Lecture 13
Causal models

Jochen Braun

Otto-von-Guericke-Universität Magdeburg,
Cognitive Biology Group

Theoretical Neuroscience II, SS 2020

Credits:
Dayan & Abbott (2001) "Theoretical Neuroscience", Chapter 10

13. Causal models

How can a brain 'understand' its environment? What would such an 'understanding' involve? **Causal models** offer plausible answers. External causes (true, physical) produce sensations (observations). Brains try to explain observed sensations in terms of internal causes (hypothetical, mental). A causal model has two sides: **generative formulation** computes probability of observation (conditional, given one cause, and marginal, given all causes), **recognition formulation** computes likelihood of causes (conditional, given one observation, and marginal, given all observations). Causal models are improved iteratively, by considering the same observations again and again: **expectation step** assigns to observations probability of causes, **maximization step** recomputes the generative model for causes (mean, variance of associated observations). We illustrate iterative improvement of causal models for **K-means** and for **expectation-maximization** algorithm. Causal models move in a 'hermeneutic circle'!

Outline

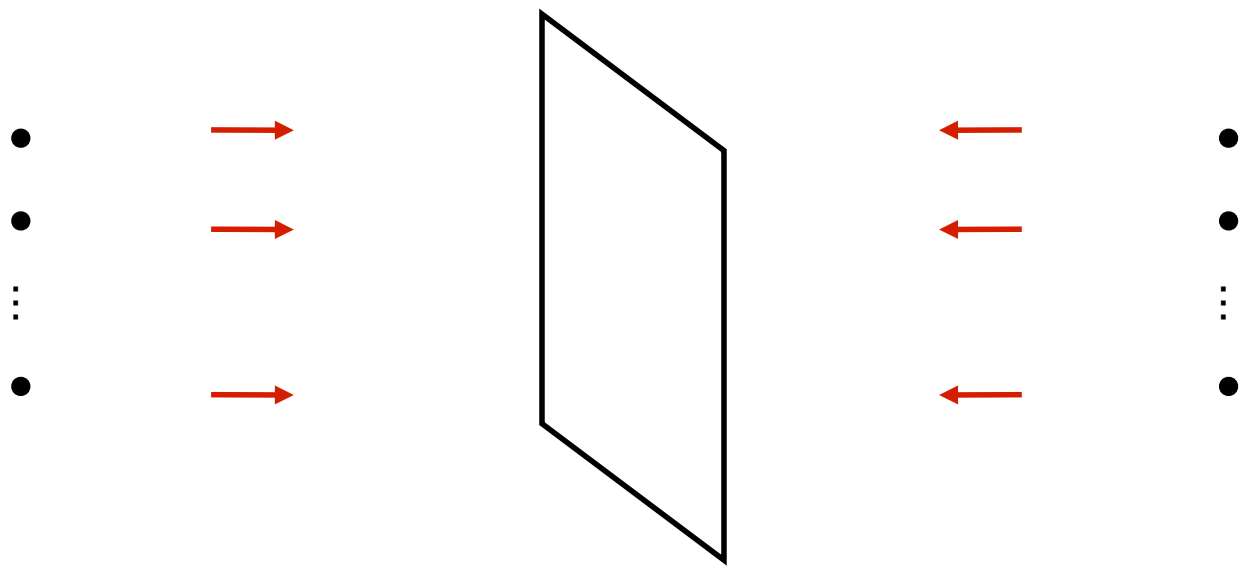
- 1 Introduction to causal models
- 2 Improving causal models
- 3 K-means
- 4 Expectation maximization (EM)
- 5 Continuous causes (optional)

1 Introduction to causal models

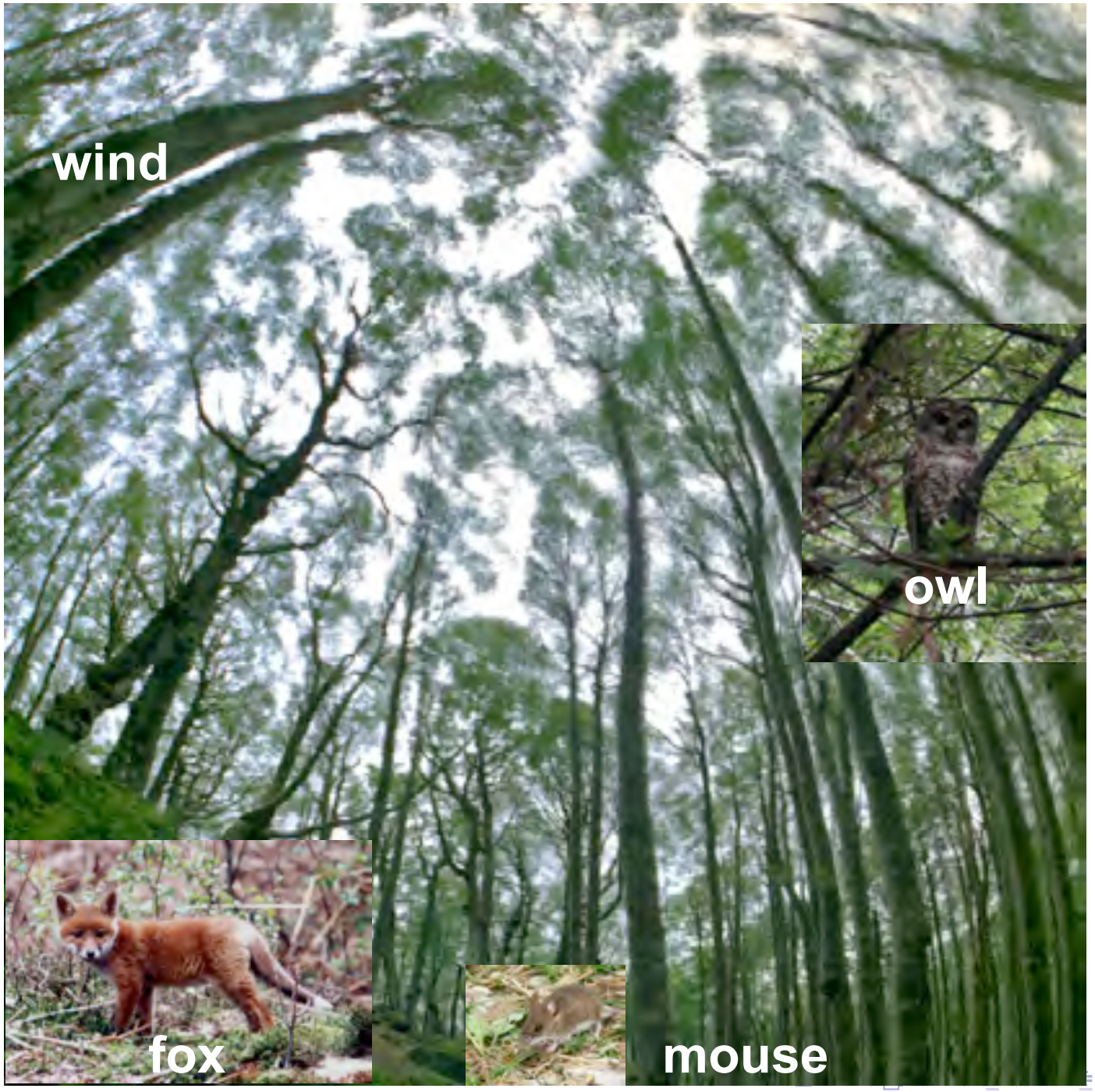
External world
(true causes)

Observation
(e.g., image)

Internal model
(model causes)



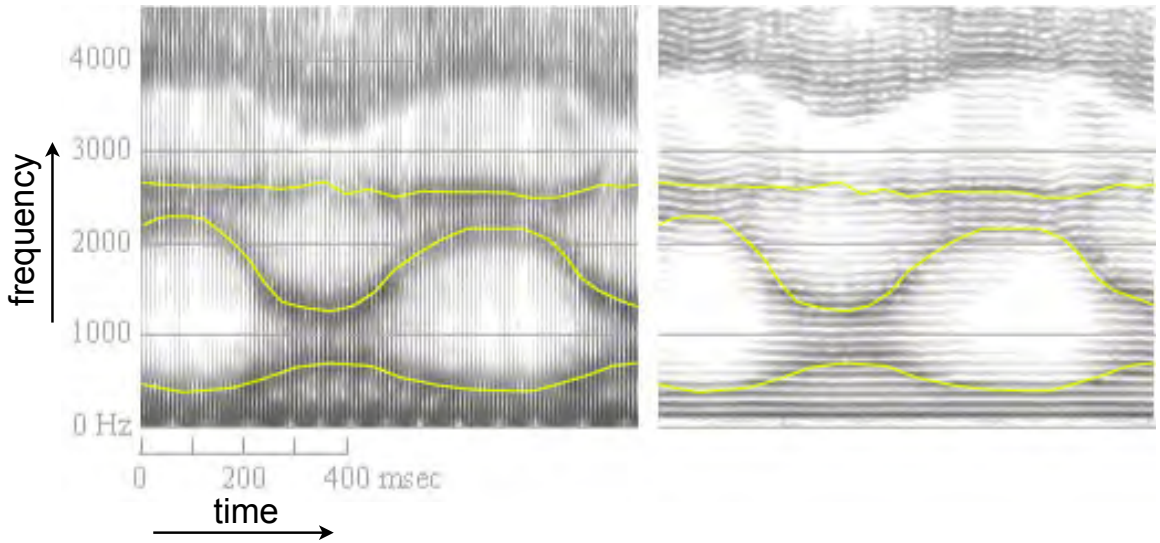
We wish to understand complex observations (sensations) in terms of few causes (ideally 'true' causes).


wind
owl
fox
mouse

Features of sound spectrogram

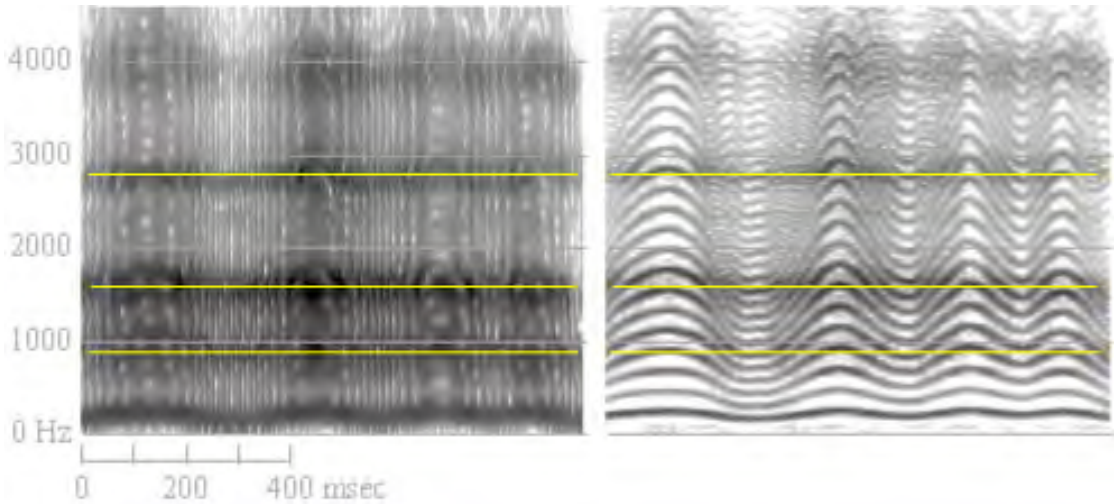
pitch constant

formant variable



formant constant

pitch variable

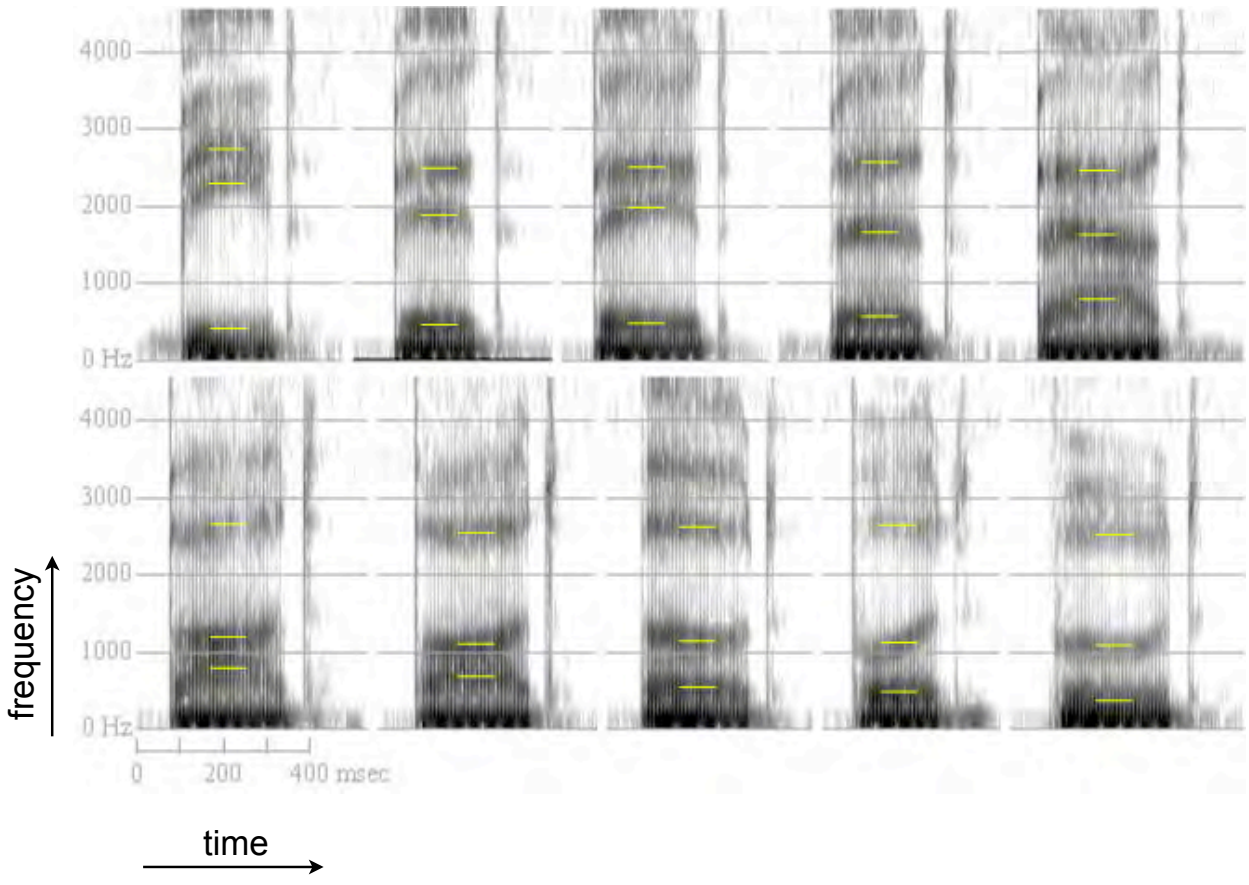


Pitch: fundamental frequency and overtones. Formant: shaping of amplitude.

Features of sound spectrogram, ctd

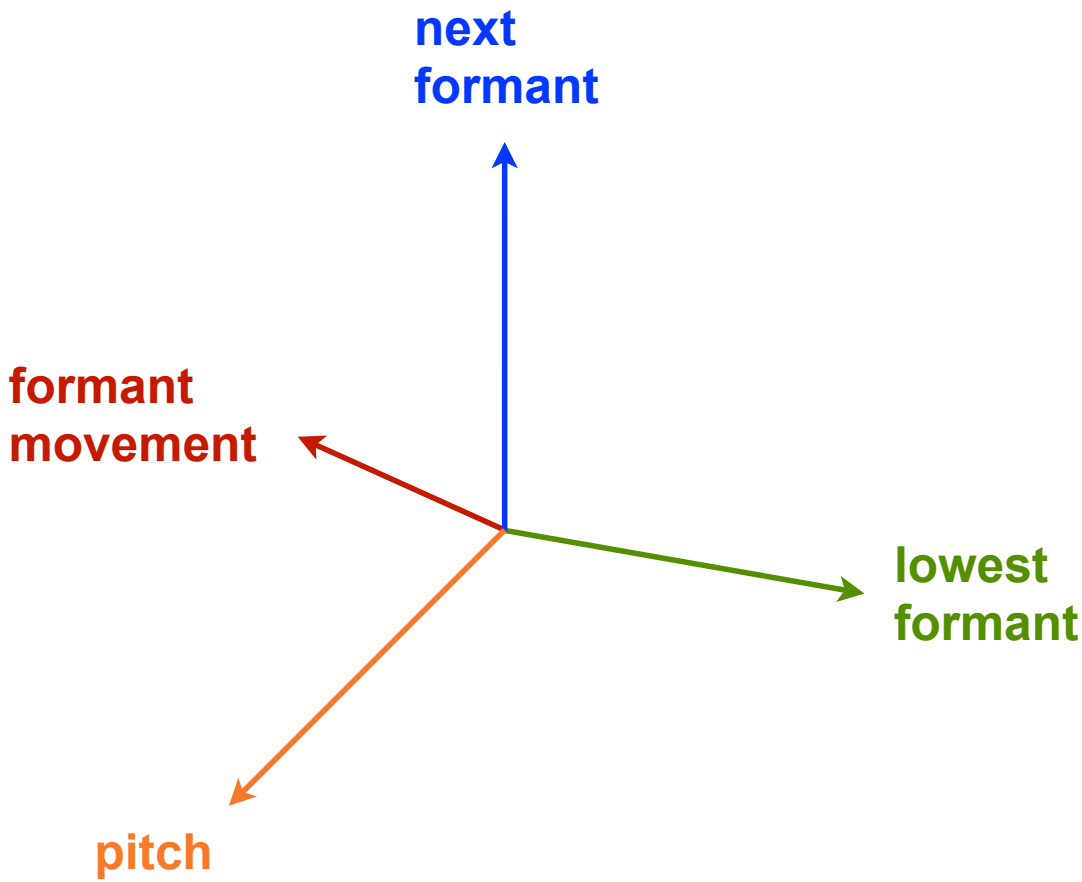
formant spacing

formant movement

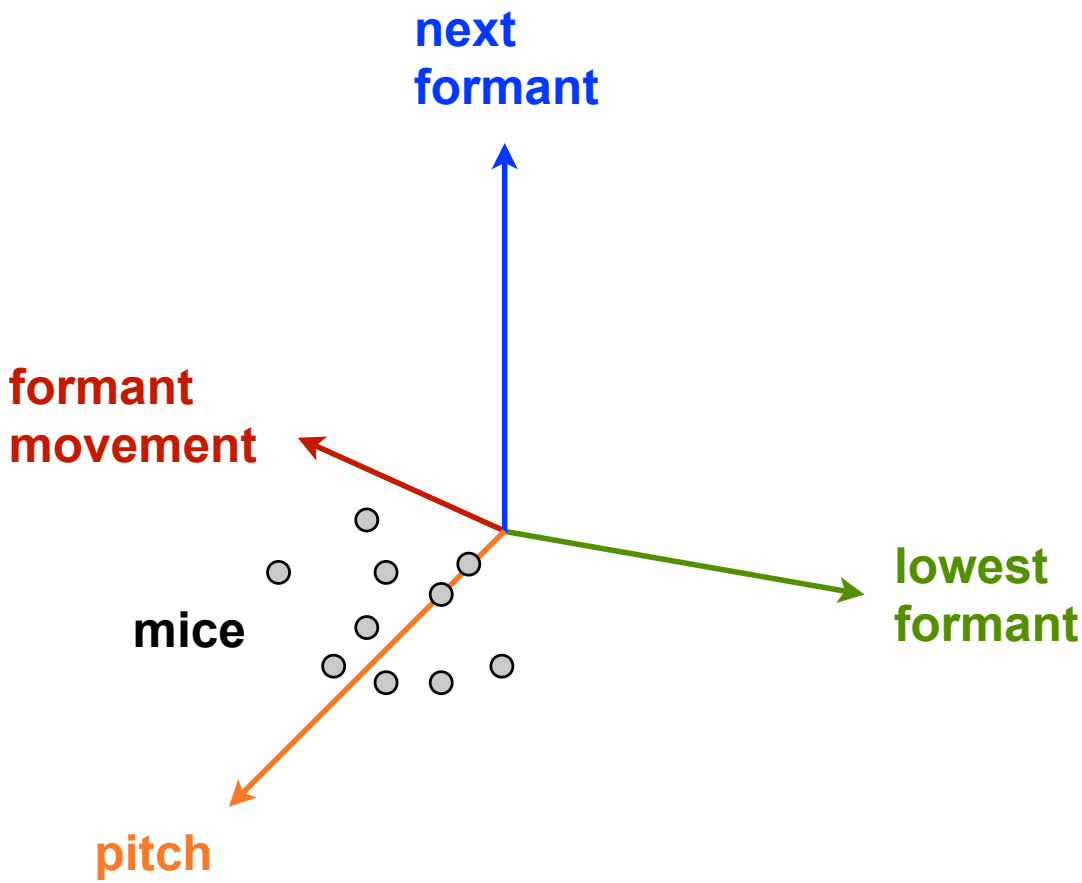


Pitch: fundamental frequency and overtones. Formant: shaping of amplitude.

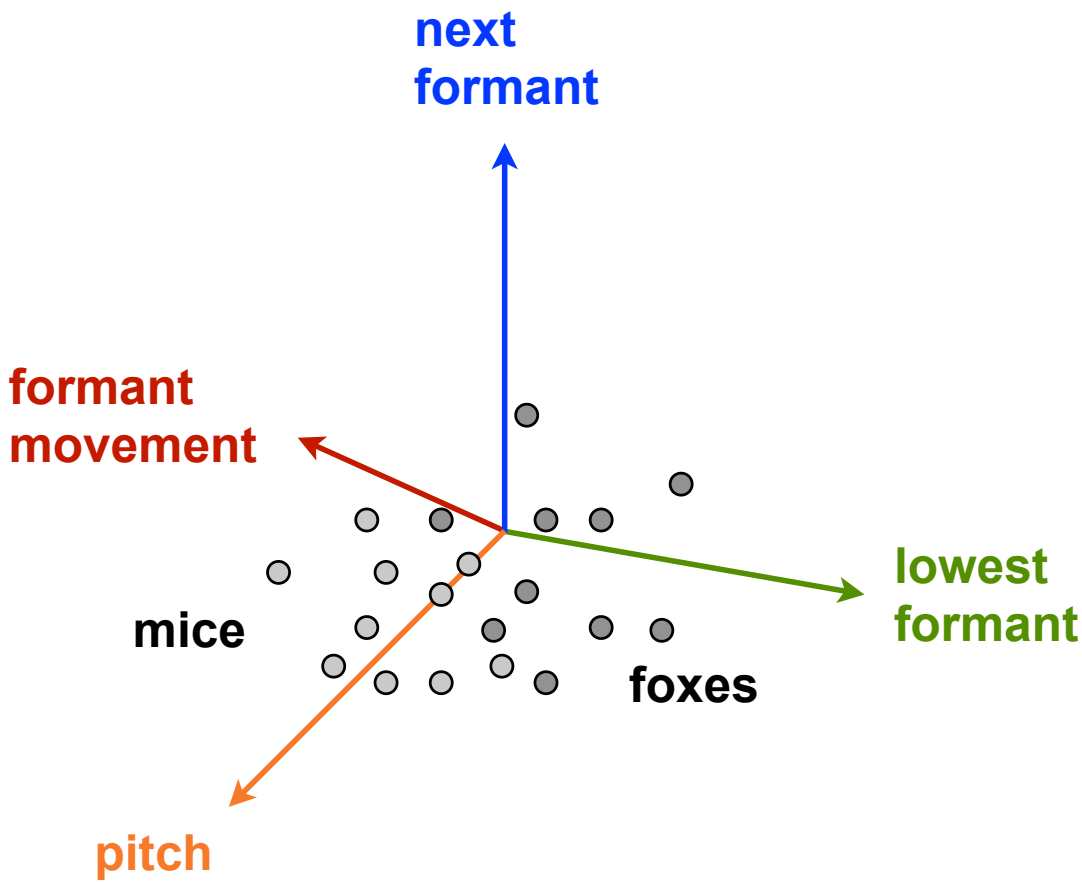
Auditory feature space (invented!)



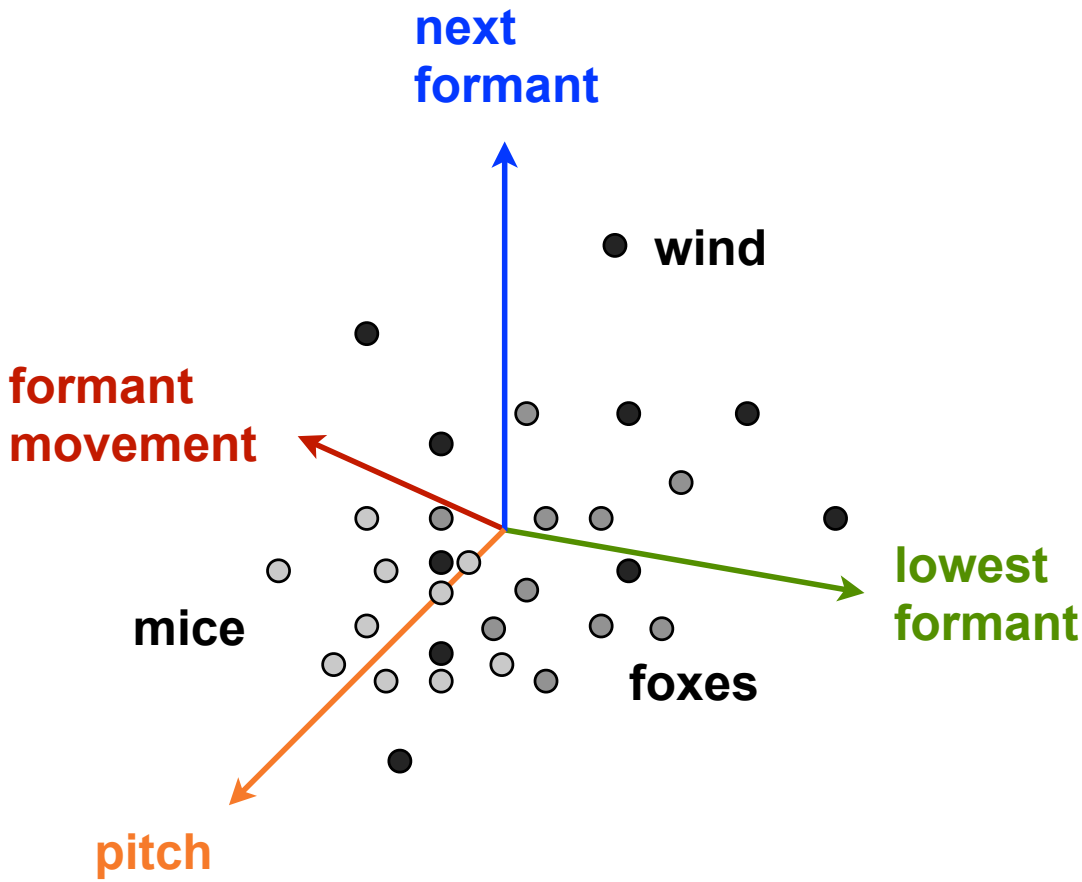
Auditory feature space



Auditory feature space

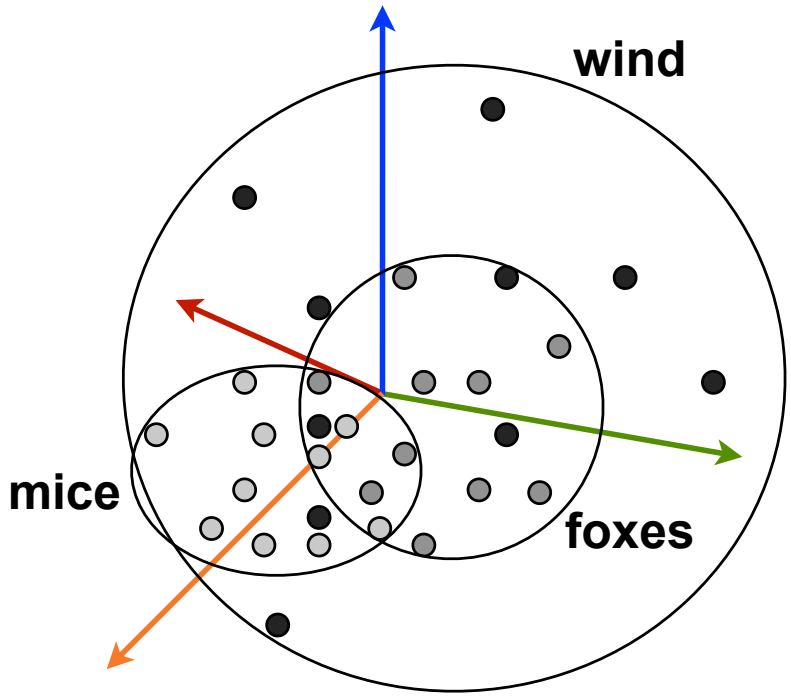


Auditory feature space



Different causes produce different sound features.

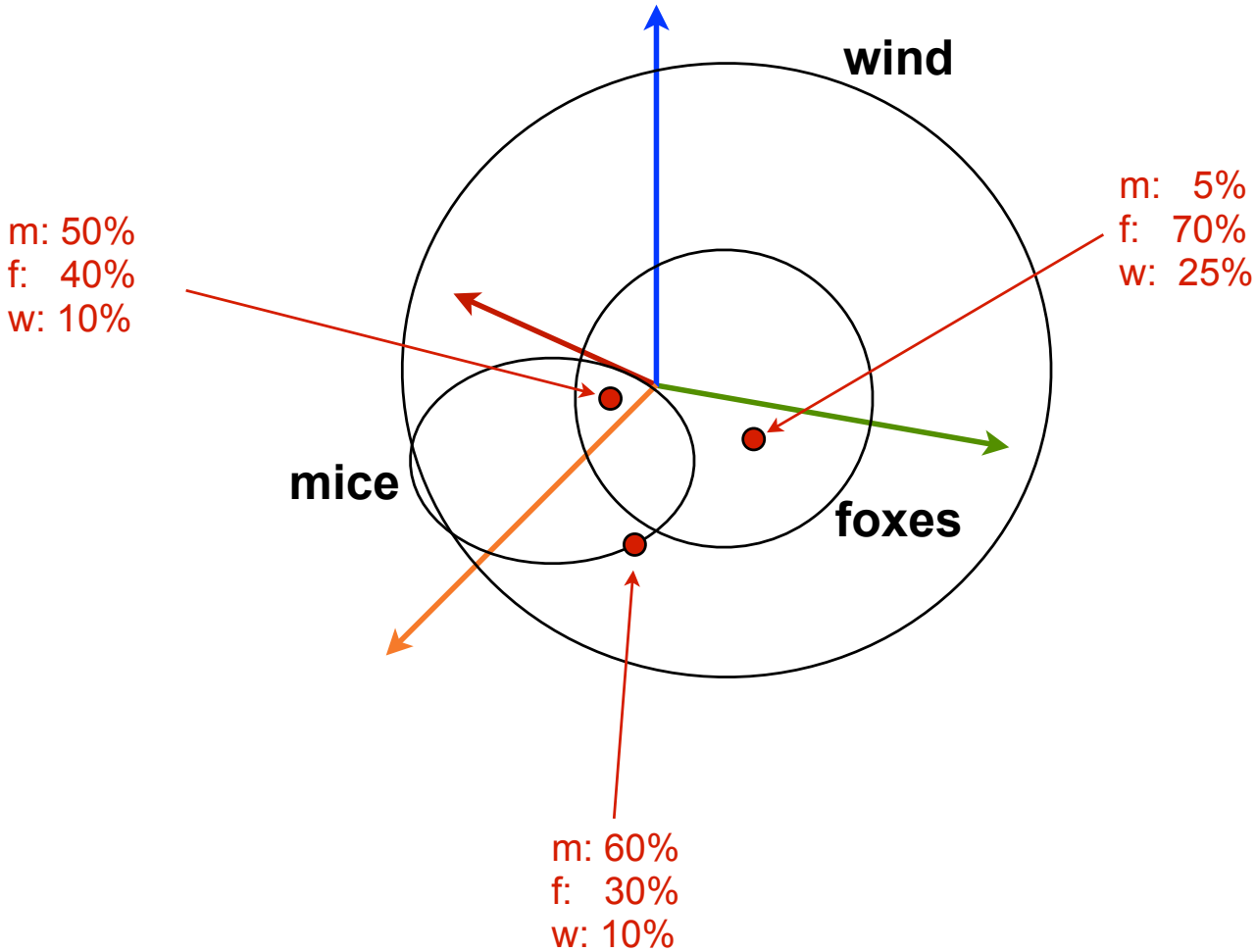
Auditory feature space



Statistically, the sound features produced by different causes are distributed differently! Ellipses indicate statistical distributions.

One stormy night ...

The owl perceives three distinct sounds. What might they be?



Exploit prior experience (statistical distributions) for 'probabilistic recognition'! Probability of cause, given observed features.

Generative formulation I:

Consider sensory events $\boldsymbol{u}$ (a vector) and their probability distribution $P(\boldsymbol{u})$. The structure of such distributions often suggests discrete underlying causes (*e.g.*, mice, foxes, wind).

A *causal description* G of such a sensory world involves two levels.

Firstly, a *causal variable* v, which takes discrete values (*e.g.*, M, F, or W) corresponding to distinct environmental causes, and the associated probabilities $P(v; G)$:

$$v \in \{M, F, W\}, \qquad P(v; G) \in \{\gamma_M, \gamma_F, \gamma_W\}$$

$$\gamma_M + \gamma_F + \gamma_W = 1$$

Generative formulation II:

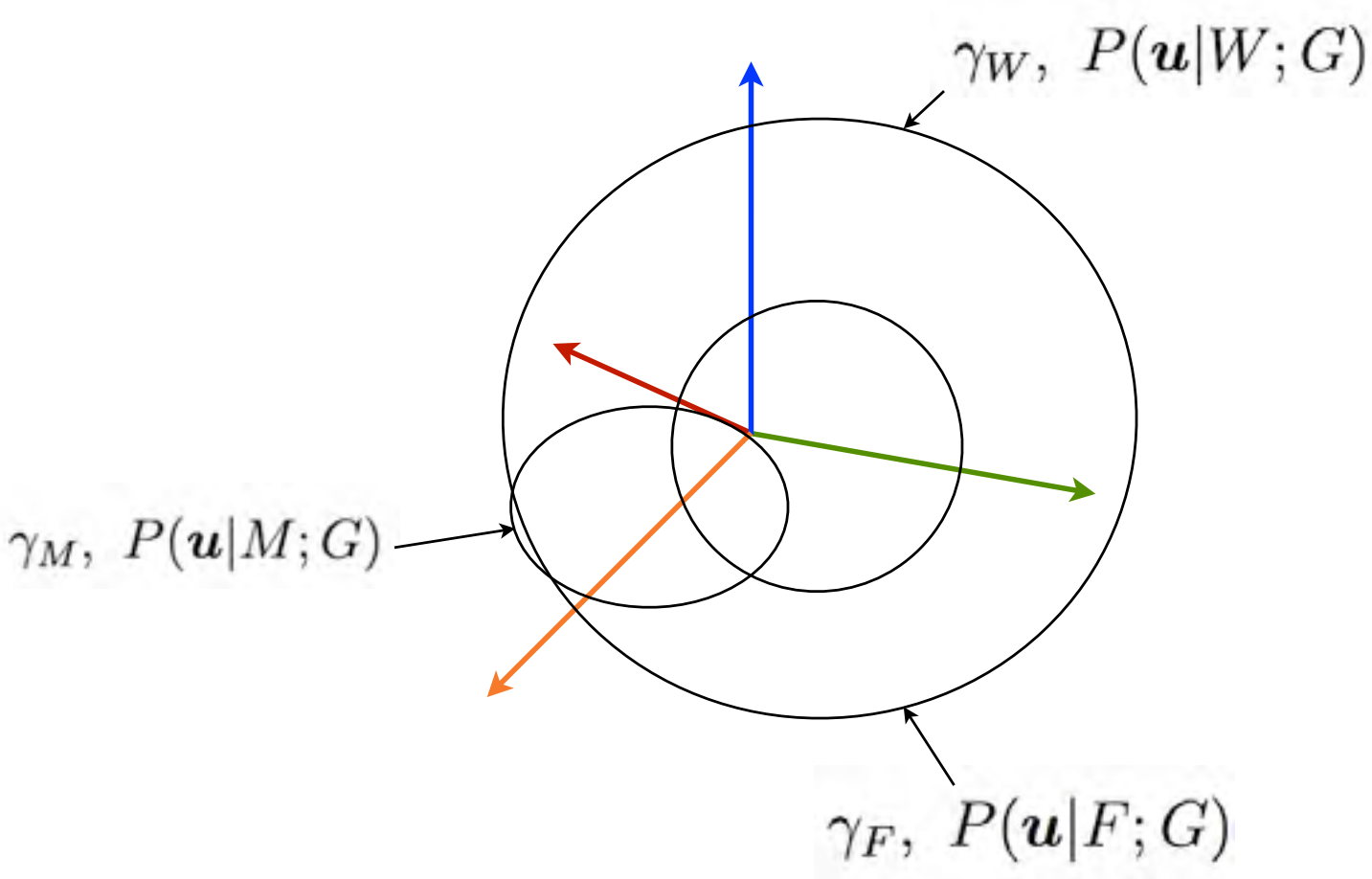
Secondly, *generative distributions* specifying the conditional probability of an observation $\boldsymbol{u}$ for each cause v. These distributions describe which sounds each cause is likely to make:

$$P(\boldsymbol{u}|M;G), \qquad P(\boldsymbol{u}|F;G), \qquad P(\boldsymbol{u}|W;G)$$

Typically, these generative distributions overlap. Thus, a given observation $\mathbf{u}$ may originate from one of several possible causes v. For this reason, the 'true' value of v cannot be determined unambiguously.

Generative formulation III:

Probability of causes γ_X and generative distributions $P(\boldsymbol{u}|X;G)$.



Probabilistic or deterministic recognition

In *probabilistic recognition*, we wish to know the probability that a given data point $\boldsymbol{u}$ was caused by cause A or by cause B. In *deterministic recognition*, we simply need to know the most likely cause, A or B.

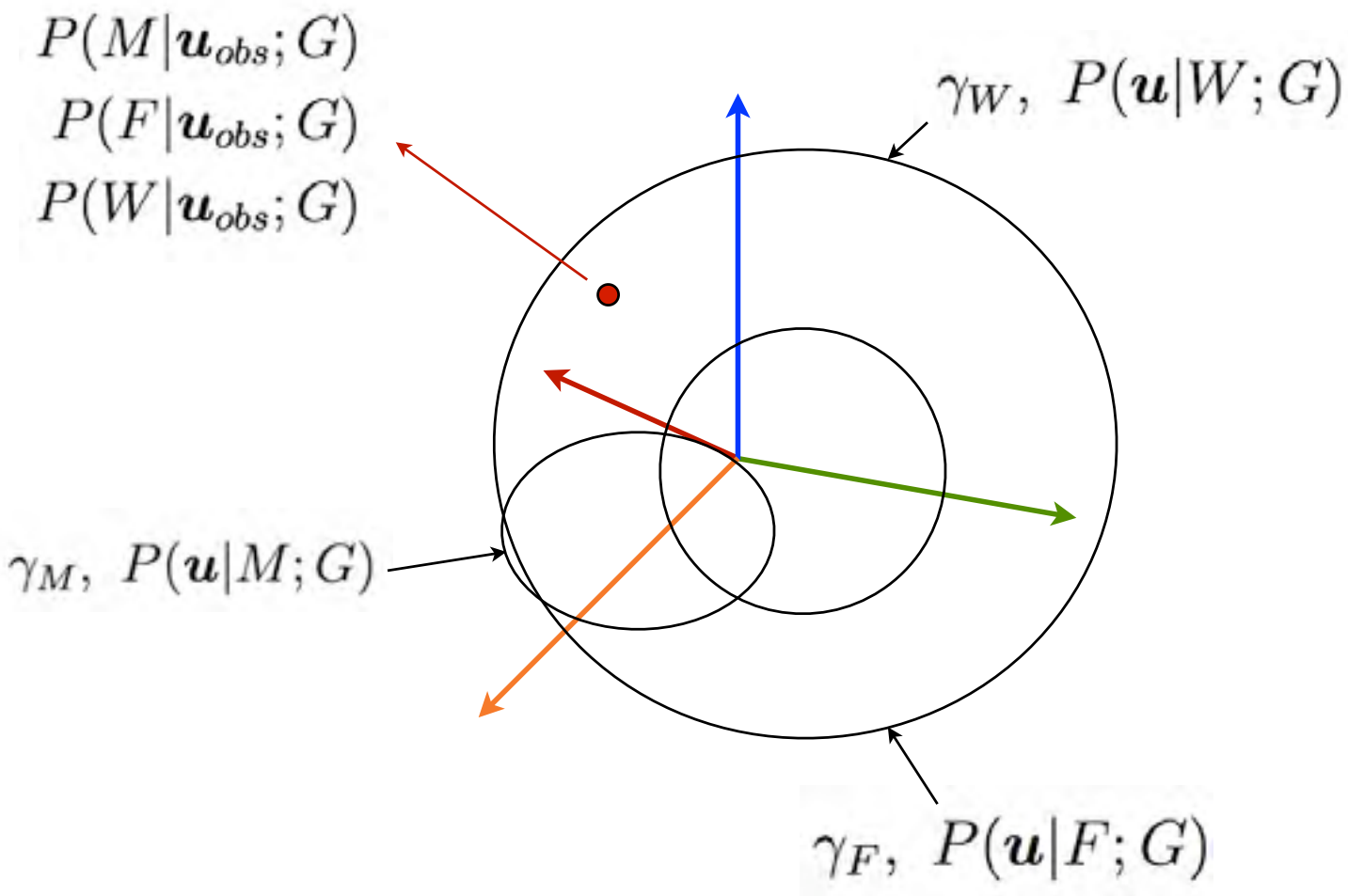
$\boldsymbol{u}_{obs}$ $\qquad P(M|\boldsymbol{u}_{obs}; G), \qquad P(F|\boldsymbol{u}_{obs}; G), \qquad P(W|\boldsymbol{u}_{obs}; G)$

For both purposes, we need a *recognition model*:

$$P(v|\boldsymbol{u}; G) = \frac{p(\boldsymbol{u}|v; G)\, P(v; G)}{p(\boldsymbol{u}; G)}$$

Such a model is obtained by applying Bayes' rule to our *generative model* G.

Recognition formulation



How likely did causes M, F, W produce a given observation $\mathbf{u}_{obs}$?

Points to note

- Causal models G describe a sensory scene compactly in terms of causes X.
- Generative formulation proceeds from hypothetical causes X to predicted observations $\boldsymbol{u}$:

$$P(\boldsymbol{u}|X;G)$$

- Inverting a generative formulation with Bayes' rule yields a recognition formulation.
- Recognition formulation proceeds from actual observations $\boldsymbol{u}_{obs}$ to inferred causes X.

$$P(X|\boldsymbol{u};G)$$

2 Improving causal models

To develop a causal model G, we need enough observations to reveal their statistical structure. We judge the quality of a causal model by its ability to reproduce (and thus 'explain') this structure.

The structure of a causal model reflects heuristic information (prejudices, assumptions, analogies, ...).

Given such a structure, we consider the **same observations** again and again, to iteratively adjust the model and improve the match between observation and reproduction.

Data vectors	$\boldsymbol{u} = (u_1, u_2, \ldots)$
Scalar causes	$v \in \{A, B\}$
Generative formulation	$P(\boldsymbol{u}\|v; G)$
Recognition formulation	$P(v\|\boldsymbol{u}; G)$

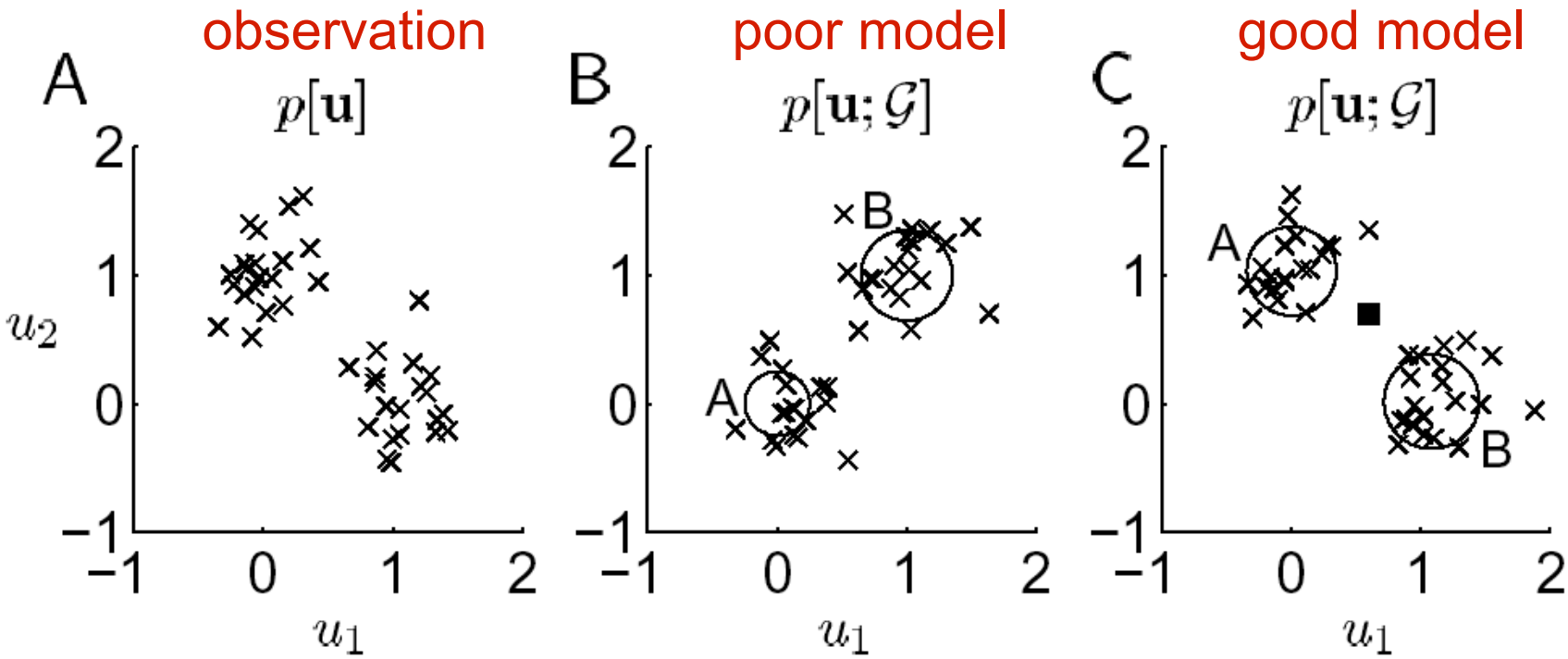


Illustration from Dayan & Abbot.

Graphical convention

The 'teardrop' represents a conditional distribution of events **u** being generated from a parameter set A.

Generative model (two Gaussian causes)

Prior distribution of causes:

$$v \in \{A, B\}, \qquad P[v; G] = \gamma_v, \qquad \gamma_A + \gamma_B = 1$$

Generative distribution:

$$p[\boldsymbol{u}|v; G] = \frac{1}{2\pi\,\Sigma_v}\,\exp\left[-\frac{|\boldsymbol{u} - \boldsymbol{g}_v|^2}{2\Sigma_v}\right]$$

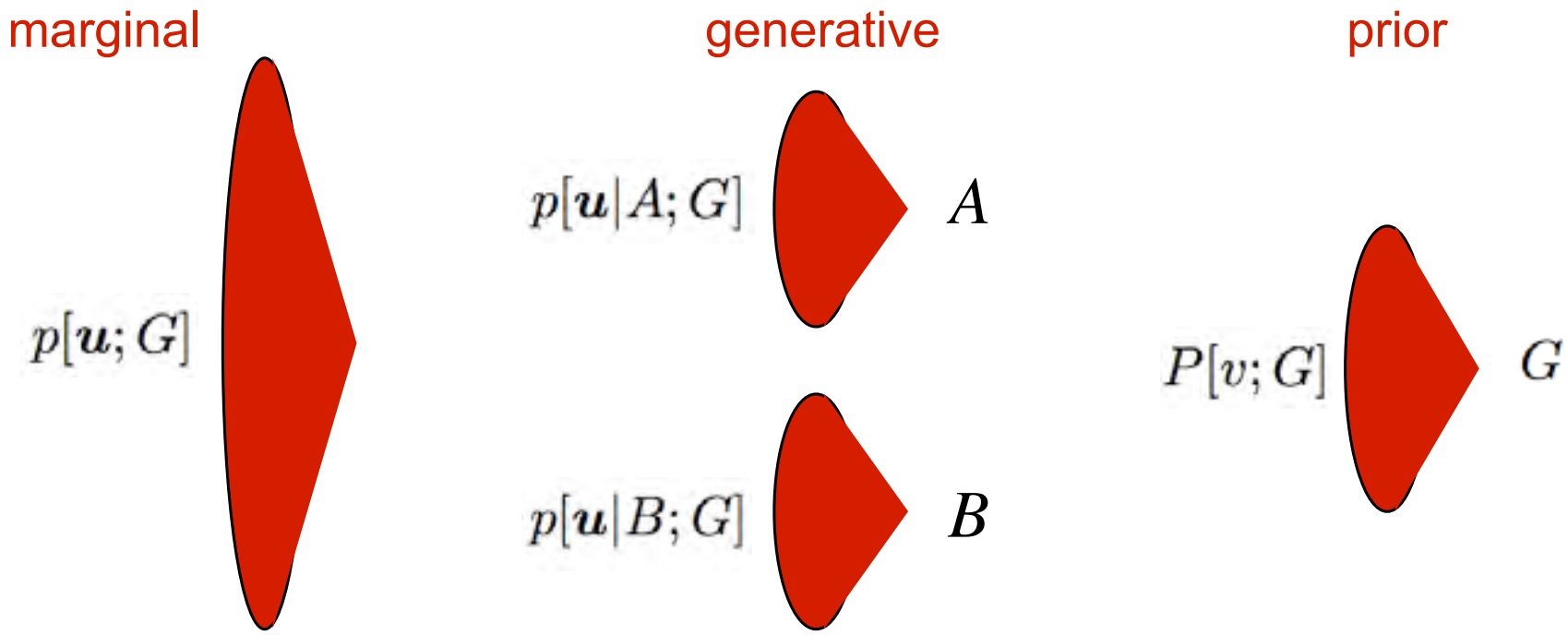
Parameter set:

$$G = \{\gamma_A, \gamma_B, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\}$$

Marginal distribution:

$$p[\boldsymbol{u}; G] = \sum_v p[\boldsymbol{u}|v; G]\,P[v; G]$$

From cause to observation



Recognition model

A *recognition model* classifies new observations in terms of their probable causes. In some cases, we can use Bayes' theorem to compute the most likely cause, given a particular observation:

Observations

$$p[\boldsymbol{u}]$$

Recognition

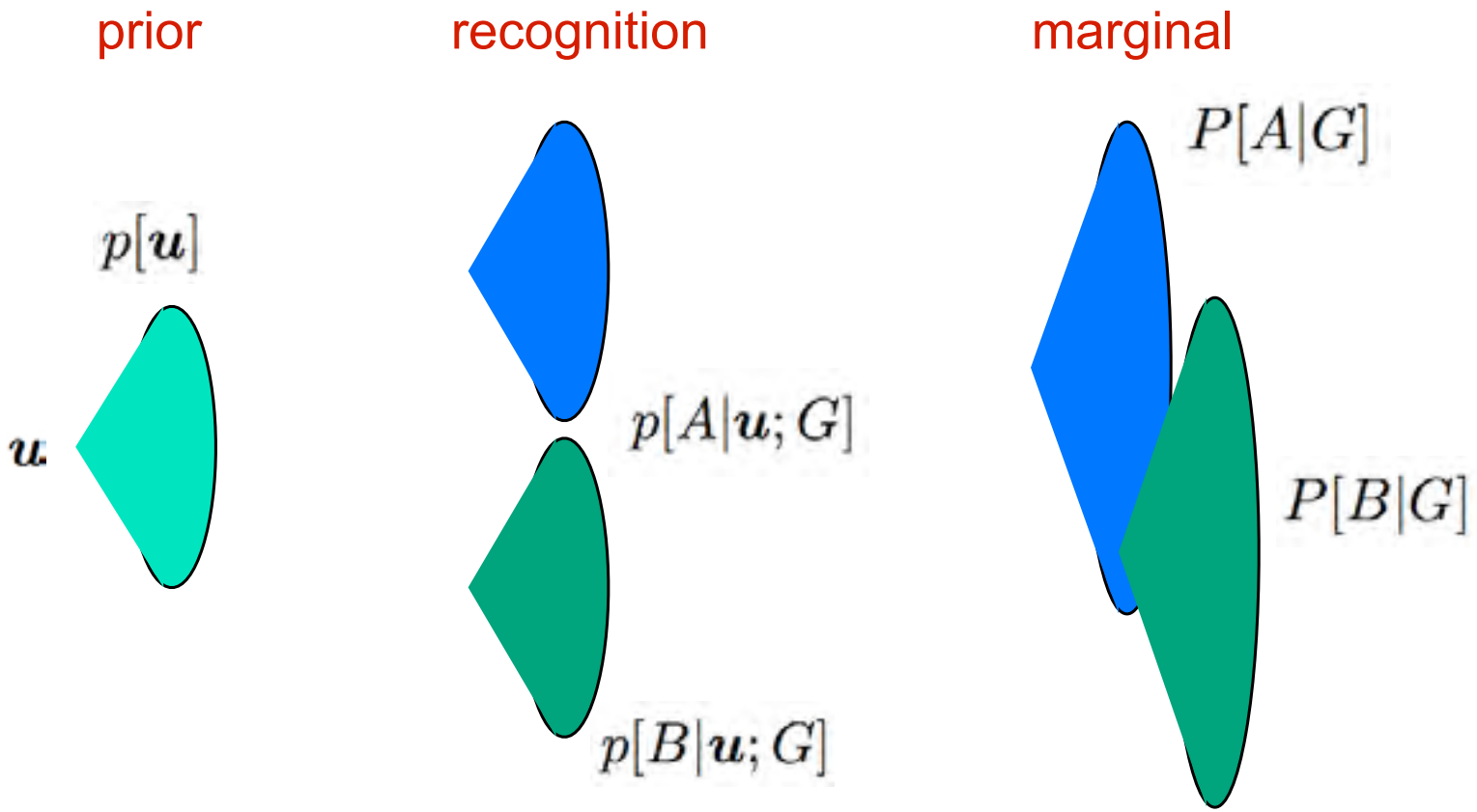
$$P[v|\boldsymbol{u}; G] = \frac{p[\boldsymbol{u}|v; G]\, P[v; G]}{p[\boldsymbol{u}; G]}$$

Marginal

$$P[v|G] = \sum_{\boldsymbol{u}} p[v|\boldsymbol{u}; G]\, p[\boldsymbol{u}]$$

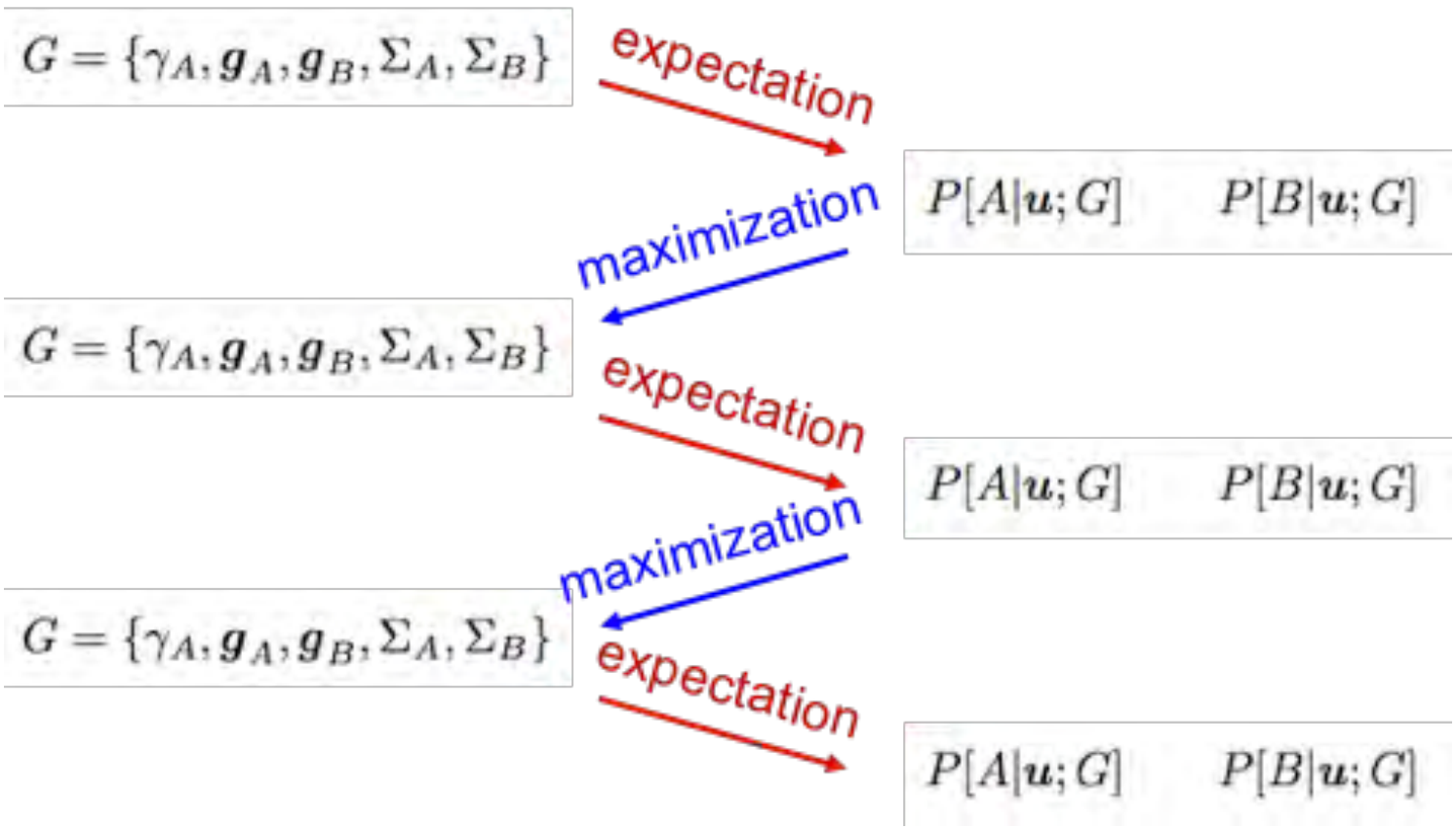
Only some situations are *invertible* in this way!

From observation to cause



Iterative improvement

Generation and recognition are *not* symmetric! In fact, alternation gradually improves generative model. Deepening understanding by considering again and again the same observations is a 'hermeneutic cycle'!



Expectation step

Given ensemble of observations $\boldsymbol{u}_i$, and generative model G, assign expected probabilities of alternative causes:

$$P[v|\boldsymbol{u};G] = \frac{p[\boldsymbol{u}|v;G]\,P[v;G]}{p[\boldsymbol{u};G]}$$

Obtain ensemble of assigned causes

$$\boldsymbol{u}_i \quad \rightarrow \quad P[A|\boldsymbol{u}_i;G], P[B|\boldsymbol{u}_i;G], \ldots$$

Maximization step

Given the ensemble of assigned causes, recompute the statistics of causes

$$\gamma_v = \langle P[v|\boldsymbol{u}; G] \rangle_{\boldsymbol{u}}$$

$$\boldsymbol{g}_v = \langle \boldsymbol{u}\, P[v|\boldsymbol{u}; G] \rangle_{\boldsymbol{u}}$$

$$\Sigma_v^2 = \left\langle |\boldsymbol{u} - \boldsymbol{g}_v|^2\, P[v|\boldsymbol{u}; G] \right\rangle_{\boldsymbol{u}}$$

Obtain improved generative model

$$G = \{\gamma_A, \gamma_B, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\}$$

Points to note

- Causal models include a generative and a recognition formulation. We can improve a model by iterating between these formulations.
- Generative formulation has prior probability of causes, conditional probability of observation (given one cause), and marginal probability of observation (given all causes).
- Recognition formulation has prior distribution of observations, conditional probability of cause (given one observation), and marginal probability of cause (given all observations).
- *Expectation step:* Given observations, we assign probabilities to different causes.
- *Maximization step:* Given probabilities of causes, we can better estimate cause distributions (*e.g.*, mean, variance, ...).

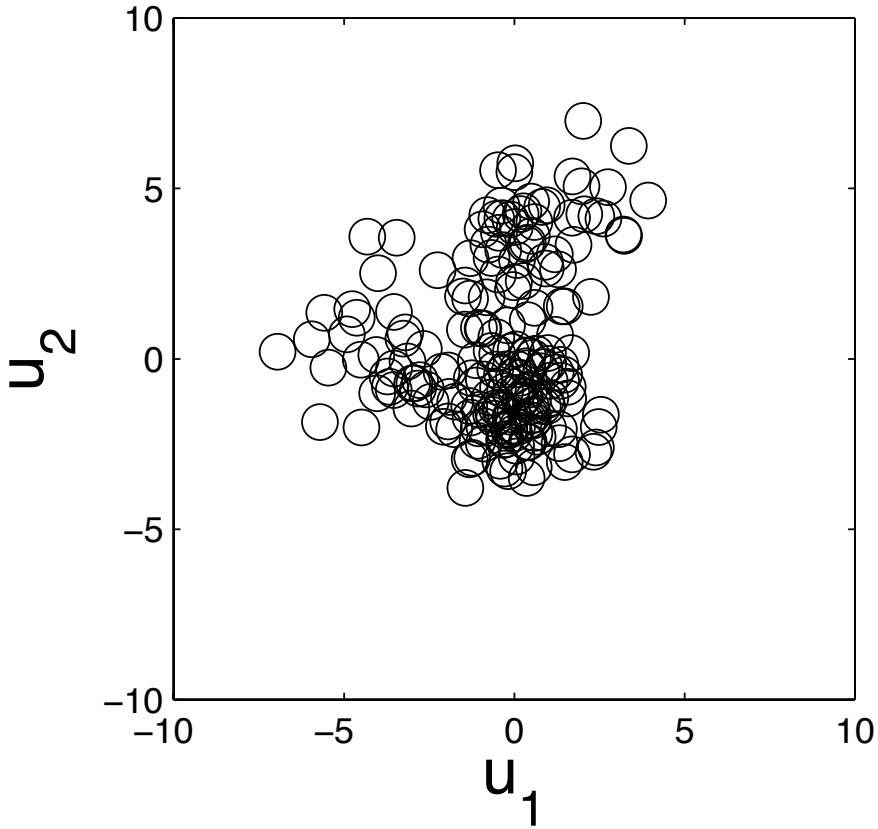
3 K-Means

We illustrate iterative convergence with a particularly simple method termed *K-means*.

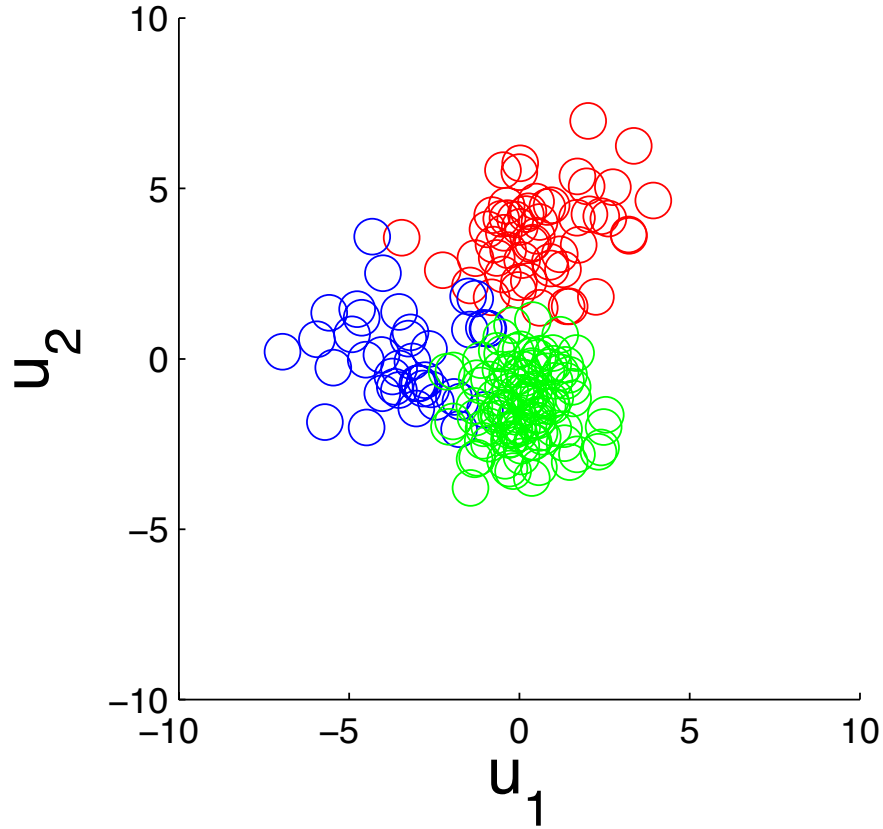
This method assumes that observations $\boldsymbol{u} = \{u_1, u_2\}$ are the product of N_{class} distinct causes. Each class is described by one 'prototypical' observation $\boldsymbol{K}_i$ $(i = 1, \ldots, N_{class})$.

Observations are assigned to causes by proximity: each observation is assumed to be caused by the nearest 'prototype'.

Observations



True causes



Initialization

Given two-dimensional observations $\boldsymbol{u} = \{u_1, u_2\}$, we choose the number N_{class} of presumed causes (based on heuristics) ....

... and randomly choose the prototypical observations $\boldsymbol{K}_i$.

Next, we iteratively improve the initial prototypes.

Deterministic classification

We compute the Euclidean distances between observations $\boldsymbol{u}$ and prototypes $\boldsymbol{K}_i$

$$d_i = |\boldsymbol{u} - \boldsymbol{K}_i|$$

and classify each observation with the nearest prototype.

Maximization

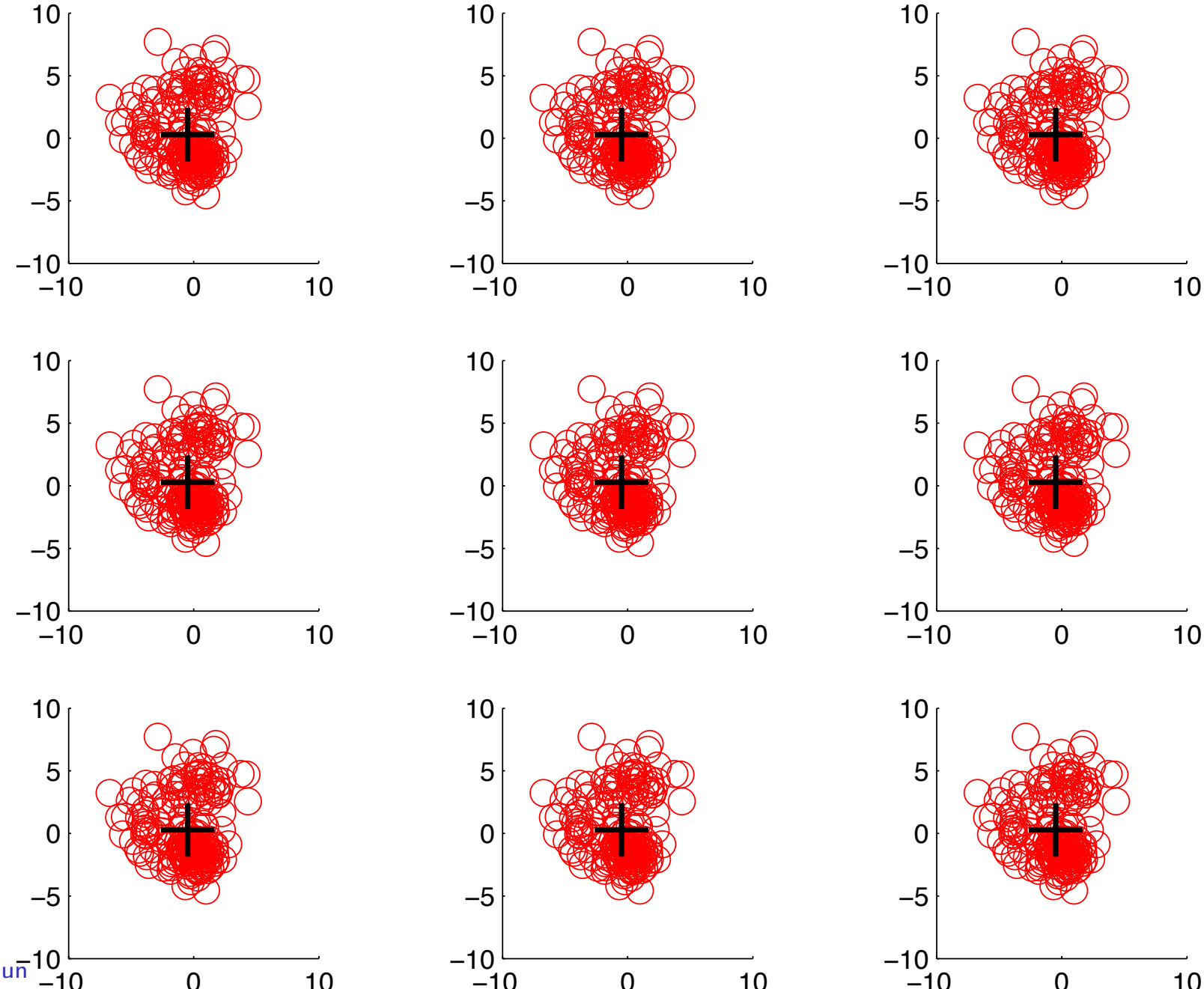
For each class, we compute the centroid of all associated observations:

$$\boldsymbol{K}_i = \frac{1}{N_i} \sum_{\boldsymbol{u} \in i} \boldsymbol{u}$$

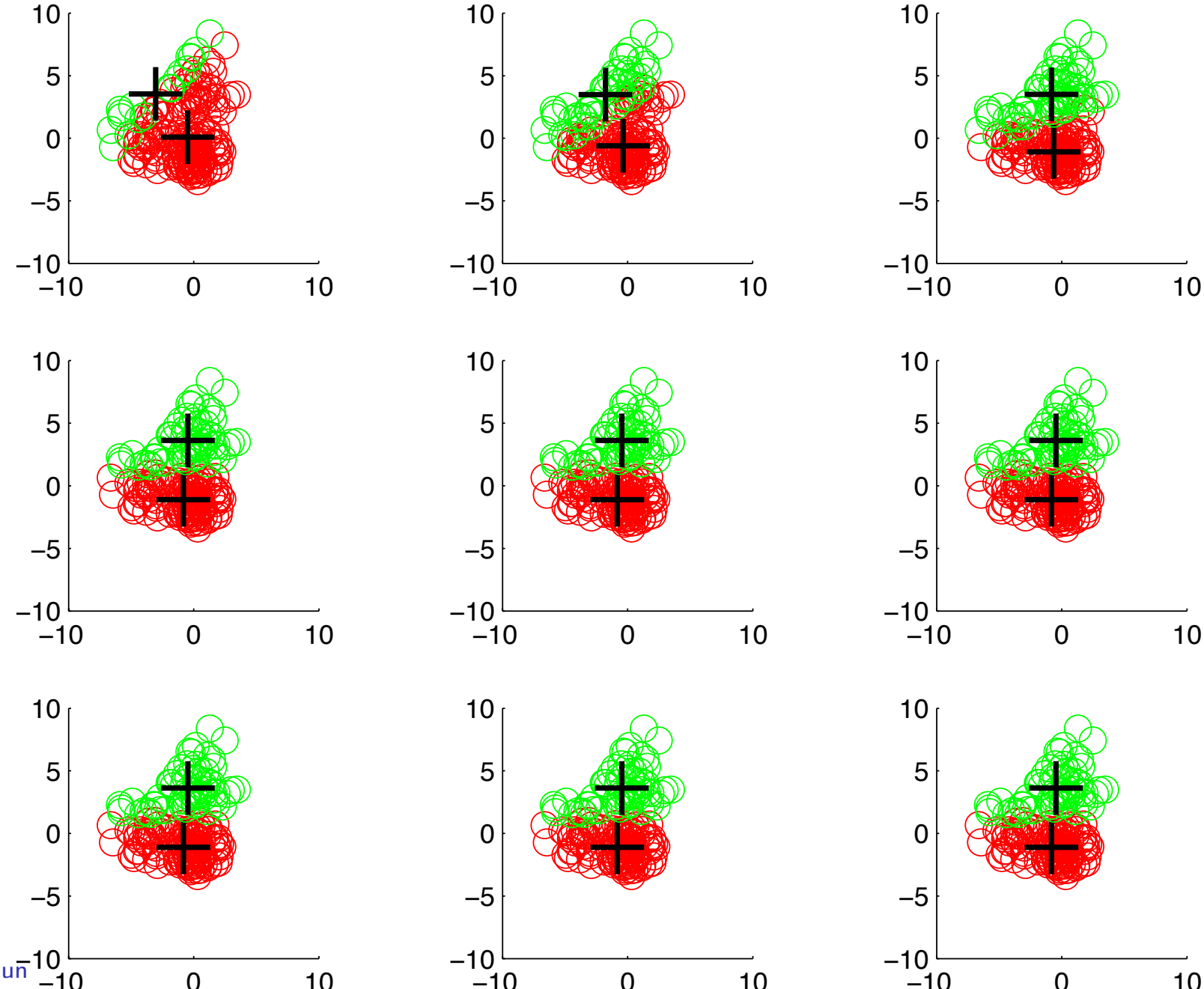
Iteration

We alternate between classification and maximization until the outcome converges.

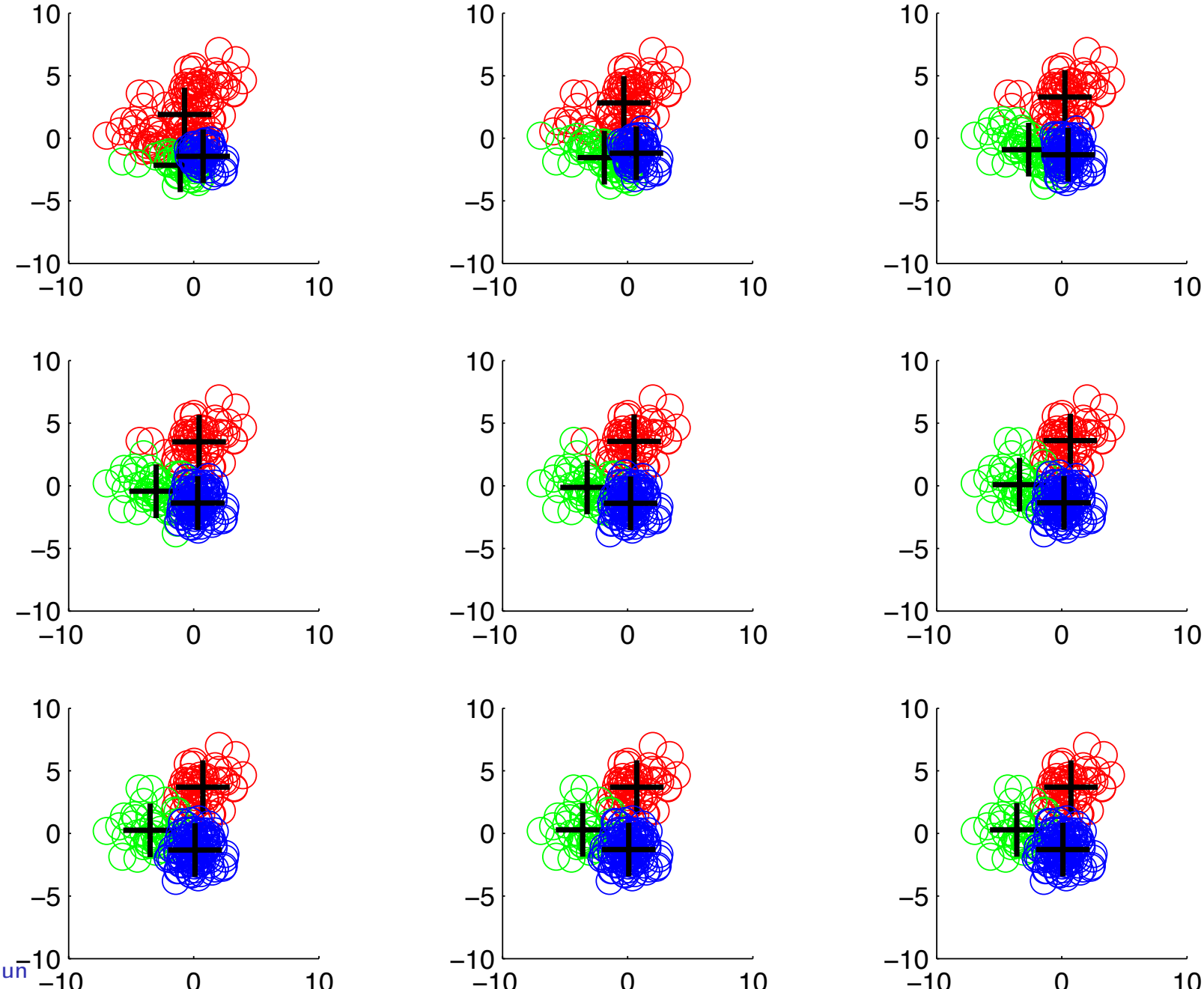
Assume one class



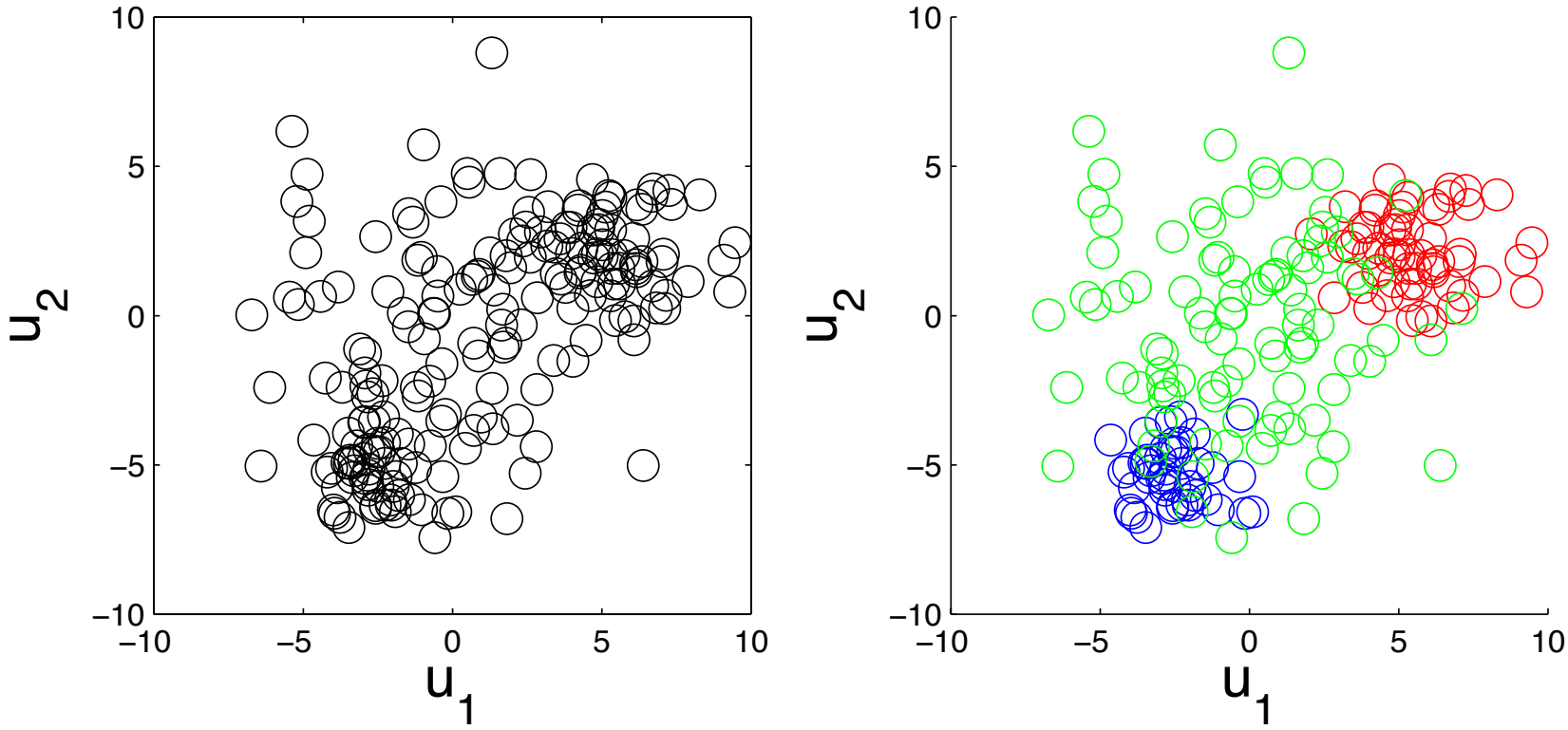
Assume two classes



Assume three classes

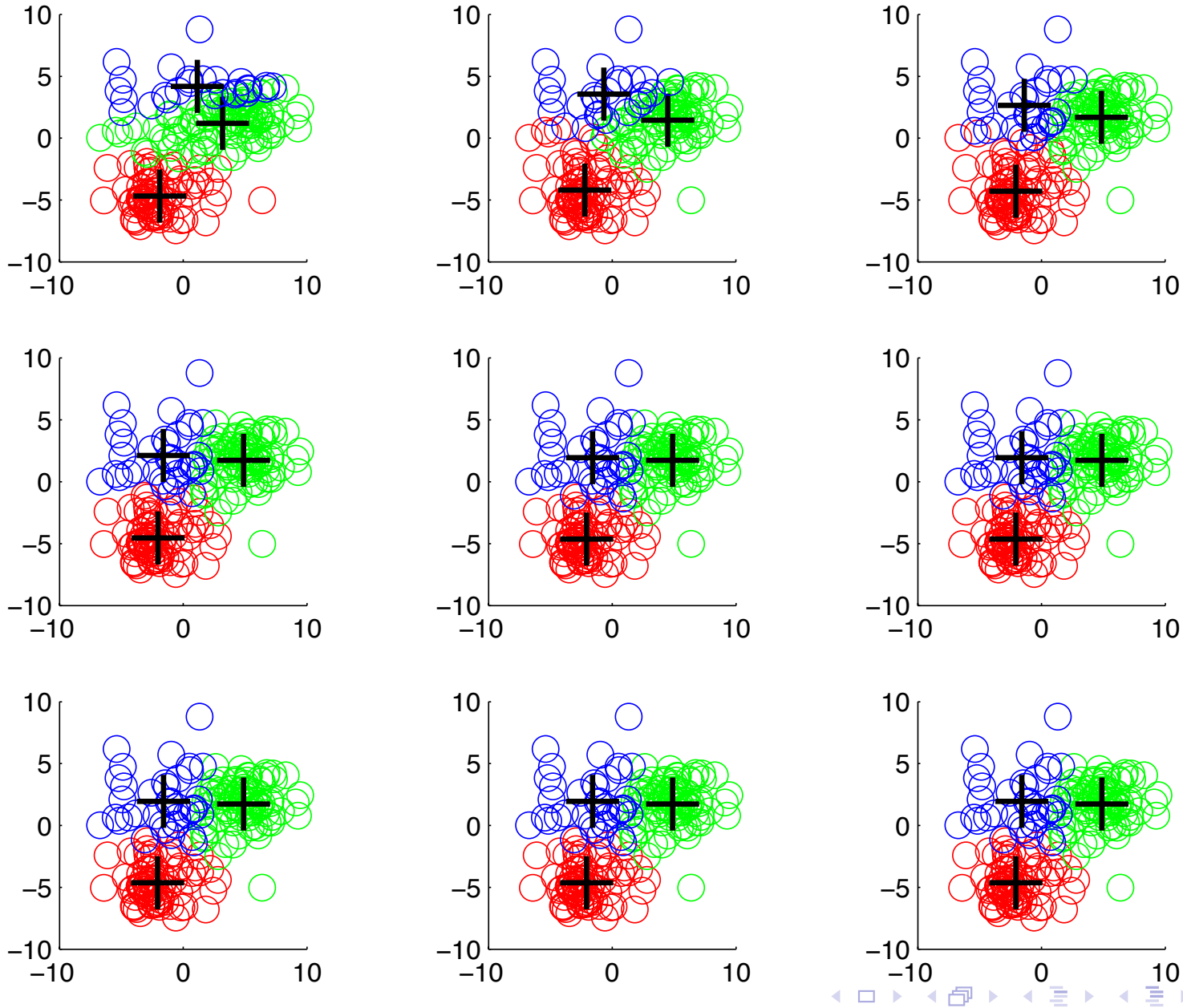


Width of generative distribution



K-means algorithm fails for generative distributions of different width. It cannot recognize green *between* blue and red!

Mistaken assignments



Points to note

- The K-means algorithm groups observations into a chosen number of classes.
- The classification criterion is distance from a prototype.
- Prototype locations are adjusted iteratively, until the overall sample-prototype distance is minimal.

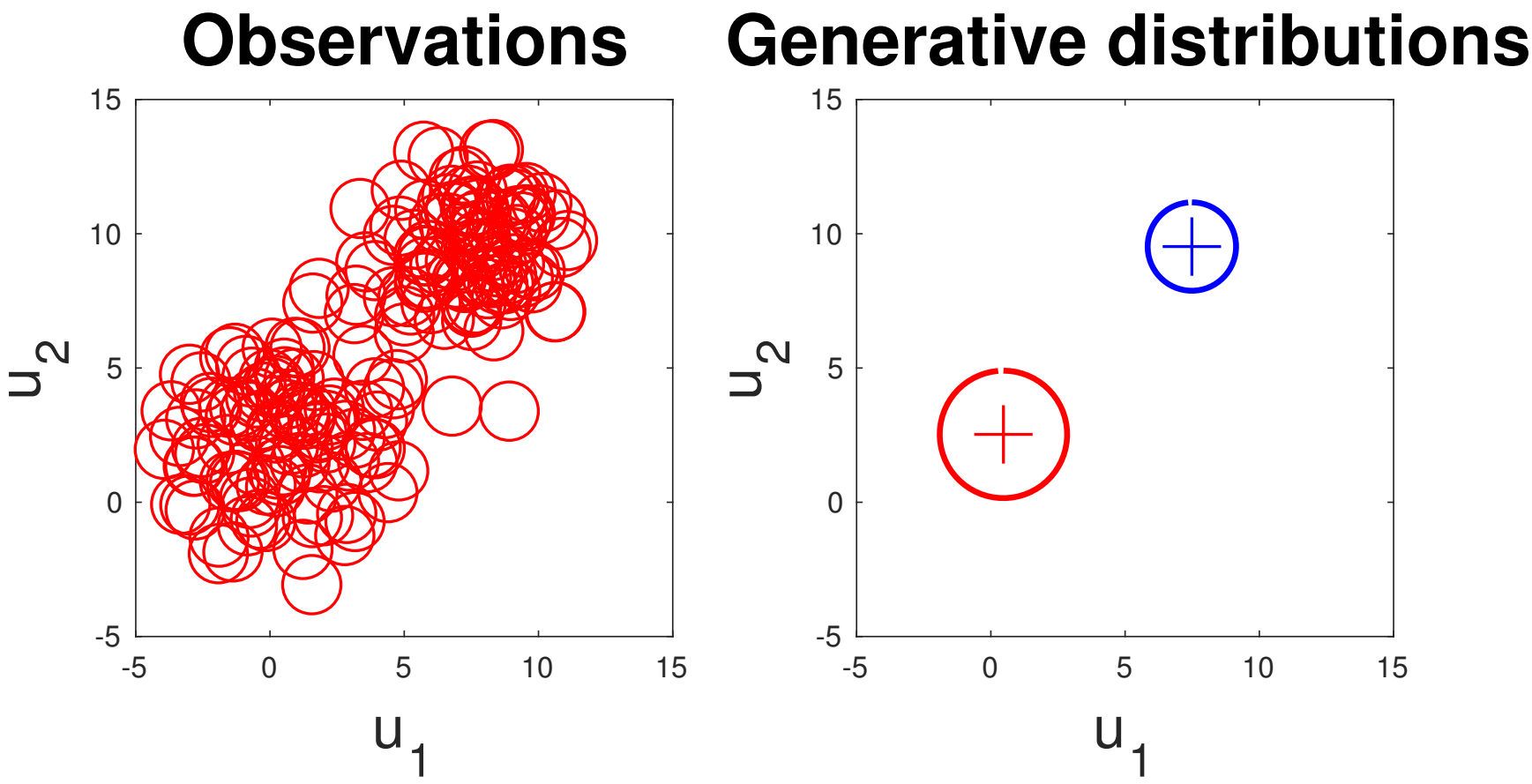
4 Expectation maximization

We return to *expectation maximization* (*EM*) to iteratively a improve more sophisticated generative model G.

The generative model of Section 2 involved two causes, each producing Gaussian-distributed observations. In this case, G comprises the means $\boldsymbol{g}_{A,B}$ and variances $\Sigma_{A,B}$ of the conditional distributions $p[u|v;G]$ and the prior probabilities $p[A|G] = \gamma_A$ and $p[B|G] = \gamma_B$.

$$G = \{\gamma_A, \gamma_B, \boldsymbol{g}_A, \boldsymbol{g}_B, \Sigma_A, \Sigma_B\}$$

2D events with two Gaussian causes



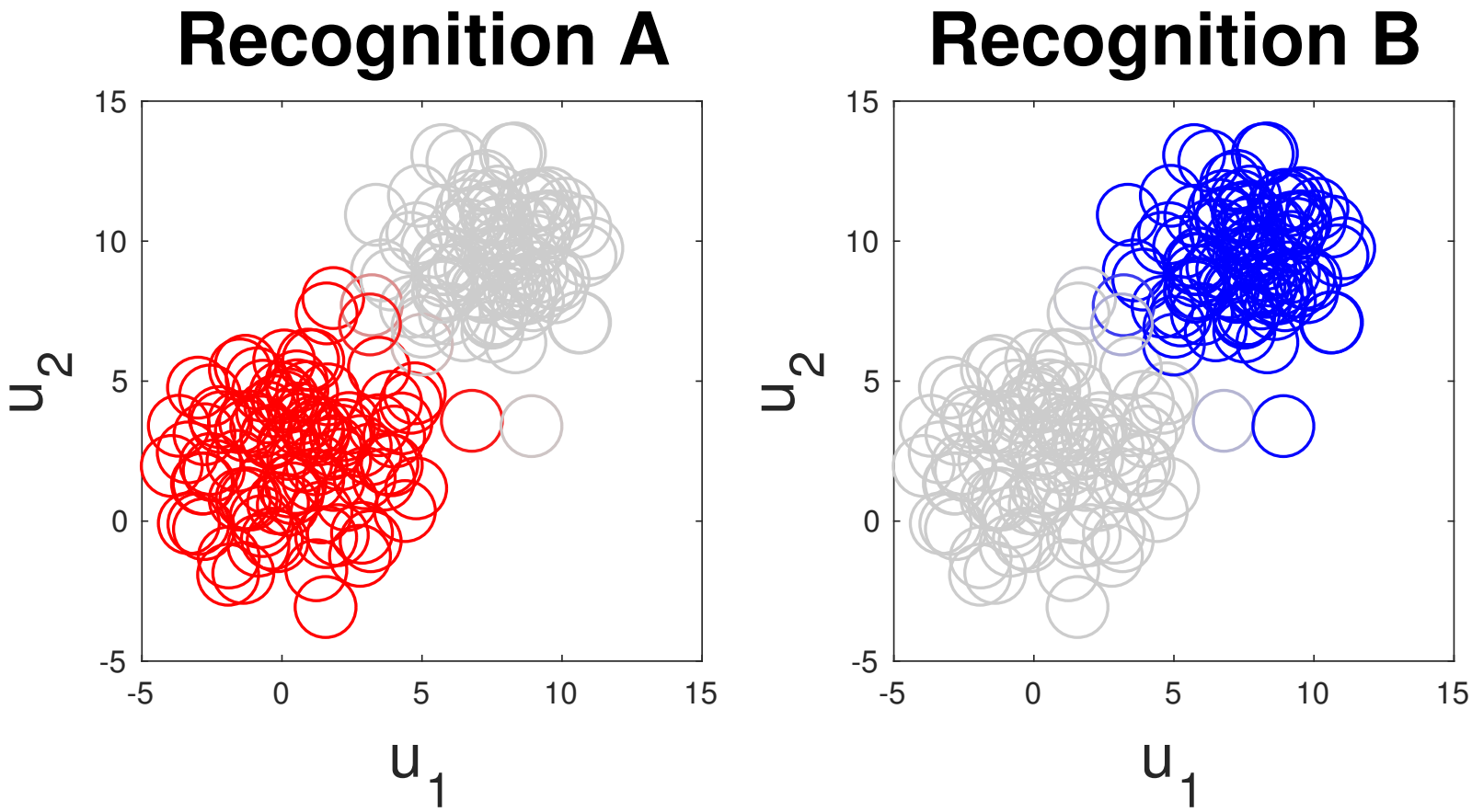
In right frame, crosses and circles mark, respectively, means and standard deviations of the two generative distributions.

If we knew the cause of each observation, we could readily compute the parameters in G. As we do not have this information (yet!), we instead use the recognition distribution $P[v|\boldsymbol{u}; G]$ as our current best guess.

The EM algorithm consists of two alternating steps, the E (expectation) step of inferring "probable causes" from the classification distribution $P[v|\boldsymbol{u}; G]$, and the M (maximization) step of computing parameters G from weighted averages over observations $\boldsymbol{u}$.

It is far from obvious that this (suspiciously circular) process of considering the **same observations** again and again will converge to an optimal generative model G!

True recognition distributions



The true cause of each observation is indicated by color. Note that every circle is *either* red or green.

Expectation step

Conditional probability of observing $\boldsymbol{u}$, given cause A or cause B:

$$p(\boldsymbol{u}|A) = \frac{1}{\sqrt{2\pi\,\Sigma_A^2}}\,\exp\left(-\frac{|\boldsymbol{u}-\boldsymbol{g}_A|^2}{2\,\Sigma_A^2}\right) \qquad 1 = \int p(\boldsymbol{u}|A)\,d\boldsymbol{u}$$

$$p(\boldsymbol{u}|B) = \frac{1}{\sqrt{2\pi\,\Sigma_B^2}}\,\exp\left(-\frac{|\boldsymbol{u}-\boldsymbol{g}_B|^2}{2\,\Sigma_B^2}\right) \qquad 1 = \int p(\boldsymbol{u}|B)\,d\boldsymbol{u}$$

Joint probability of observing $\boldsymbol{u}$ *and* it being due to cause A or cause B:

$$p(\boldsymbol{u},A) = p(\boldsymbol{u}|A)\,p(A) = \gamma_A\,p(\boldsymbol{u}|A) \qquad \gamma_A = \int p(\boldsymbol{u},A)\,d\boldsymbol{u}$$

$$p(\boldsymbol{u},B) = p(\boldsymbol{u}|B)\,p(B) = \gamma_B\,p(\boldsymbol{u}|B) \qquad \gamma_B = \int p(\boldsymbol{u},B)\,d\boldsymbol{u}$$

Total probability of observing $\boldsymbol{u}$ (due to either cause).

$$p(\boldsymbol{u}) = p(\boldsymbol{u}, A) + p(\boldsymbol{u}, B) \qquad 1 = \int p(\boldsymbol{u})\, d\boldsymbol{u}$$

Conditional probability of cause A or B, given an observation $\boldsymbol{u}$:

$$p(A|\boldsymbol{u}) = \frac{p(\boldsymbol{u}, A)}{p(\boldsymbol{u})} = \frac{p(\boldsymbol{u}, A)}{p(\boldsymbol{u}, A) + p(\boldsymbol{u}, B)}$$

$$p(B|\boldsymbol{u}) = \frac{p(\boldsymbol{u}, B)}{p(\boldsymbol{u})} = \frac{p(\boldsymbol{u}, B)}{p(\boldsymbol{u}, A) + p(\boldsymbol{u}, B)}$$

Maximization step

Given a sufficiently large set of observations $\boldsymbol{u}$, any average taken over this set will be weighted by the probability density of observations $p(\boldsymbol{u})$, as more probable observations will contribute more samples to the set. We can use these weighted averages to estimate model parameters.

Weighted averages over conditional probability of causes:

$$\gamma_A = \int p(\boldsymbol{u}|A)\, d\boldsymbol{u} \approx \langle\, p(A|\boldsymbol{u})\,\rangle_{\boldsymbol{u}}$$

$$\gamma_B = \int p(\boldsymbol{u}|B)\, d\boldsymbol{u} \approx \langle\, p(B|\boldsymbol{u})\,\rangle_{\boldsymbol{u}}$$

Looking again at the same observations, but with 'new eyes', that is, with an improved model! Moving in a 'hermeneutic circle'!

Mean of distribution due to cause A or B:

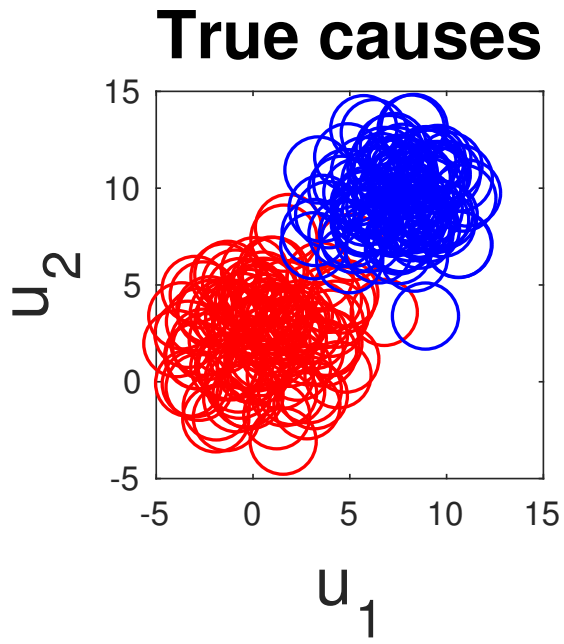
$$\boldsymbol{g}_A = \int \boldsymbol{u}\, p(\boldsymbol{u}|A)\, d\boldsymbol{u} \approx \frac{1}{\gamma_A} \left\langle \boldsymbol{u}\, p(A|\boldsymbol{u}) \right\rangle_{\boldsymbol{u}}$$

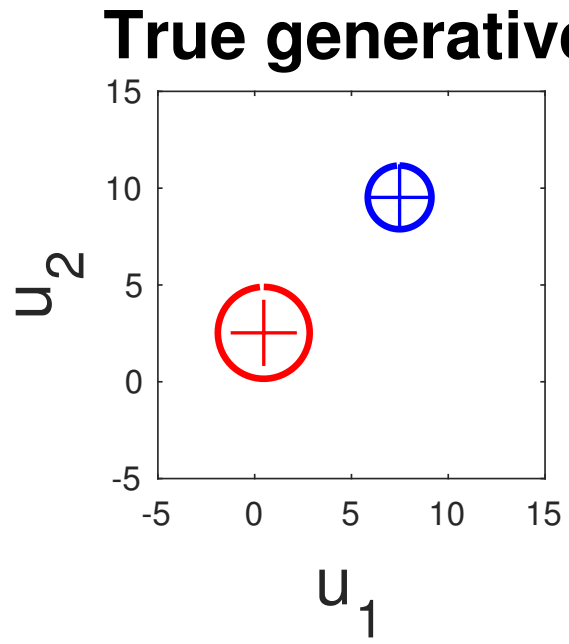
$$\boldsymbol{g}_B = \int \boldsymbol{u}\, p(\boldsymbol{u}|B)\, d\boldsymbol{u} \approx \frac{1}{\gamma_B} \left\langle \boldsymbol{u}\, p(B|\boldsymbol{u}) \right\rangle_{\boldsymbol{u}}$$

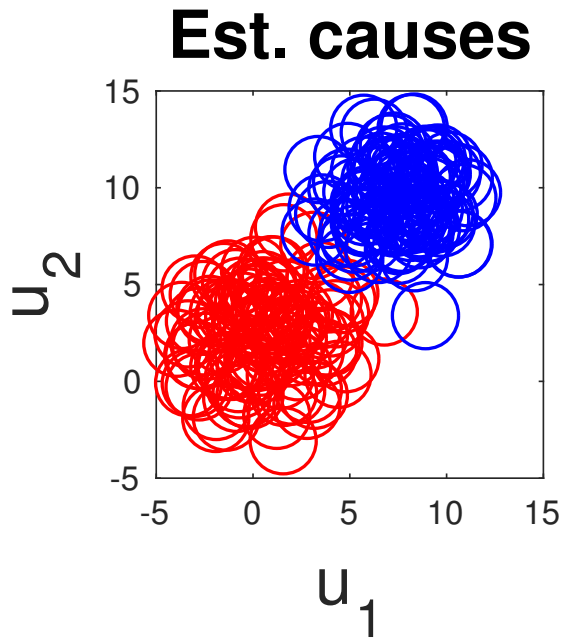
Variance of distribution due to cause A or B:

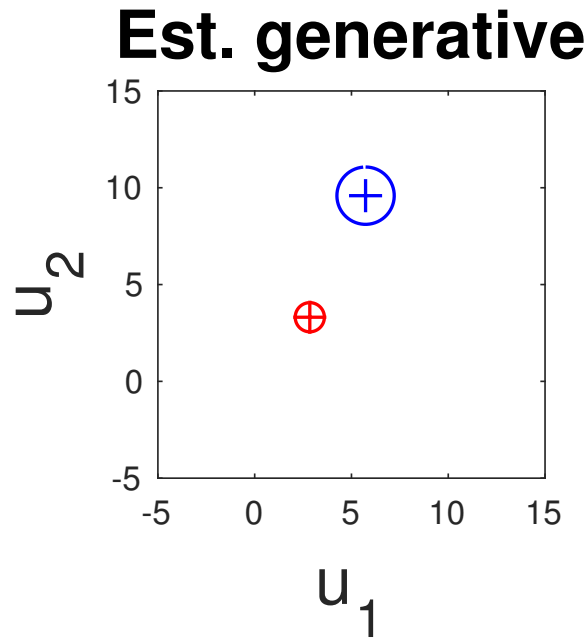
$$\Sigma_A^2 = \int |\boldsymbol{u} - \boldsymbol{g}_A|^2\, p(\boldsymbol{u}|A)\, d\boldsymbol{u} \approx \frac{1}{\gamma_A} \left\langle |\boldsymbol{u} - \boldsymbol{g}_A|^2\, p(A|\boldsymbol{u}) \right\rangle_{\boldsymbol{u}}$$

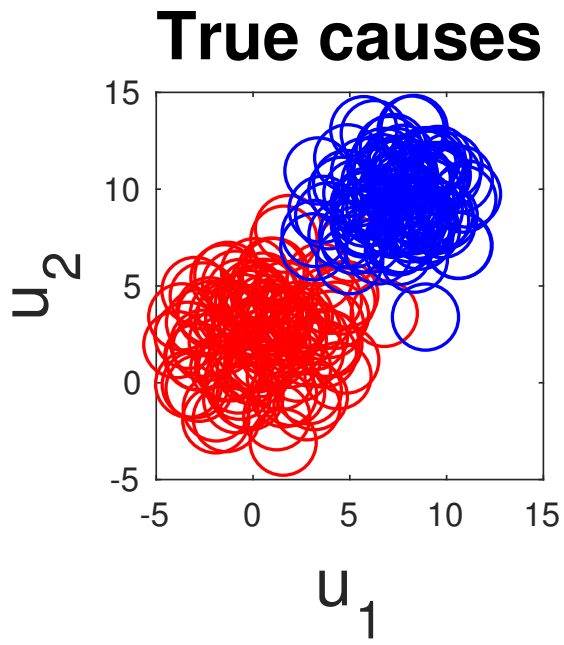
$$\Sigma_B^2 = \int |\boldsymbol{u} - \boldsymbol{g}_B|^2\, p(\boldsymbol{u}|B)\, d\boldsymbol{u} \approx \frac{1}{\gamma_B} \left\langle |\boldsymbol{u} - \boldsymbol{g}_B|^2\, p(B|\boldsymbol{u}) \right\rangle_{\boldsymbol{u}}$$

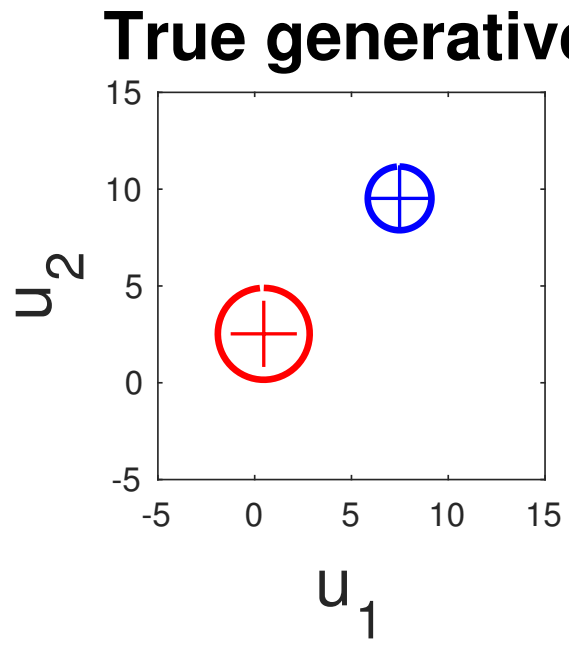

True causes
15
10
5
0
-5
-5
0
5
10
15
u2
u1

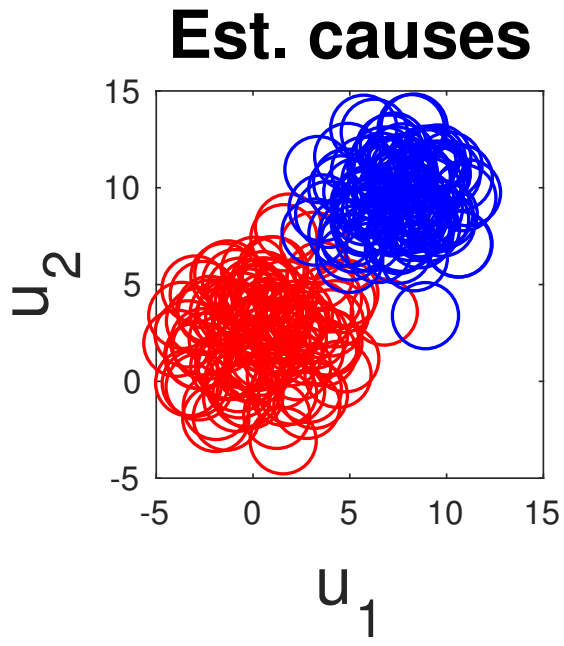

True generative
15
10
5
0
-5
-5
0
5
10
15
u2
u1

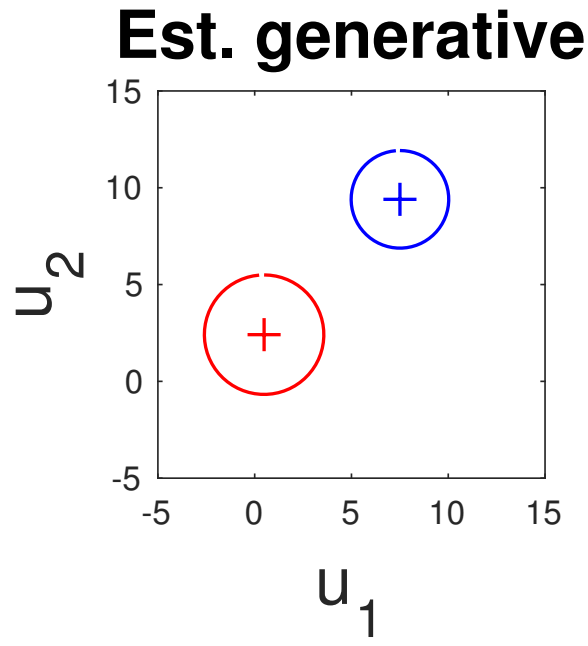

Est. causes
15
10
5
0
-5
-5
0
5
10
15
u2
u1

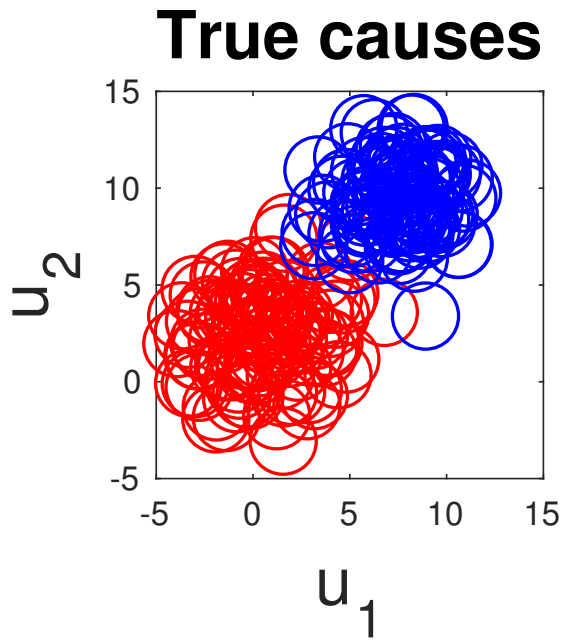

Est. generative
15
10
5
0
-5
-5
0
5
10
15
u2
u1

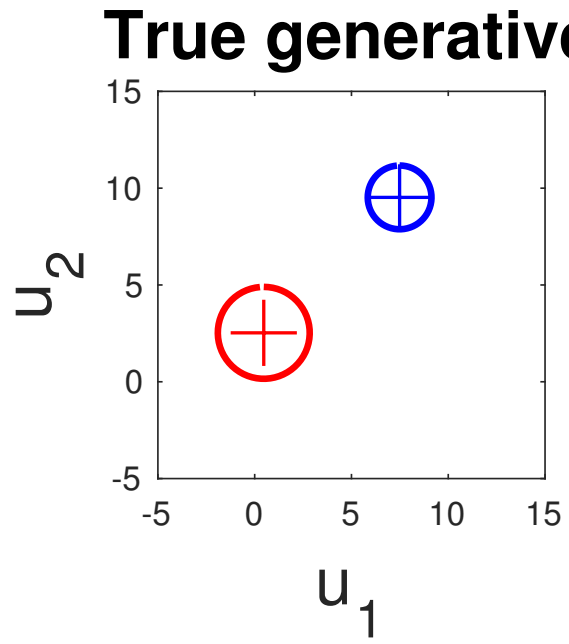

True causes
15
10
5
0
-5
-5
0
5
10
15
u2
u1

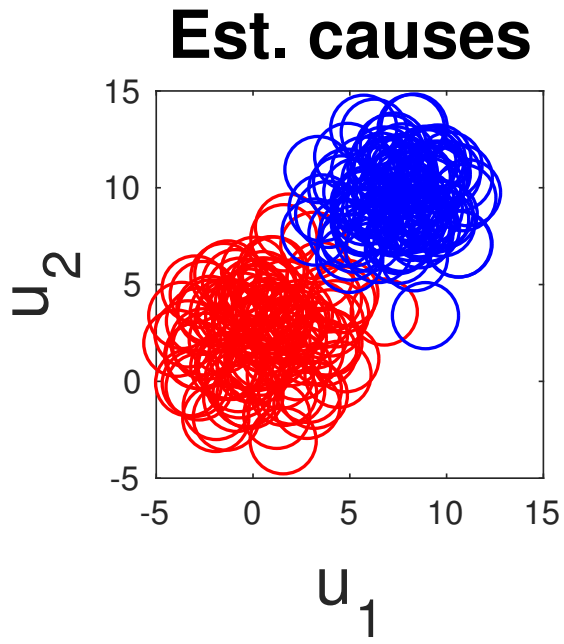

True generative
15
10
5
0
-5
-5
0
5
10
15
u2
u1

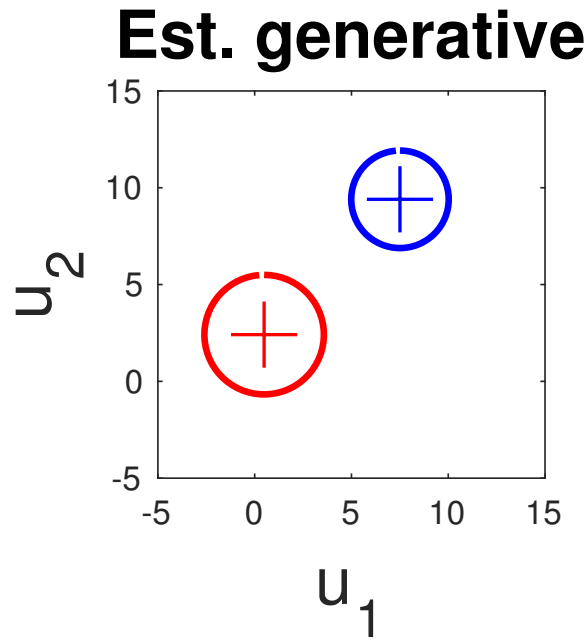

Est. causes
15
10
5
0
-5
-5
0
5
10
15
u2
u1


Est. generative
15
10
5
0
-5
-5
0
5
10
15
u2
u1


True causes
15
10
5
0
-5
-5
0
5
10
15
u2
u1


True generative
15
10
5
0
-5
-5
0
5
10
15
u2
u1


Est. causes
15
10
5
0
-5
-5
0
5
10
15
u2
u1


Est. generative
15
10
5
0
-5
-5
0
5
10
15
u2
u1

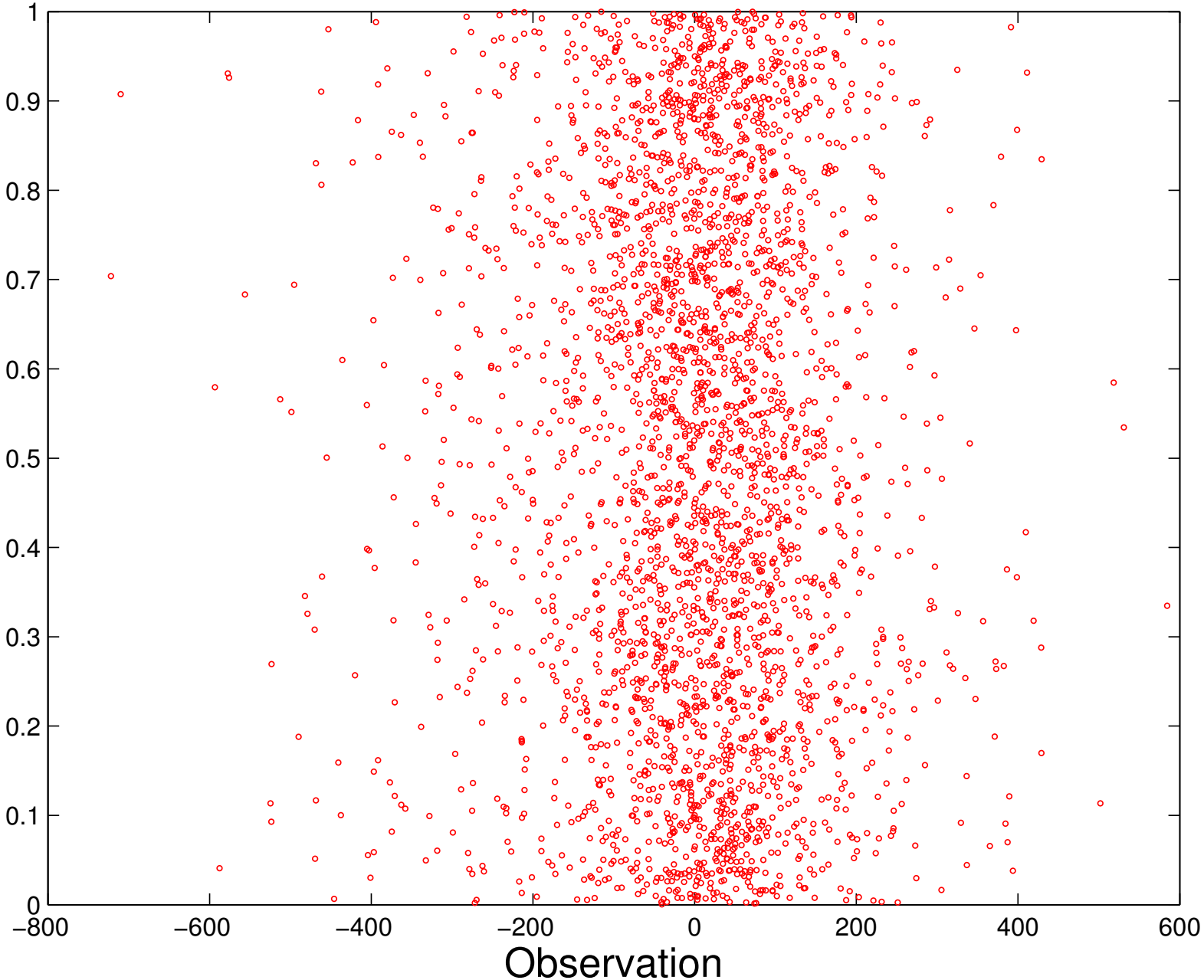
Points to note

- Expectation-maximization (EM) is a general procedure for improving causal models.
- We have illustrated the principle for a simple model with two Gaussian causes.
- Given observations, we estimate the respective probabilities of alternative causes.
- To improve the model, we compute cause-probability-weighted averages over our observations:

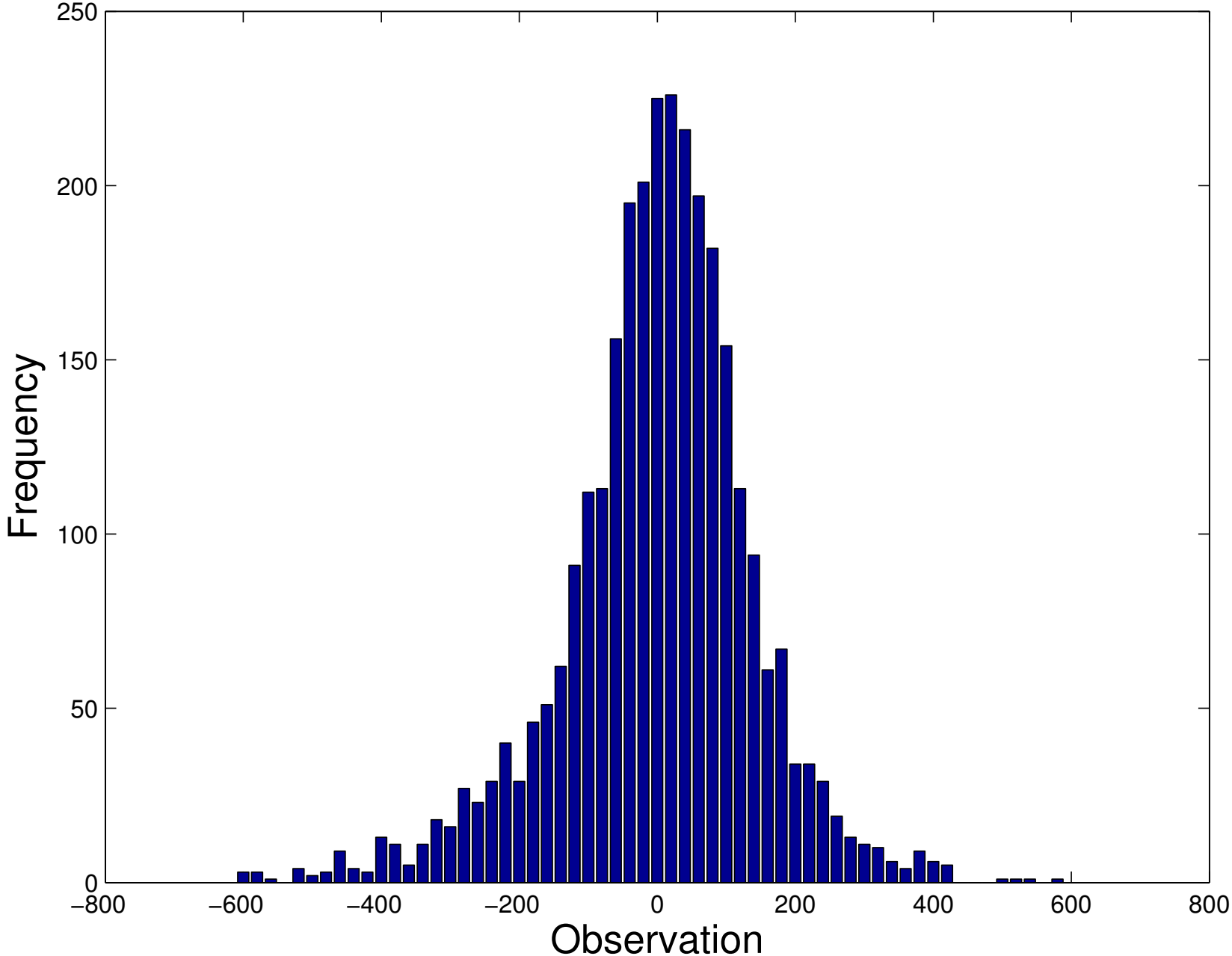
$$\langle S(\boldsymbol{u}) \rangle = \int S(\boldsymbol{u})\, p(A|u)\, p(\boldsymbol{u})\, d\boldsymbol{u} \approx \frac{1}{\gamma_A} \left\langle S(\boldsymbol{u})\, p(A|\boldsymbol{u}) \right\rangle_{\boldsymbol{u}}$$

- As the model improves, the estimated probability approaches the actual probability of causes.

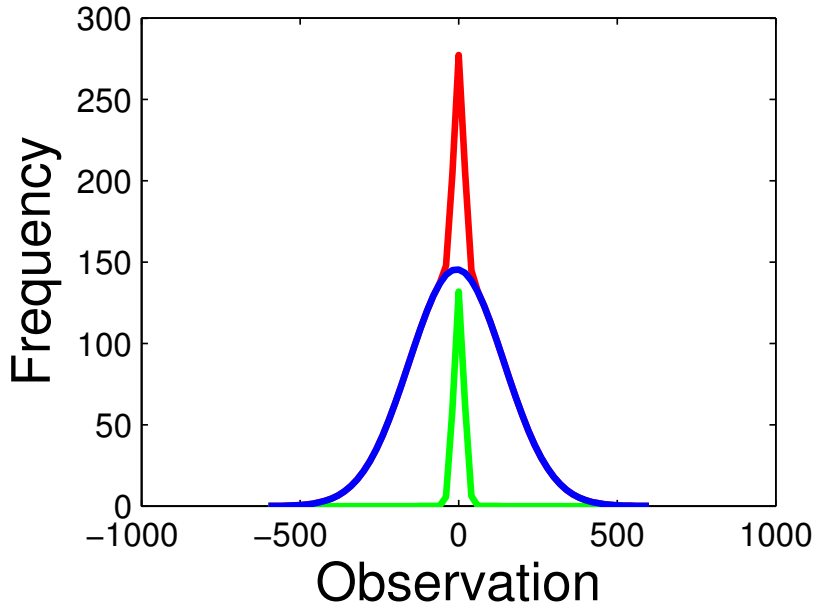
Example: 1D observations with two Gaussian causes

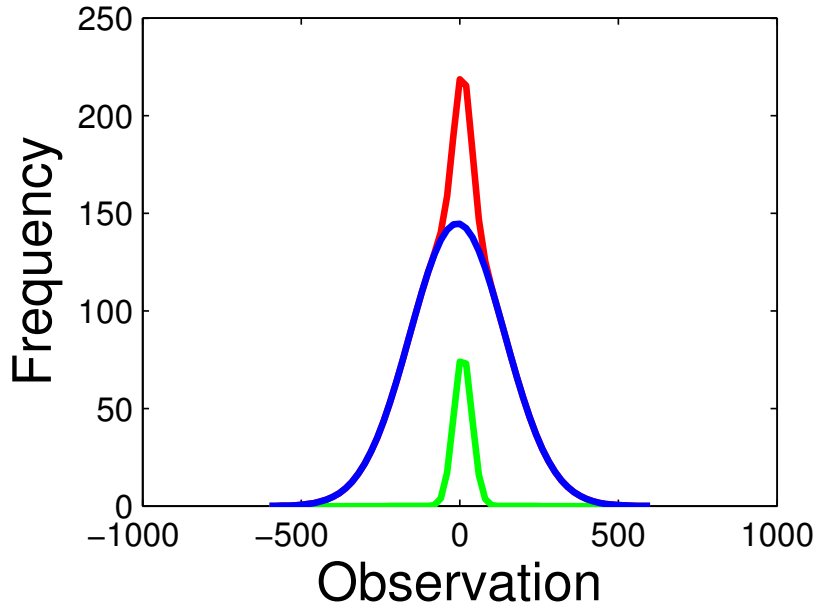


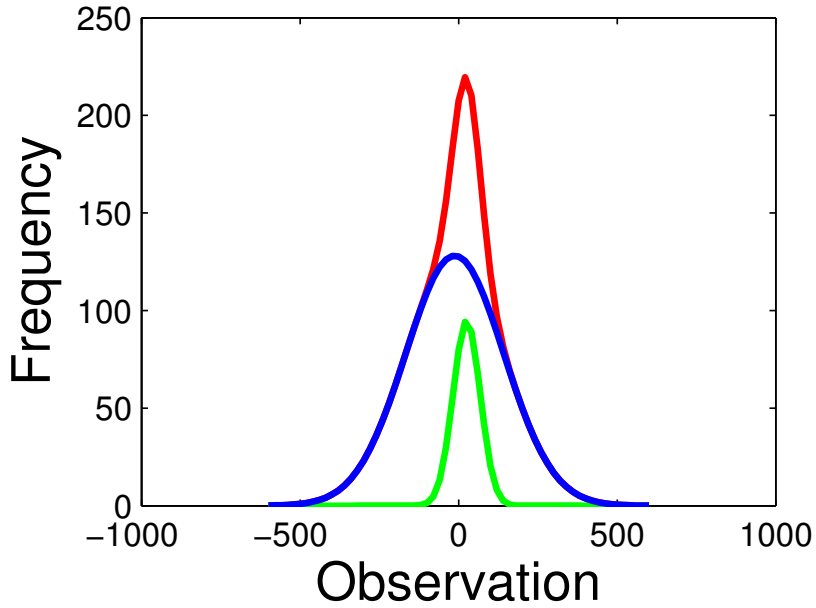
Observed distribution

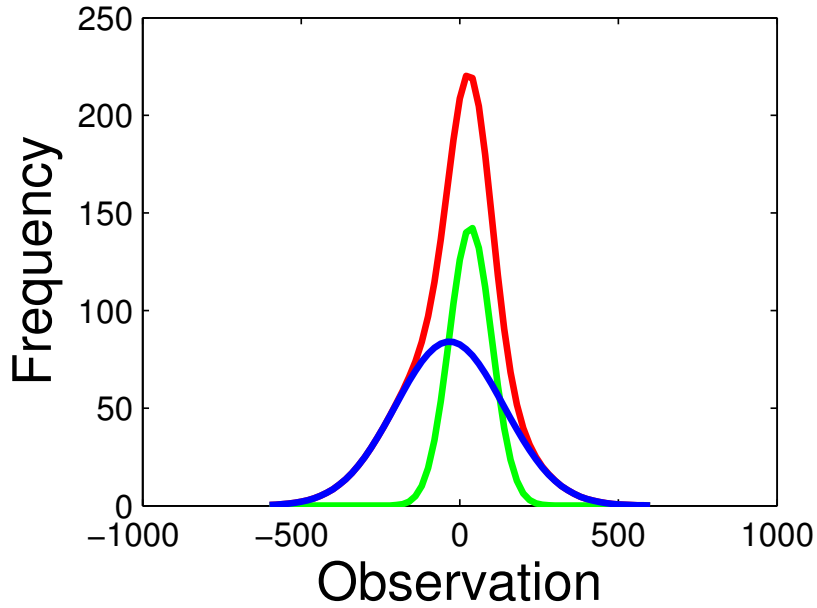


Iteration

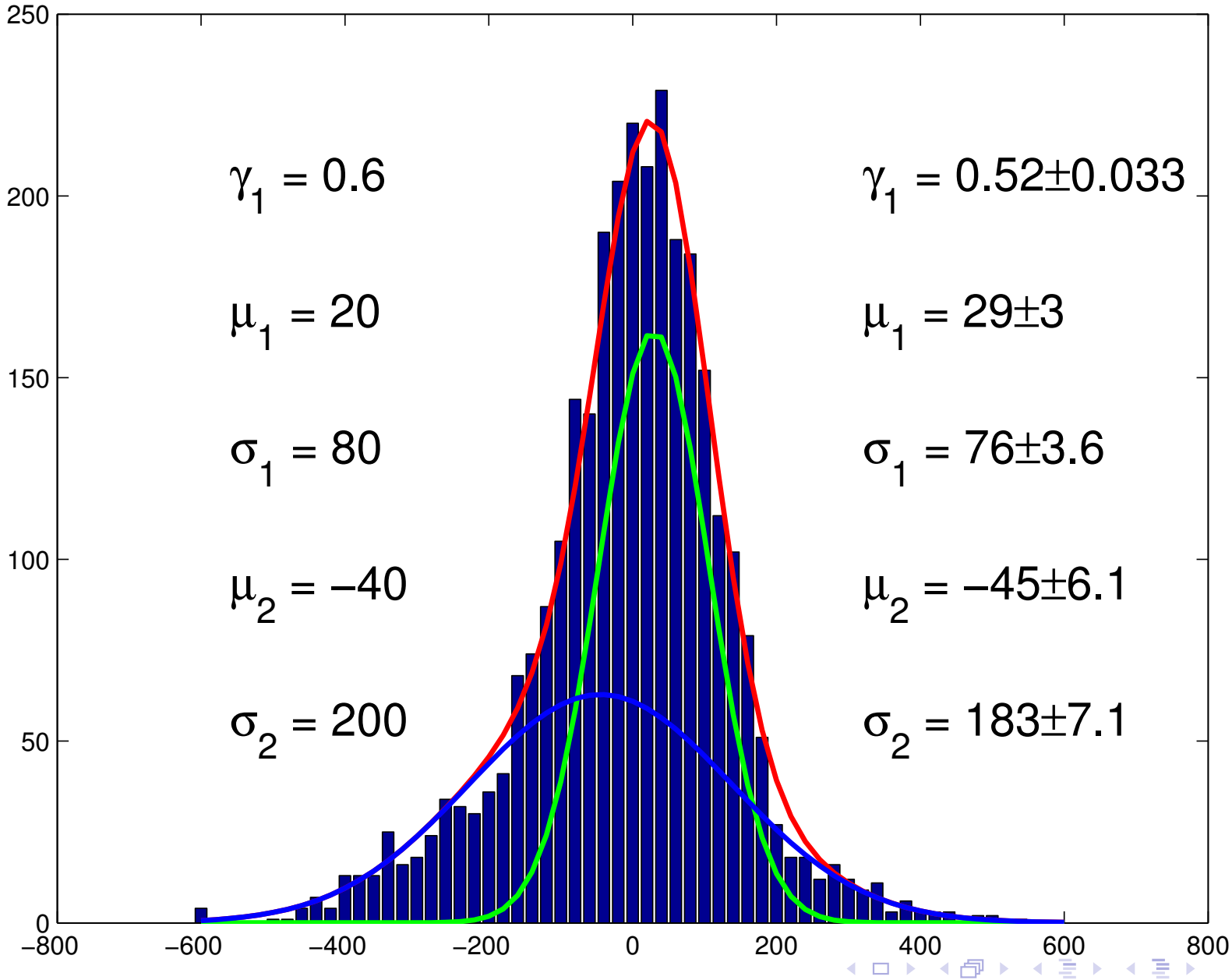

300
250
200
150
100
50
0
Frequency
−1000
−500
0
500
1000
Observation


250
200
150
100
50
0
Frequency
−1000
−500
0
500
1000
Observation


250
200
150
100
50
0
Frequency
−1000
−500
0
500
1000
Observation


250
200
150
100
50
0
Frequency
−1000
−500
0
500
1000
Observation

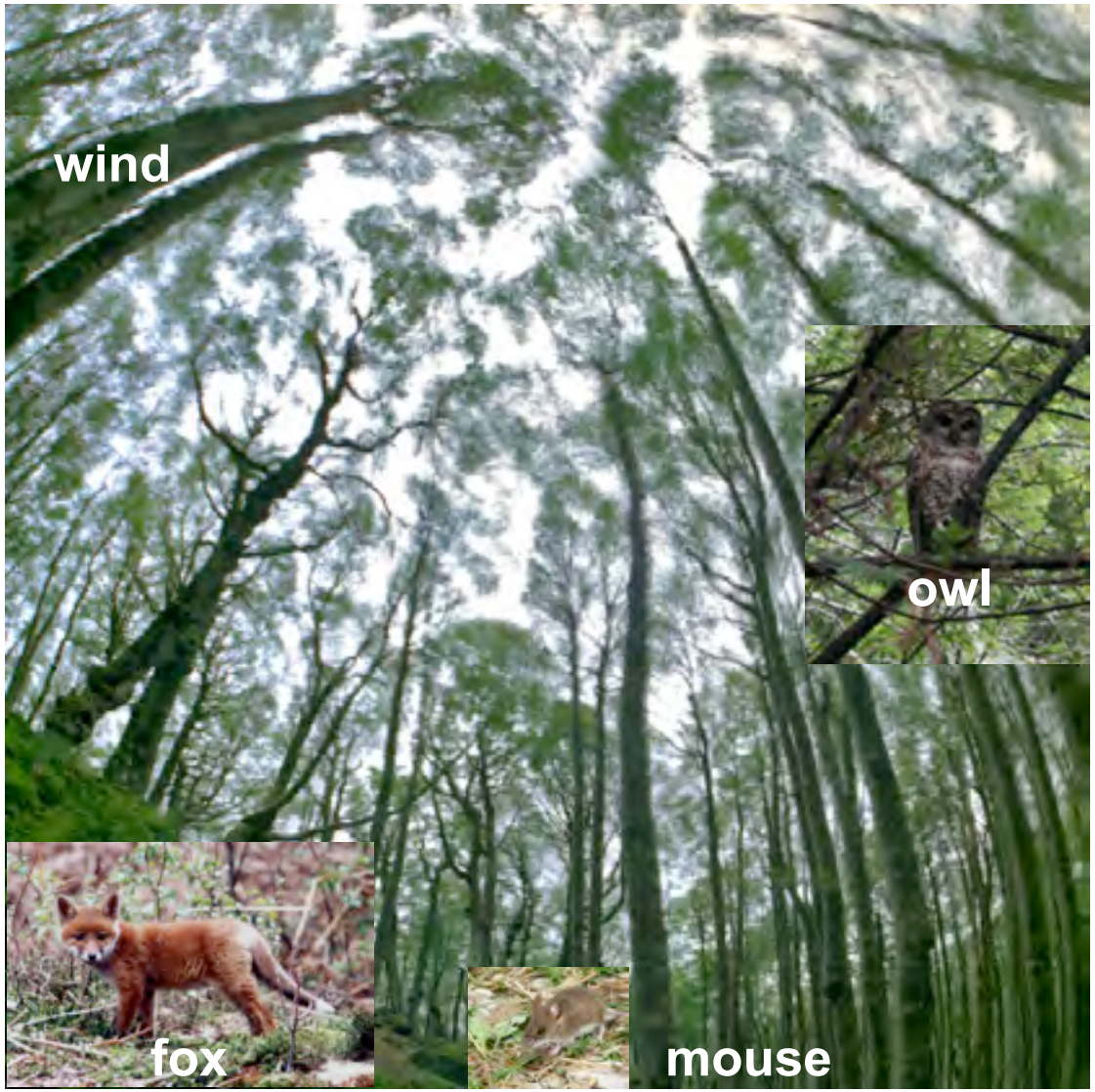
Estimated causes



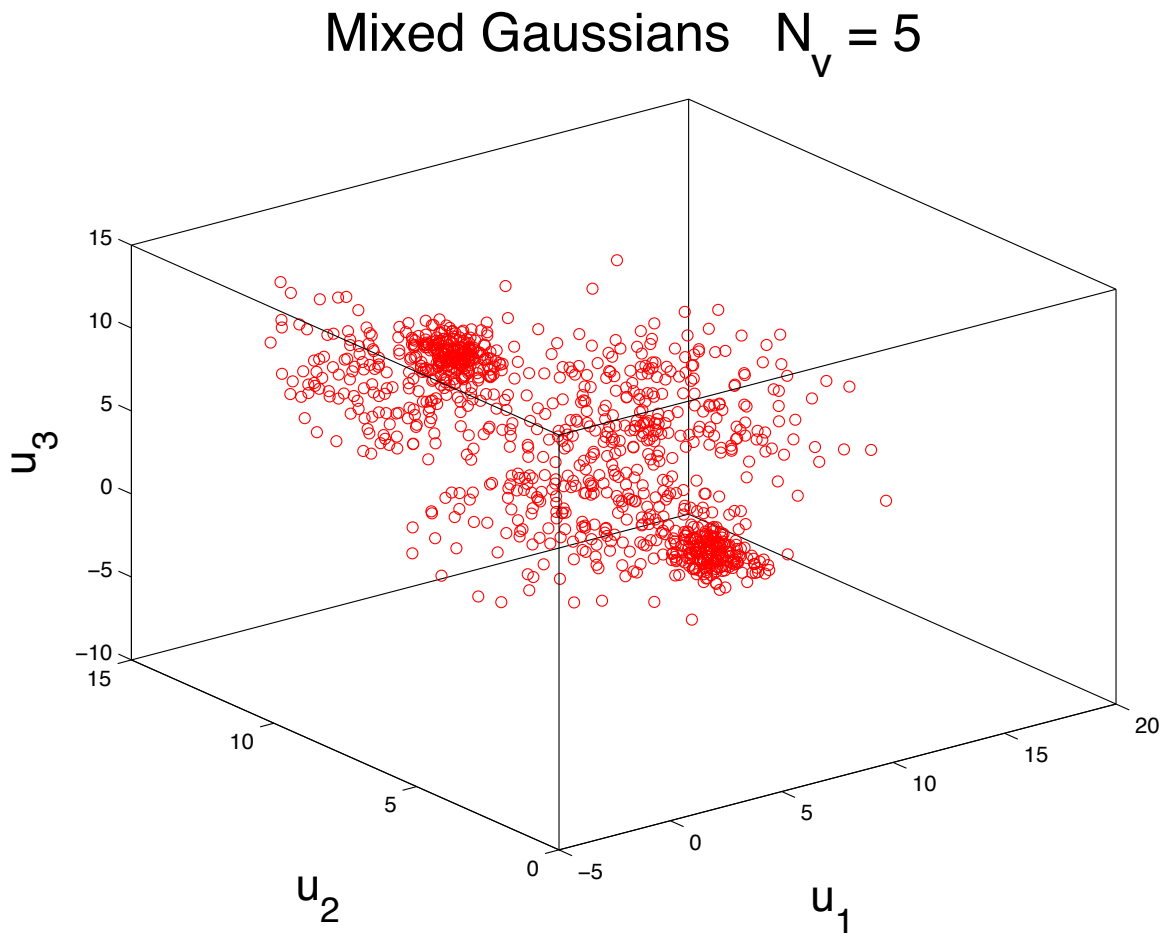
Points to note

- Expectation-maximization (EM) is useful in many situations.
- Here we have fitted observations with two Gaussian causes.
- EM is conceptually subtle, but computationally simple.

5 Continuous causes (optional)

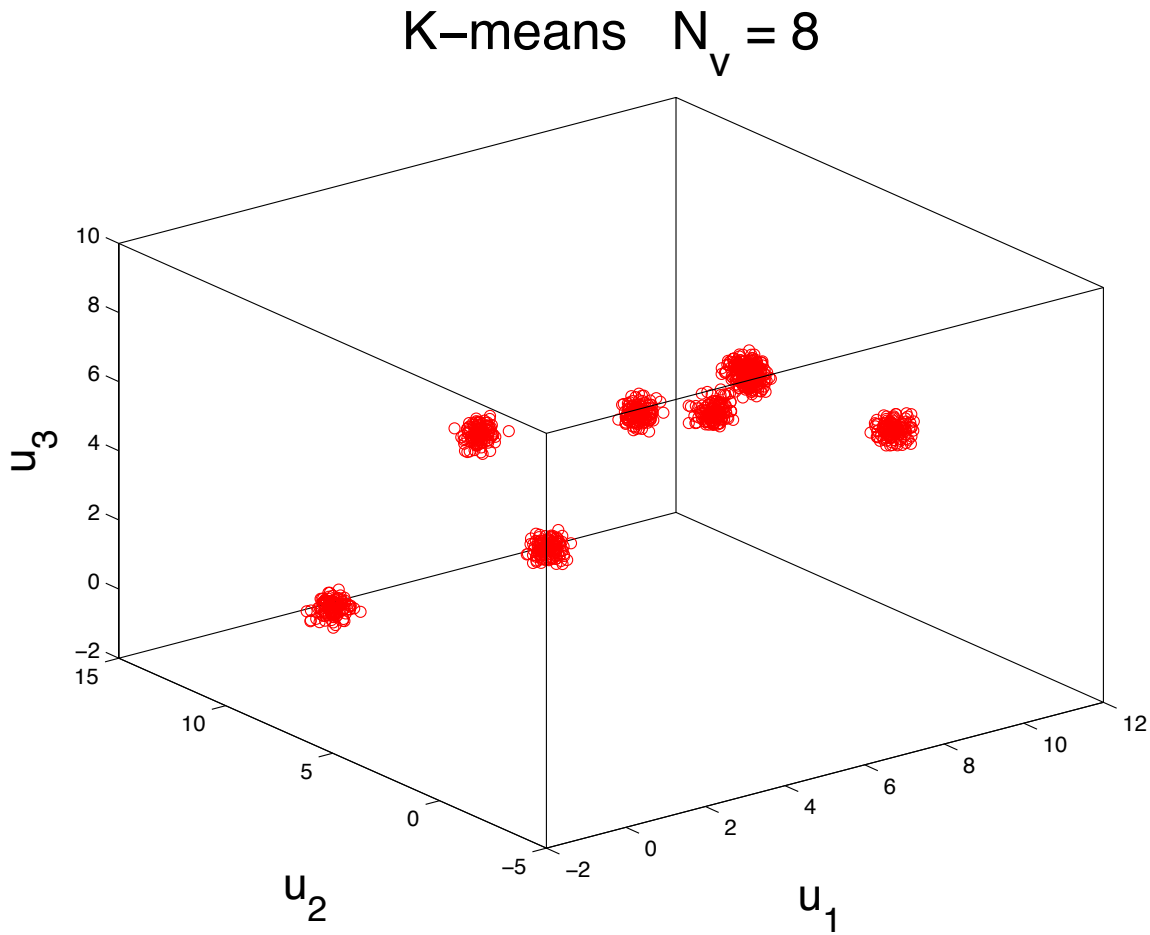


A causal model with N_v separate Gaussian distributions is called a "mixture of Gaussians".



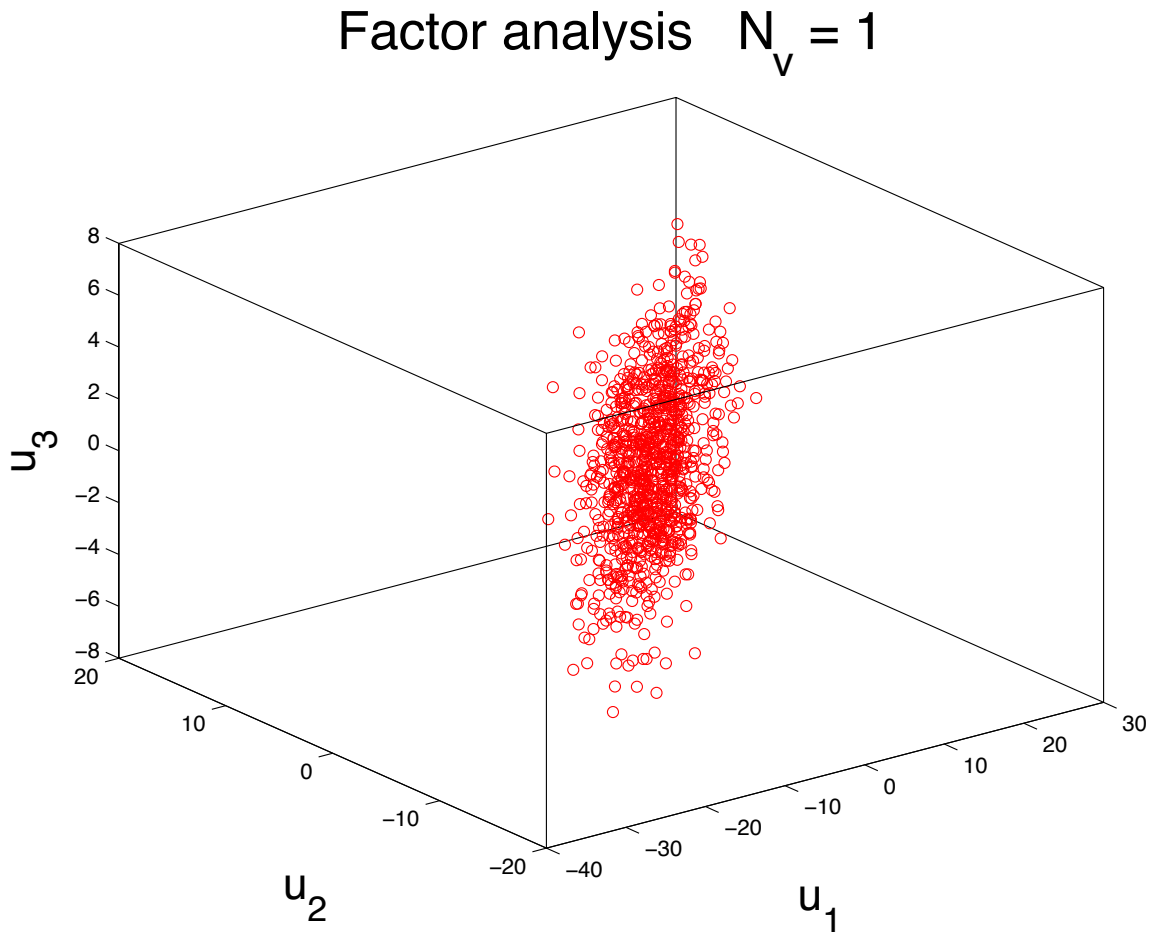
K-means

In the degenerate case of zero variance, the recognition formulation becomes deterministic. This corresponds to the K-means algorithm.



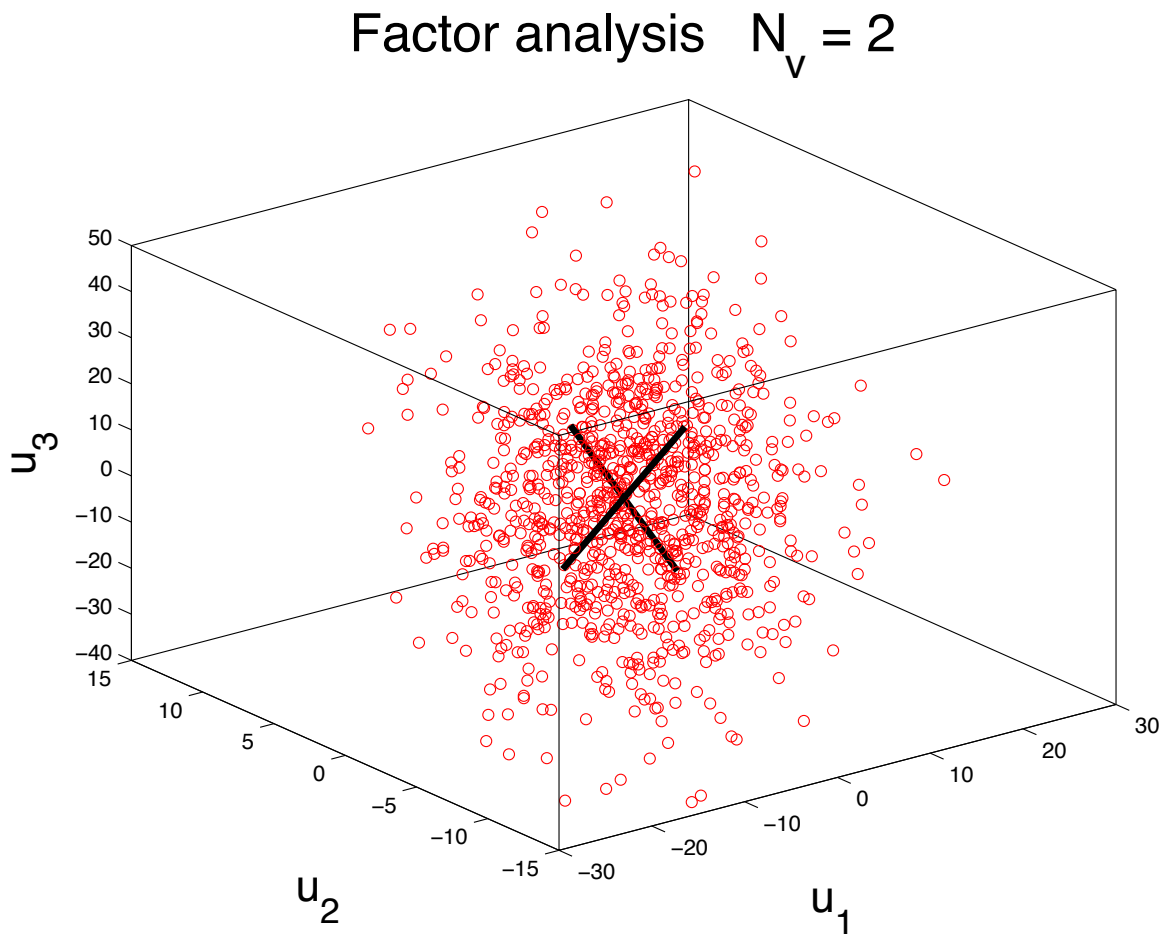
Factor analysis

When causes are distributed continuously, we speak of "factors". In the illustrated "factor", the means of causes are distributed normally along a line.



Factor analysis

In the "factor" illustrated here, the means of causes are Gaussian distributed over a plane.



Points to note

- More realistic causes exhibit variable statistics.
- This can be captured with *factor analysis*.
- Special cases of *factor analysis* are *Principle components analysis (PCA)* and *Independent components analysis (ICA)*

Next:
Principal component analysis