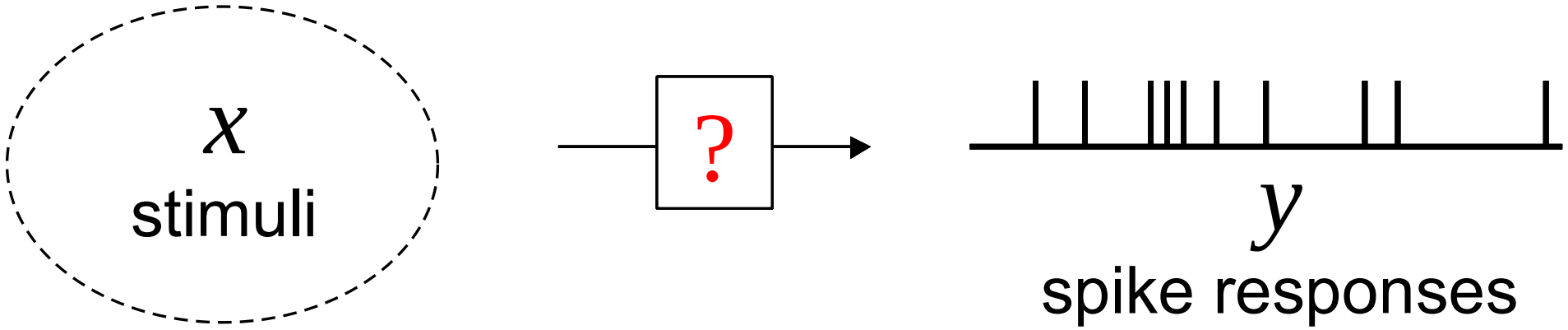
Statistical methods for neural decoding

Liam Paninski

Gatsby Computational Neuroscience Unit
University College London
http://www.gatsby.ucl.ac.uk/∼liam
liam@gatsby.ucl.ac.uk

November 9, 2004

Review...



...discussed encoding: $p(spikes \mid \vec{x})$

Decoding

Turn problem around: given spikes, estimate input $\vec{x}$.

What information can be extracted from spike trains

— by "downstream" areas?

— by experimenter?

Optimal design of neural prosthetic devices.

Decoding examples

Hippocampal place cells: how is location coded in populations of cells?

Retinal ganglion cells: what information is extracted from a visual scene and sent on to the brain? What information is discarded?

Motor cortex: how can we extract as much information from a collection of MI cells as possible?

Discrimination vs. decoding

Discrimination: distinguish between one of two alternatives

— e.g., detection of "stimulus" or "no stimulus"

General case: estimation of continuous quantities

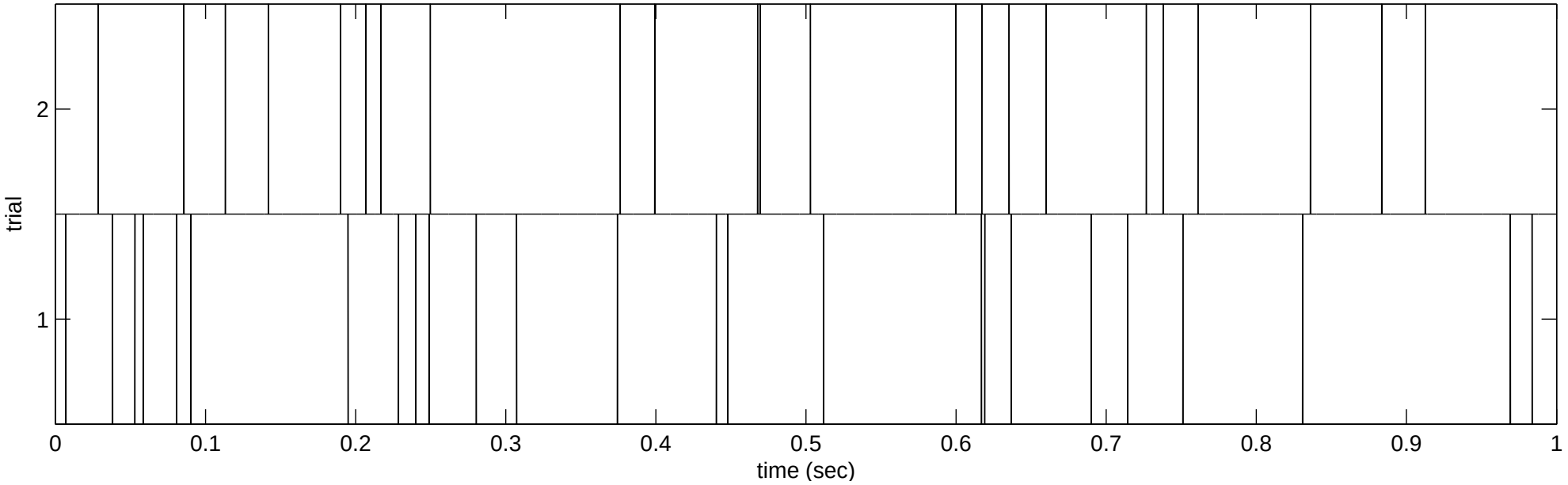
— e.g., stimulus intensity

Same basic problem, but slightly different methods...

Decoding methods: discrimination

Classic problem: stimulus detection.

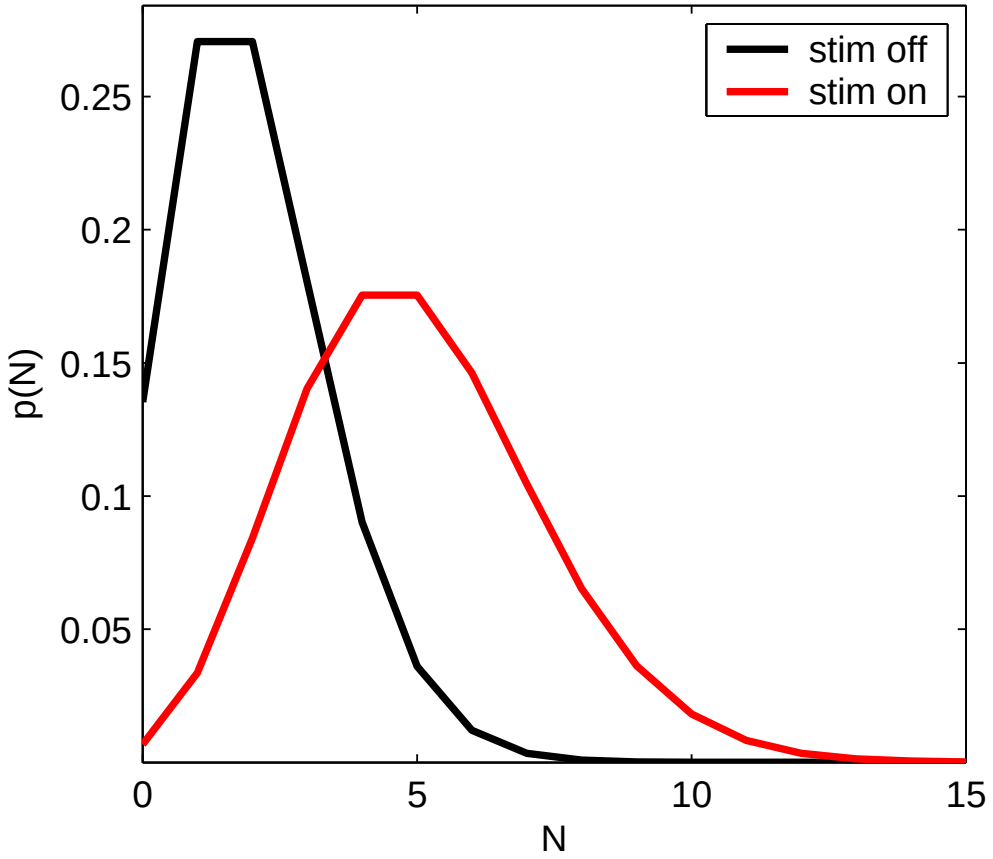
Data:



Was stimulus on or off in trial 1? In trial 2?

Decoding methods: discrimination

Helps to have encoding model $p(N\ spikes|stim)$:



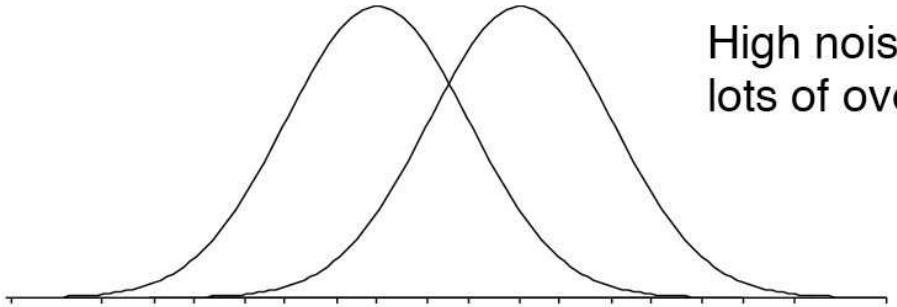
Discriminability

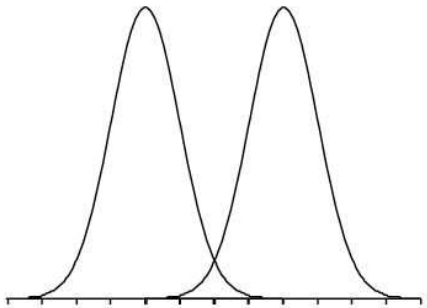
discriminability depends on two factors:

— noise in two conditions

— separation of means

"d' " = separation / spread = signal / noise





Poisson example

Discriminability

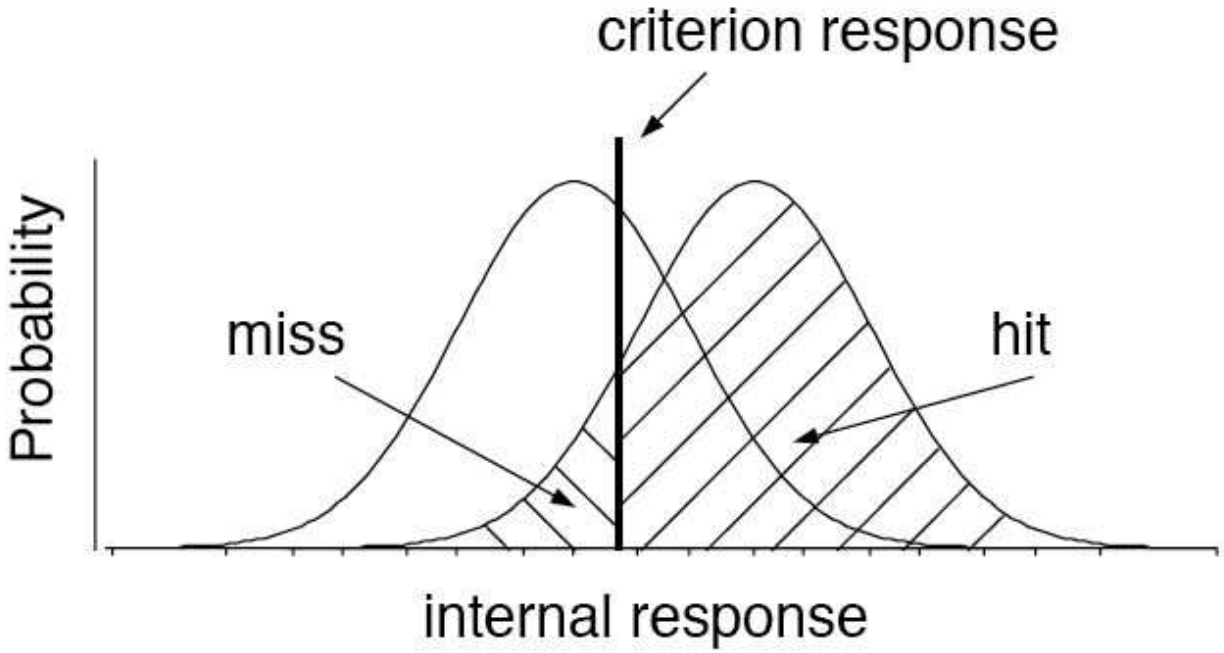
$$d' = \frac{|\mu_0 - \mu_1|}{\sigma} \sim \frac{|\lambda_0 - \lambda_1|}{\sqrt{\lambda}}$$

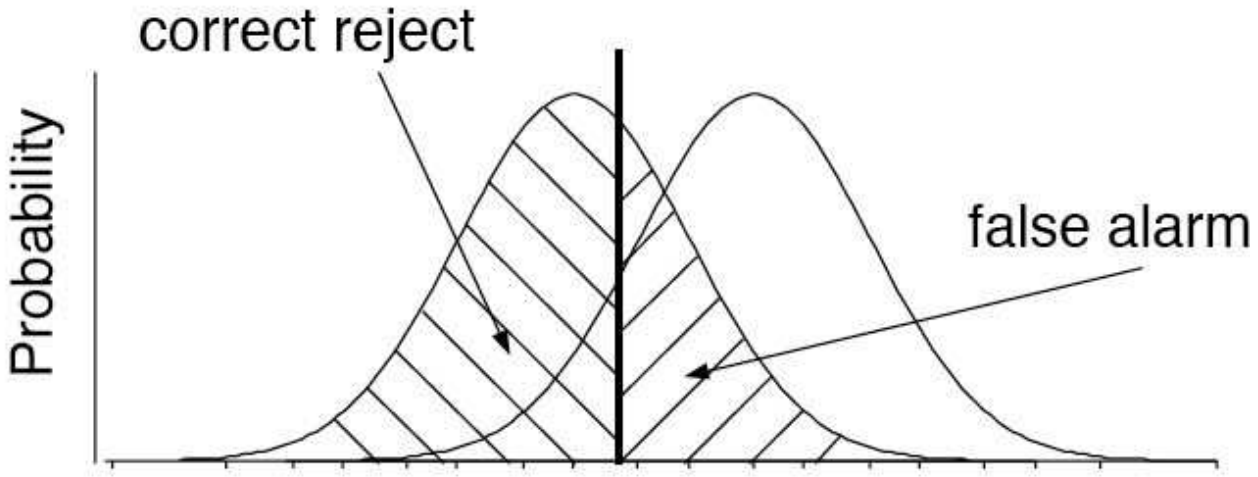
Much easier to distinguish $Poiss(1)$ from $Poiss(2)$ than $Poiss(99)$ from $Poiss(100)$.

— $d' \sim \frac{1}{\sqrt{1}} = 1$ vs. $d' \sim \frac{1}{\sqrt{100}} = .1$

Signal to noise increases like $\frac{|T\lambda_0 - T\lambda_1|}{\sqrt{T\lambda}} \sim \sqrt{T}$: speed-accuracy tradeoff

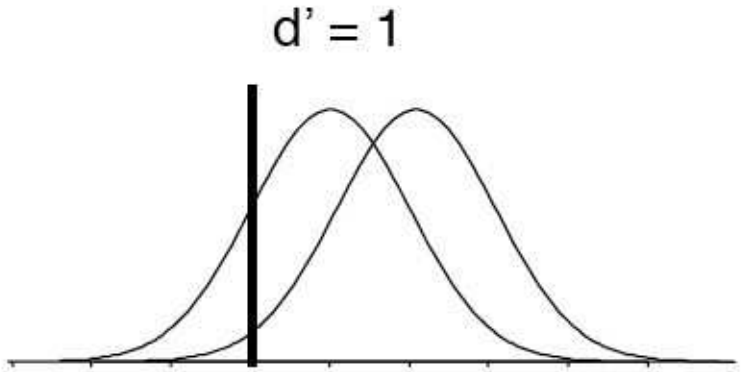
Discrimination errors



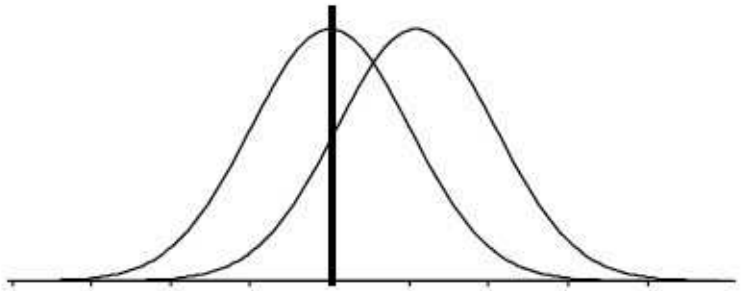


Optimal discrimination

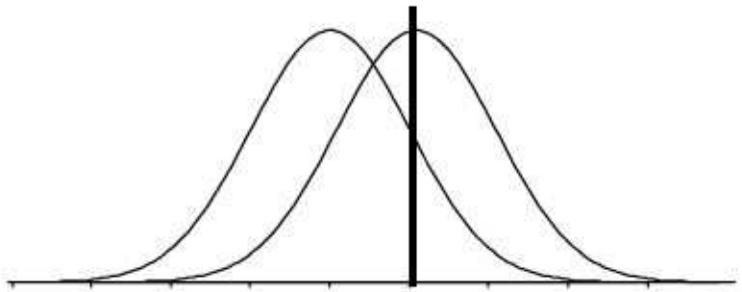
What is optimal? Maximize hit rate or minimize false alarm?



Hits = 97.5%
False alarms = 84%



Hits = 84%
False alarms = 50%



Hits = 50%
False alarms = 16%

Optimal discrimination: decision theory

Write down explicit loss function, choose behavior to minimize expected loss

Two-choice loss $L(\theta, \hat{\theta})$ specified by four numbers:

— $L(0,0)$: correct, $\theta = \hat{\theta} = 0$

— $L(1,1)$: correct, $\theta = \hat{\theta} = 1$

— $L(1,0)$: missed stimulus

— $L(0,1)$: false alarm

Optimal discrimination

Denote $q(data) = p(\hat{\theta} = 1|data)$.

Choose $q(data)$ to minimize $E_{p(\theta|data)}(L(\theta, \hat{\theta})) \sim$

$$\sum_{\theta} p(\theta)p(data|\theta)\Big(q(data)L(\theta, 1) + (1 - q(data))L(\theta, 0)\Big)$$

(Exercise: compute optimal $q(data)$; prove that optimum exists and is unique.)

Optimal discrimination: likelihood ratios

It turns out that optimal

$$q_{opt}(data) = 1\left(\frac{p(data|\theta = 1)}{p(data|\theta = 0)} > T\right) :$$

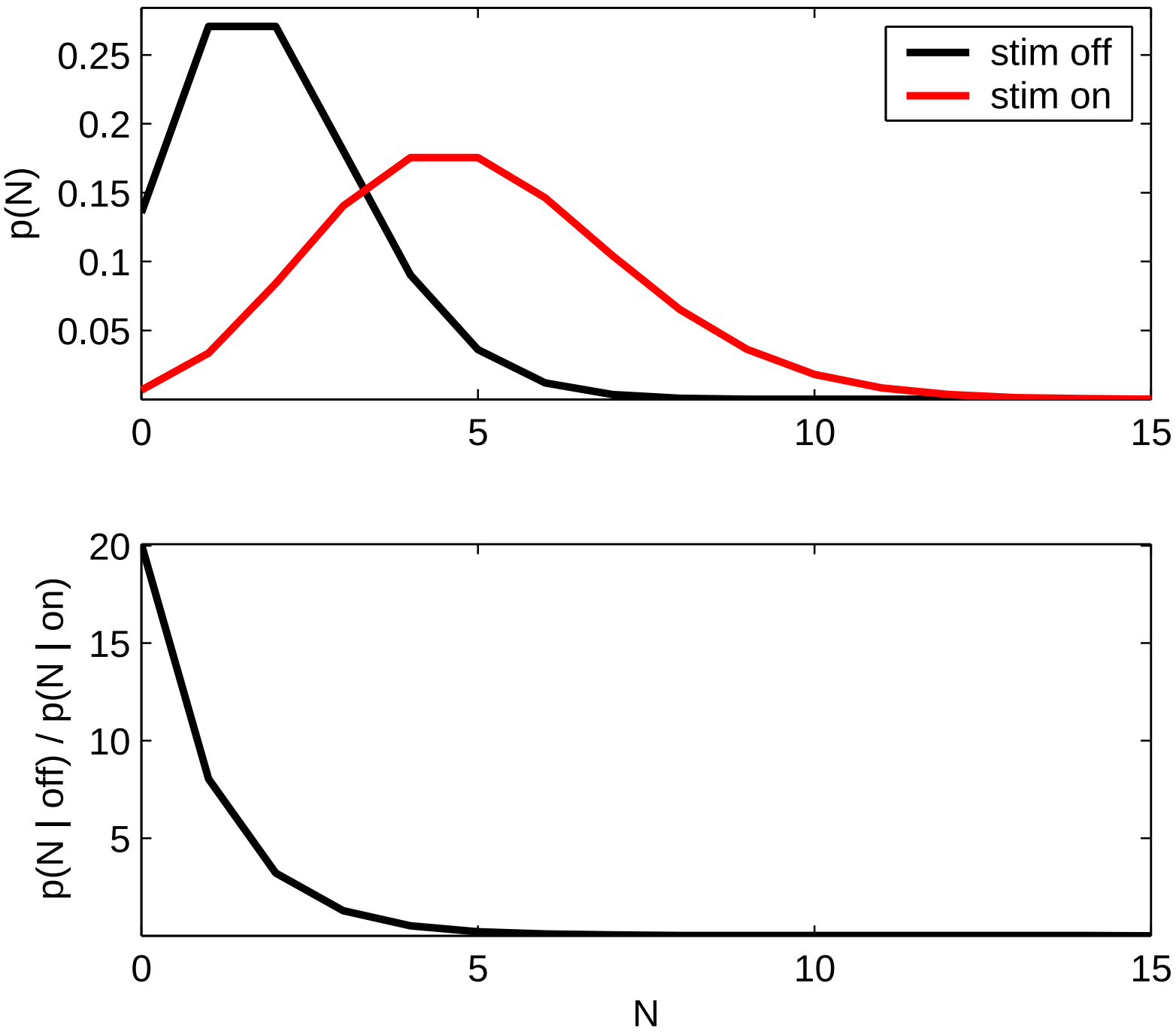
likelihood-based thresholding. Threshold T depends on prior $p(\theta)$ and loss $L(\theta, \hat{\theta})$.

— Deterministic solution: always pick the stimulus with higher weighted likelihood, no matter how close

— All information in data is encapsulated in likelihood ratio.

— Note relevance of encoding model $p(spikes|stim) = p(data|\theta)$

Likelihood ratio



Poisson case: full likelihood ratio

Given spikes at $\{t_i\}$,

$$likelihood = e^{-\int \lambda_{stim}(t)dt} \prod_i \lambda_{stim}(t_i)$$

Log-likelihood ratio:

$$\int (\lambda_1(t) - \lambda_0(t))dt + \sum_i \log \frac{\lambda_0}{\lambda_1}(t_i)$$

Poisson case

$$\int (\lambda_1(t) - \lambda_0(t))dt + \sum_i \log \frac{\lambda_0}{\lambda_1}(t_i)$$

Plug in homogeneous case: $\lambda_j(t) = \lambda_j$.

$$K + N \log \frac{\lambda_0}{\lambda_1}$$

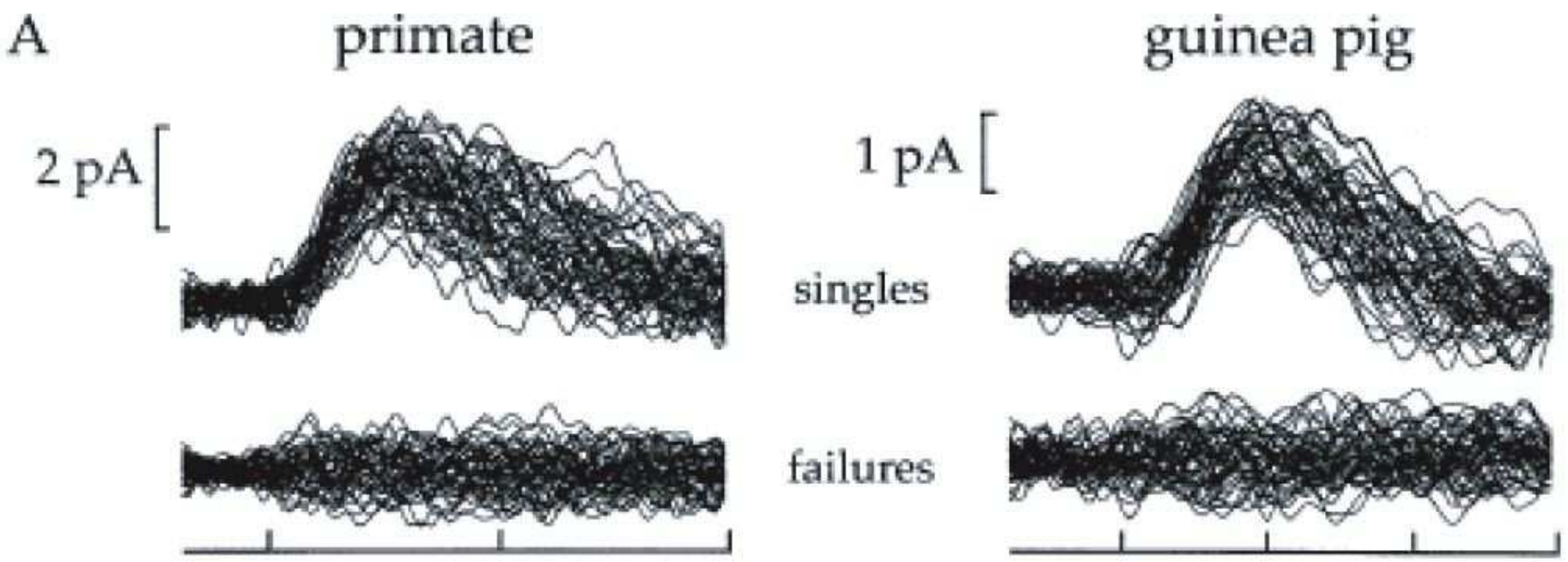
Counting spikes not a bad idea if spikes are really a homogeneous Poisson process; here N = a "sufficient statistic."

— But in general, good to keep track of when spikes arrive.

(The generalization to multiple cells should be clear.)

Discriminability: multiple dimensions

Examples: synaptic failure, photon capture (Field and Rieke, 2002), spike clustering



1-D: threshold separates two means. > 1 D?

Multidimensional Gaussian example

Look at log-likelihood ratio:

$$\log \frac{\frac{1}{Z}\exp[-\frac{1}{2}(\vec{x}-\vec{\mu}_1)^t C^{-1}(\vec{x}-\vec{\mu}_1)]}{\frac{1}{Z}\exp[-\frac{1}{2}(\vec{x}-\vec{\mu}_0)^t C^{-1}(\vec{x}-\vec{\mu}_0)]}$$

$$= \frac{1}{2}[(\vec{x}-\vec{\mu}_0)^t C^{-1}(\vec{x}-\vec{\mu}_0) - (\vec{x}-\vec{\mu}_1)^t C^{-1}(\vec{x}-\vec{\mu}_1)]$$

$$= C^{-1}(\vec{\mu}_1-\vec{\mu}_0)\cdot\vec{x}$$

Likelihood ratio depends on $\vec{x}$ only through projection $C^{-1}(\vec{\mu}_1-\vec{\mu}_0)\cdot\vec{x}$; thus, threshold just looks at this projection, too

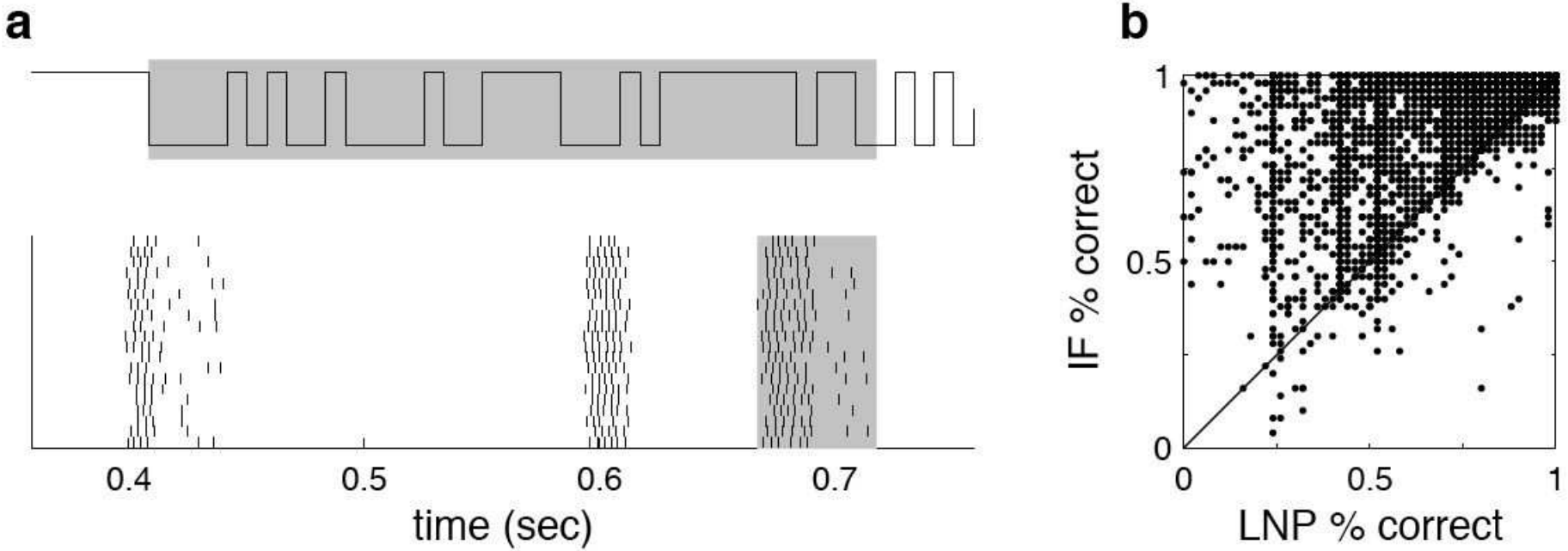
— same regression-like formula we're used to.

C white: projection onto differences of means

What happens when covariance in two conditions is different? (exercise)

Likelihood-based discrimination

Using correct model is essential (Pillow et al., 2004):



— ML methods are only optimal if model describes data well

Nonparametric discrimination

(Eichhorn et al., 2004) examines various classification algorithms from machine learning (SVM, nearest neighbor, Gaussian processes).

Reports significant improvement over "optimal" Bayesian approach under simple encoding models

— errors in estimating encoding model?

— errors in specifying encoding model (not flexible enough)?

Decoding

Continuous case: different cost functions

— mean square error: $L(r, s) = (r - s)^2$

— mean absolute error: $L(r, s) = |r - s|$

Minimizing "mistake" probability makes less sense...

...however, likelihoods will still play an important role.

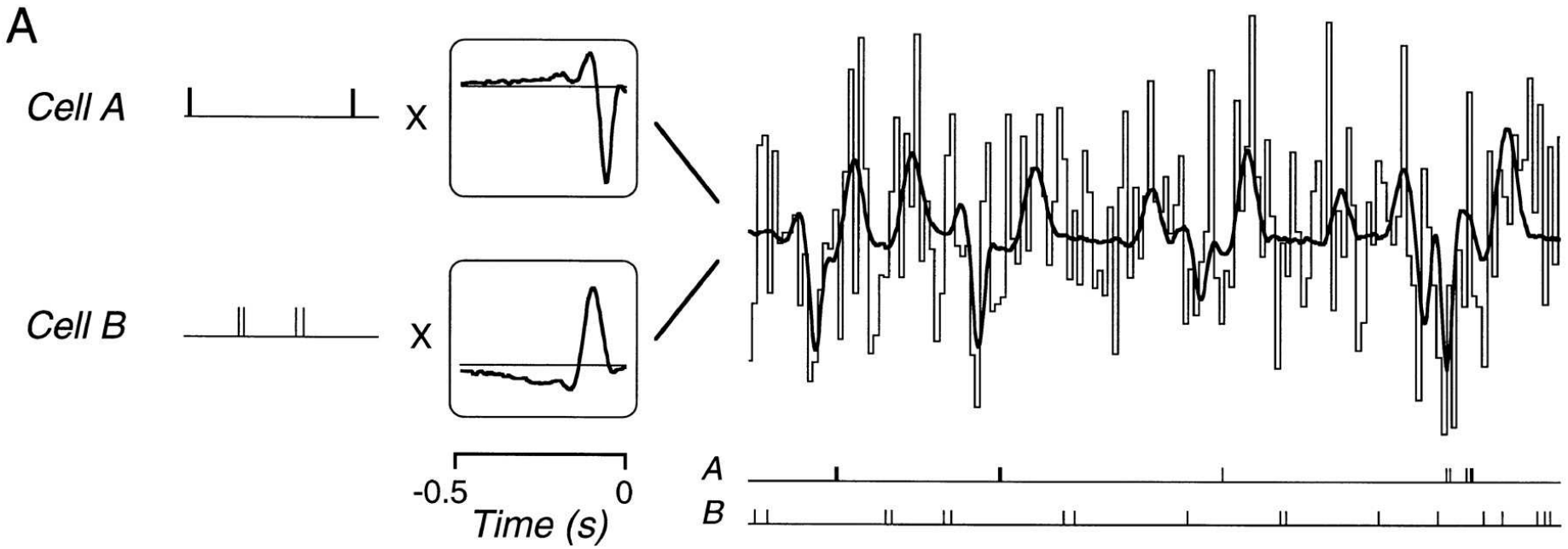
Decoding methods: regression

Standard method: linear decoding.

$$\hat{x}(t) = \sum_i \vec{k}_i * spikes_i(t) + b;$$

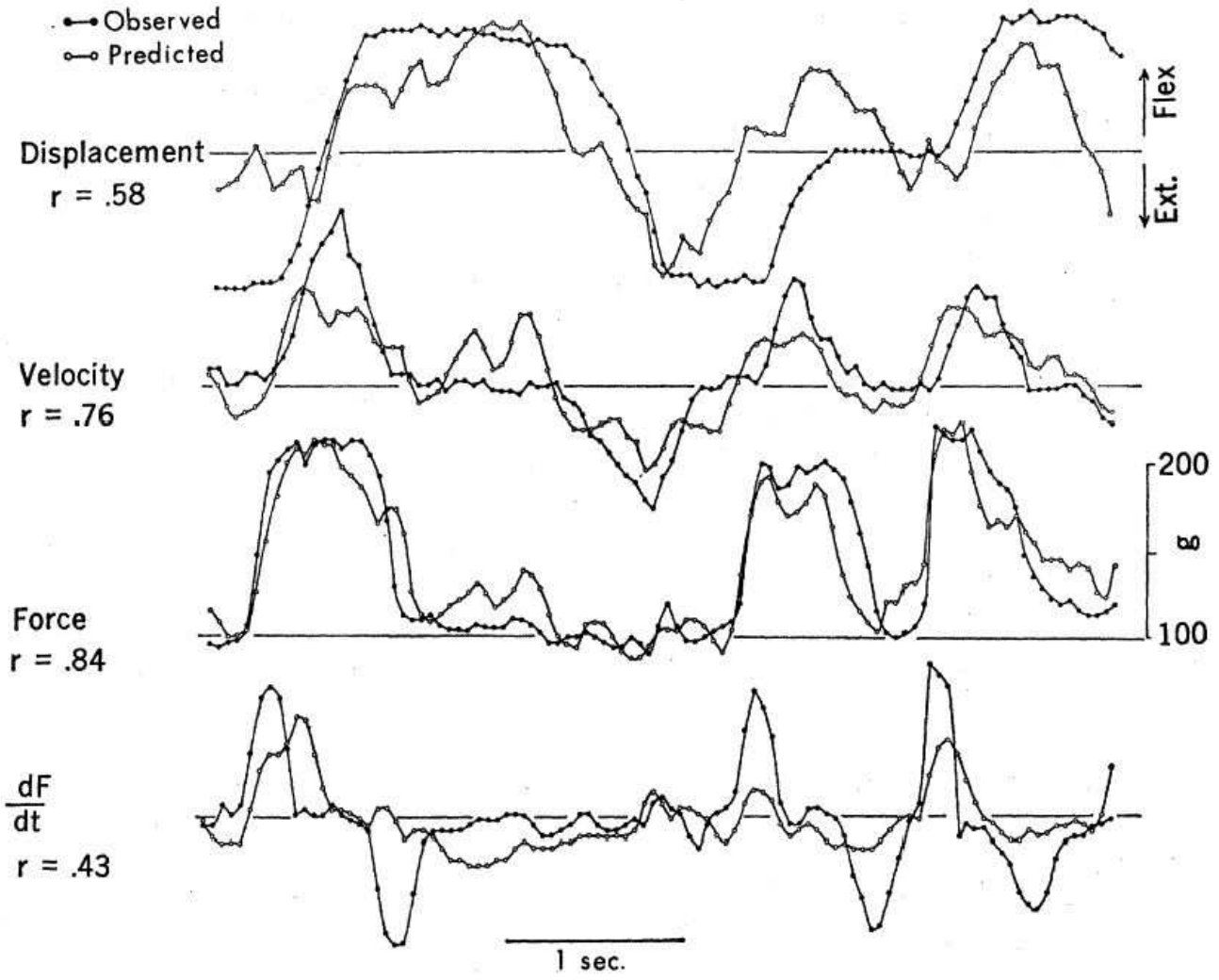
one filter $\vec{k}_i$ for each cell; all chosen together, by regression (with or without regularization)

Decoding sensory information



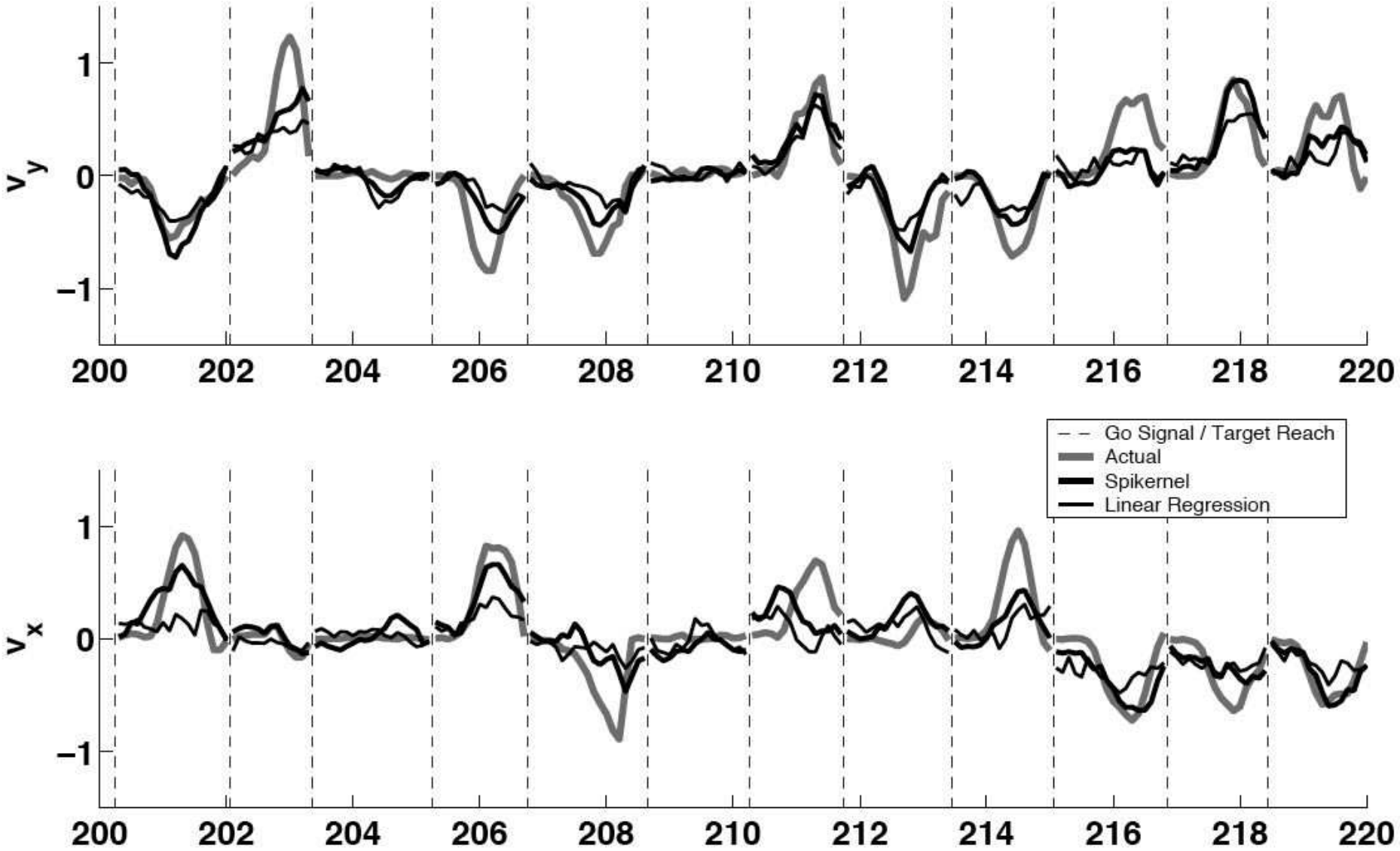
(Warland et al., 1997; Rieke et al., 1997)

Decoding motor information



(Humphrey et al., 1970)

Decoding methods: nonlinear regression



(Shpigelman et al., 2003): 20% improvement by SVMs over linear methods

Bayesian decoding methods

Let's make more direct use of

1) our new, improved neural encoding models, and

2) any prior knowledge about the signal we want to decode

Good encoding model $\implies$ good decoding (Bayes)

Bayesian decoding methods

To form optimal least-mean-square Bayes estimate, take posterior mean given data

(Exercise: posterior mean = LMS optimum. Is this optimum unique?)

Requires that we:

— compute $p(\vec{x}|spikes)$

— perform integral $\int p(\vec{x}|spikes)\vec{x}d\vec{x}$

Computing $p(\vec{x}|spikes)$

Bayes' rule:

$$p(\vec{x}|spikes) = \frac{p(spikes|\vec{x})p(\vec{x})}{p(spikes)}$$

— $p(spikes|\vec{x})$: encoding model

— $p(\vec{x})$: experimenter controlled, or can be modelled (e.g. natural scenes)

— $p(spikes) = \int p(spikes|\vec{x})p(\vec{x})d\vec{x}$

Computing Bayesian integrals

Monte Carlo approach for conditional mean:

— draw samples $\vec{x}_j$ from prior $p(\vec{x})$

— compute likelihood $p(spikes|\vec{x}_j)$

— now form average:

$$\hat{x} = \frac{\sum_j p(spikes|\vec{x}_j)\vec{x}_j}{\sum_j p(spikes|\vec{x}_j)}$$

— confidence intervals obtained in same way

Special case: hidden Markov models

Setup: $x(t)$ is Markov; $\lambda(t)$ depends only on $x(t)$

Examples:

— place cells ($x =$ position)

— IF and escape-rate voltage-firing rate models ($x =$ subthreshold voltage)

Special case: hidden Markov models

How to compute optimal hidden path $\hat{x}(t)$?

Need to compute $p(x(t)|\{spikes(0,t)\})$

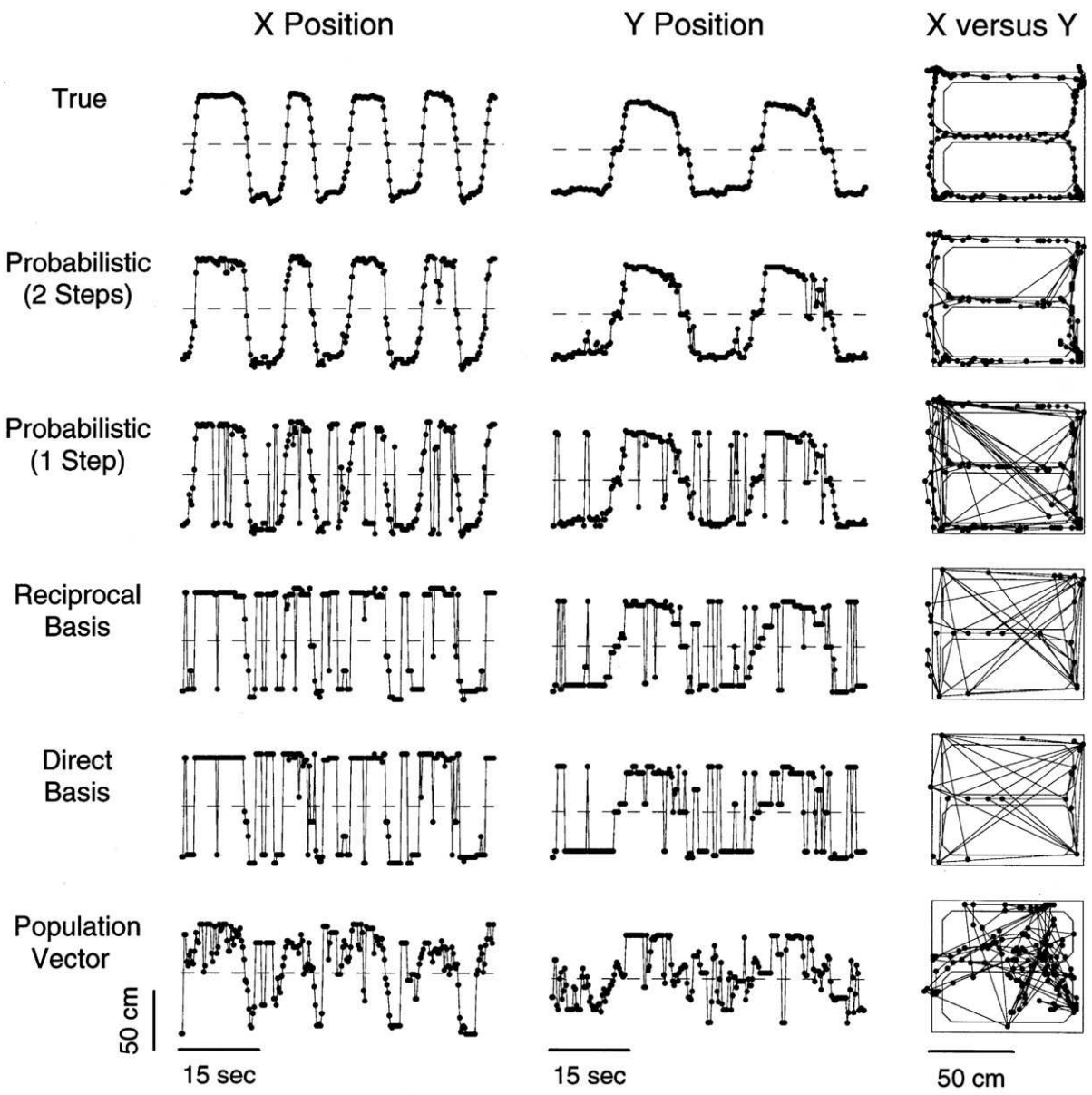
$$p\Big(\{x(0,t)\}\Big|\{spikes(0,t)\}\Big) \sim p\Big(\{spikes(0,t)\}\Big|\{x(0,t)\}\Big)p\Big(\{x(0,t)\}\Big)$$

$$p\Big(\{spikes(0,t)\}\Big|\{x(0,t)\}\Big) = \prod_{0<s<t} p\Big(spikes(s)\Big|x(s)\Big)$$

$$p\Big(\{x(0,t)\}\Big) = \prod_{0<s<t} p\Big(x(s)\Big|x(s-dt)\Big)$$

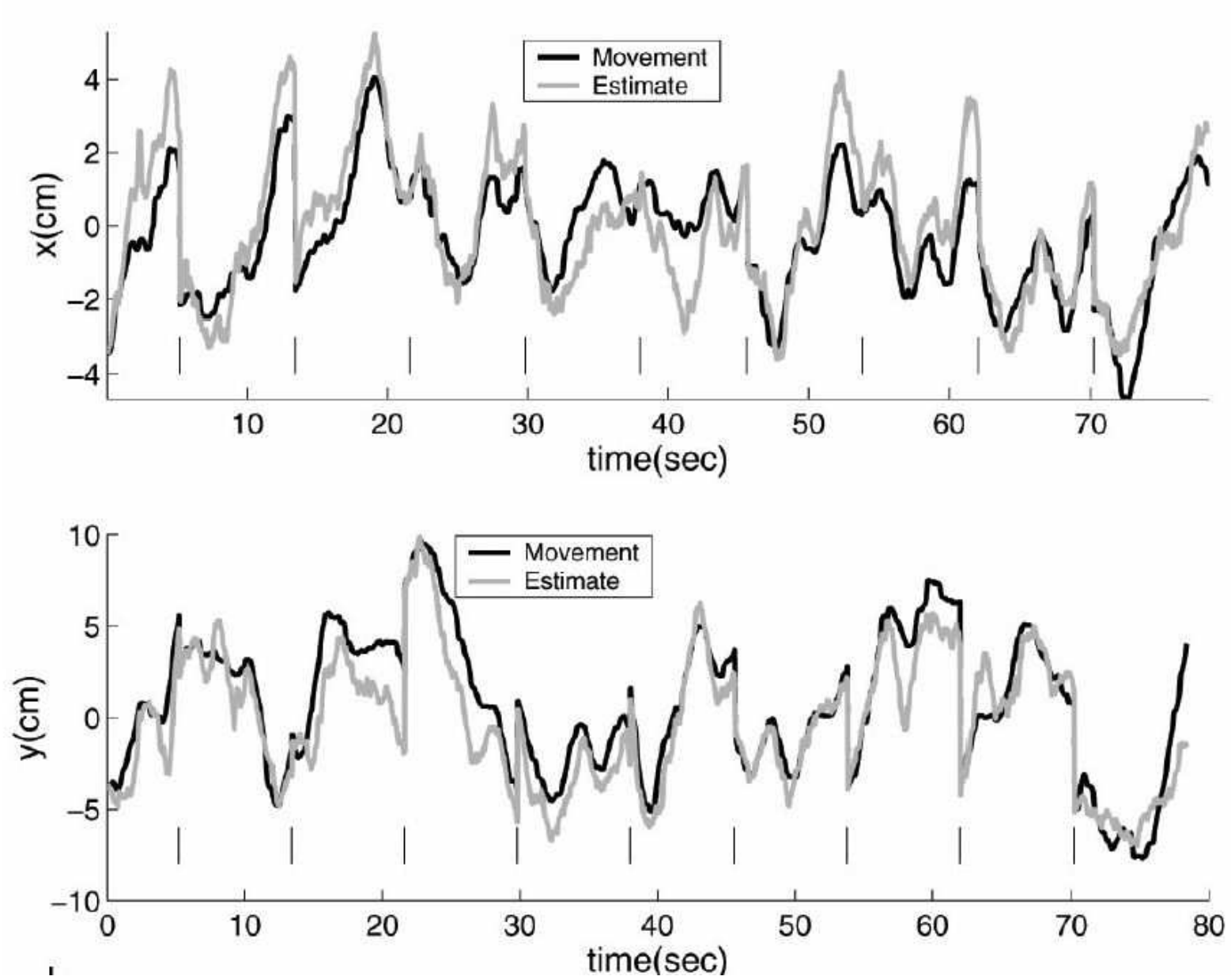
Product decomposition $\Longrightarrow$ fast, efficient recursive methods

Decoding location from HC ensembles



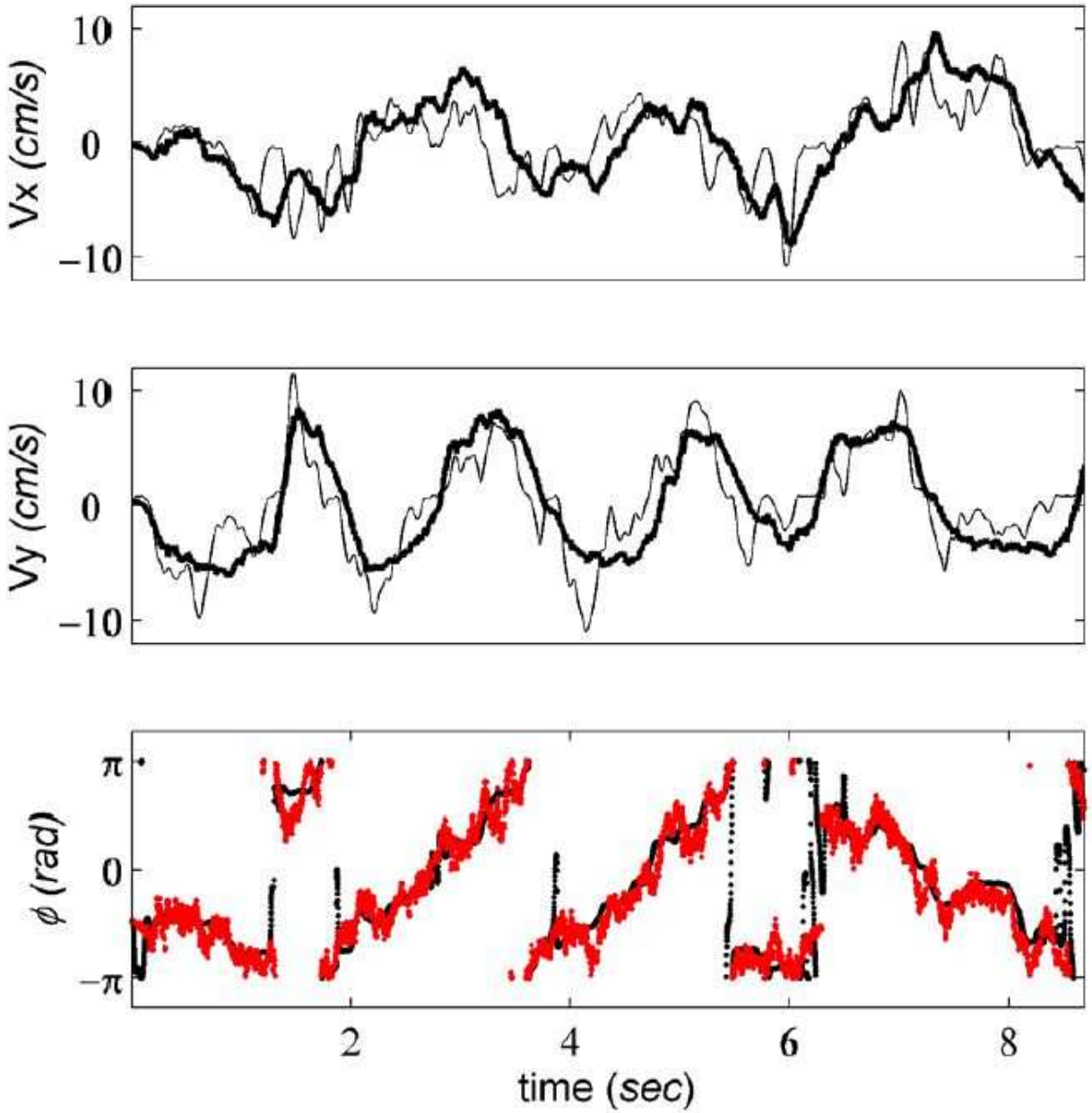
(Zhang et al., 1998; Brown et al., 1998)

Decoding hand position from MI ensembles



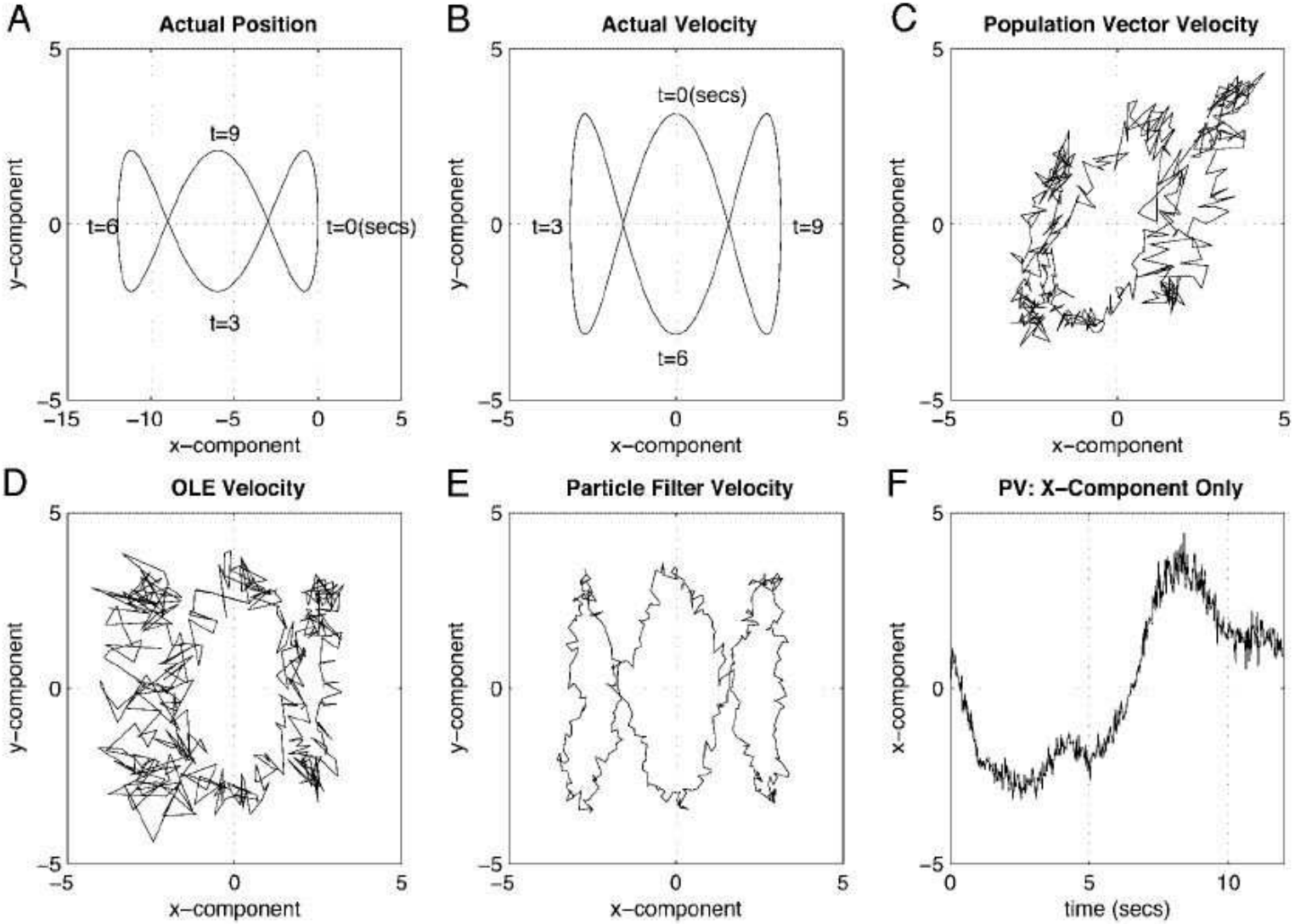
(17 units); (Shoham et al., 2004)

Decoding hand velocity from MI ensembles



(Truccolo et al., 2003)

Comparing linear and Bayes estimates



(Brockwell et al., 2004)

Summary so far

Easy to decode spike trains, once we have a good model of encoding process

Can we get a better analytical handle on these estimators' quality?

— How many neurons do we need to achieve 90% correct?

— What do error distributions look like?

— What is the relationship between neural variability and decoding uncertainty?

Theoretical analysis

Can answer all these questions in asymptotic regime.

Idea: look at case of lots of conditionally independent neurons given stimulus $\vec{x}$. Let the number of cells $N \to \infty$.

We'll see that:

— Likelihood-based estimators are asymptotically Gaussian

— Maximum likelihood solution is asymptotically optimal

— Variance $\approx cN^{-1}$; c set by "Fisher information"

Theoretical analysis

Setup:

— True underlying parameter / stimulus θ.

— Data: lots of cells' firing rates, $\{n_i\}$

— Corresponding encoding models: $p(n_i|\theta)$

Posterior likelihood, given $\vec{n} = \{n_i\}$:

$$\begin{aligned} p(\theta|\vec{n}) &\sim p(\theta)p(\vec{n}|\theta) \\ &= p(\theta)\prod_i p(n_i|\theta) \end{aligned}$$

— Taking logs,

$$\log p(\theta|\vec{n}) = K + \log p(\theta) + \sum_i \log p(n_i|\theta)$$

(note use of conditional independence given θ)

Likelihood asymptotics

$$\log p(\theta|\vec{n}) = K + \log p(\theta) + \sum_i \log p(n_i|\theta)$$

We have a sum of independent r.v.'s $\log p(n_i|\theta)$. Apply law of large numbers:

$$\begin{aligned} \frac{1}{N} \log p(\theta|\vec{n}) &\sim \frac{1}{N} \log p(\theta) + \frac{1}{N} \sum_i \log p(n_i|\theta) \\ &\to 0 + E_{\theta_0} \log p(n|\theta) \end{aligned}$$

Kullback-Leibler divergence

$$\begin{aligned} E_{\theta_0} \log p(n|\theta) &= \int p(n|\theta_0) \log p(n|\theta) \\ &= \int p(n|\theta_0) \log \frac{p(n|\theta)}{p(n|\theta_0)} + K \\ &= -D_{KL}(p(n|\theta_0); p(n|\theta)) + K \end{aligned}$$

$$D_{KL}(p;q) = \int p(n) \log \frac{p(n)}{q(n)}$$

$D_{KL}(p;q)$ is positive unless $p = q$. To see this, use Jensen's inequality (exercise): for any concave function $f(n)$,

$$\int p(n) f(n) \leq f\left(\int p(n) n \right)$$

Likelihood asymptotics

So

$$p(\theta|\vec{n}) \approx \frac{1}{Z} \exp(-N D_{KL}(p(n|\theta_0); p(n|\theta))) :$$

the posterior probability of any $\theta \neq \theta_0$ decays exponentially,
with decay rate =

$$D_{KL}(p(n|\theta_0); p(n|\theta)) > 0 \qquad \forall\ \theta \neq \theta_0.$$

Local expansion: Fisher information

$$p(\theta|\vec{n}) \approx \frac{1}{Z} \exp(-N D_{KL}(p(n|\theta_0); p(n|\theta))).$$

$-D_{KL}(\theta_0; \theta)$ has unique maximum at θ_0

$$\implies \nabla D_{KL}(\theta_0; \theta)\Big|_{\theta=\theta_0} = 0.$$

Second-order expansion:

$$\frac{\partial^2}{\partial \theta^2} D_{KL}(\theta_0; \theta)\Big|_{\theta=\theta_0} = J(\theta_0)$$

$J(\theta_0)$ = curvature of D_{KL} = "Fisher information" at θ_0

Fisher info sets asymptotic variance

So expanding around θ_0,

$$\begin{aligned} p(\theta|\vec{n}) &\approx \frac{1}{Z}\exp(-ND_{KL}(\theta_0;\theta)) \\ &= \frac{1}{Z}\exp(-\frac{N}{2}((\theta-\theta_0)^t J(\theta_0)(\theta-\theta_0)+h.o.t.)) \end{aligned}$$

i.e., posterior likelihood $\approx$ Gaussian with mean θ_0, covariance

$$\frac{1}{N}J(\theta_0)^{-1}.$$

More expansions

What about mean? Depends on $h.o.t.$, but we know it's close to θ_0, because posterior decays exponentially everywhere else.

How close? Try expanding $f_i(\theta) = \log p(n_i|\theta)$:

$$\begin{aligned}
&\sum_i f_i(\theta) \\
&\quad\approx \quad \sum_i f_i(\theta_0) + \sum_i \nabla f_i(\theta)\Big|_{\theta_0}^t (\theta - \theta_0) + \frac{1}{2}\sum_i (\theta-\theta_0)^t \frac{\partial^2 f_i(\theta)}{\partial\theta^2}\Big|_{\theta_0} (\theta-\theta_0) \\
&\quad\approx \quad K_N + \sum_i \nabla f_i(\theta)\Big|_{\theta_0}^t (\theta - \theta_0) - \frac{1}{2} N(\theta-\theta_0)^t J(\theta_0)(\theta-\theta_0)
\end{aligned}$$

Likelihood asymptotics

Look at

$$\nabla f_i(\theta)\Big|_{\theta_0}$$

Random vector with mean zero and covariance $J(\theta_0)$ (exercise).

So apply central limit theorem:

$$G_N = \sum_i \nabla f_i(\theta)\Big|_{\theta_0}$$

is asymptotically Gaussian, with mean zero and covariance $NJ(\theta_0)$.

Likelihood asymptotics

So $\log p(n_i|\theta)$ looks like a random upside-down bowl-shaped function:

$$\log p(n_i|\theta) \approx K_N + G_N^t(\theta - \theta_0) - \frac{1}{2}N(\theta - \theta_0)^t J(\theta_0)(\theta - \theta_0).$$

Curvature of bowl is asymptotically deterministic: $-\frac{N}{2}J(\theta_0)$.

Bottom of bowl (i.e., MLE) is random, asymptotically Gaussian with mean θ_0 and variance $(NJ(\theta_0))^{-1}$ (exercise)

MLE optimality: Cramer-Rao bound

MLE is asymptotically unbiased (mean $= \theta_0$), with variance $(NJ(\theta_0))^{-1}$.

It turns out that this is the best we can do.

Cramer-Rao bound: any unbiased estimator $\hat{\theta}$ has variance

$$V(\hat{\theta}) \geq (NJ(\theta_0))^{-1}.$$

So MLE is asymptotically optimal.

Summary of asymptotic analysis

Quantified how much information we can expect to extract from neuronal populations

Introduced two important concepts: D_{KL} and Fisher info

Obtained a clear picture of how MLE and posterior distributions (and by extension Bayesian estimators — minimum mean-square, minimum absolute error, etc.) behave

Coming up...

Are populations of cells optimized for maximal (Fisher) information?

What about correlated noise? Interactions between cells?

Broader view (non estimation-based): information theory

References

Brockwell, A., Rojas, A., and Kass, R. (2004). Recursive Bayesian decoding of motor cortical signals by particle filtering. *Journal of Neurophysiology*, 91:1899–1907.

Brown, E., Frank, L., Tang, D., Quirk, M., and Wilson, M. (1998). A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. *Journal of Neuroscience*, 18:7411–7425.

Eichhorn, J., Tolias, A., Zien, A., Kuss, M., Rasmussen, C., Weston, J., Logothetis, N., and Schoelkopf, B. (2004). Prediction on spike data using kernel algorithms. *NIPS*, 16.

Field, G. and Rieke, F. (2002). Mechanisms regulating variability of the single photon responses of mammalian rod photoreceptors. *Neuron*, 35:733–747.

Humphrey, D., Schmidt, E., and Thompson, W. (1970). Predicting measures of motor performance from multiple cortical spike trains. *Science*, 170:758–762.

Pillow, J., Paninski, L., Uzzell, V., Simoncelli, E., and Chichilnisky, E. (2004). Accounting for timing and variability of retinal ganglion cell light responses with a stochastic integrate-and-fire model. *SFN Abstracts*.

Rieke, F., Warland, D., de Ruyter van Steveninck, R., and Bialek, W. (1997). *Spikes: Exploring the neural code*. MIT Press, Cambridge.

Shoham, S., Fellows, M., Hatsopoulos, N., Paninski, L., Donoghue, J., and Normann, R. (2004). Optimal decoding for a primary motor cortical brain-computer interface. *Under review, IEEE Transactions on Biomedical Engineering*.

Shpigelman, L., Singer, Y., Paz, R., and Vaadia, E. (2003). Spikernels: embedding spike neurons in inner product spaces. *NIPS*, 15.

Truccolo, W., Eden, U., Fellows, M., Donoghue, J., and Brown, E. (2003). Multivariate conditional intensity models for motor cortex. *Society for Neuroscience Abstracts*.

Warland, D., Reinagel, P., and Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *Journal of Neurophysiology*, 78:2336–2350.

Zhang, K., Ginzburg, I., McNaughton, B., and Sejnowski, T. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *Journal of Neurophysiology*, 79:1017–1044.