How to win friends and influence people with PCA

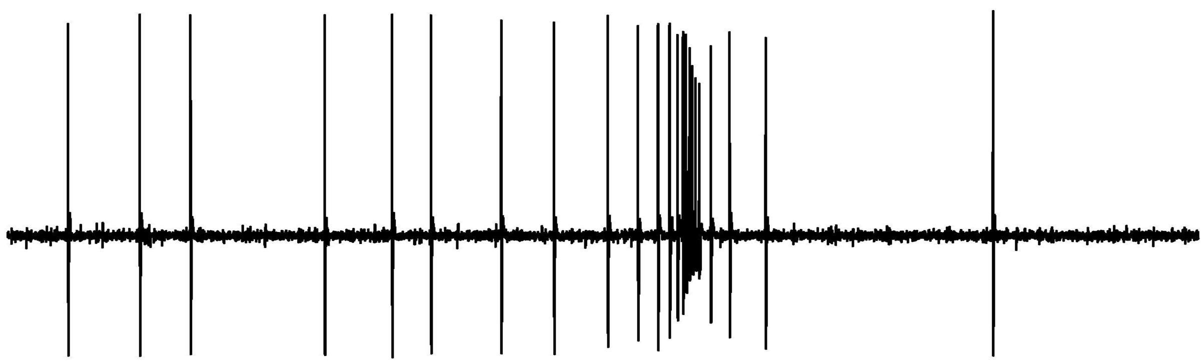
Goals:
Qualitative introduction to PCA
Spike sorting
Behavioral analysis

Resource:
http://www.snl.salk.edu/~shlens/pca.pdf

Sam Sober
samuel.j.sober@emory.edu

We want a quantitative criterion for deciding whether a recording is single- or multiunit.

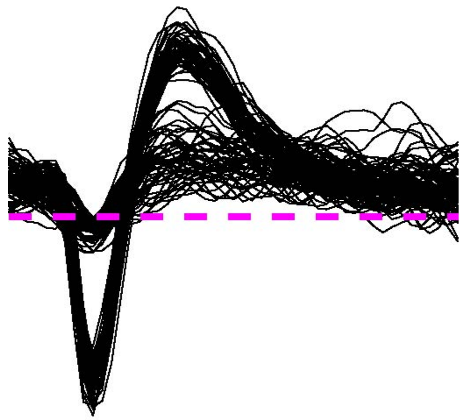
Single unit:



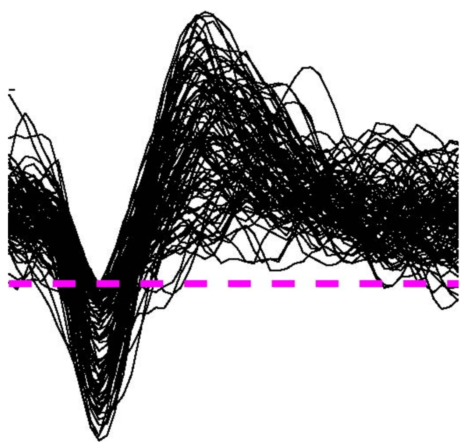
?

We want a quantitative criterion for deciding whether a recording is single- or multiunit.

Single unit:



?



An ad-hoc method:

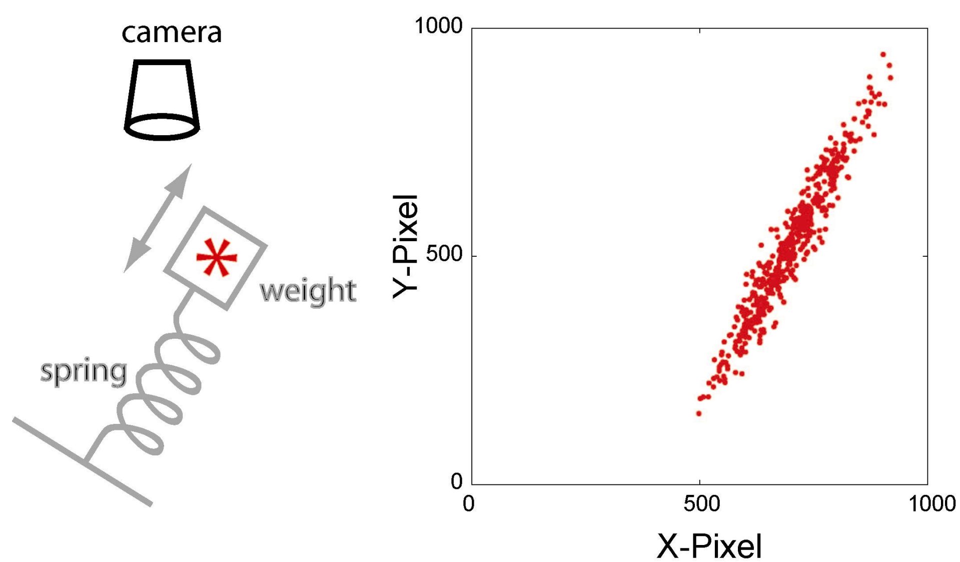
Starts with Principal Components Analysis (PCA)

A math-free intro to PCA:

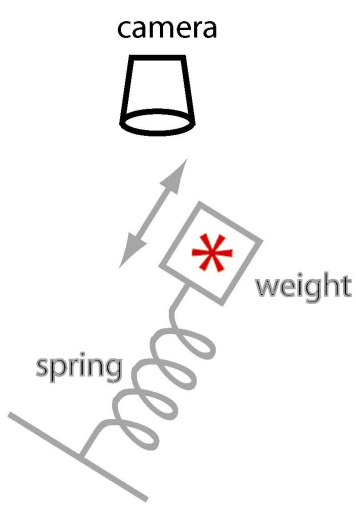
(Based on http://www.keck.ucsf.edu/~sam/PCA_tutorial_Shlens.pdf)

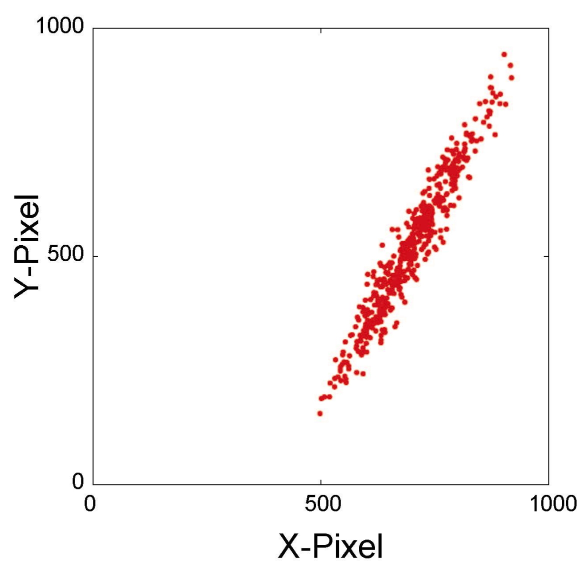
Measurements aren't always in the "right" coordinates:

-Axes don't correspond to anything meaningful
-System is 1-D, data are 2-D.



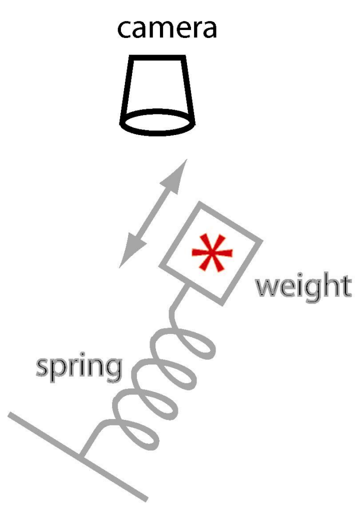
We want a new coordinate system:

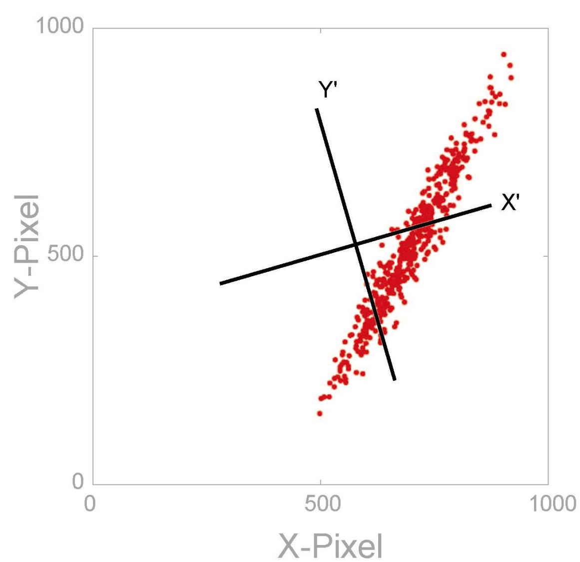




We want a new coordinate system:

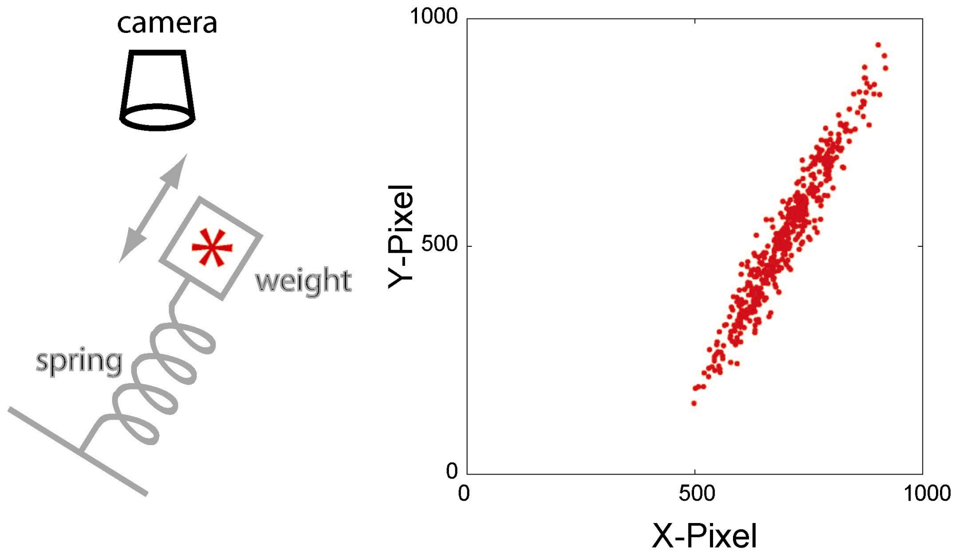
For example:





PCA: Change coordinate system so that axes reflect important* directions of variability

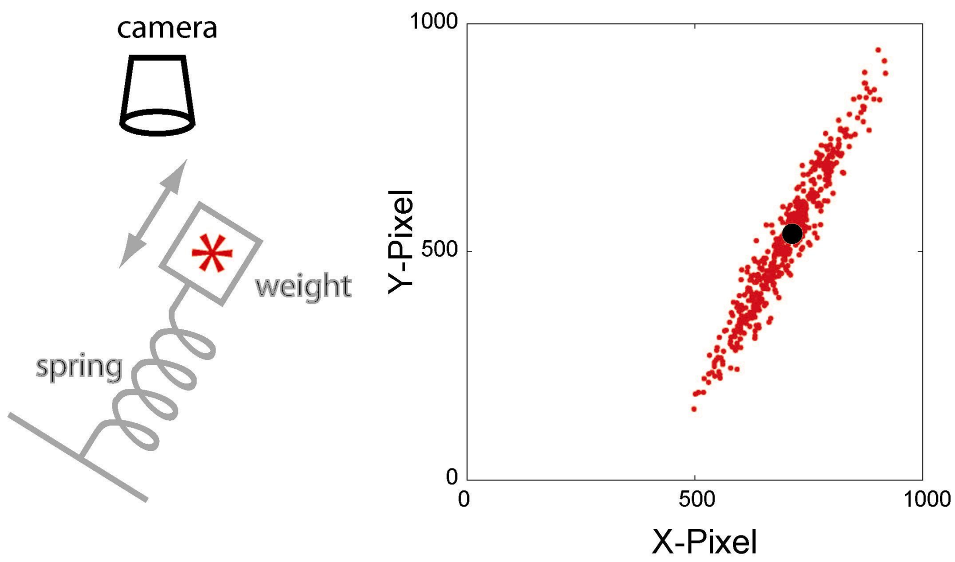
*assumption: important directions are ones with greatest variability: $Var_{signal} >> Var_{noise}$



PCA: Change coordinate system so that axes reflect important* directions of variability

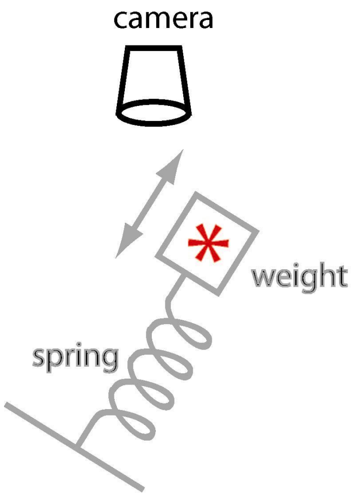
*assumption: important directions are ones with greatest variability: $\mathrm{Var}_{signal} \gg \mathrm{Var}_{noise}$

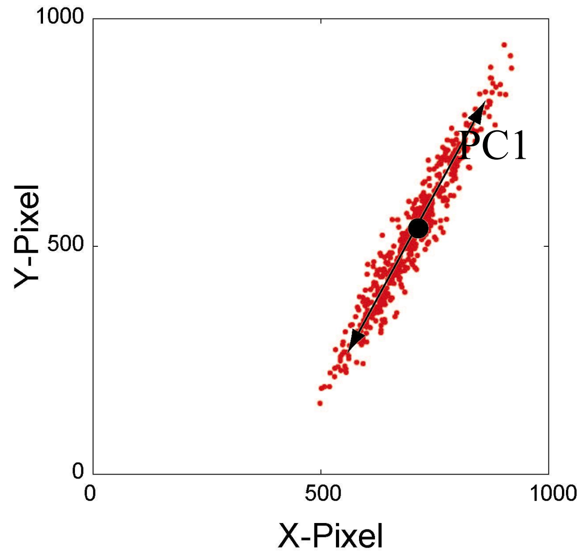
1. Place origin at mean of all data



PCA: Change coordinate system so that axes reflect important* directions of variability

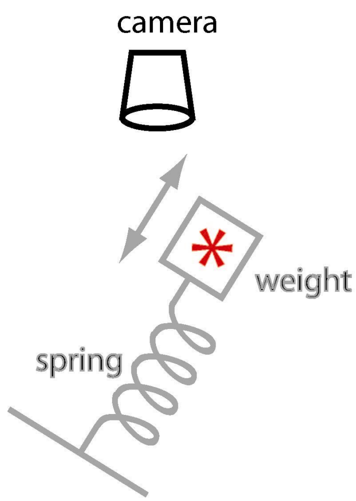
1. Place origin at mean of all data
2. Find direction with biggest variance – 1st principal component

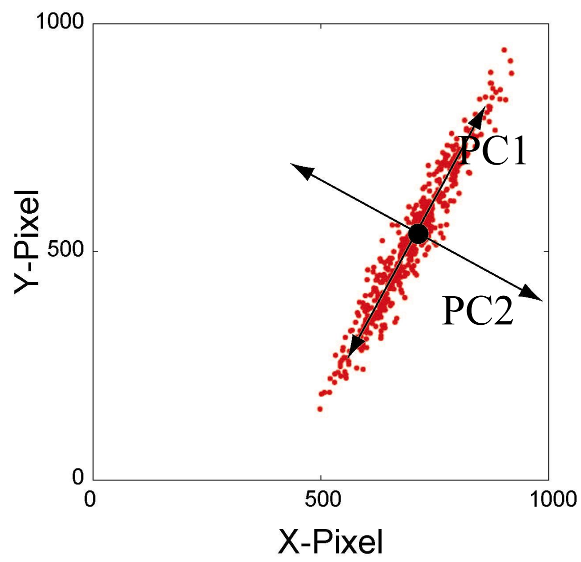




PCA: Change coordinate system so that axes reflect important* directions of variability

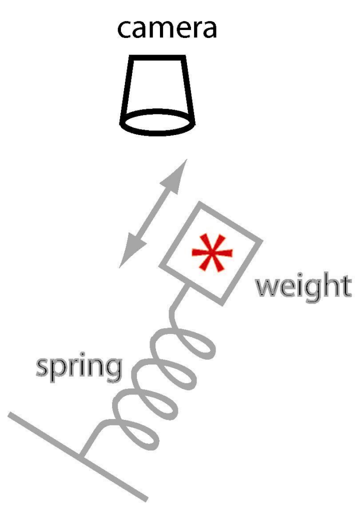
1. Place origin at mean of all data
2. Find direction with biggest variance – 1st principal component
3. Find orthogonal direction with next biggest variance – 2nd principal component

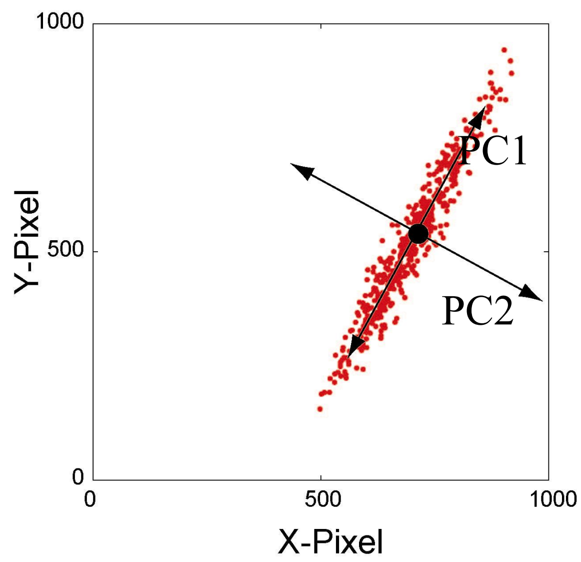




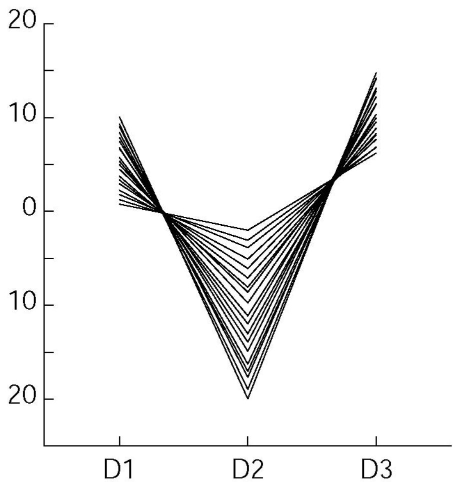
PCA: Change coordinate system so that axes reflect important* directions of variability

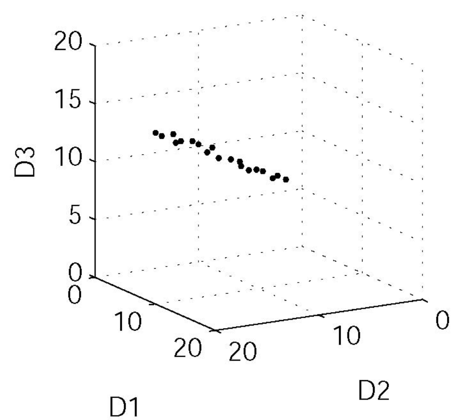
1. Place origin at mean of all data
2. Find direction with biggest variance – 1st principal component
3. Find orthogonal direction with next biggest variance – 2nd principal component
4. Keep going through *n* dimensions to get *n* principal components



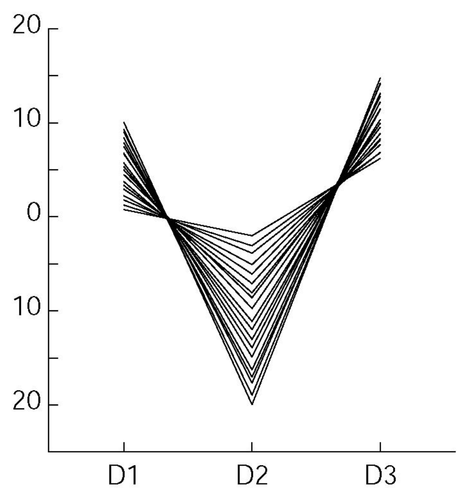


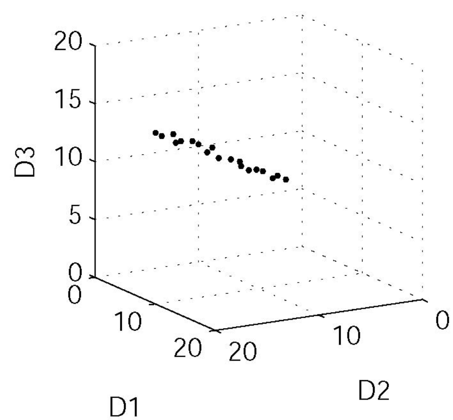
1. Place origin at mean of all data
2. Find direction with biggest variance – 1st principal component
3. Find orthogonal direction with next biggest variance – 2nd principal component

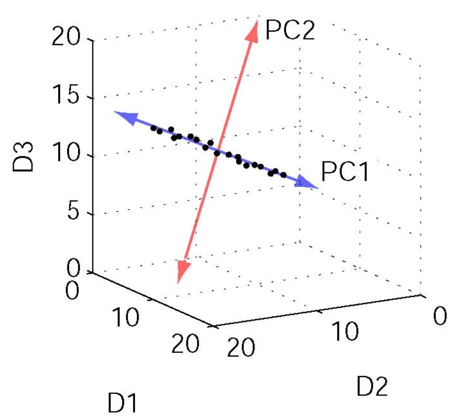




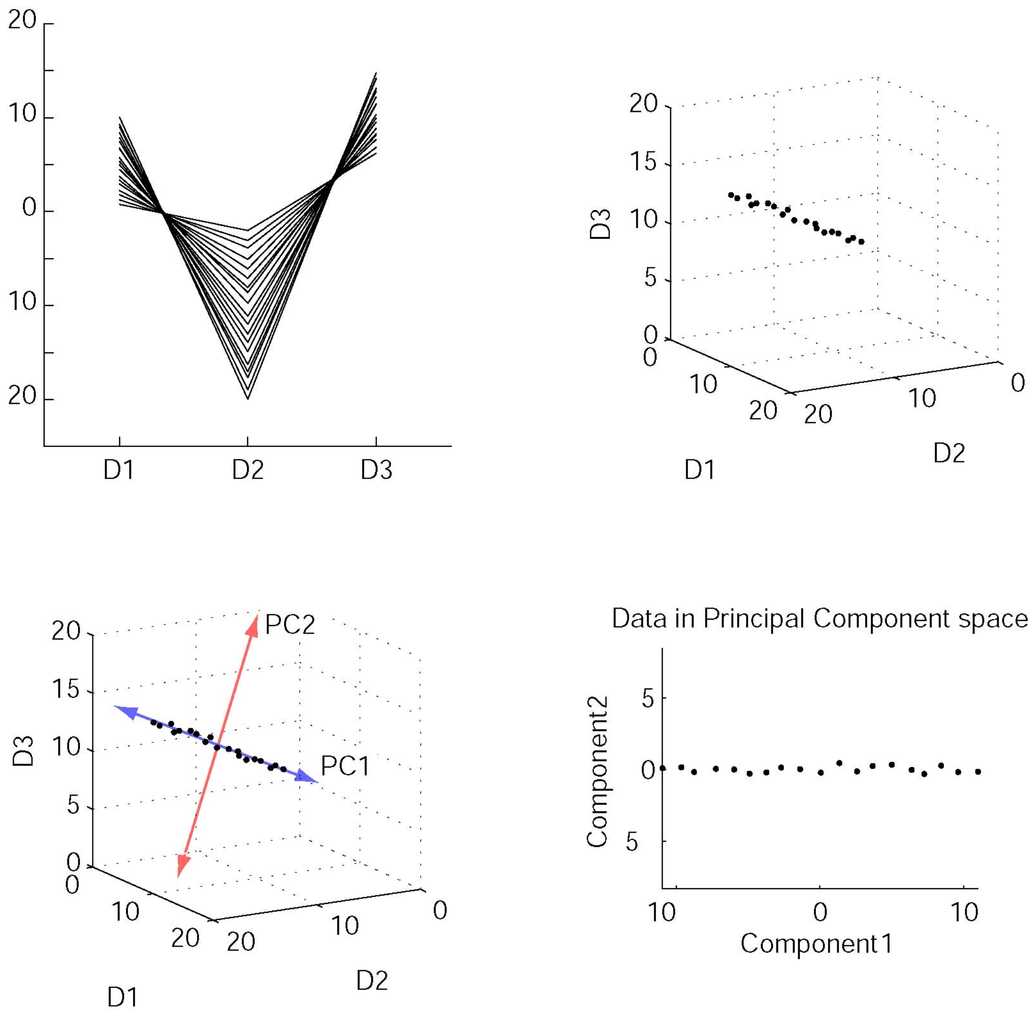
1. Place origin at mean of all data
2. Find direction with biggest variance – 1st principal component
3. Find orthogonal direction with next biggest variance – 2nd principal component





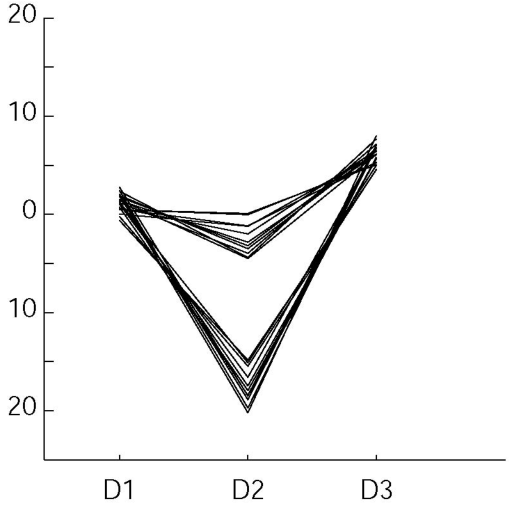


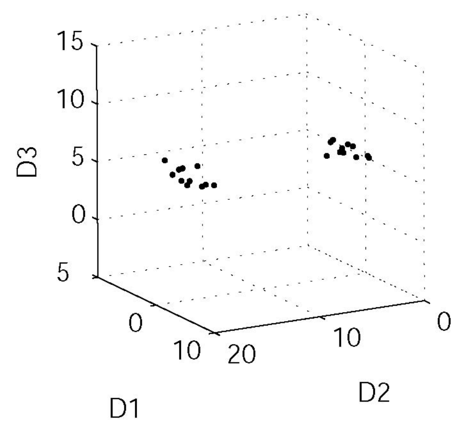
1. Place origin at mean of all data
2. Find direction with biggest variance – 1st principal component
3. Find orthogonal direction with next biggest variance – 2nd principal component

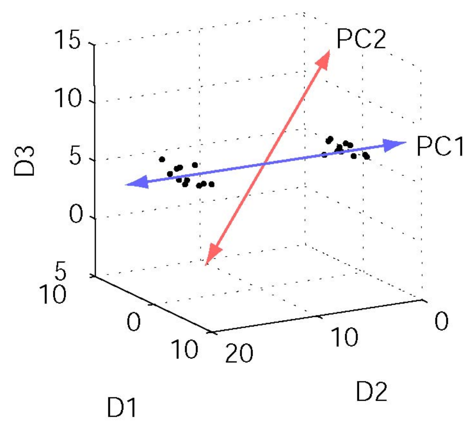


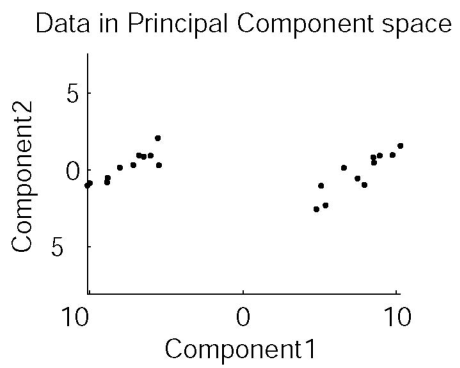
This a 2-D view of 3-D, data, looking at the 2 most "important" (variable) directions.

1. Place origin at mean of all data
2. Find direction with biggest variance – 1st principal component
3. Find orthogonal direction with next biggest variance – 2nd principal component



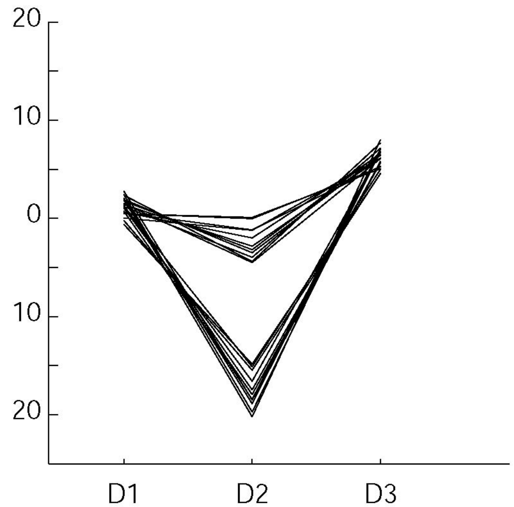


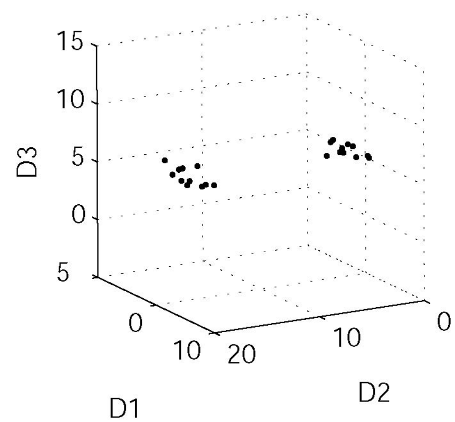


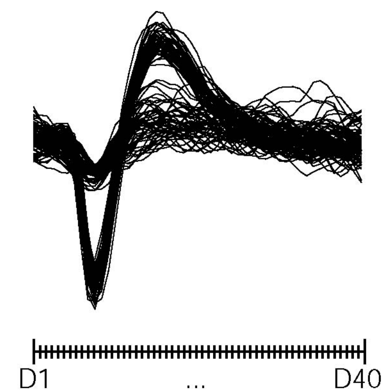


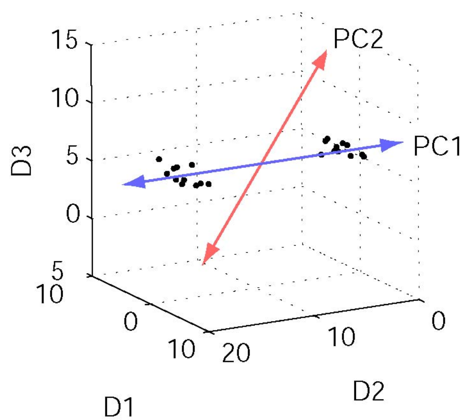
This a 2-D view of 3-D, data, looking at the 2 most "important" (variable) directions.

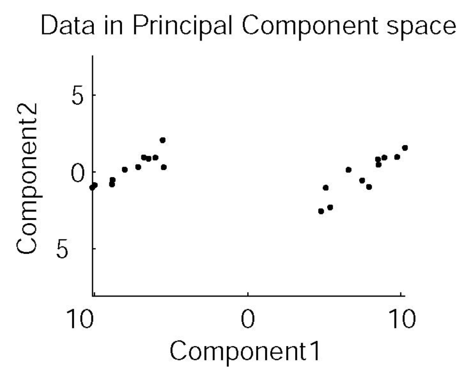
1. Place origin at mean of all data
2. Find direction with biggest variance – 1st principal component
3. Find orthogonal direction with next biggest variance – 2nd principal component





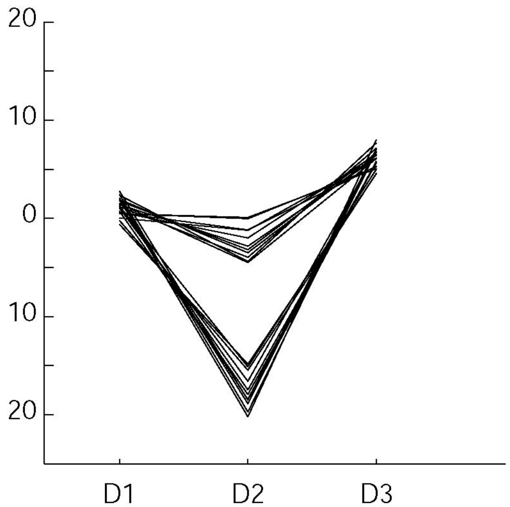


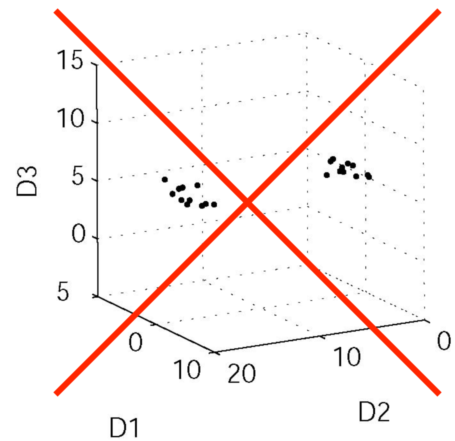


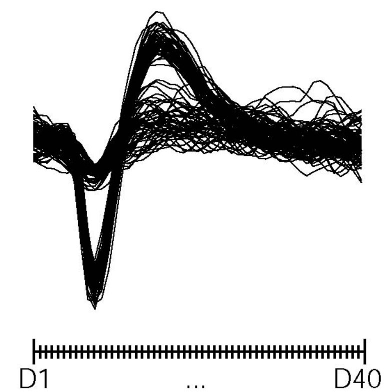


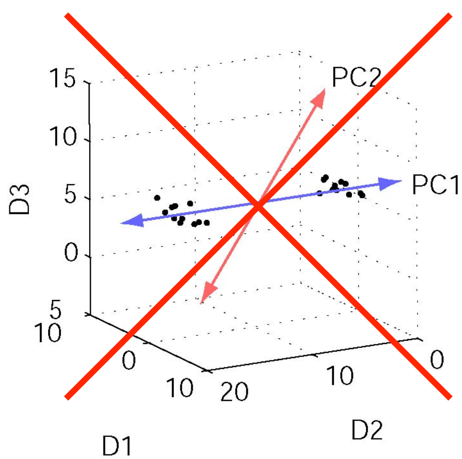
This a 2-D view of 3-D, data, looking at the 2 most “important” (variable) directions.

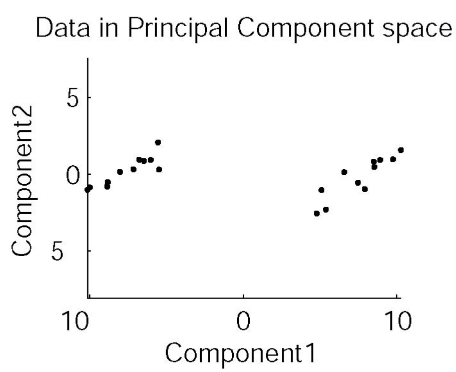
1. Place origin at mean of all data
2. Find direction with biggest variance – 1st principal component
3. Find orthogonal direction with next biggest variance – 2nd principal component





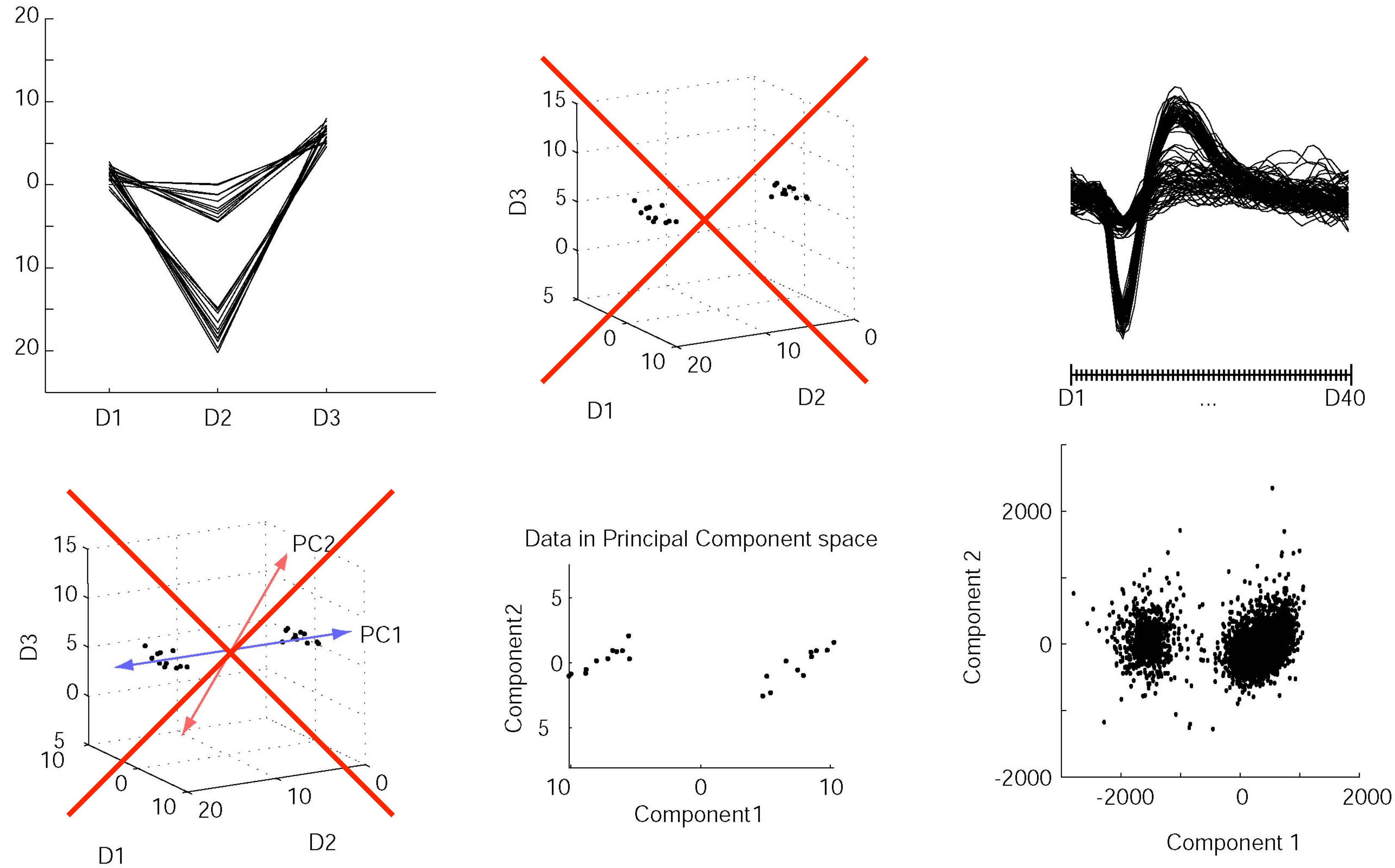




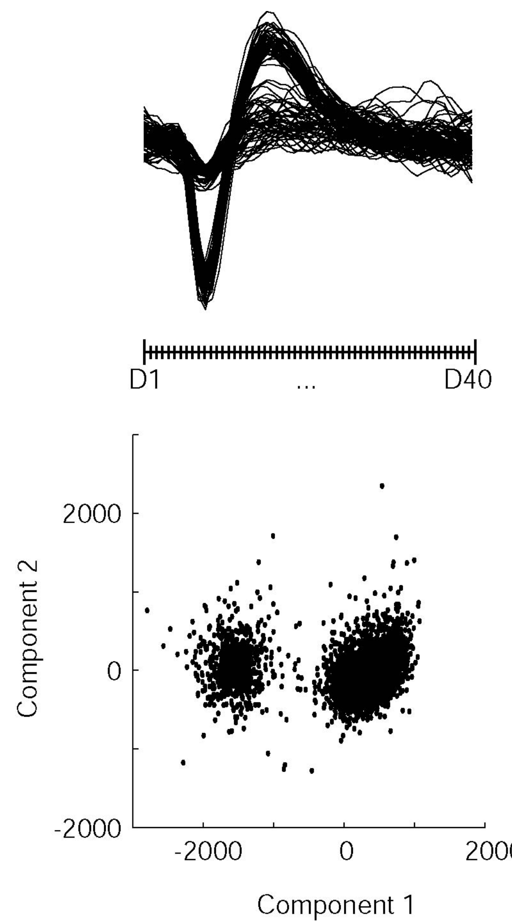


This a 2-D view of 3-D, data, looking at the 2 most "important" (variable) directions.

1. Place origin at mean of all data
2. Find direction with biggest variance – 1st principal component
3. Find orthogonal direction with next biggest variance – 2nd principal component

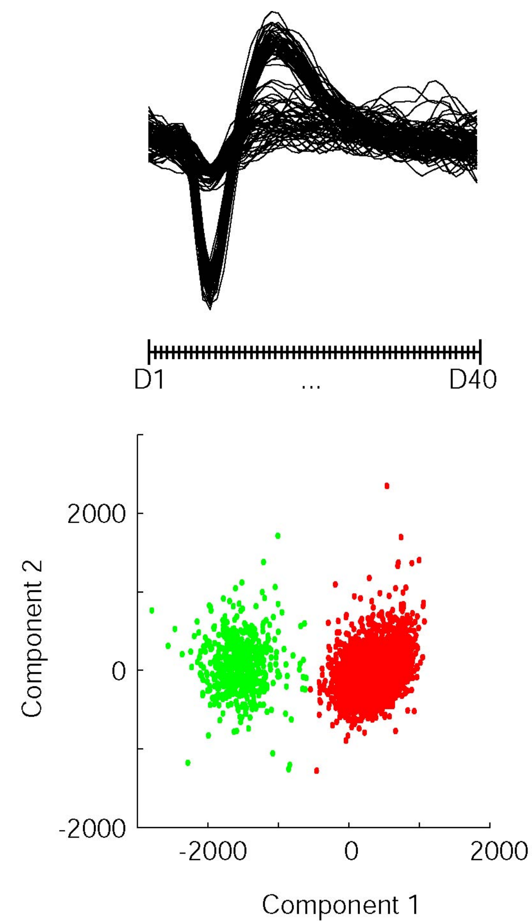


How can we quantify isolation?



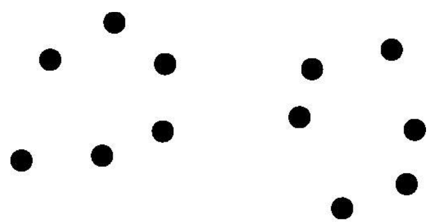
How can we quantify isolation?

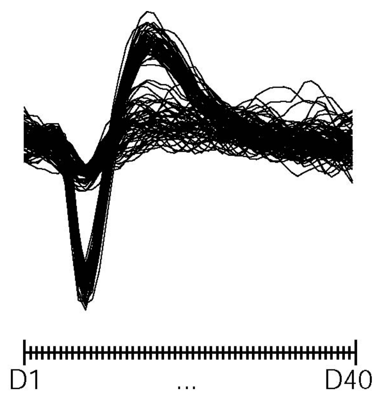
1. Cluster with kmeans.m
- Set cluster number manually.
- Clustering based on distance from center

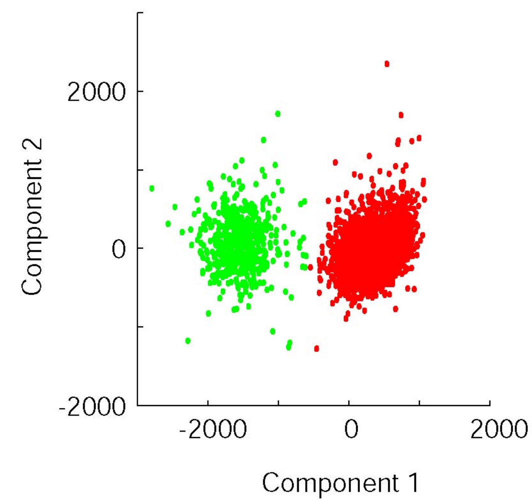


How can we quantify isolation?

1. Cluster with kmeans.m
- Set cluster number manually.
- Clustering based on distance from center

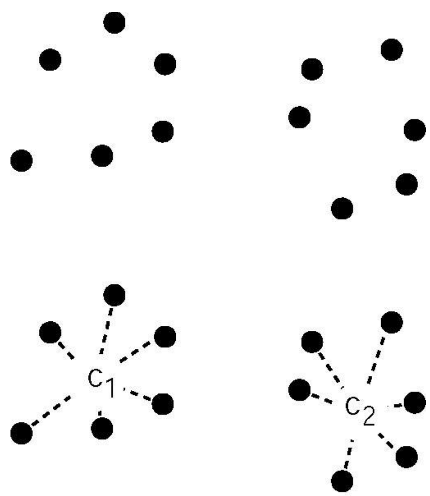


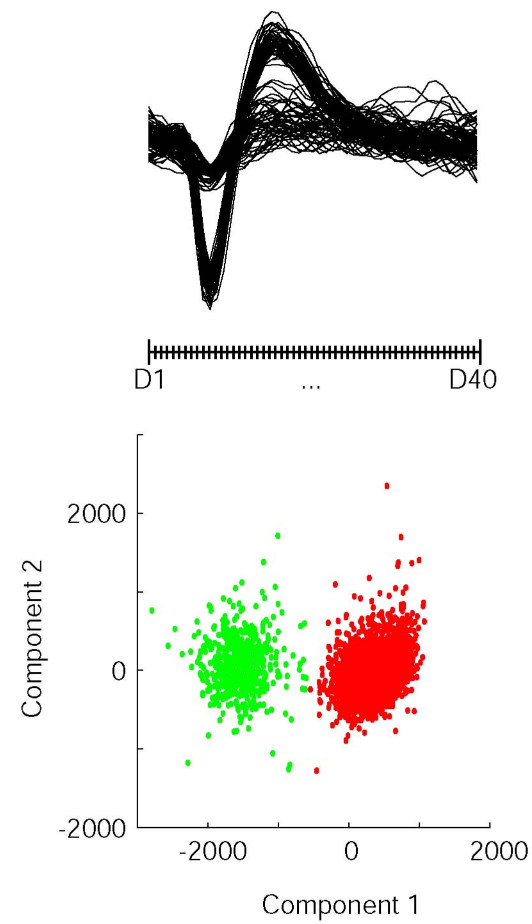




How can we quantify isolation?

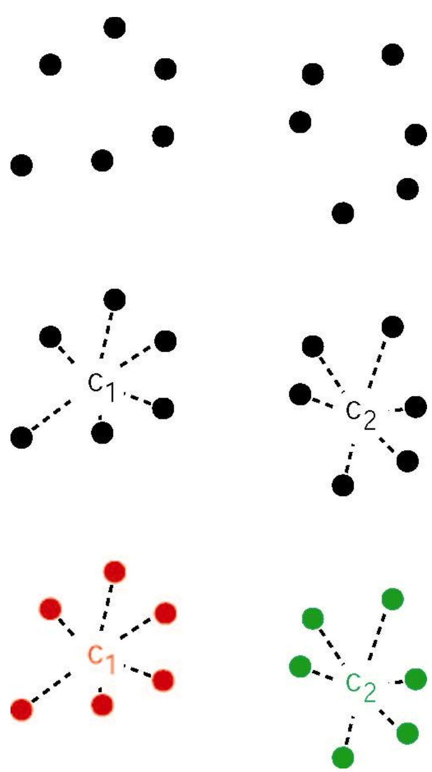
1. Cluster with kmeans.m
- Set cluster number manually.
- Clustering based on distance from center

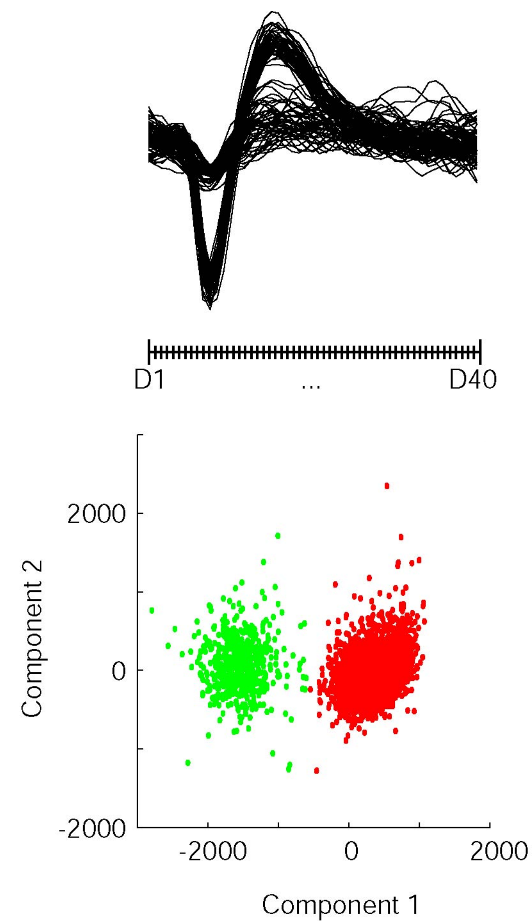




How can we quantify isolation?

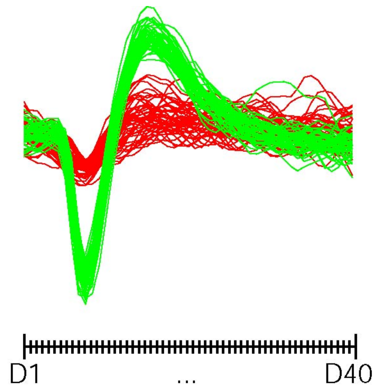
1. Cluster with kmeans.m
- Set cluster number manually.
- Clustering based on distance from center

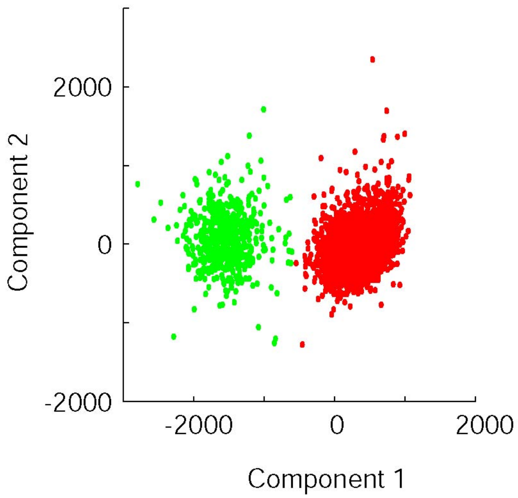




How can we quantify isolation?

1. Cluster with kmeans.m
- Set cluster number manually.
- Clustering based on distance from center

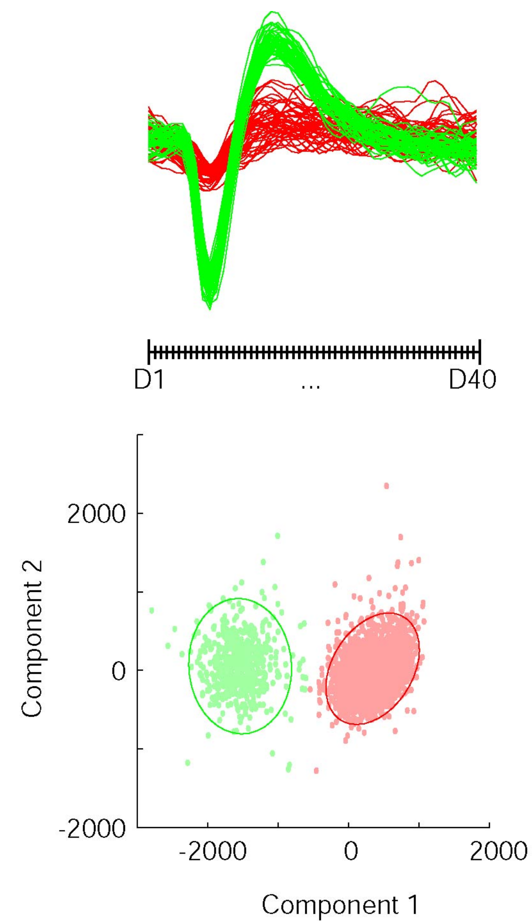




How can we quantify isolation?

1. Cluster with kmeans.m
- Set cluster number manually.
- Clustering based on distance from center

• Describe clusters as 2D gaussians

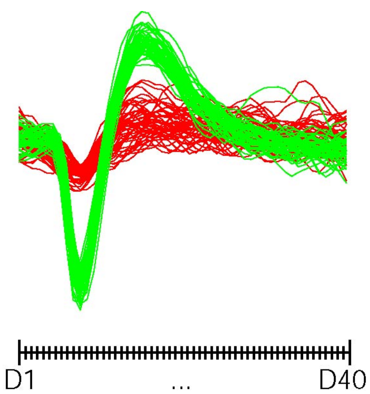


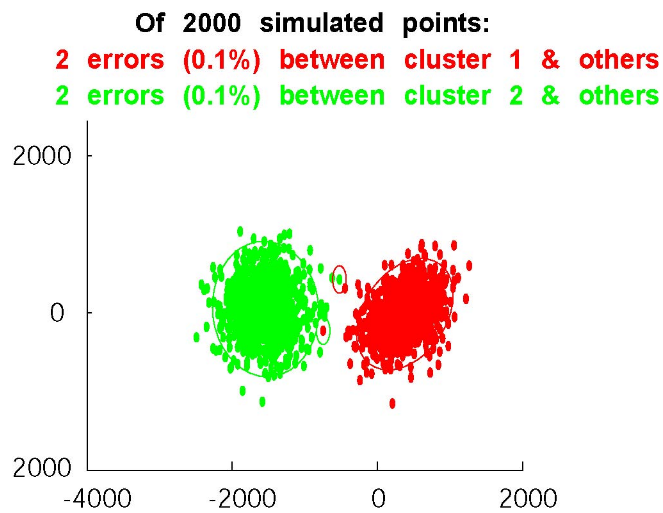
How can we quantify isolation?

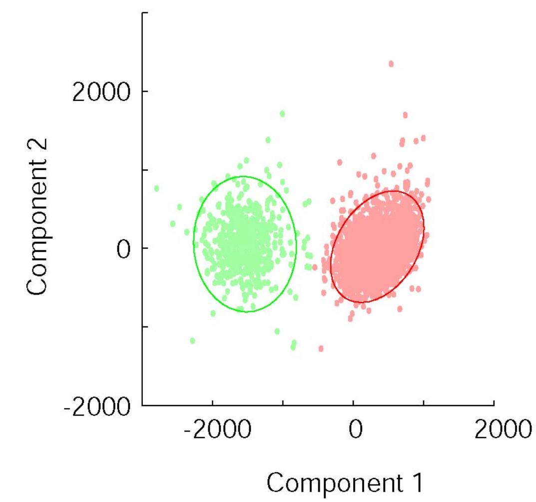
1. Cluster with kmeans.m
- Set cluster number manually.
- Clustering based on distance from center

• Describe clusters as 2D gaussians

• Simulate distributions to estimate error rate

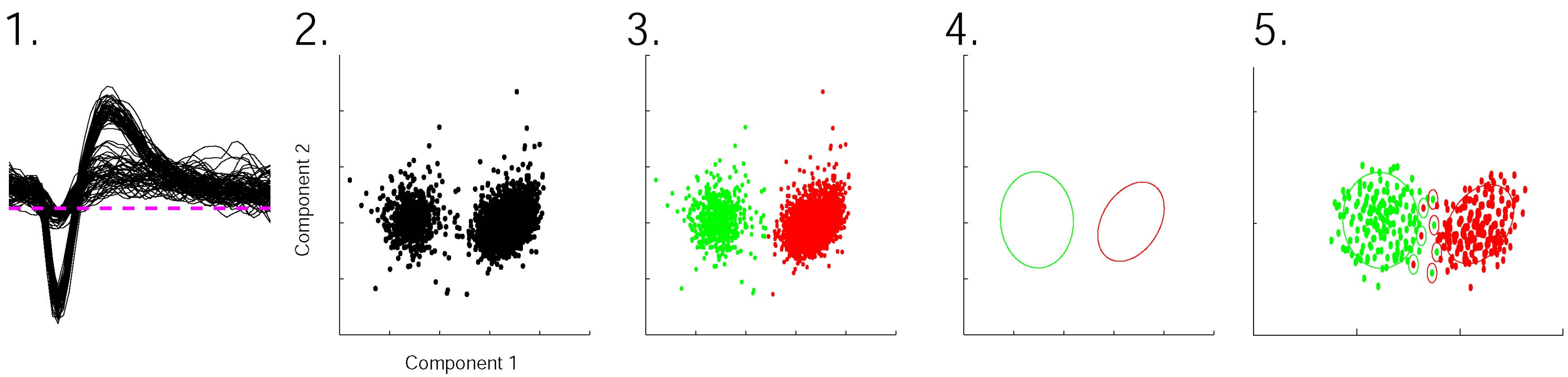




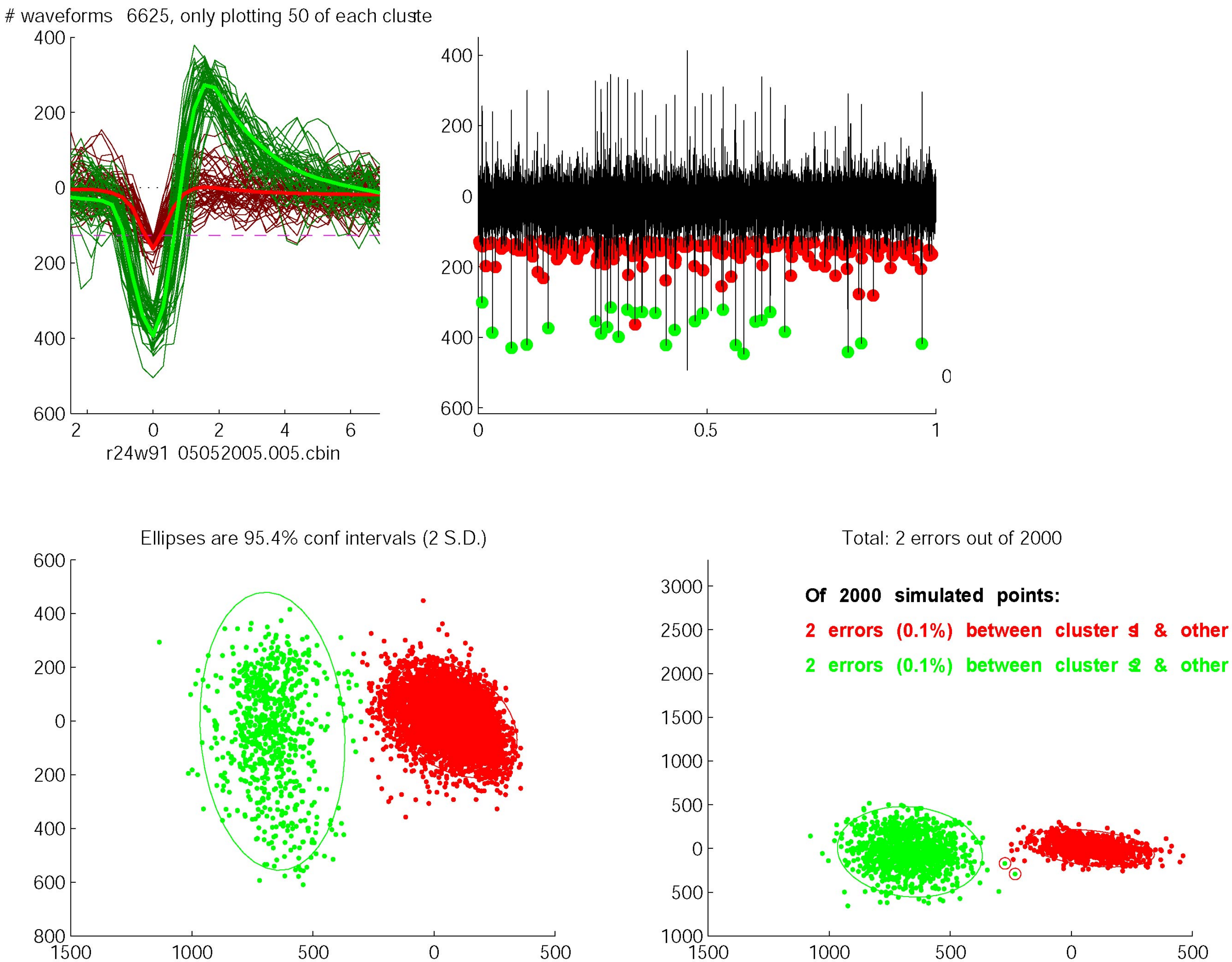


Summary of **S**am's **Ad**-hoc **U**nit **C**lassifier (S.A.U.C.Y.)

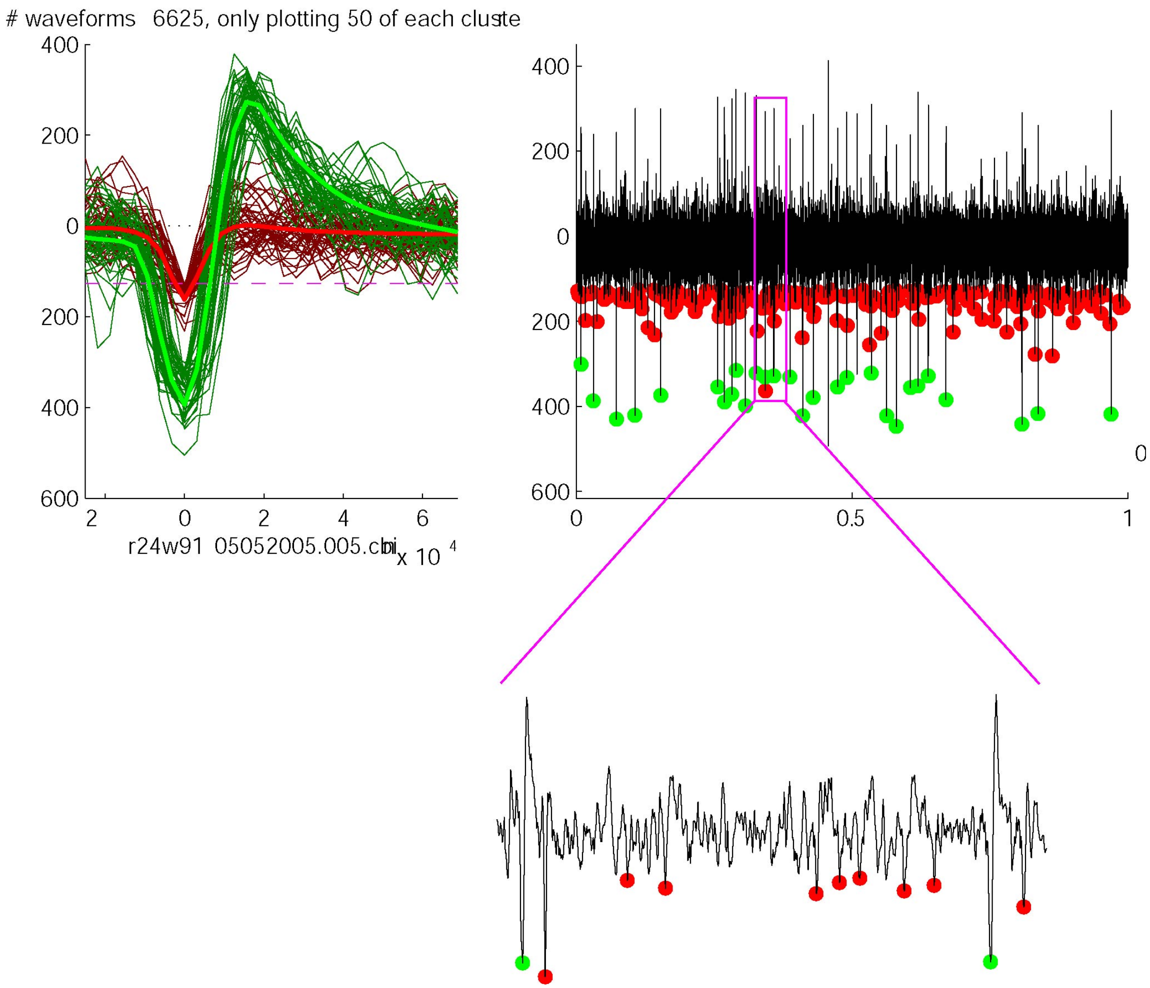
1. Set threshold to get waveforms
2. Run PCA
3. Use kmeans to cluster based on PC1+2
4. Find mean+var of clusters
5. Simulate 2D gaussians to estimate error rate.



VTA data (Ritu)



VTA data (Ritu)



(note classification based on shape, not just amplitude)

Applying PCA to behavioral analysis: example from birdsong

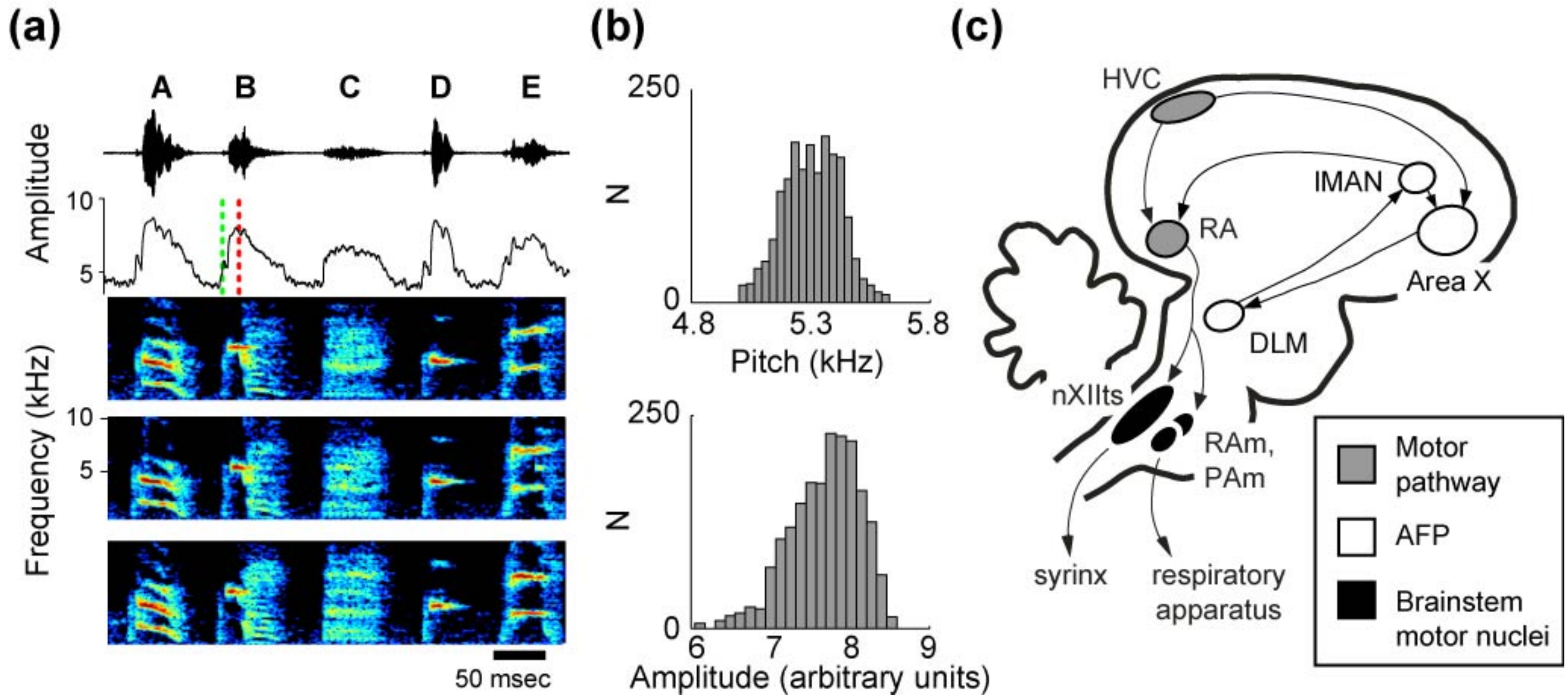
10370 • The Journal of Neuroscience, October 8, 2008 • 28(41):10370–10379

Behavioral/Systems/Cognitive

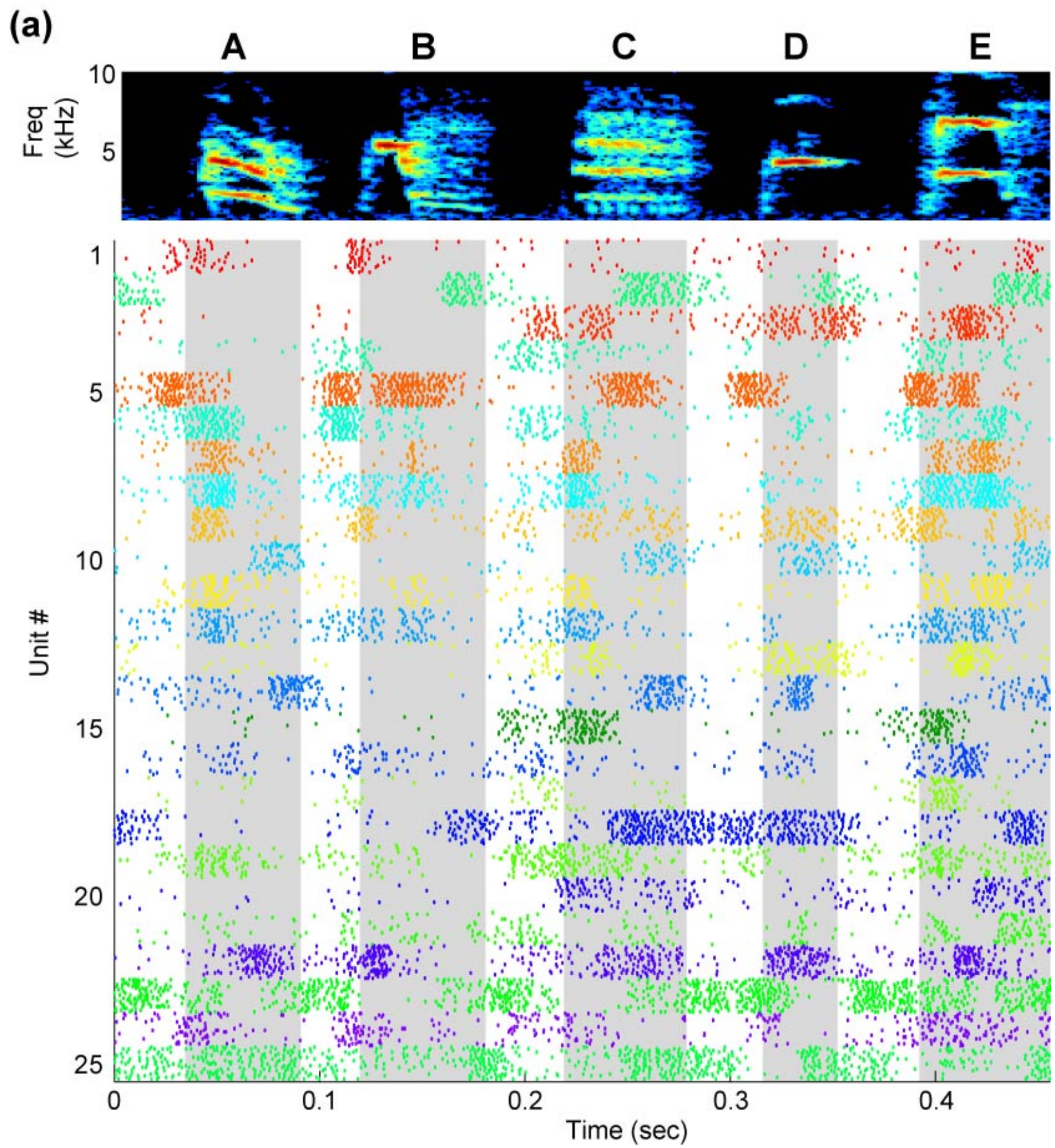
Central Contributions to Acoustic Variation in Birdsong

Samuel J. Sober,* Melville J. Wohlgemuth,* and Michael S. Brainard
Department of Physiology, W. M. Keck Center for Integrative Neuroscience, San Francisco, California 94143-0444

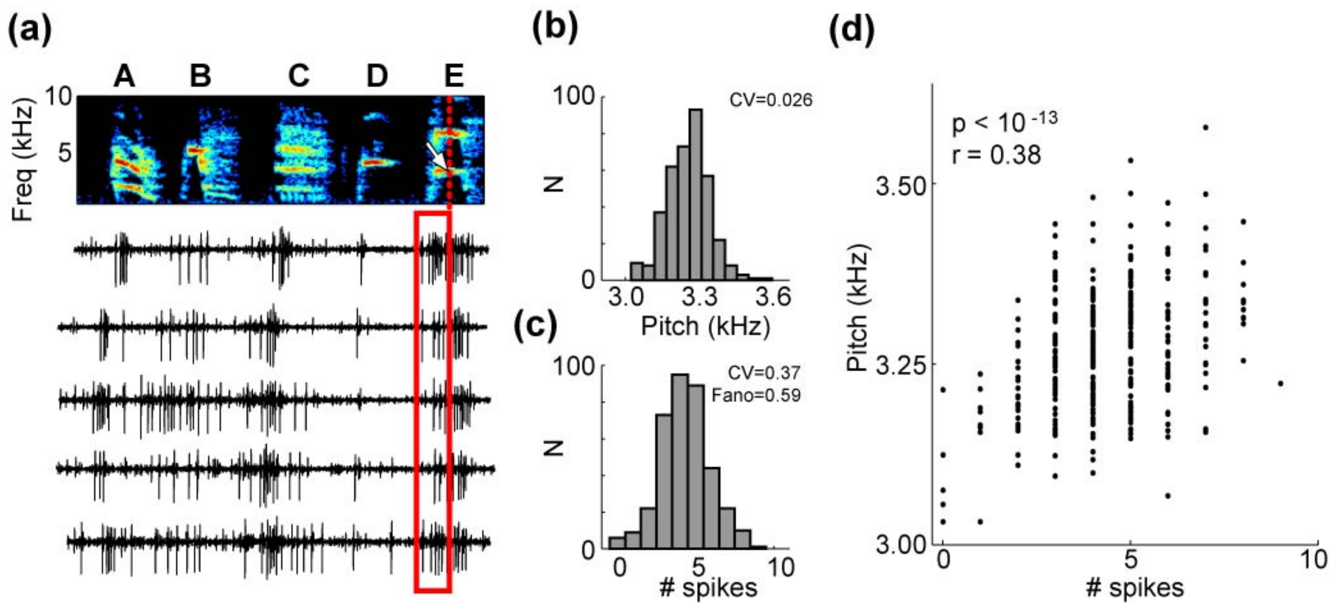
The question: how is acoustic variation encoded by RA?



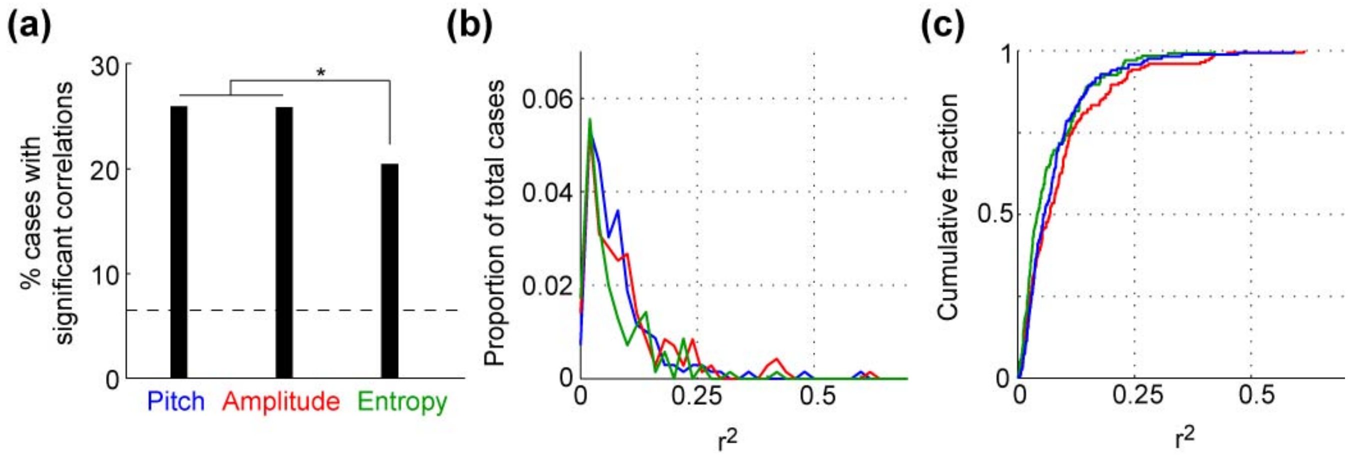
Record a bunch of neurons…



Look for correlations between activity and acoustic features…



…and find a bunch of them.



Why choose pitch, amplitude, and entropy?

- Refined over learning
- Functional anatomy of syrinx/respiratory system

The reviewer weighs in:

1. The three acoustic properties chosen for the analyses are insufficient for capturing the complexity of syllables. It is, thus, unclear whether the magnitude of the effect of RA response variation on syllable variation is quantified accurately.

To capture the complexity of songs, it is possible to break down the waveforms of a motif's syllables into a compact linear combination of (linearly) independent - complex - components (consider ICA, PCA, or wavelet analysis). **The trial-to-trial variation of the syllables (or motifs) can be represented as variations along the specified basis set. These variations can be correlated with RA response variation.**

The benefit of this method is twofold; (A) improvement of the accuracy and simplification of the paper's conclusions, (B) proper testing of the complex patterns for the contribution of neural response variation to the song variation in different syllables.

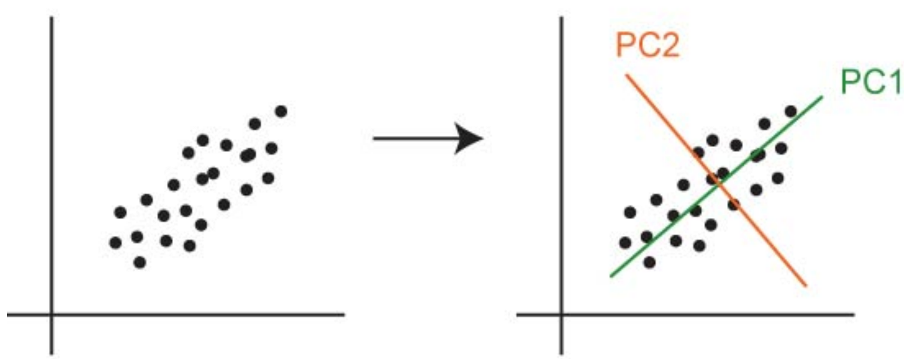
Our approach:

Use PCA as a (relatively) assumption-free tool to identify important dimensions of acoustic variation.

Describe song variation along these dimensions (princpal components) rather than as measured values of pitch, amplitude, or entropy.

Correlate RA activity with PCA-based measures of behavior.

Analyzing acoustic variation with PCA:



PCs are:
- Centered on mean
- Describe deviations from mean
- PC1 describes deviations along most-variable dimensions

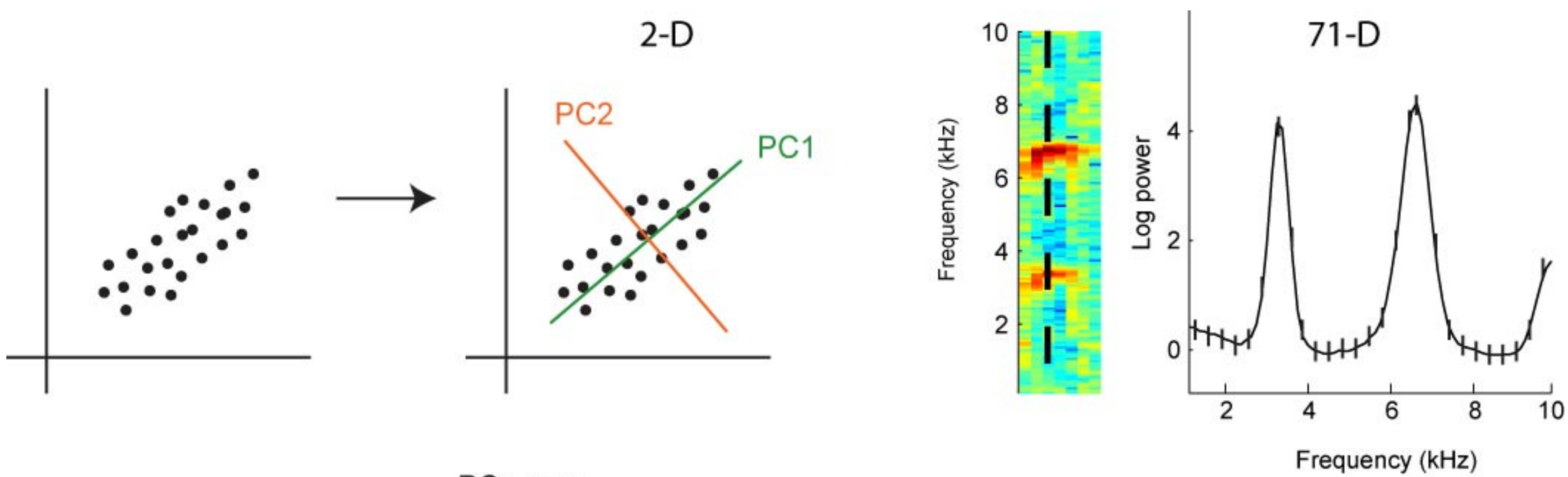
Our approach:

Use PCA as a (relatively) assumption-free tool to identify important dimensions of acoustic variation.

Describe song variation along these dimensions (princpal components) rather than as measured values of pitch, amplitude, or entropy.

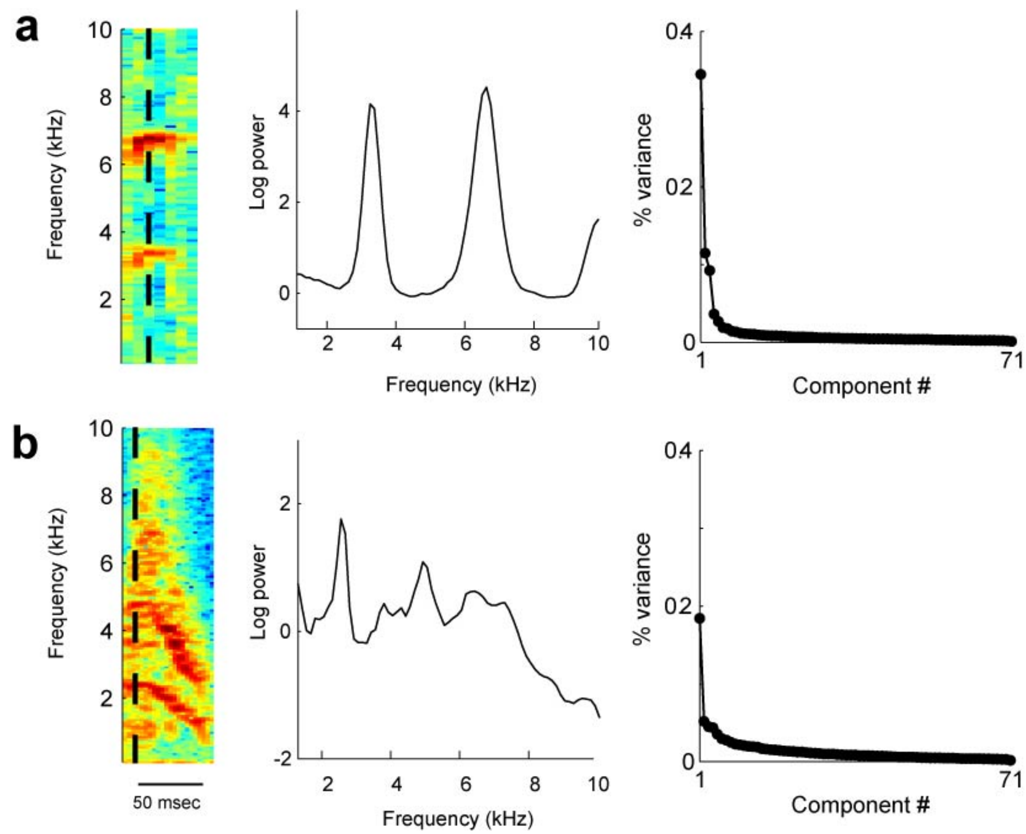
Correlate RA activity with PCA-based measures of behavior.

Analyzing acoustic variation with PCA:

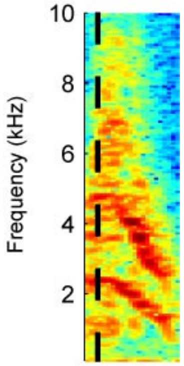


PCs are:
- Centered on mean
- Describe deviations from mean
- PC1 describes deviations along most-variable dimensions

PCA results:
"A few important dimensions of variation in each syllable"

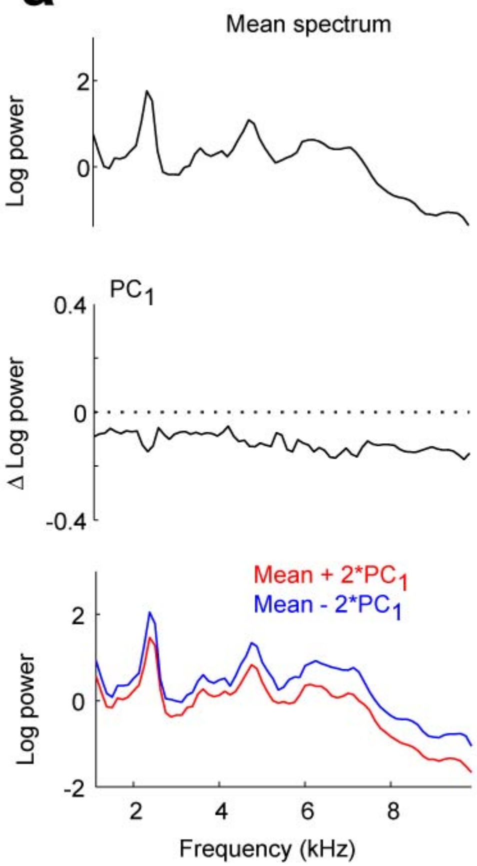


What do these components look like?

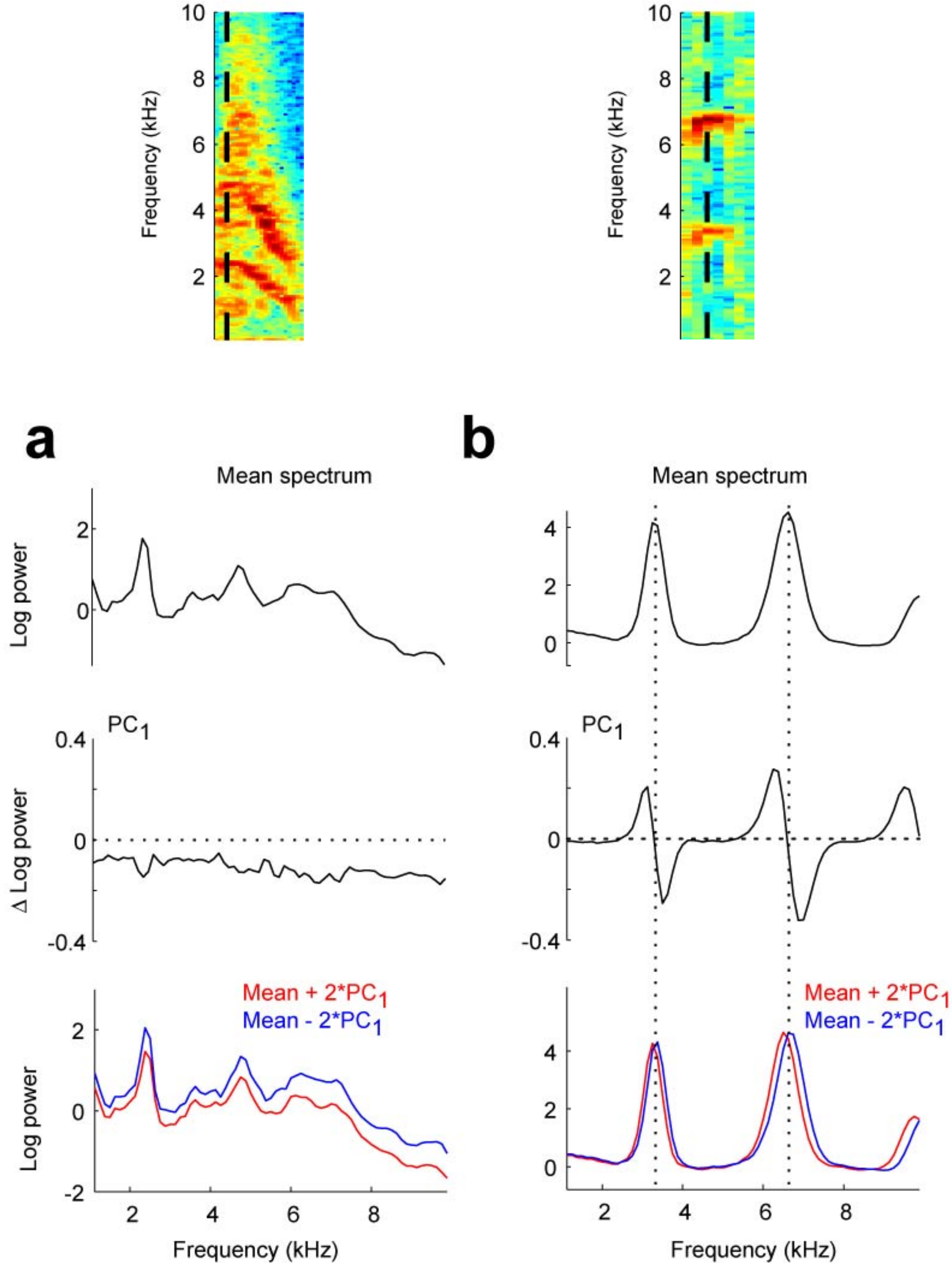


10
8
6
4
2
Frequency (kHz)

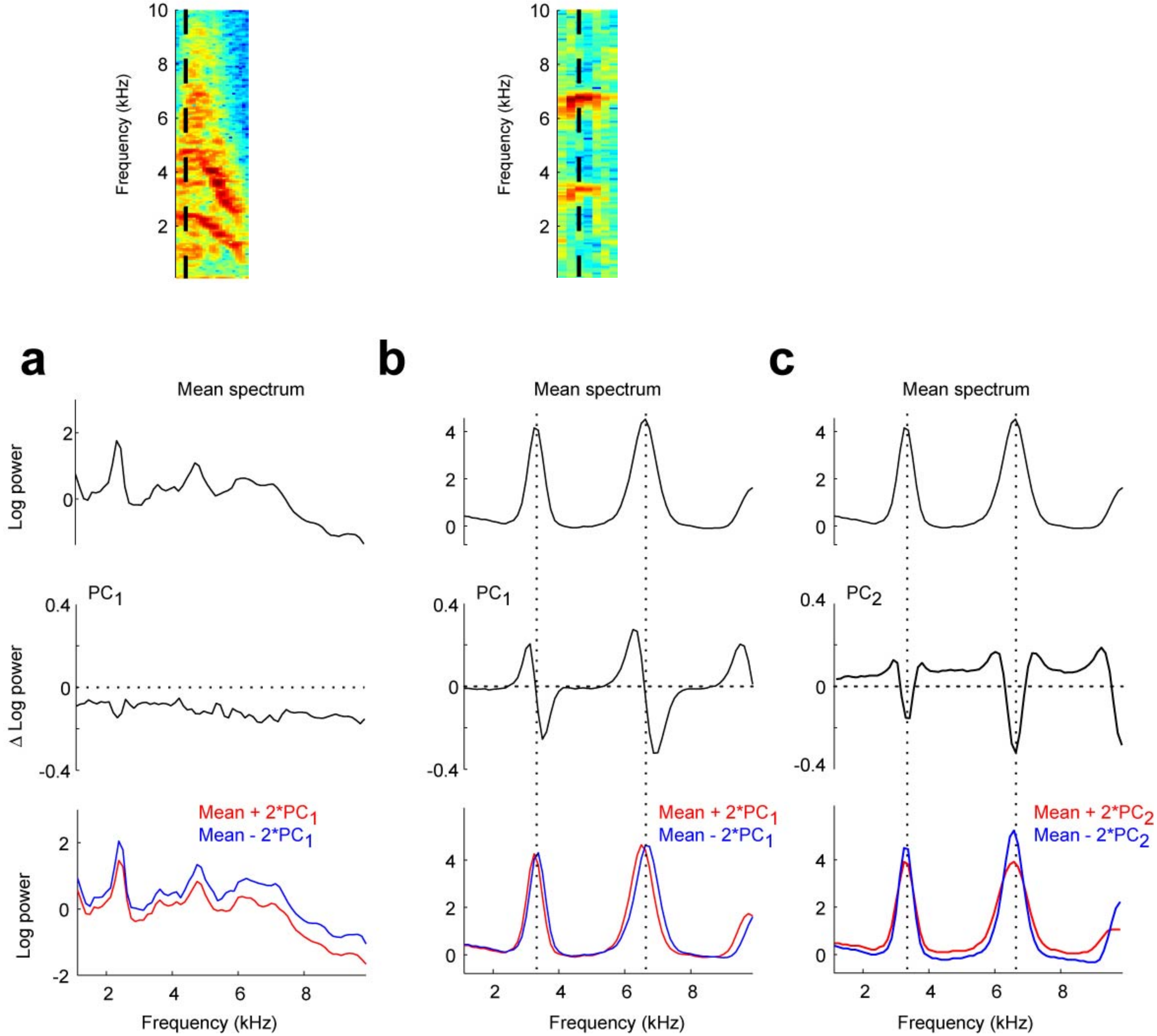
a



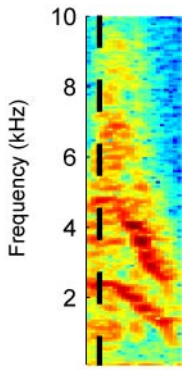
Mean spectrum
Log power
2
0
PC1
0.4
0
-0.4
Δ Log power
Mean + 2*PC1
Mean - 2*PC1
Log power
2
0
-2
2
4
6
8
Frequency (kHz)

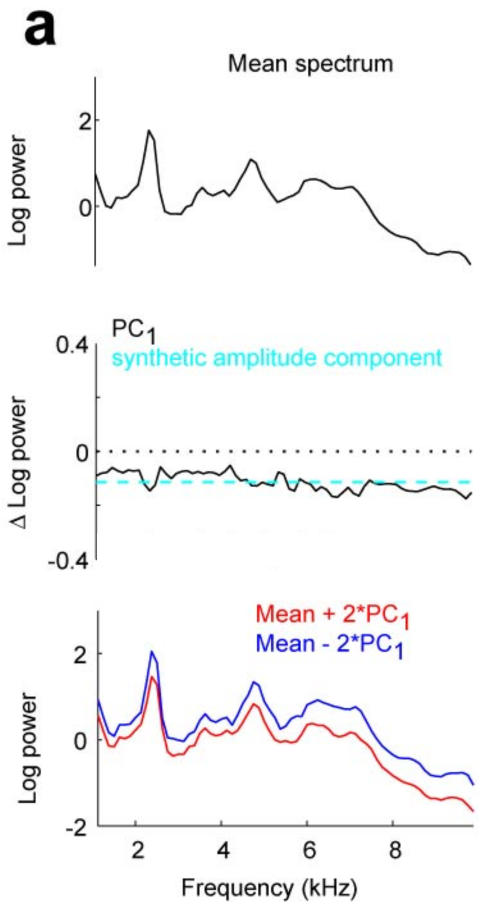


Frequency (kHz)
10
8
6
4
2
Frequency (kHz)
10
8
6
4
2
a
Mean spectrum
Log power
2
0
PC1
0.4
0
-0.4
Δ Log power
Mean + 2*PC1
Mean - 2*PC1
Log power
2
0
-2
2
4
6
8
Frequency (kHz)
b
Mean spectrum
4
2
0
PC1
0.4
0
-0.4
Mean + 2*PC1
Mean - 2*PC1
4
2
0
2
4
6
8
Frequency (kHz)

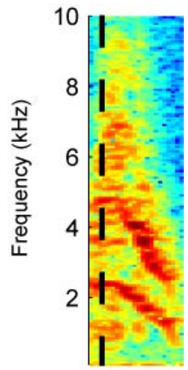


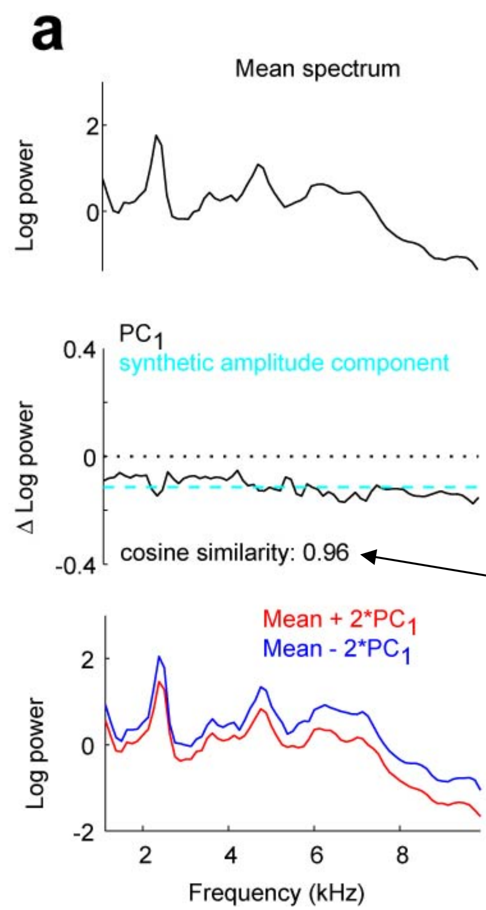
Frequency (kHz)
10
8
6
4
2
a
b
c
Mean spectrum
Log power
PC1
PC2
Δ Log power
0.4
0
-0.4
Mean + 2*PC1
Mean - 2*PC1
Mean + 2*PC2
Mean - 2*PC2
Frequency (kHz)





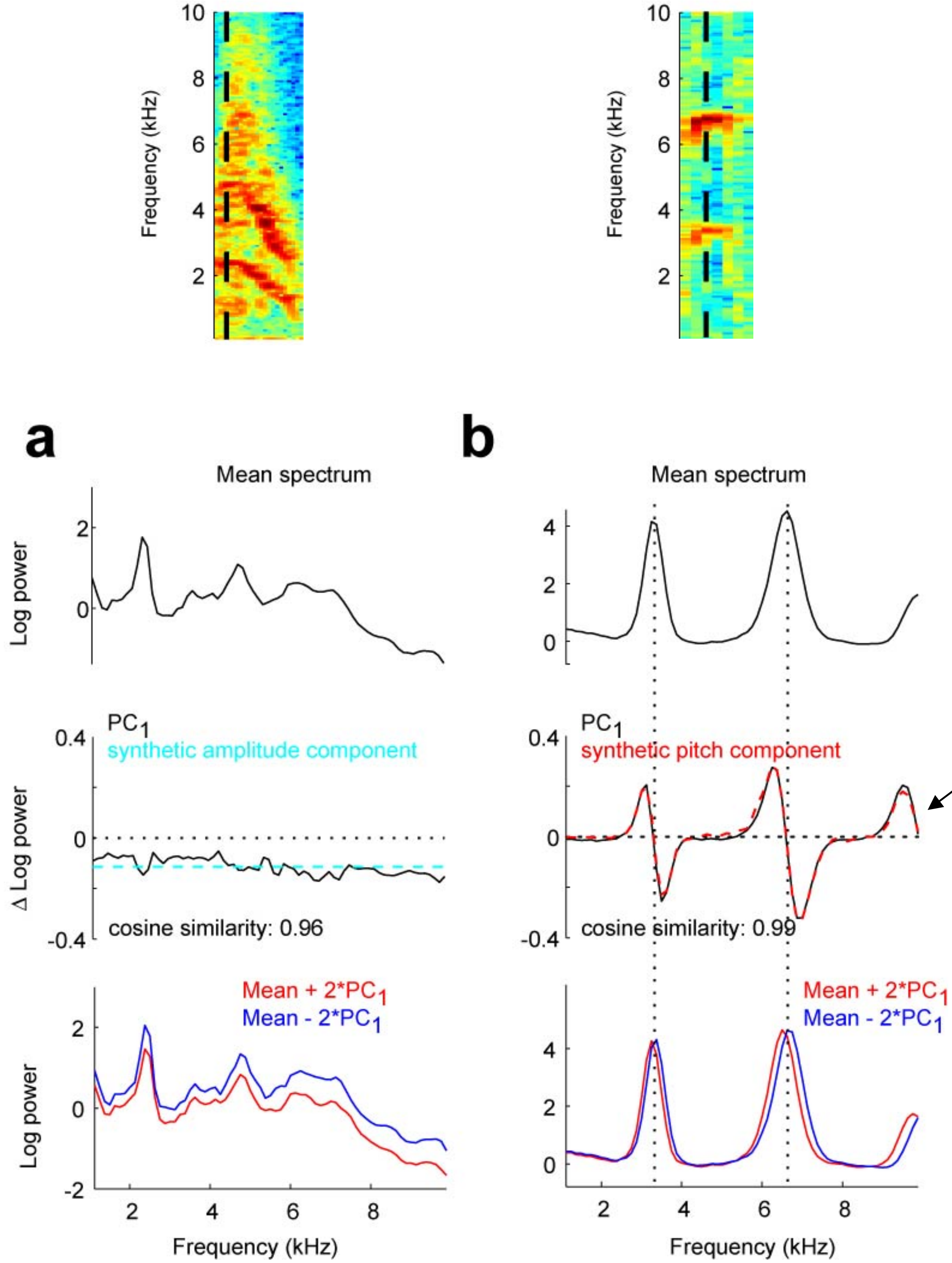
"Synthetic amplitude component" = Best-fit scalar offset



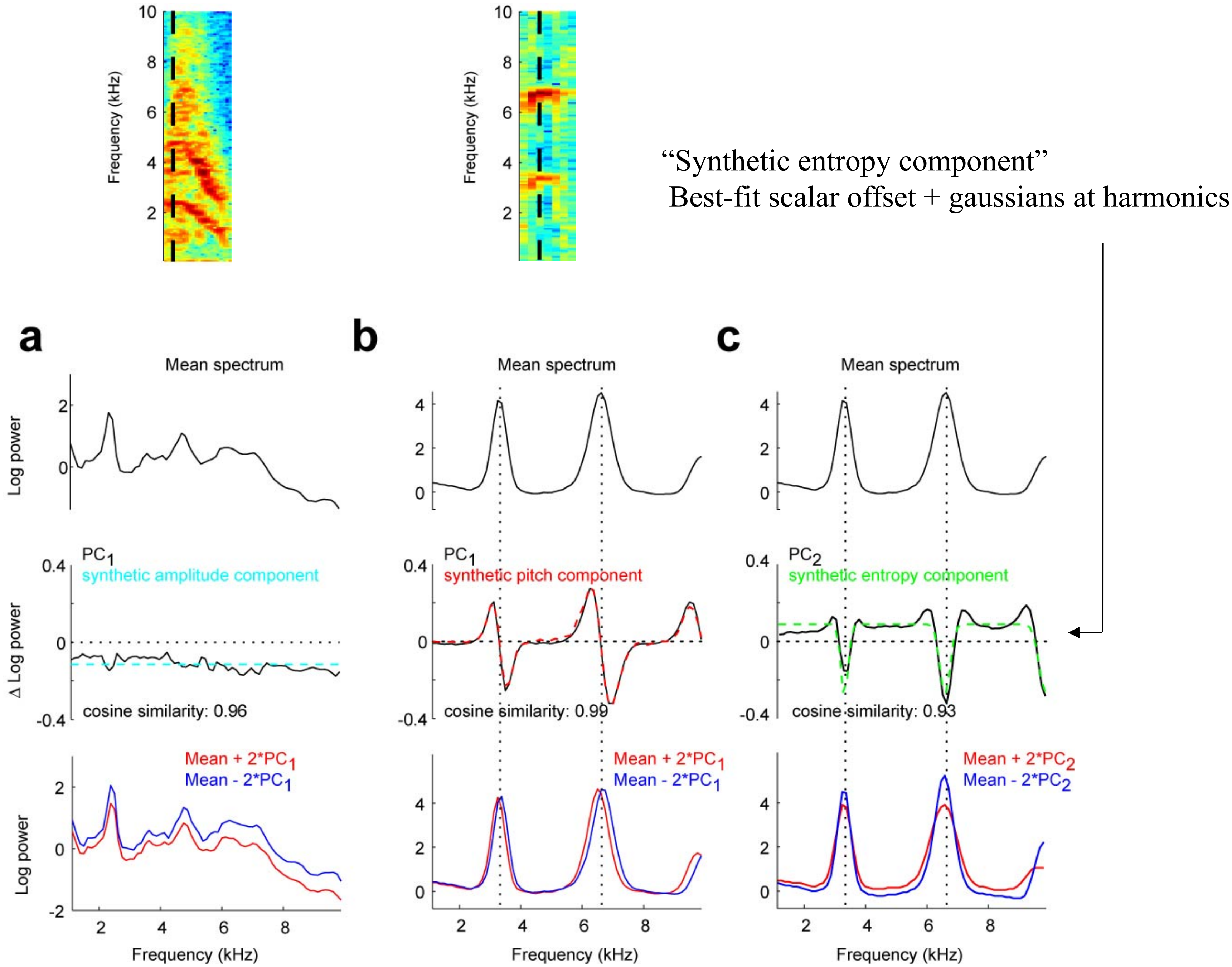


"Synthetic amplitude component" = Best-fit scalar offset

Quantify similarity of synthetic component and PC1

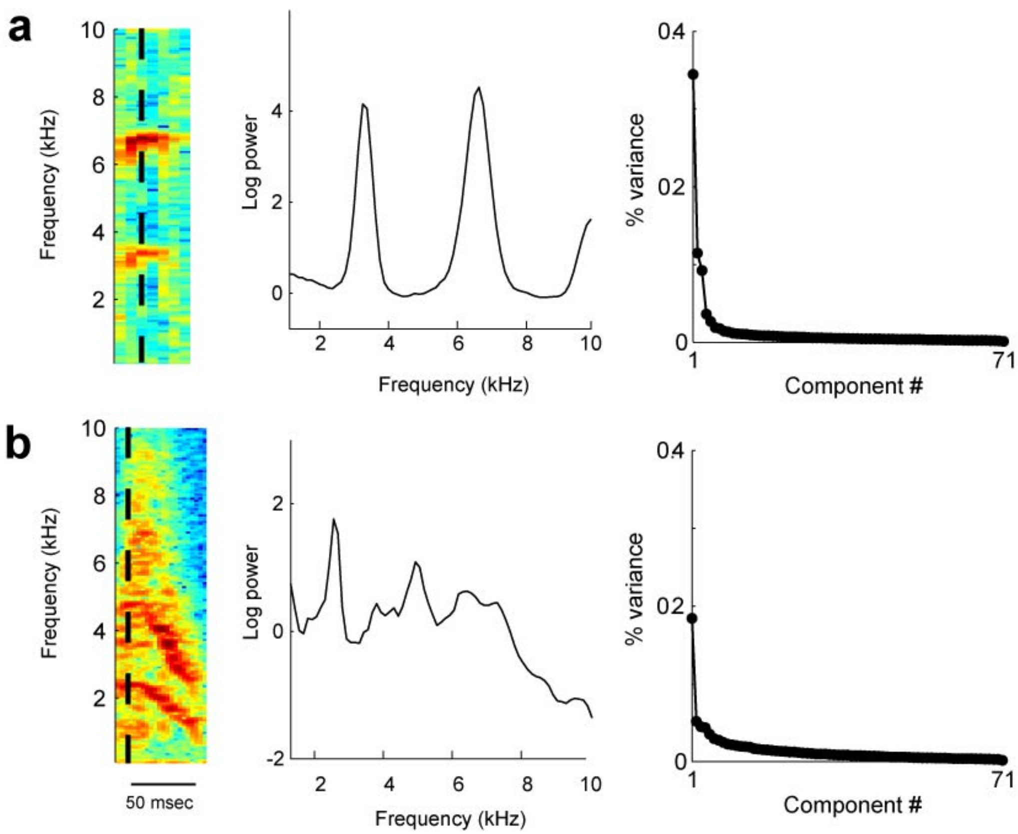


“Synthetic pitch component”
Best-fit shift+subtract



10
8
6
4
2
Frequency (kHz)
“Synthetic entropy component”
Best-fit scalar offset + gaussians at harmonics
a
b
c
Mean spectrum
Log power
PC1
synthetic amplitude component
synthetic pitch component
PC2
synthetic entropy component
Δ Log power
cosine similarity: 0.96
cosine similarity: 0.99
cosine similarity: 0.93
Mean + 2*PC1
Mean - 2*PC1
Mean + 2*PC2
Mean - 2*PC2
Frequency (kHz)

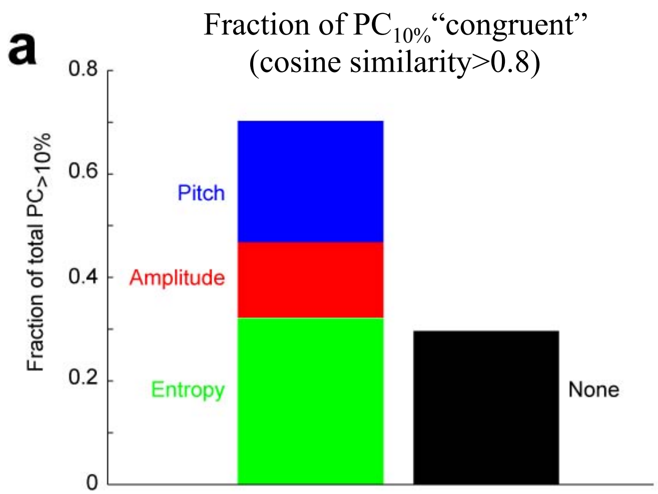
PCA results:
"A few important dimensions of variation in each syllable"



How to quantify which components are "important"?

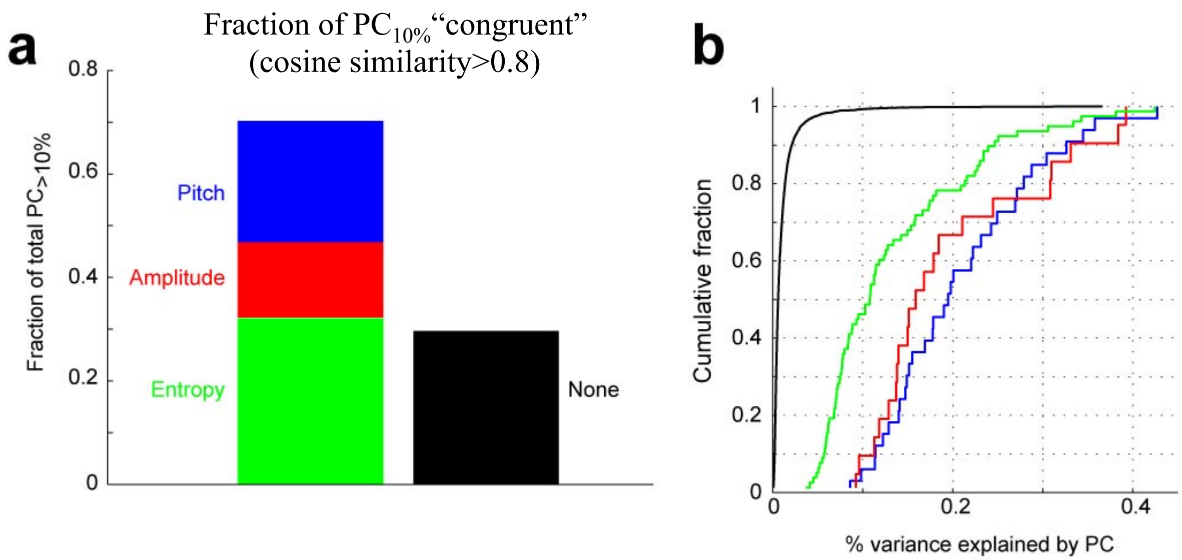
Define "important" dimensions as $PC_{10\%}$: each syllable has 1-3

Our response to the reviewer:



(a) Most important dimensions ($PC_{10\%}$) are congruent with pitch, amplitude, or entropy.

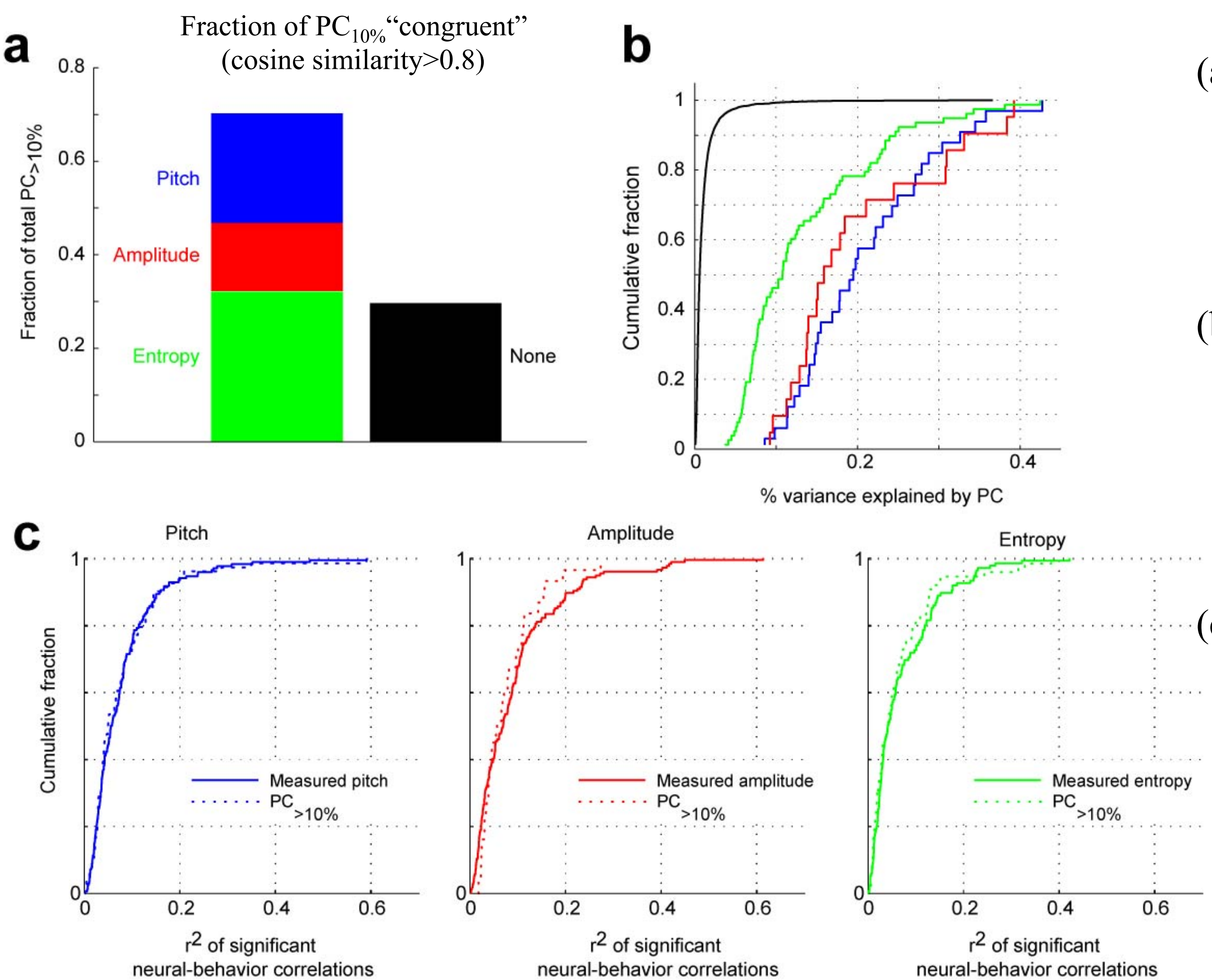
Our response to the reviewer:



(a) Most important dimensions ($PC_{10\%}$) are congruent with pitch, amplitude, or entropy.

(b) Dimensions that are congruent with pitch, amplitude, or entropy are more important than other dimensions.

Our response to the reviewer:

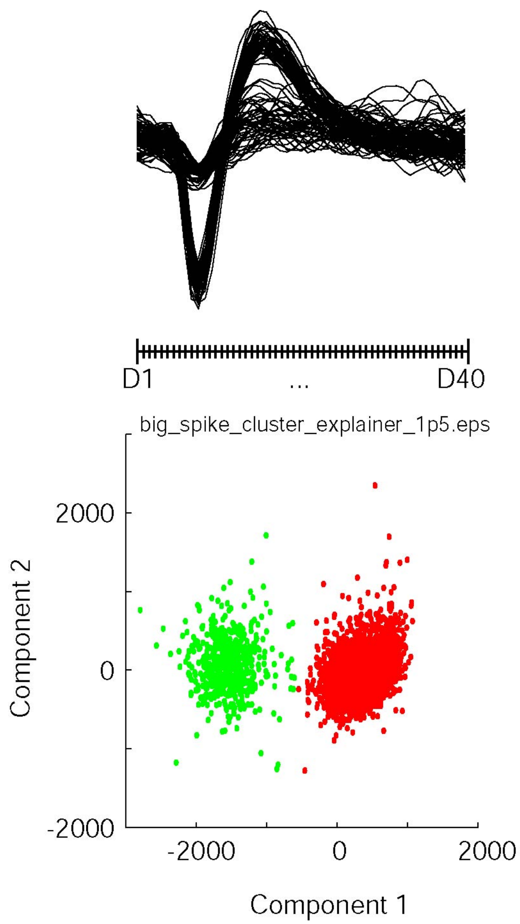


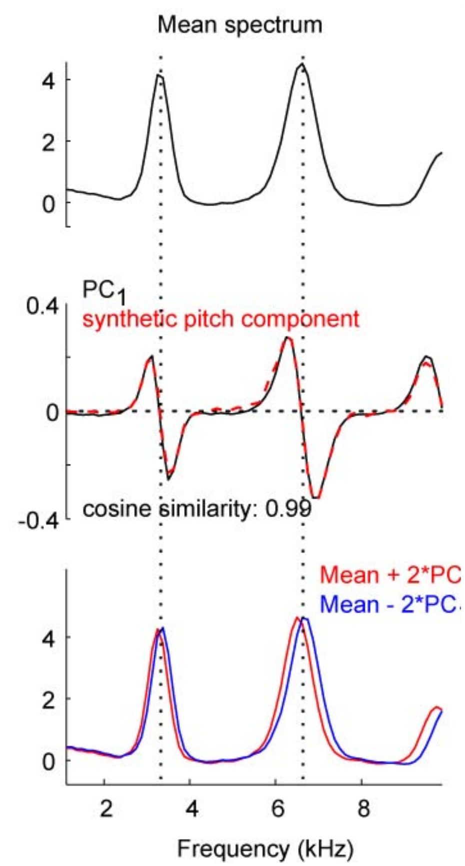
(a) Most important dimensions ($PC_{10\%}$) are congruent with pitch, amplitude, or entropy.

(b) Dimensions that are congruent with pitch, amplitude, or entropy are more important than other dimensions.

(c) Strength of neural-behavior correlations aren't different when behavior is described as PCs or as measured variations in p,a,e.

So:
PCA is a great tool for dimensionality reduction





for clustering... ... or identifying key features

Warning: PCA rests on some key assumptions

1. Assumes high SNR (larger variance = important dimension)
2. PCs are orthogonal

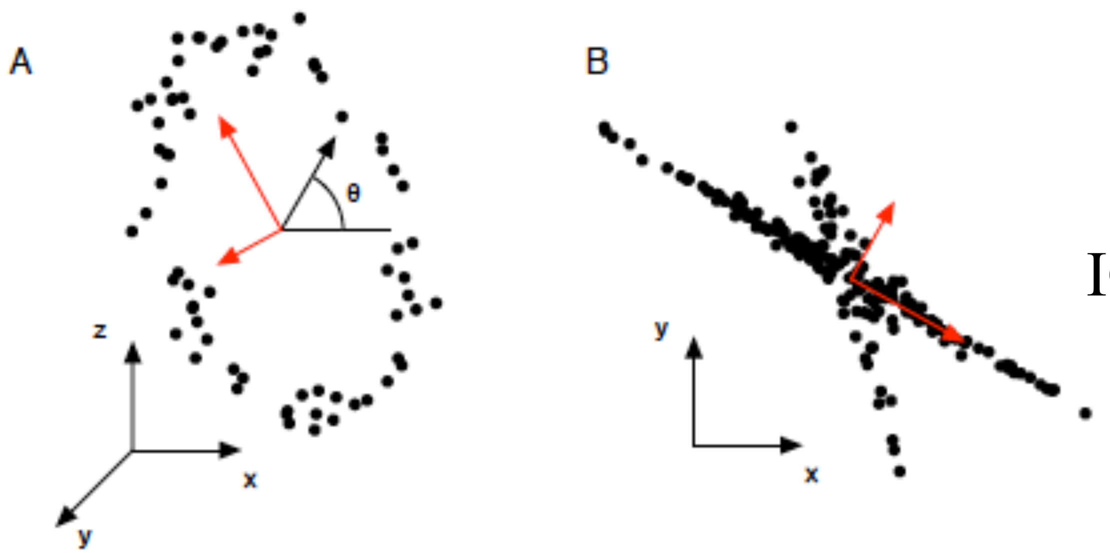


FIG. 6 Example of when PCA fails (red lines). (a) Tracking a person on a ferris wheel (black dots). All dynamics can be described by the phase of the wheel θ, a non-linear combination of the naive basis. (b) In this example data set, non-Gaussian distributed data and non-orthogonal axes causes PCA to fail. The axes with the largest variance do not correspond to the appropriate answer.

Other techniques:
ICA, non-negative matrix factorization, wavelet analysis