Statistical inference for errorfully observed graphs

Carey E. Priebe
Department of Applied Mathematics & Statistics
Johns Hopkins University

2012 Graph Exploitation Symposium
MIT Lincoln Laboratory
April 2012

Connectome Inference

Motivation

The connections made by cortical brain cells are anatomically nanoscopic, yet each cell in the cortex has several centimeters of local anatomical wiring. This wiring packs the cortical volume essentially completely. Bock et al. [Nature, 2011] recently characterized the in vivo responses of a group of cells in mouse visual cortex, then imaged a volume of brain containing the cells using a custom-built high throughput electron microscopy (EM) camera array. Each voxel in the resulting data set occupies about 4 x 4 x 45 cubic nanometers of brain; the 10 teravoxel volume spans 450 x 350 x 50 cubic micrometers. The imaged volume is of sufficient size and resolution that they were able to trace the local connectivity of the physiologically characterized cells. One can therefore record what cells in the brain are doing and then trace their connectivity - a combination which could enable a new level of understanding of cortical circuits to be achieved.

Davi Bock & Joshua Vogelstein

PVB2011 available at http://arxiv.org/pdf/1108.6271

PVB2011

Abstract

We demonstrate a meaningful prospective power analysis for an (admittedly idealized) illustrative connectome inference task. Modeling neurons as vertices and synapses as edges in a simple random graph model, we optimize the trade-off between the number of (putative) edges identified and the accuracy of the edge identification procedure. We conclude that explicit analysis of the quantity/quality trade-off is imperative for optimal neuroscientific experimental design. In particular, identifying edges faster/more cheaply, but with more error, can yield superior inferential performance.

This page deliberately left blank.

Connectome Inference

Model & Hypotheses

Block model structure: $\mathbb{G}$

Vertices represent neurons; edges represent synapses.

$\mathcal{E}$: the collection of n_E excitatory neurons

$\mathcal{I}$: the collection of n_I inhibitory neurons.

Let $n = |V| = |\mathcal{E}| + |\mathcal{I}| = n_E + n_I$; $n_E = \lambda n$ and $n_I = (1-\lambda)n$.

$$\begin{aligned} P[u \sim v] &= p_{EE} \quad \text{for} \quad u, v \in \mathcal{E}, \\ P[u \sim v] &= p_{II} \quad \text{for} \quad u, v \in \mathcal{I}, \\ P[u \sim v] &= p_{EI} = p_{IE} \quad \text{otherwise.} \end{aligned}$$

Hypotheses:

$$\begin{aligned} H_0 &: \quad p_{EE} = p_{II} = p_{EI} \\ H_A &: \quad p_{EE} = p_{II} < p_{EI} \ . \end{aligned}$$

Connectome Inference

Data

For $i = 1, \cdots, z$, we define the random variable X_i representing a perfect edge observation via the "tracing algorithm" given by

(1) a neuron: choose a vertex v_i uniformly at random from V.

(2) a synapse: choose an edge $v_i \sim \cdot$ uniformly at random from among edges incident to v_i.

(3) the post-synaptic neuron: identify vertex w_i for $v_i \sim w_i$.

(4) the nature of the synapse: $X_i = I\{v_i, w_i \in \mathcal{E} \quad \text{or} \quad v_i, w_i \in \mathcal{I}\}$.

Connectome Inference

Data

For $i = 1, \cdots, z$, we define the random variable X_i representing a putative edge observation via the "tracing algorithm" given by

(1) a neuron: choose a vertex v_i uniformly at random from V.

(2) a synapse: choose an edge $v_i \sim \cdot$ uniformly at random from among edges incident to v_i.

(3') the post-synaptic neuron: identify $\widetilde{w}_i$ for $v_i \sim \widetilde{w}_i$; with probability $(1 - \varepsilon)$ $\widetilde{w}_i = w_i$, otherwise $\widetilde{w}_i$ is random.

(4') the nature of the synapse: $\widetilde{X}_i = I\{v_i, \widetilde{w}_i \in \mathcal{E} \quad \text{or} \quad v_i, \widetilde{w}_i \in \mathcal{I}\}$.

Connectome Inference

Inference

$$
\begin{aligned}
P[\widetilde{X}_i = 1] = p_{\widetilde{X}} &= p_{\widetilde{X}}(n, \lambda, p_{EE}, p_{EI}, \varepsilon) \\
&= (1-\varepsilon)\left(\frac{\lambda p_{EE} n_E}{p_{EE} n_E + p_{EI} n_I} + \frac{(1-\lambda) p_{II} n_I}{p_{II} n_I + p_{IE} n_E}\right) \\
&\quad + \varepsilon(2\lambda^2 - 2\lambda + 1)
\end{aligned}
$$

Since we have (approximately) independent random variables $\widetilde{X}_i \sim Bernoulli(p_{\widetilde{X}})$, we reject for small values of the test statistic $\overline{\widetilde{X}}_z = \frac{1}{z}\sum_{i=1}^{z} \widetilde{X}_i$ based on having observed z errorful edges.

$$
\begin{aligned}
P[\overline{\widetilde{X}}_z < c_\alpha | H_A] = \beta_{z,\varepsilon} &= \beta_{z,\varepsilon}(n, \lambda, p_{EE}, p_{EI}; \alpha) \\
&= \Phi\left(\frac{p^0_{\widetilde{X}}(1 - p^0_{\widetilde{X}})\Phi^{-1}(\alpha) + \sqrt{z}(p^0_{\widetilde{X}} - p^A_{\widetilde{X}})}{p^A_{\widetilde{X}}(1 - p^A_{\widetilde{X}})}\right)
\end{aligned}
$$

Connectome Inference

Example

With parameter values $n = 10000$, $\lambda = 0.9$, $p_{EE} = p_{II} = 0.1$, and $p_{EI} = 0.2$ (H_A holds) for the random graph model $\mathbb{G}$, testing at level $\alpha = 0.05$ yields

$$\begin{aligned} \beta_{50,0} &\approx 0.429, \\ \beta_{50,0.5} &\approx 0.196, \\ \beta_{250,0.5} &\approx 0.488. \end{aligned}$$

Connectome Inference

Example

The power $\beta(\varepsilon)$ obtained when using the edge tracing algorithm engineered to produce $z = h(\varepsilon)$ putative edges with edge tracing error ε is given by $\beta(\varepsilon) = \Phi(g(\varepsilon))$ where

$$g(\varepsilon) = \frac{p^0_{\widetilde{X}}(\varepsilon)(1 - p^0_{\widetilde{X}}(\varepsilon))\Phi^{-1}(\alpha) + h(\varepsilon)^{1/2}(p^0_{\widetilde{X}}(\varepsilon) - p^A_{\widetilde{X}}(\varepsilon))}{p^A_{\widetilde{X}}(\varepsilon)(1 - p^A_{\widetilde{X}}(\varepsilon))}.$$

Connectome Inference

Example

Assuming that h is differentiable with respect to ε on $[0, 1)$, we obtain

$$\frac{\partial \beta}{\partial \varepsilon} = \phi(g(\varepsilon))g'(\varepsilon).$$

- $\frac{\partial \beta}{\partial \varepsilon}|_{\varepsilon=\varepsilon_0} > 0$ implies less expensive more errorful (larger ε) edge tracing (resulting in larger z) will yield increased power.

- $\frac{\partial \beta}{\partial \varepsilon}|_{\varepsilon=\varepsilon_0} < 0$ implies that inference will improve with more accurate but more expensive edge tracing (resulting in fewer putative edges).

- Finding $\varepsilon^\star$ such that $\frac{\partial \beta}{\partial \varepsilon}|_{\varepsilon=\varepsilon^\star} = 0$ will (after checking appropriate side conditions) yield optimal power $\beta^\star = \beta(\varepsilon^\star)$.

Connectome Inference

Example

Illustration:

$$z = h(\varepsilon) = 50 + \frac{200}{\sin(\pi/4)} \sin(\varepsilon\pi/2),$$

designed to give $h(0) = 50$, $\beta(0) \approx 0.429$ and $h(1/2) = 250$, $\beta(1/2) \approx 0.488$ for consistency with our running example.

Connectome Inference

Example

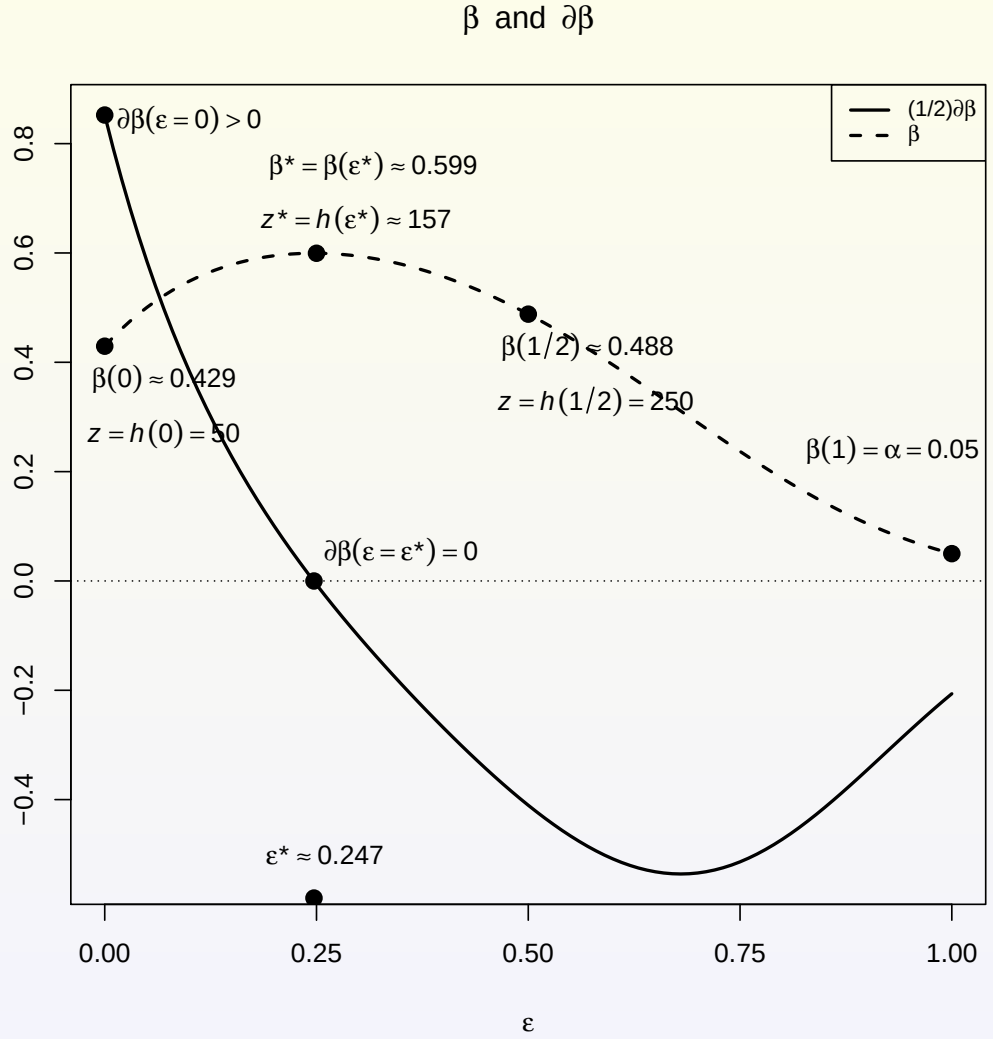


Figure: Power and its derivative

Connectome Inference

Extension

The significance of our "admittedly idealized" illustrative setting is a simple version of a general question of scientific interest: how does connectivity probability depend on the neurons in question? Real scientific interest lies in more elaborate graph models and hypotheses – $K > 2$ kinds of cells and K^2 connection probabilities, or even an unknown number of cell types. The method described here can be generalized to these more realistic settings – some maintaining analytic tractability, but many realistic complex generalizations will of course require us to resort to numerical approximation methods.

Daniel Sussman & Minh Tang & Donniell Fishkind

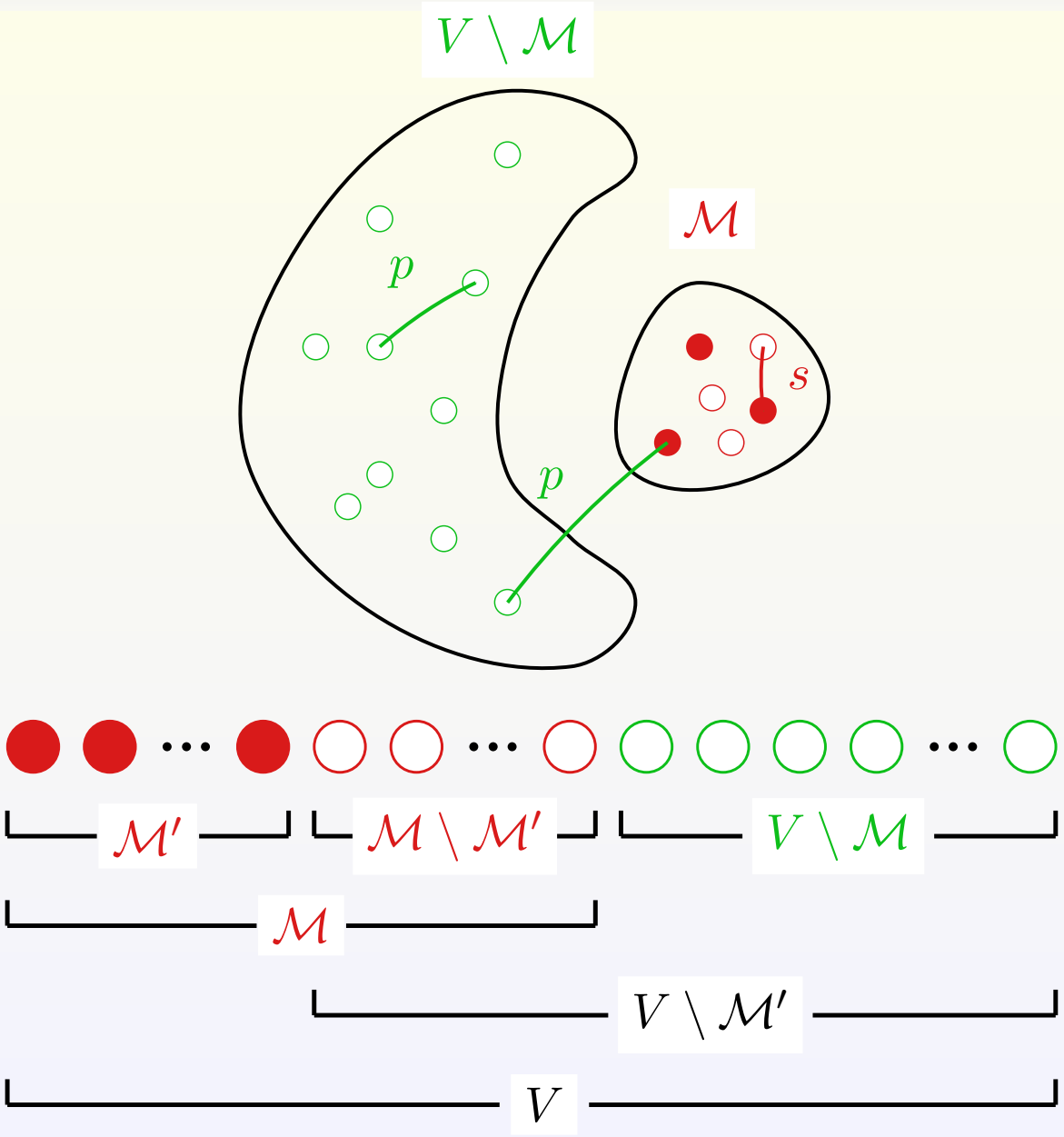
STFP2011 available at http://arxiv.org/pdf/1108.2228

Vertex Assignment

Theorem:
Let $\mathbb{G}$ be an identifiable SBM. (or RDP LPM such that ...)
Then STFP2011 yields consistent vertex assignment.

For $v \in V$, $\tau(v) : \Omega \to [K]$; STFP2011 yields $\widehat{\tau}(v)$.
$P[\widehat{\tau}(v) \neq \tau(v)] \to 0$ as $n \to \infty$.

Vertex Nomination



Vertex Nomination

Corollary:
Let $\mathbb{G}$ be an identifiable SBM. (or RDP LPM such that . . .)
Then STFP2011+ yields consistent vertex nomination.

Vertex Nomination

Normalized Sum of Reciprocal Ranks

$$\text{NSRR} = \left(\sum_{v \in \mathcal{M} \setminus \mathcal{M}'} \frac{1}{\text{rank}(v)} \right) / \left(\sum_{i=1}^{m-m'} \frac{1}{i} \right)$$

Connectome Inference

Extension: Errorful Observation

$$SBM(B, \pi)$$

$$E_{B,\pi}[size(G)] = ((n\pi)^T B(n\pi) - 1^T diag(B)(n\pi))/2$$

$$\rho_{B,\pi} = E_{B,\pi}[size(G)]/\binom{n}{2} = (n\pi^T B\pi - 1^T diag(B)\pi)/(n-1)$$

$$h : [0,1] \to [0,1] \text{ increasing } ; \varepsilon \in [0,1)$$

$$B_\varepsilon = h(\varepsilon)\left[(1-\varepsilon)\rho_{B,\pi}^{-1} B + \varepsilon J\right]$$

$$SBM(B_\varepsilon, \pi)$$

HLT Content & Context Inference

Extension: Errorful Observation

$$SBM(B,\pi);\mathbb{G}_{B,\pi}=(V,E)$$

$$\binom{V}{2}=E\sqcup\overline{E}$$

$$p_k(c)=P_{c,k}\left[uv\in\widetilde{E}\mid uv\in E\right]=1-F_{E,k}(c)$$

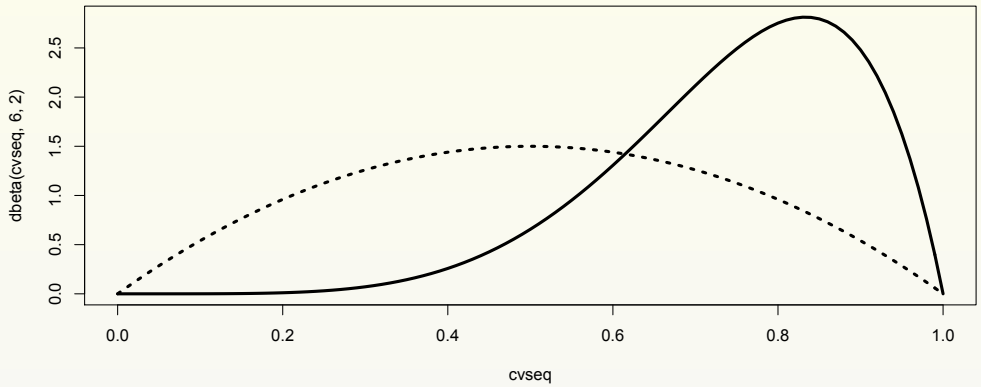
$$\overline{p}_k(c)=P_{c,k}\left[uv\in\widetilde{E}\mid uv\in\overline{E}\right]=1-F_{\overline{E},k}(c)$$

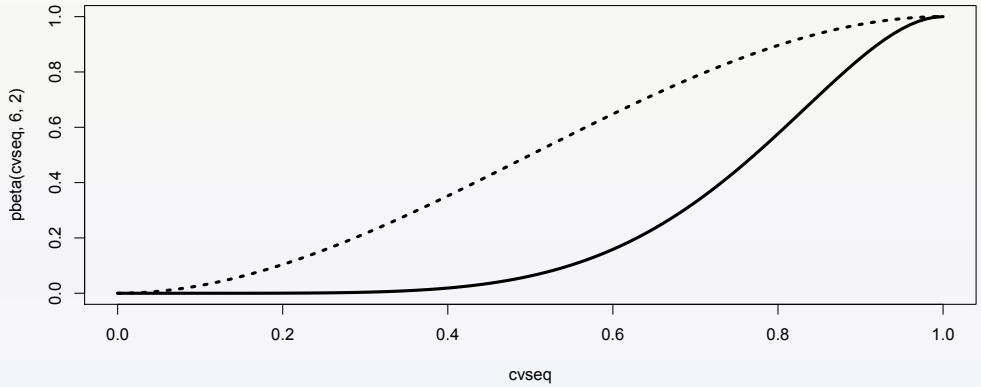
$$h:[0,1]\to[0,1]\text{ decreasing };k\in(0,\infty)$$

$$B_{k,c}=h(k)\left[\left(p_k(c)-\overline{p}_k(c)\right)B+\overline{p}_k(c)J\right]$$

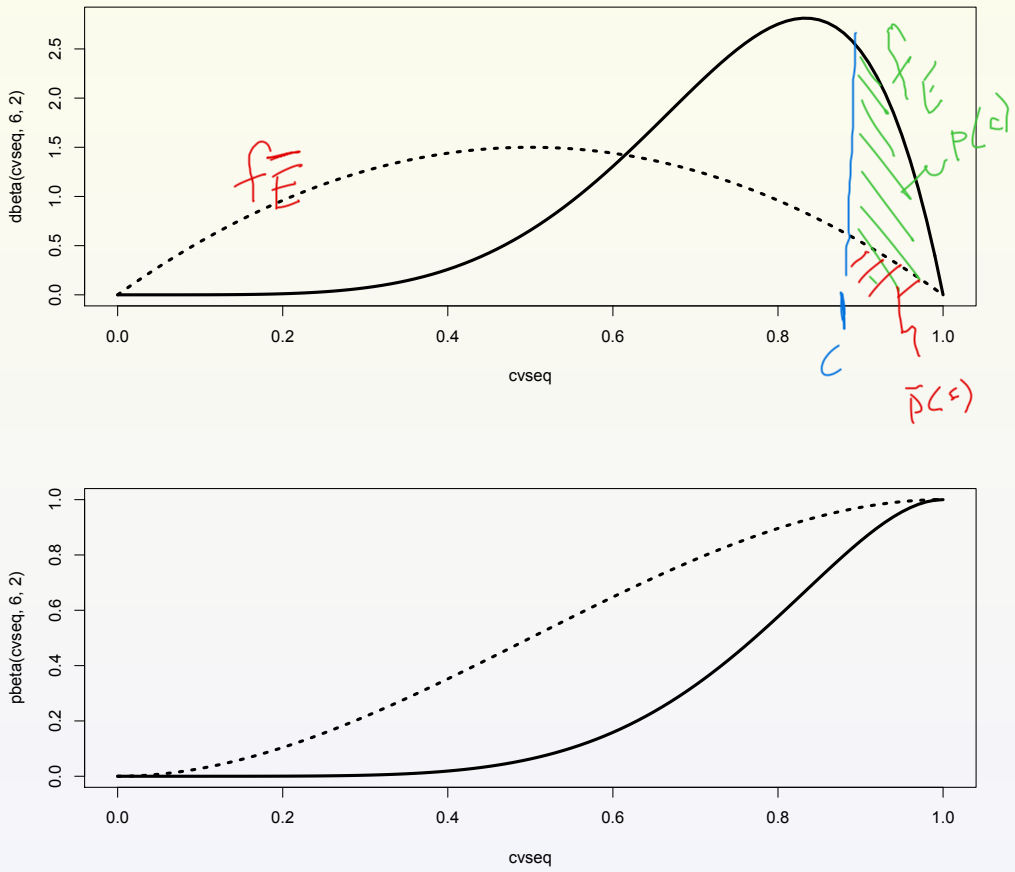
$$SBM(B_{k,c},\pi);\widetilde{\mathbb{G}}_{B_{k,c},\pi}=(V,\widetilde{E})$$

Vertex Assignment for Errorful Observation

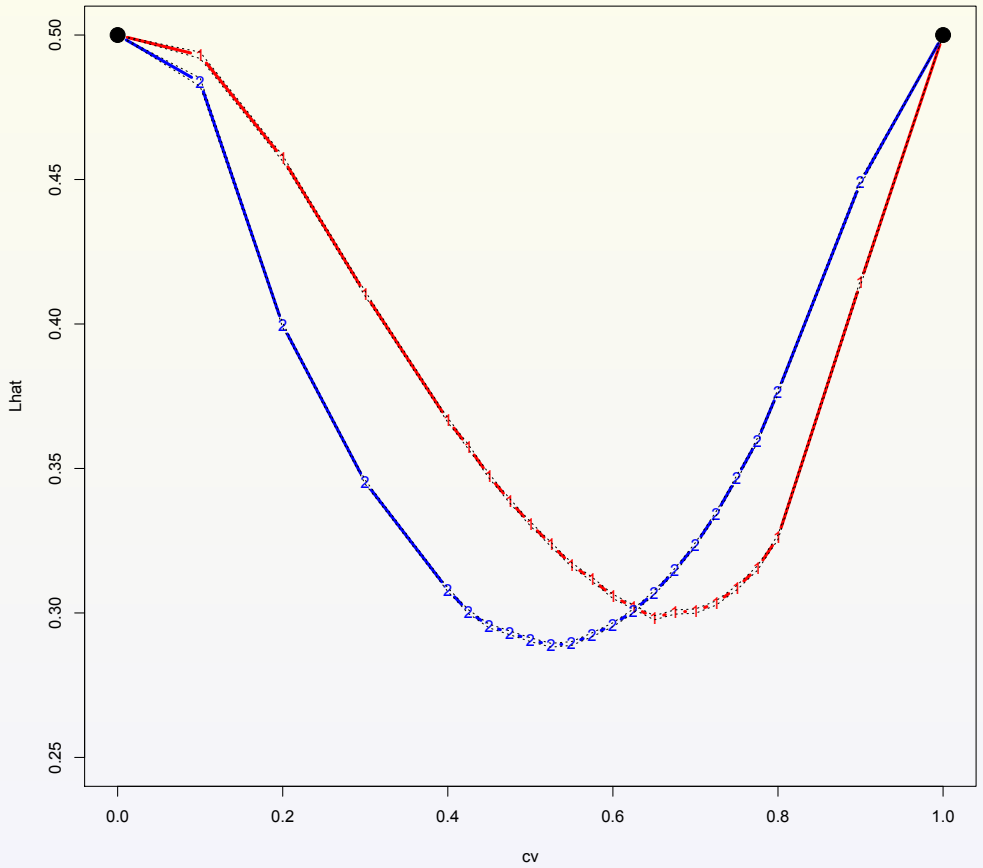




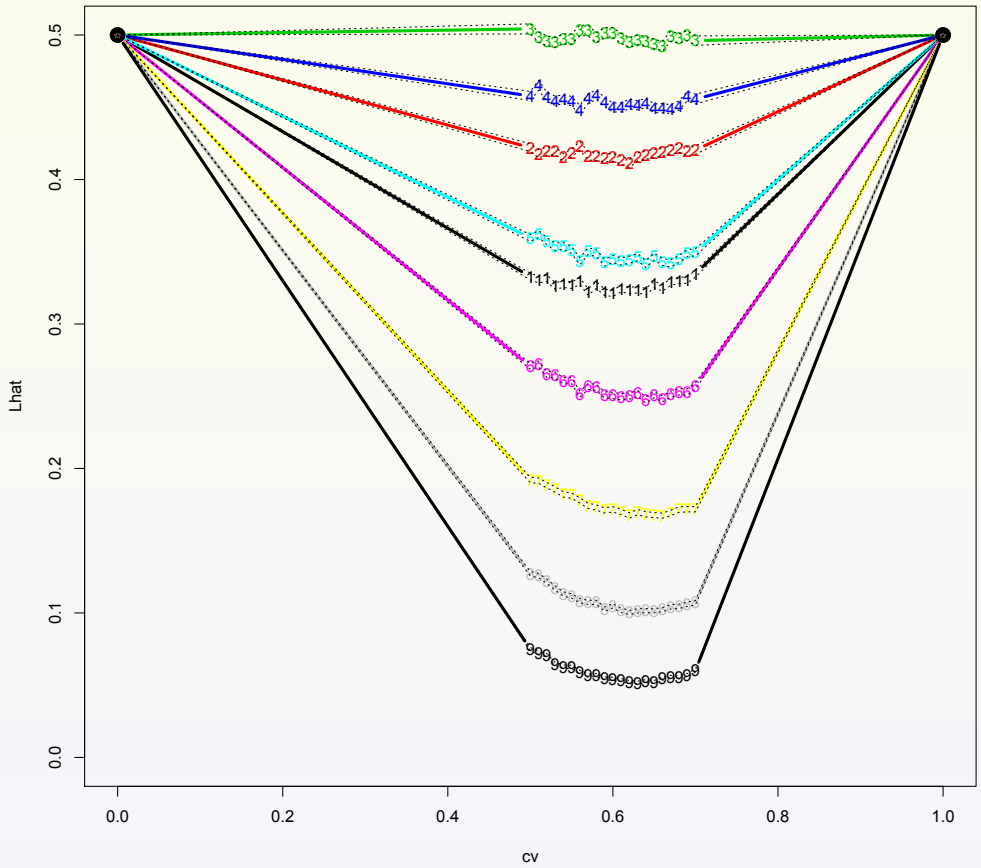
Vertex Assignment for Errorful Observation



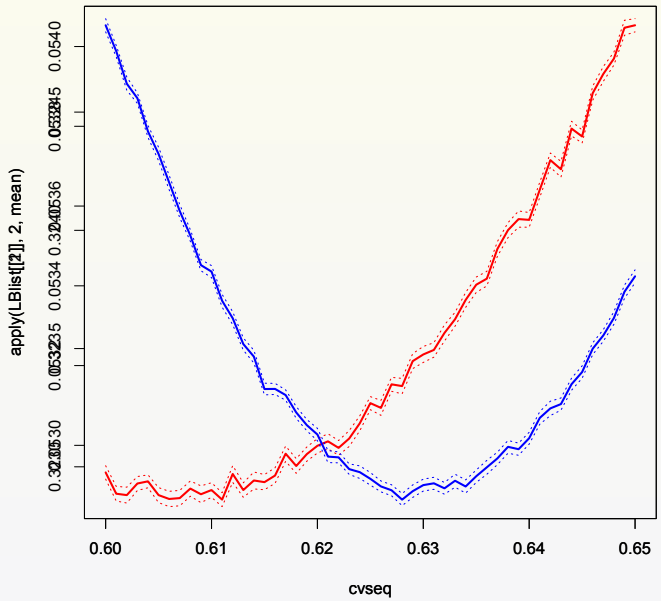
Vertex Assignment for Errorful Observation



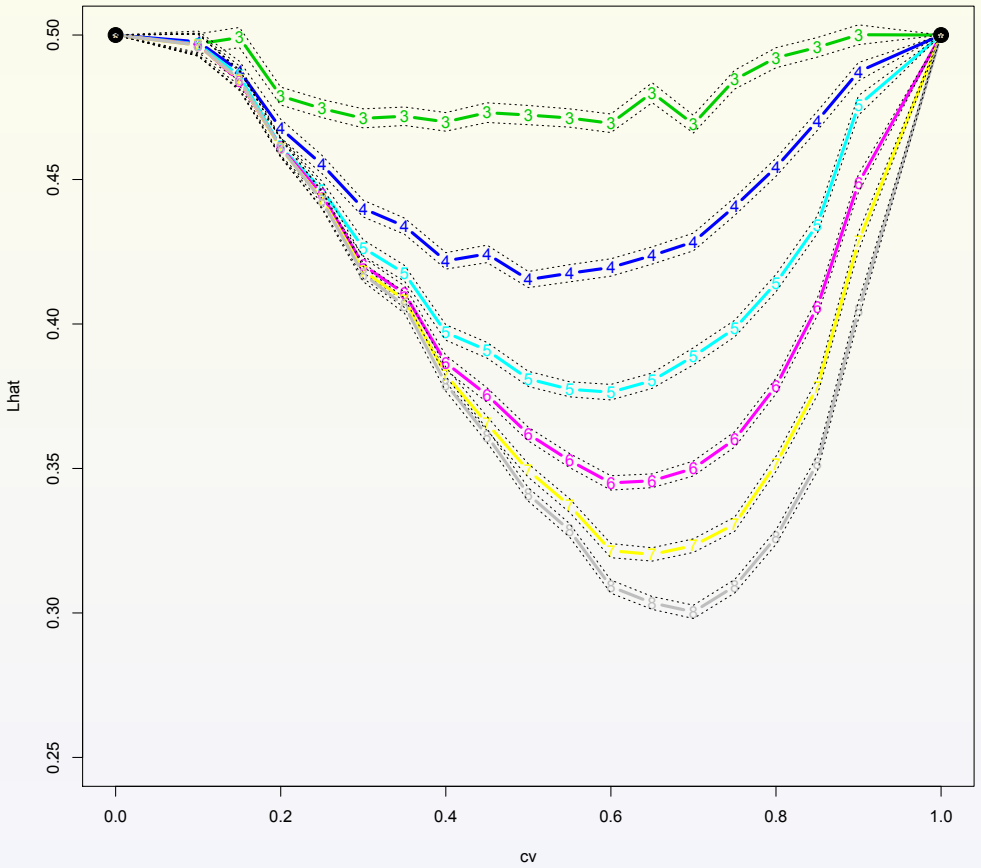
Vertex Assignment for Errorful Observation



Vertex Assignment for Errorful Observation



Vertex Assignment for Errorful Observation



Leopold Kronecker to Hermann von Helmholtz:

"The wealth of your practical experience
with sane and interesting problems
will give to mathematics
a new direction and a new impetus."