian matrices, eigenvalues, and eigenvectors for the analysis in Sections 2.1 and 2.2.

A.1. L_2 cost

$$
\begin{aligned}
C_{L_2}(\theta_1,\theta_2,\theta_3)|_{\theta_1,\theta_3=\frac{\pi}{2}} &= 2 \\
\frac{\partial C_{L_2}(\theta_1,\theta_2,\theta_3)}{\partial\vec{\theta}}|_{\theta_1,\theta_3=\frac{\pi}{2}} &= \begin{pmatrix}0 & 0 & 0\end{pmatrix} \\
H(C_{L_2})|_{\theta_1,\theta_3=\frac{\pi}{2}} &= \begin{pmatrix}2 & 0 & 2\cos 2\theta_2 \\ 0 & 0 & 0 \\ 2\cos 2\theta_2 & 0 & 2\end{pmatrix} \\
\text{EVal.}(H_{L_2})|_{\theta_1,\theta_3=\frac{\pi}{2}} &= \begin{pmatrix}0 \\ 4\sin^2\theta_2 \\ 4\cos^2\theta_2\end{pmatrix} \\
\text{EVec.}(H_{L_2})|_{\theta_1,\theta_3=\frac{\pi}{2}} &= \begin{pmatrix}0\\1\\0\end{pmatrix}, \begin{pmatrix}-1\\0\\1\end{pmatrix}, \begin{pmatrix}1\\0\\1\end{pmatrix}
\end{aligned}
\tag{28}
$$

A.2. The second eigenvalue of the L_2 cost

For the second eigenvalue and eigenvector, $4\sin^2\theta_2$ and (-1, 0, 1), centered at $\theta_1 = \theta_3 = \pi/2$, $\theta_2 = 0$, the cost along the direction of the eigenvector is

$$
\begin{aligned}
C(\text{EVal}_2,\Delta\theta) = \frac{1}{2}\big(&2\cos^2(\pi/2-\Delta\theta) + 2\cos^2 0 + 2\cos^2(0+\pi/2+\Delta\theta) \\
&+ 2(\cos(\pi/2-\Delta\theta)\cos 0 + \sin(\pi/2-\Delta\theta)\sin 0)^2 \\
&+ 2(\cos(\pi/2-\Delta\theta)\cos(0+\pi/2+\Delta\theta) \\
&+ \sin(\pi/2-\Delta\theta)\sin(0+\pi/2+\Delta\theta))^2 \\
&+2(\cos 0\cos(0+\pi/2+\Delta\theta) + \sin 0\sin(0+\pi/2+\Delta\theta))^2\big) \\
= 1 &+ 4\sin^2\Delta\theta + \sin^4\Delta\theta + \cos^4\Delta\theta - 2\sin^2\Delta\theta\cos^2\Delta\theta
\end{aligned}
\tag{29}
$$

which Taylor-expanded around $\Delta\theta = 0$ gives

$$
\begin{aligned}
C(\text{EVal}_2,\Delta\theta) &= 1 + 4(\Delta\theta^2 + \frac{\Delta\theta^4}{3}) + (\Delta\theta^4) + (1 - 2\Delta\theta^2 + \frac{5\Delta\theta}{3}) \\
&\quad - 2(\Delta\theta^2 - \frac{4\Delta\theta}{3}) + O(\Delta\theta^5) \\
&= 2 + \frac{4\Delta\theta^4}{3} + O(\Delta\theta^5).
\end{aligned}
\tag{30}
$$

This shows that although the second (and third) derivative is zero, the fourth derivative is positive, meaning this point is a minimum.

A.3. L_4 cost

$$
\begin{aligned}
C_{L_4}(\theta_1,\theta_2,\theta_3)|_{\theta_1,\theta_3=\frac{\pi}{2}} &= \cos^4\theta_2 + \sin^4\theta_2 \\
\frac{\partial C_{L_4}(\theta_1,\theta_2,\theta_3)}{\partial\vec{\theta}}|_{\theta_1,\theta_3=\frac{\pi}{2}} &= \frac{1}{2}\begin{pmatrix}\sin 4\theta_2 & -2\sin 4\theta_2 & -\sin 4\theta_2\end{pmatrix} \\
H(C_{L_4})|_{\theta_1,\theta_3=\frac{\pi}{2}} &= \begin{pmatrix} -2\cos 4\theta_2 & 2\cos 4\theta_2 & \cos 2\theta_2 + \cos 4\theta_2 \\ 2\cos 4\theta_2 & -4\cos 4\theta_2 & -2\cos 4\theta_2 \\ \cos 2\theta_2 + \cos 4\theta_2 & -2\cos 4\theta_2 & -2\cos 4\theta_2 \end{pmatrix} \\
\text{EVal.}(H_{L_4})|_{\theta_1,\theta_3=\frac{\pi}{2}} &= \begin{pmatrix} \cos 2\theta_2 - \cos 4\theta_2 \\ -\frac{1}{2}\cos 2\theta_2 - \frac{7}{2}\cos 4\theta_2 - \dots \\ \dots \frac{1}{2\sqrt{2}}\sqrt{34 - 2\cos 2\theta_2 + \cos 4\theta_2 - 2\cos 6\theta_2 + 33\cos 8\theta_2} \\ -\frac{1}{2}\cos 2\theta_2 - \frac{7}{2}\cos 4\theta_2 + \dots \\ \dots \frac{1}{2\sqrt{2}}\sqrt{34 - 2\cos 2\theta_2 + \cos 4\theta_2 - 2\cos 6\theta_2 + 33\cos 8\theta_2} \end{pmatrix} \\
\text{EVec.}(H_{L_4})|_{\theta_1,\theta_3=\frac{\pi}{2}} &= \begin{pmatrix}1\\0\\1\end{pmatrix}, \\
&\quad \begin{pmatrix} -1 \\ (\frac{\sqrt{2}}{8}\sqrt{\begin{matrix}2\cos 2\theta_2 + \cos 4\theta_2 - \dots \\ \dots 2\cos 6\theta_2 + 33\cos 8\theta_2 + 34\end{matrix}} - \dots \\ \dots - 2\cos 2\theta_2)\sec 4\theta_2 + \frac{1}{4} \\ 1 \end{pmatrix}, \\
&\quad \begin{pmatrix} -1 \\ \frac{1}{4} - (2\cos\frac{1}{4}\theta_2 + \dots \\ \dots \frac{\sqrt{2}}{8}\sqrt{\begin{matrix}-2\cos 2\theta_2 + \cos 4\theta_2 - 2\cos 6\theta_2 + \dots \\ \dots 33\cos 8\theta_2 + 34\end{matrix}})\sec 4\theta_2 \\ 1 \end{pmatrix}
\end{aligned}
\tag{31}
$$

Appendix B. Proofs of Theorems 1 and 2

B.1. L_2 cost minima and equiangular tight-frames: proof of Theorem 1

We can prove Theorem 1 in two ways. The first way is by showing that a dictionary of concatenated identity matrices is at the global minimum of the C_{L_2} cost and is a short proof. The second is longer, but makes an explicit connection to the C_{L_2} cost optimizing one but not both of the conditions of an equiangular tight-frame (Eqs 15 and 16).

B.2. Proof 1

Here we prove Theorem 1 by showing that a dictionary of stacked identity matrices is a global minimum of the C_{L_2} cost.

Proof [Proof 1 of Theorem 1] For an overcomplete dictionary $W \in \mathbb{R}^{L\times D}$, the expression for the minimum possible coherence is known (Strohmer and Heath Jr, 2003)

$$\text{Coherence}_{\min}(W) = \sqrt{\frac{L-D}{D(L-1)}}. \tag{32}$$

The C_{L_2} cost is the sum of the off-diagonal elements of the squared Gram matrix. The minimum possible value for each element is given by the square of Eq 32 and so the minimum of the C_{L_2} cost will be achieved if all (L^2-L) off-diagonal elements are equal to the square of the value in Eq 32

$$C_{L_2,\ \min} = (L^2-L)\sqrt{\frac{L-D}{D(L-1)}}^2 = \frac{L(L-D)}{D}. \tag{33}$$

For $W = W_0$ constructed as an identity matrix concatenated L/D times, the sum of the off-diagonal elements of the squared Gram matrix is

$$C_{L_2}(W_0) = (\frac{L}{D}^2 - \frac{L}{D})D = \frac{L(L-D)}{D} \tag{34}$$

which is the minimum value. However, W_0 has coherence = 1 since the stacked identity matrices contain identical elements. ■

B.3. Proof 2

Here we prove Theorem 1 in two steps: first we can show the equivalence, up to an additive constant, of minimizing the L_2 cost and minimizing the L_2 norm of the error of Eq 16. Then we show that the pathological solution (Section 2.1) is at the global minimum of this cost.

Proof [Proof 2 of Theorem 1] For a normalized ($\sum_k W_{ik}^2 = 1,\ \forall\ i$) matrix, W

$$\begin{aligned}
C_{L_2} &= \sum_{ij}(\sum_k W_{ik}W_{jk} - \delta_{ij})^2 \\
&= \sum_{ij}(\sum_k W_{ik}W_{jk} - \delta_{ij})(\sum_l W_{il}W_{jl} - \delta_{ij}) \\
&= \sum_{ijkl} W_{ik}W_{jk}W_{il}W_{jl} - 2\sum_{ijk} W_{ik}W_{jk}\delta_{ij} + \sum_{ij}\delta_{ij}^2 \\
&= \sum_{ijkl} W_{ik}W_{jk}W_{il}W_{jl} - 2\sum_{ik} W_{ik}^2 + \text{const.}(L) \\
&= \sum_{ijkl} W_{ik}W_{jk}W_{il}W_{jl} + \text{const.}(L)
\end{aligned} \tag{35}$$

$$
\begin{aligned}
C_{\text{Eq 16}} &= \sum_{kl}(\sum_{i} W_{ik}W_{il} - \frac{L}{D}\delta_{kl})^2 \\
&= \sum_{kl}(\sum_{i} W_{ik}W_{il} - \frac{L}{D}\delta_{kl})(\sum_{j} W_{jk}W_{jl} - \frac{L}{D}\delta_{kl}) \\
&= \sum_{ijkl} W_{ik}W_{il}W_{jk}W_{jl} - 2\sum_{ikl}\frac{L}{D}W_{ik}W_{il}\delta_{kl} + \sum_{kl}(\frac{L}{D}\delta_{kl})^2 \\
&= \sum_{ijkl} W_{ik}W_{il}W_{jk}W_{jl} - 2\frac{L}{D}\sum_{ik} W_{ik}^2 + \text{const.}(L, D) \\
&= \sum_{ijkl} W_{ik}W_{il}W_{jk}W_{jl} + \text{const.}(L, D)
\end{aligned}
\tag{36}
$$

where $\sum_k W_{ik}^2 = 1,\ \forall\ i$ is used extensively and the index letters were initially chosen to make the comparison of the final lines more clear. Le et al. (2011) show this first equivalence and that the L_2 cost is also shown to be equivalent to the reconstruction cost with whitened data (Lemmas 3.1 and 3.2 in Le et al. (2011)).

Now we can show that the same dictionary that was described in Section 2.1: W_0, an integer overcomplete dictionary where each set of complete bases is an orthonormal basis, exactly satisfies Eq 16 and so is a minimum of the L_2 cost. This solution is very far away from an ETF in the sense of Eq 15. A dictionary of this form, $W \in \mathbb{R}^{L\times D}$, can be constructed as $W_{ij} = \delta_{(i \bmod D)j}$ with $L = n \times D,\ n > 1,\ \in \mathbb{Z}$, that is, a D dimensional identity matrix tiled n times.

This construction satisfies Eq 16 and therefore has a value of 0 for $C_{\text{Eq 16}}$. Since $C_{\text{Eq 16}}$ is a sum of quadratic, and therefore non-negative, terms, this construction is a global minimum of $C_{\text{Eq 16}}$ and the L_2 cost

$$
\begin{aligned}
\sum_k W_{ki}W_{kj} &= \sum_k \delta_{(k \bmod D)i}\delta_{(k \bmod D)j} \\
&= n\delta_{ij} \\
&= \frac{L}{D}\delta_{ij} \\
&\Rightarrow C_{\text{Eq 16}} = 0
\end{aligned}
\tag{37}
$$

as $k \bmod D = i$ a total of n times when $i = j$.

However, this construction has off-diagonal Gram matrix elements that are either 0 or 1:

$$
\begin{aligned}
\cos\theta_{ij} &= \sum_k W_{ik}W_{jk} \\
&= \sum_k \delta_{(i \bmod D)k}\delta_{(j \bmod D)k} \\
&= \delta_{(i \bmod D)(j \bmod D)},
\end{aligned}
\tag{38}
$$

which is not equal or close to an equiangular solution, that is, $\cos\theta_{ij} = \cos\alpha,\ \forall i \neq j$. ■

B.4. Invariance to continuous transformations: proof of Theorem 2

Here we prove Theorem 2: the L_2 cost, initialized from the pathological solution, is invariant to transformations, Φ, constructed as orthogonal rotations applied to any basis subset and an identity transformation on the remaining bases. This shows that low coherence and high coherence configurations are both global minima of the L_2 cost.

Proof [Proof of Theorem 2] For an D dimensional space with an n times overcomplete dictionary, with n an integer greater than 1, the pathological dictionary configuration is a orthonormal basis tiled n times. The dictionary elements can be labels as the sequential subsets of orthornormal subsets $W_1, \ldots, W_D, \ldots, W_{2D}, \ldots, W_{n\times D}$. So, bases W_1 through W_D form a full-rank, orthonormal basis and this basis is tiled n times.

Consider the following partition of the bases: partition $\mathcal{A}$ is the first orthonormal set, bases W_1 through W_D, and partition $\mathcal{B}$ the remainder of the bases, W_{D+1} through $W_{n\times D}$. Let P be a projection operator for $\mathcal{A}$ and P^C its complement projection operator for $\mathcal{B}$, that is, $P^C W_i = W_i,\ PW_i = 0\ \forall\ W_i \in \mathcal{B}$ and $PW_j = W_j,\ P^C W_j = 0\ \forall\ W_j \in \mathcal{A}$. Let $R \in \mathrm{O}(L)$ be any rotation and PR a rotation that only acts on the $\mathcal{A}$ subspace. The operator $\Phi = PR + P^C$ is a rotation applied to all elements of $\mathcal{A}$ which leaves elements of $\mathcal{B}$ unchanged. Under its action, only terms in the cost between elements of $\mathcal{A}$ and $\mathcal{B}$ will change. It is straightforward to show that the terms in the cost that have both elements within $\mathcal{A}$ or both within $\mathcal{B}$ are constant since the rotation does not alter the relative pairwise angles.

For $W_i \in \mathcal{B}$, we can write down the terms in the L_2 cost which contain itself and elements from $\mathcal{A}PR$

$$\begin{aligned} C_{W_i}(\mathcal{A}\Phi) &= \sum_{W_j\in\mathcal{A}} (R^T P^T W_j^T W_i)^2 + (W_i^T W_j PR)^2 \\ &= \sum_{W_j\in\mathcal{A}} (R^T W_j^T W_i)^2 + (W_i^T W_j R)^2 \\ &= 2\sum_{W_j\in\mathcal{A}} \mathrm{Proj}_{W_j R}(W_i)^2 \\ &= 2|W_i|^2 \\ &= C_{W_i}(\mathcal{A}). \end{aligned} \tag{39}$$

Since the $W_j \in \mathcal{A}$ remain an orthonormal basis under a rotation, the sum of the projections-squared is the L_2 norm-squared of W_i which is constant. Since this is true for every $W_i \in \mathcal{B}$, the entire cost is constant under this transformation. This argument holds for any subset which forms an orthonormal basis and so all orthonormal subsets can rotate arbitrarily with respect to each other without changing the value of the L_2 cost, but the coherence of the matrix does depend on the transformation, Φ. This shows that the L_2 global minimum contains dictionaries with coherence $= 1$ and < 1 which can be continuously transformed into each other. ■

Appendix C. Additional figures

C.1. Extended Fig 2

Fig C1 is identical analysis as Fig 2 with all cost functions included.

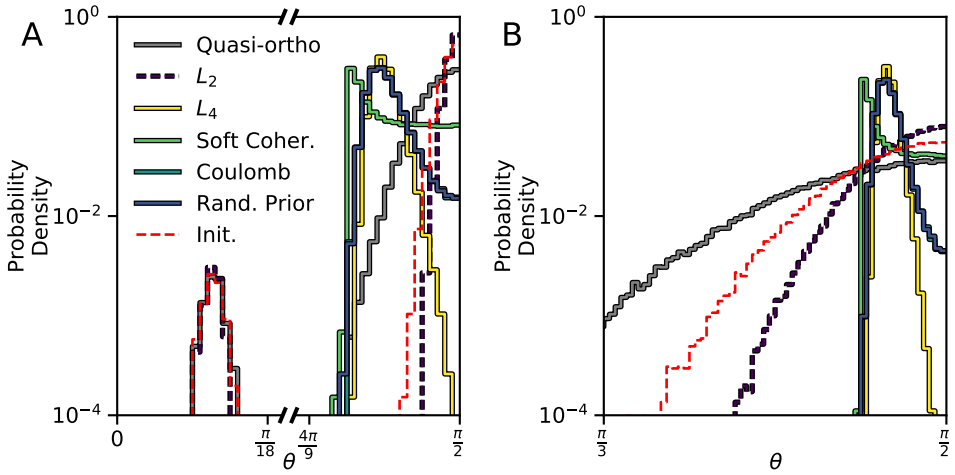

Figure C1: Coherence control costs have minima with varying coherence which can depend on initialization. Color legend is preserved across panels. For both panels a 2 times overcomplete dictionary with a data dimension of 64 was used. **A** Distribution of pairwise angles (log scale) obtained by numerically minimizing a subset of the coherence cost functions for the pathological dictionary initialization. Red dotted line indicates the initial distribution of pairwise angles. Note that the horizontal axis is broken at 10 and 80 degrees. **B** Angle distributions obtained (as in **A**) from a uniform random dictionary initialization. Note that the horizontal axis only includes $\frac{\pi}{3}$ to $\frac{\pi}{2}$.

C.2. Pairwise distributions for different powers

Fig C2 shows parwise angle distrubutions for cost functions based on powers of the difference between the gram matrix and the identity matrix for powers from 1 to 6 (skipping 2).

C.3. Local minima and saddle points are rare

Fig C3A shows the distribution of the mean-centered final cost values normalized by the average cost value at initialization

$$\text{Normalized Cost} = \frac{C_{\min} - \langle C_{\min} \rangle}{\langle C_{\text{init}} \rangle} \tag{40}$$

over 1000 uniform n-sphere initializations. All optimized cost functions are tighly bunched (1 part in $\sim 10^{-3}$). Minimizing the L_2 cost finds minima that are exactly equal to single precision floating point. This means that all costs are empirically converging to the same

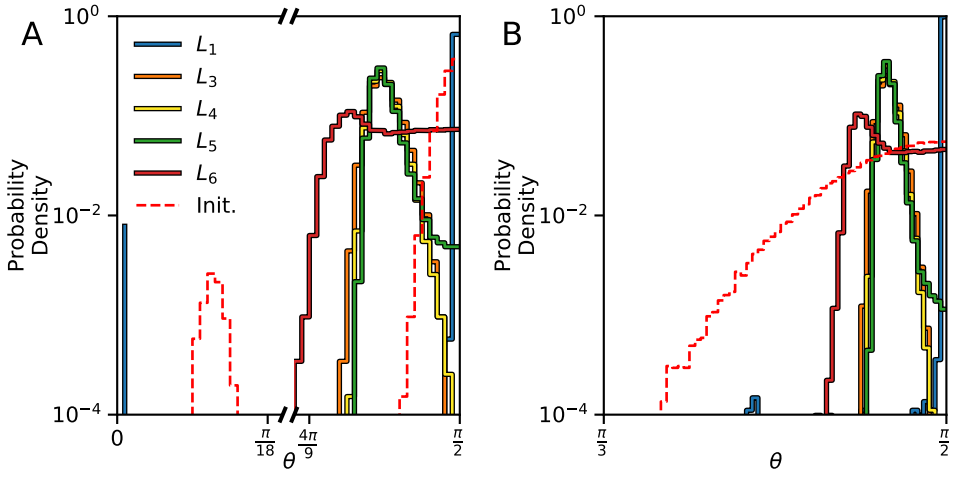

Figure C2: Coherence control costs based on powers of the difference between the Gram matrix and the identity matrix have highest coherence for powers 4 and 5. **A** Distribution of pairwise angles (log scale) obtained by numerically minimizing power-based coherence cost functions for the pathological dictionary initialization. Red dotted line indicates the initial distribution of pairwise angles. Note that the horizontal axis is broken at 10 and 80 degrees. **B** Angle distributions obtained (as in **A**) from a uniform random dictionary initialization. Note that the horizontal axis only includes 65 to 90 degrees.

value across many random re-initializations implying that they are not getting stuck in local minima or saddle points.

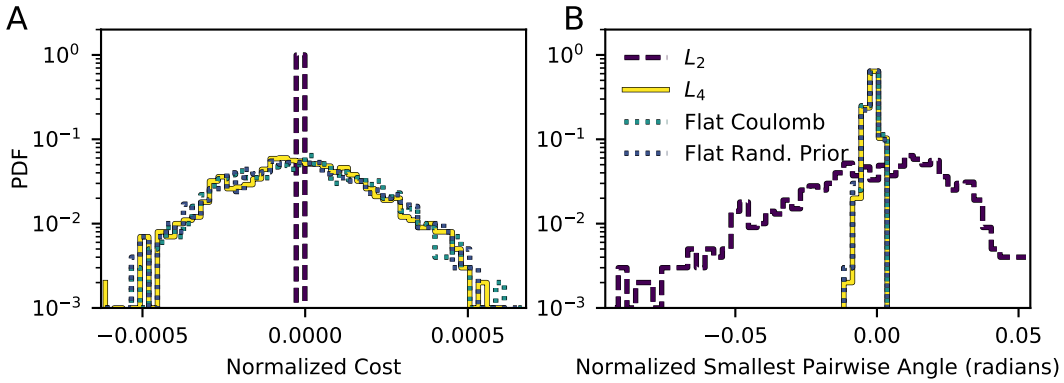

Figure C3: Local minima and saddle points do not impede optimization. **A** The distribution of the mean-centered final cost values normalized by the average cost value at initialization over 1000 uniform n-sphere initializations. **B** The distribution of the mean-centered final smallest pariwise angles over 1000 uniform n-sphere initializations.

Fig C3B shows the distribution of mean-centered final smallest pariwise angles

$$\text{Normalized Smallest Pairwise Angle} = \Theta_{\min} - \langle \Theta_{\min} \rangle \tag{41}$$

over 1000 uniform n-sphere initializations. Although all cost functions minimize to consistent values, they do not all minimize to consistent values of coherence (cosine of the smallest pairwise angle). The L_4, Flat Coulomb and Flat Random Prior costs all minimize to have tightly bunched coherence ($\sim .002$ radians, ~ 0.1 degrees) compared to the L_2 cost (~ 0.025 radians, ~ 1.5 degrees). This shows that all costs do not have problems with local minima or saddle points, but that minimizing the L_2 cost does not lead to consistently low coherence.

C.4. Extended Fig 4

Fig C4 is identical analysis as Fig 4 with all cost functions included.

C.5. Extended Fig 5

Fig C5 is identical analysis as Fig 5 with all cost functions included.

C.6. Supplemented Fig 3D.

Fig C6 is similar to Fig 3D for the Coulomb and Random Prior costs and their flattened versions.

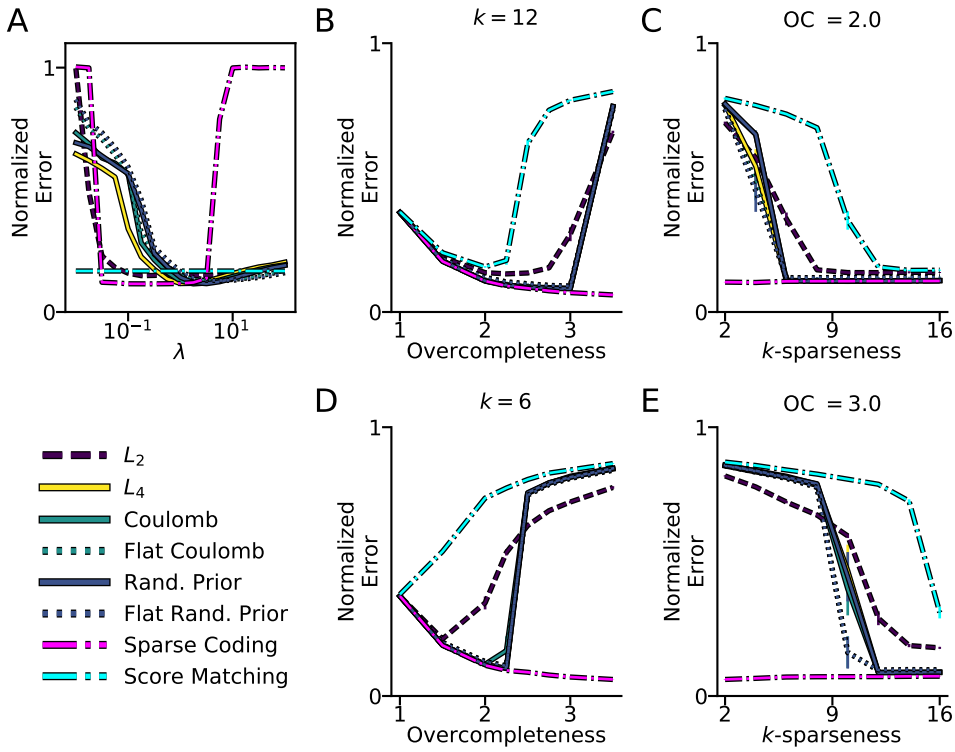

Figure C4: Coherence control costs do not all recover mixing matrices well. All ground truth mixing matrices were generated from the Soft Coherence cost and had a data dimension of 32. Color and line style legend are preserved across panels. **A** The normalized recovery error (see Section 3 for details) for a 2-times overcomplete mixing matrix and $k = 12$ as a function of the sparsity prior weight (λ). Since score matching does not have a λ parameter, it is plotted at a constant. **B** Recovery performance ($\pm$ s.e.m., $n = 10$) at the best value of λ as a function of overcompleteness at $k = 12$. **C** Recovery performance ($\pm$ s.e.m., $n = 10$) at the best value of λ as a function of k-sparseness at 2-times overcompleteness. **D, E** Same plots as **B** and **C** at a point where methods do not perform as well: $k = 6$ and 3-times overcomplete. In **B-E**, the L_4, Coulomb, Flattened Coulomb, Random, and Flattened Random lines are largely overlapping.

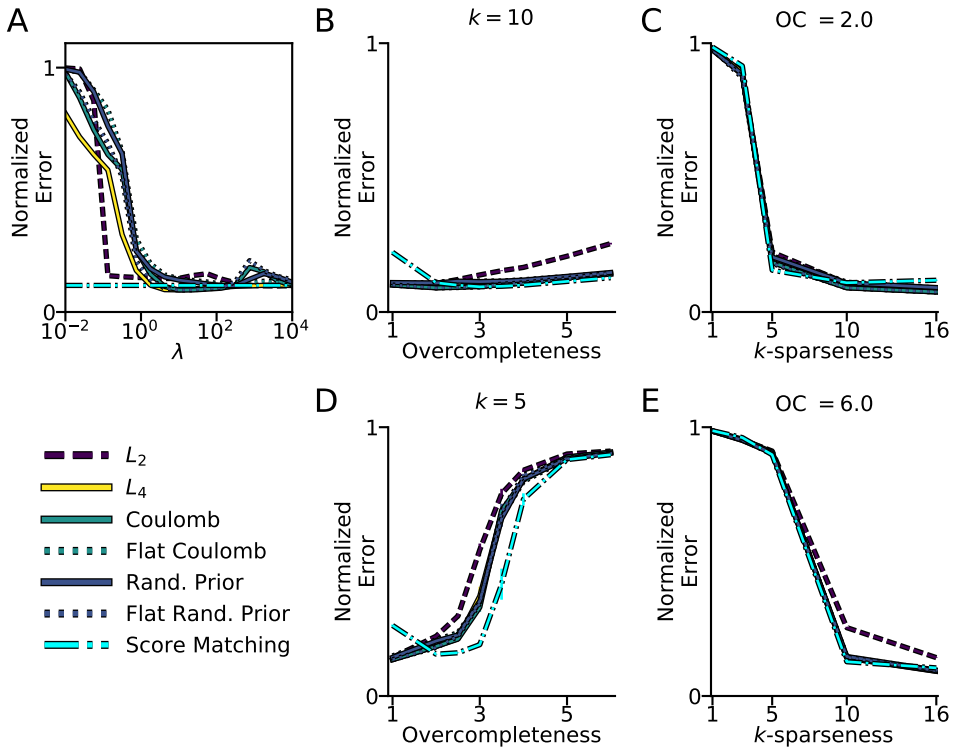

Figure C5: Coherence control costs do not all recover analysis matrices well. All ground truth analysis matrices were generated from the Soft Coherence cost and had a data dimension of 32. Color and line style legend are preserved across panels. **A** The normalized recovery error (see Section 3 for details) for a 2-times overcomplete analysis matrix and $k = 10$ as a function of the sparsity prior weight (λ). Since score matching does not have a λ parameter, it is plotted at a constant. **B** Recovery performance ($\pm$ s.e.m., $n = 10$) at the best value of λ as a function of overcompleteness at $k = 10$. **C** Recovery performance ($\pm$ s.e.m., $n = 10$) at the best value of λ as a function of k-sparseness at 2-times overcompleteness. **D, E** Same plots as **B** and **C** at a point where methods do not perform as well: $k = 5$ and 6-times overcomplete. In **B-E**, the L_4, Coulomb, Flattened Coulomb, Random, and Flattened Random lines are largely overlapping.

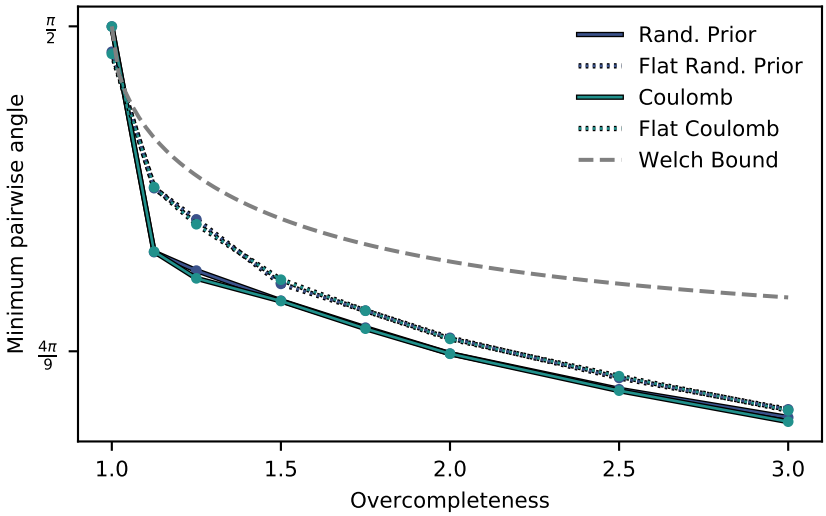

Figure C6: Quadratic terms dominate the minima of coherence control costs as a function of overcompleteness. The median minimum pairwise angle (arccosine of coherence) across 10 initializations is plotted as a function of overcompleteness for a dictionary with a data dimension of 32. The largest possible value (Welch Bound) is also shown as a function of overcompleteness.

References

Gautam Agarwal, Ian H Stevenson, Antal Berényi, Kenji Mizuseki, György Buzsáki, and Friedrich T Sommer. Spatially distributed local fields in the hippocampus encode rat position. *Science*, 344(6184):626–630, 2014.

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? In *Advances in Neural Information Processing Systems*, pages 4261–4271, 2018.

Chenglong Bao, Yuhui Quan, and Hui Ji. A convergent incoherent dictionary learning algorithm for sparse coding. In *European Conference on Computer Vision*, pages 302–316. Springer, 2014.

Horace B Barlow. The ferrier lecture, 1980: Critical limiting factors in the design of the eye and visual cortex. *Proceedings of the Royal Society of London B: Biological Sciences*, 212(1186):1–34, 1981.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Anthony J Bell and Terrence J Sejnowski. The independent components of natural scenes are edge filters. *Vision research*, 37(23):3327–3338, 1997.

Andrew Brock, Theodore Lim, James Millar Ritchie, and Nicholas J Weston. Neural photo editing with introspective adversarial networks. In *5th International Conference on Learning Representations 2017*, 2017.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Nicole L Carlson, Vivienne L Ming, and Michael Robert DeWeese. Sparse codes for speech predict spectrotemporal receptive fields in the inferior colliculus. *PLoS Comput Biol*, 8 (7):e1002594, 2012.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Il Yong Chun and Jeffrey A Fessler. Convolutional analysis operator learning: Acceleration, convergence, application, and neural networks. *arXiv preprint arXiv:1802.05584*, 2018.

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994.

Christine A Curcio and Kimberly A Allen. Topography of ganglion cells in human retina. *Journal of comparative Neurology*, 300(1):5–25, 1990.

Mark A Davenport, Marco F Duarte, Yonina C Eldar, and Gitta Kutyniok. Introduction to compressed sensing. *preprint*, 93(1):2, 2011.

Arnaud Delorme, Terrence Sejnowski, and Scott Makeig. Enhanced detection of artifacts in eeg data using higher-order statistics and independent component analysis. *Neuroimage*, 34(4):1443–1449, 2007.

Michael R DeWeese, Tomáš Hromádka, and Anthony M Zador. Reliability and representational bandwidth in the auditory cortex. *Neuron*, 48(3):479–488, 2005.

David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

Michael Elad, Peyman Milanfar, and Ron Rubinstein. Analysis versus synthesis in signal priors. *Inverse problems*, 23(3):947, 2007.

Matthew Fickus and Dustin G Mixon. Tables of the existence of equiangular tight frames. *arXiv preprint arXiv:1504.00253*, 2015.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 399–406. Omnipress, 2010.

Christopher J Hillar and Friedrich T Sommer. When can dictionary learning uniquely recover sparse data from subsamples? *IEEE Transactions on Information Theory*, 61 (11):6290–6297, 2015.

Jun-ichiro Hirayama, Takeshi Ogawa, and Aapo Hyvärinen. Unifying blind separation and clustering for resting-state eeg/meg functional connectivity analysis. *Neural computation*, 2015.

Stephen D Howard, A Robert Calderbank, and Stephen J Searle. A fast reconstruction algorithm for deterministic compressive sensing using second order reed-muller codes. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pages 11–15. IEEE, 2008.

Tao Hu, Cengiz Pehlevan, and Dmitri B Chklovskii. A hebbian/anti-hebbian network for online sparse dictionary learning derived from symmetric matrix factorization. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 613–619. IEEE, 2014.

Aapo Hyvarinen. Fast ica for noisy data using gaussian moments. In *ISCAS'99. Proceedings of the 1999 IEEE International Symposium on Circuits and Systems VLSI (Cat. No. 99CH36349)*, volume 5, pages 57–61. IEEE, 1999.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. In *Journal of Machine Learning Research*, pages 695–709, 2005.

Aapo Hyvärinen and Mika Inki. Estimating overcomplete independent component bases for image windows. *Journal of Mathematical Imaging and Vision*, 17(2):139–152, 2002.

Aapo Hyvärinen and Urs Köster. Complex cell pooling and the statistics of natural images. *Network: Computation in Neural Systems*, 18(2):81–100, 2007.

Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

Aapo Hyvärinen, Razvan Cristescu, and Erkki Oja. A fast algorithm for estimating overcomplete ica bases for image windows. In *Neural Networks, 1999. IJCNN'99. International Joint Conference on*, volume 2, pages 894–899. IEEE, 1999.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2001.

Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–2017. URL `http://www.scipy.org/`. [Online; accessed 2017-03-15].

David J Klein, Peter König, and Konrad P Körding. Sparse spectrotemporal coding of sounds. *EURASIP Journal on Advances in Signal Processing*, 2003(7):902061, 2003.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Quoc V Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Y Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in Neural Information Processing Systems*, pages 1017–1025, 2011.

Geneviève Leuba and Rudolf Kraftsik. Changes in volume, surface estimate, three-dimensional shape and total number of neurons of the human primary visual cortex from midgestation until old age. *Anatomy and embryology*, 190(4):351–366, 1994.

Michael S Lewicki and Bruno A Olshausen. Probabilistic framework for the adaptation and comparison of image codes. *JOSA A*, 16(7):1587–1601, 1999.

Michael S Lewicki and Terrence J Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

Jörg Lücke, Richard Turner, Maneesh Sahani, and Marc Henniges. Occlusive components analysis. In *Advances in Neural Information Processing Systems*, pages 1069–1077, 2009.

Boris Mailhé, Daniele Barchiesi, and Mark D Plumbley. Ink-svd: Learning incoherent dictionaries for sparse representations. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3573–3576. IEEE, 2012.

Jerry L Northern and Marion P Downs. *Hearing in children*. Lippincott Williams & Wilkins, 2002.

Bruno A Olshausen. Highly overcomplete sparse coding. In *IS&T/SPIE Electronic Imaging*, pages 86510S–86510S. International Society for Optics and Photonics, 2013.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Boaz Ophir, Michael Elad, Nancy Bertin, and Mark D Plumbley. Sequential minimal eigenvalues-an approach to analysis dictionary learning. In *2011 19th European Signal Processing Conference*, pages 1465–1469. IEEE, 2011.

Ignacio Ramırez, Federico Lecumberry, and Guillermo Sapiro. Sparse modeling with universal priors and learned incoherent dictionaries. Technical report, Citeseer, 2009.

Martin Rehn and Friedrich T Sommer. A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *Journal of computational neuroscience*, 22(2):135–146, 2007.

Dario L Ringach. Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of neurophysiology*, 88(1):455–463, 2002.

Christopher J Rozell, Don H Johnson, Richard G Baraniuk, and Bruno A Olshausen. Sparse coding via thresholding and local competition in neural circuits. *Neural computation*, 20 (10):2526–2563, 2008.

Ron Rubinstein, Tomer Peleg, and Michael Elad. Analysis k-svd: A dictionary-learning algorithm for the analysis sparse model. *IEEE Transactions on Signal Processing*, 61(3): 661–677, 2013.

Christian D Sigg, Tomas Dikk, and Joachim M Buhmann. Learning dictionaries with bounded self-coherence. *IEEE Signal Processing Letters*, 19(12):861–864, 2012.

Steve Smale. Mathematical problems for the next century. *The Mathematical Intelligencer*, 20(2):7–15, 1998.

Evan C Smith and Michael S Lewicki. Efficient auditory coding. *Nature*, 439(7079):978–982, 2006.

H Spoendlin and A Schrott. Analysis of the human auditory nerve. *Hearing research*, 43 (1):25–38, 1989.

Thomas Strohmer and Robert W Heath Jr. Grassmannian frames with applications to coding and communication. *Applied and computational harmonic analysis*, 14(3):257–275, 2003.

Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton. Energy-based models for sparse overcomplete representations. *The Journal of Machine Learning Research*, 4: 1235–1260, 2003.

Theano Development Team. Theano: A Python framework for fast computation of mathematical expressions. *arXiv e-prints*, abs/1605.02688, May 2016. URL `http://arxiv.org/abs/1605.02688`.

J Hans van Hateren and Arjen van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London B: Biological Sciences*, 265(1394):359–366, 1998.

Lloyd Welch. Lower bounds on the maximum cross correlation of signals (corresp.). *IEEE Transactions on Information theory*, 20(3):397–399, 1974.

Joel Zylberberg, Jason Timothy Murphy, and Michael Robert DeWeese. A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of v1 simple cell receptive fields. *PLoS computational biology*, 7(10):e1002250, 2011.