Mutual Information
Or, How Dependent Are They?

Dave Darmon

June 5, 2014

Measuring Dependence

Presentation goals

- Formulate dependence as a statistical problem.
- Introduce mutual information as a useful measure of dependence.
- Discuss *how to estimate* mutual information in practice.
  - The usual way *and* a better way.
- Discuss a paper exploring measures of dependence applied to co-expression data.

Measuring Dependence

The Setup – Most General Case

- 'Reality': We observe n collections of observations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ where

$$\mathbf{X}_i = (X_{1i}, X_{2i}, \ldots, X_{di}),$$

  and want to determine the overall associations amongst the X_j.
  - In our context, each X_j might be the gene expression level for gene j in an individual.
- If we think of $\mathbf{X} = (X_1, \ldots, X_d)$ as a random vector, we want to infer its joint distribution.

Measuring Dependence

The Setup – In Practice

- This is, in general, very difficult.
- Instead, we pretend we observe n pairs of measurements,

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n),$$

and want to determine if an association exists between X and Y.

  - In our context, X_i might be the gene expression level of gene x in individual i, and Y_i is the gene expression level of gene y in individual i.

- This *pairwise* is much easier, and will be the focus of this presentation.

Measuring Dependence

The Setup – A Working Example

- From *The DREAM5 Network Inference Challenge*:
  - Coexpression levels for 2810 genes measured using 160 microarrays.
  - The usual story: infer a gene regulatory network from the coexpression levels.
- From our original slide, for each microarray $i = 1, \ldots, 160$, we have

$$\mathbf{X}_i = (X_{1i}, X_{2i}, \ldots, X_{2810i}).$$

- i.e. In theory, we would like to infer a 2810-dimensional distribution from 160 data points.
  - That ain't gonna happen.
- Instead, play the usual pairwise game...

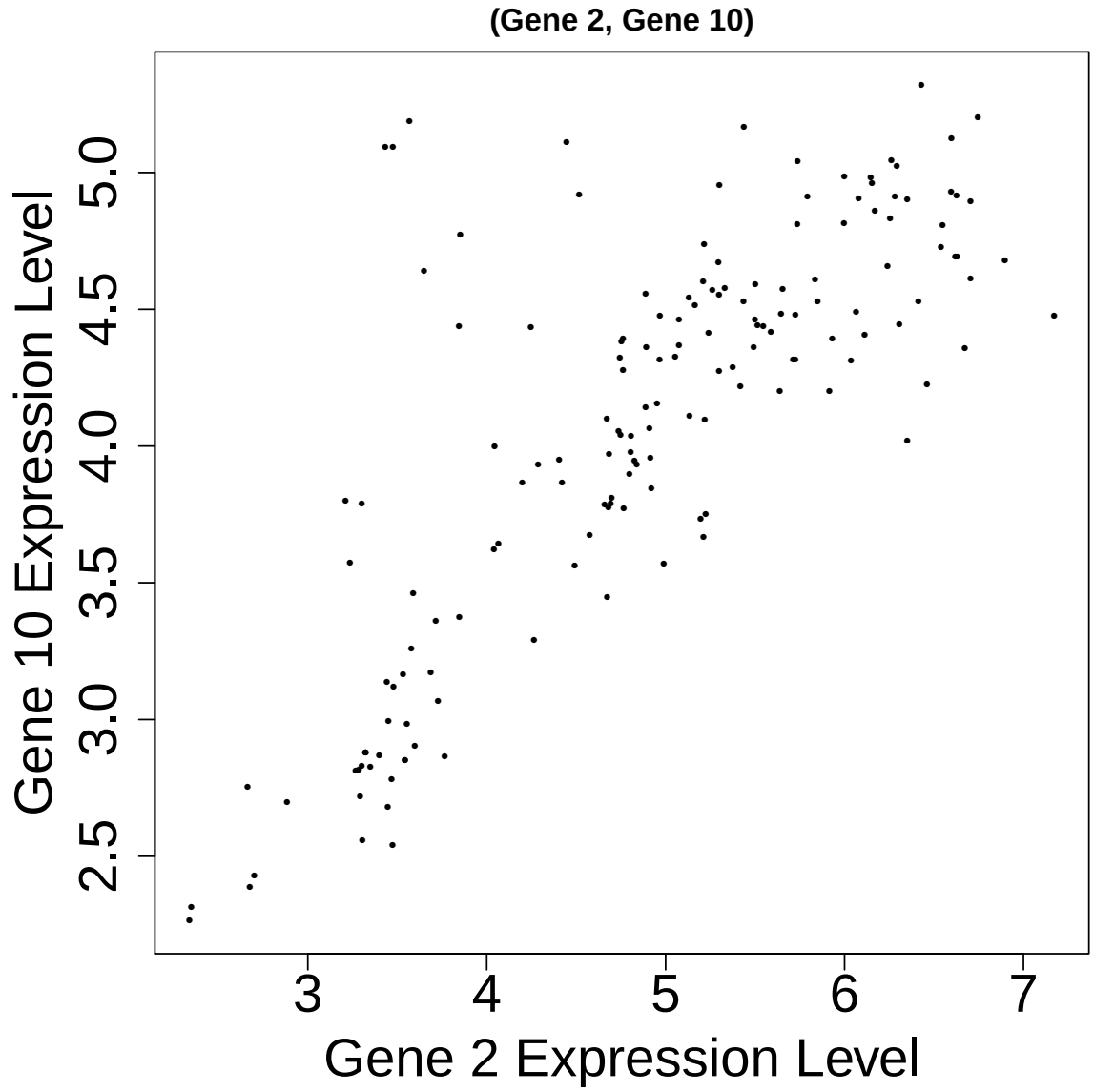

Figure: Pairwise expression levels for two genes from *The DREAM5 Network Inference Challenge*.

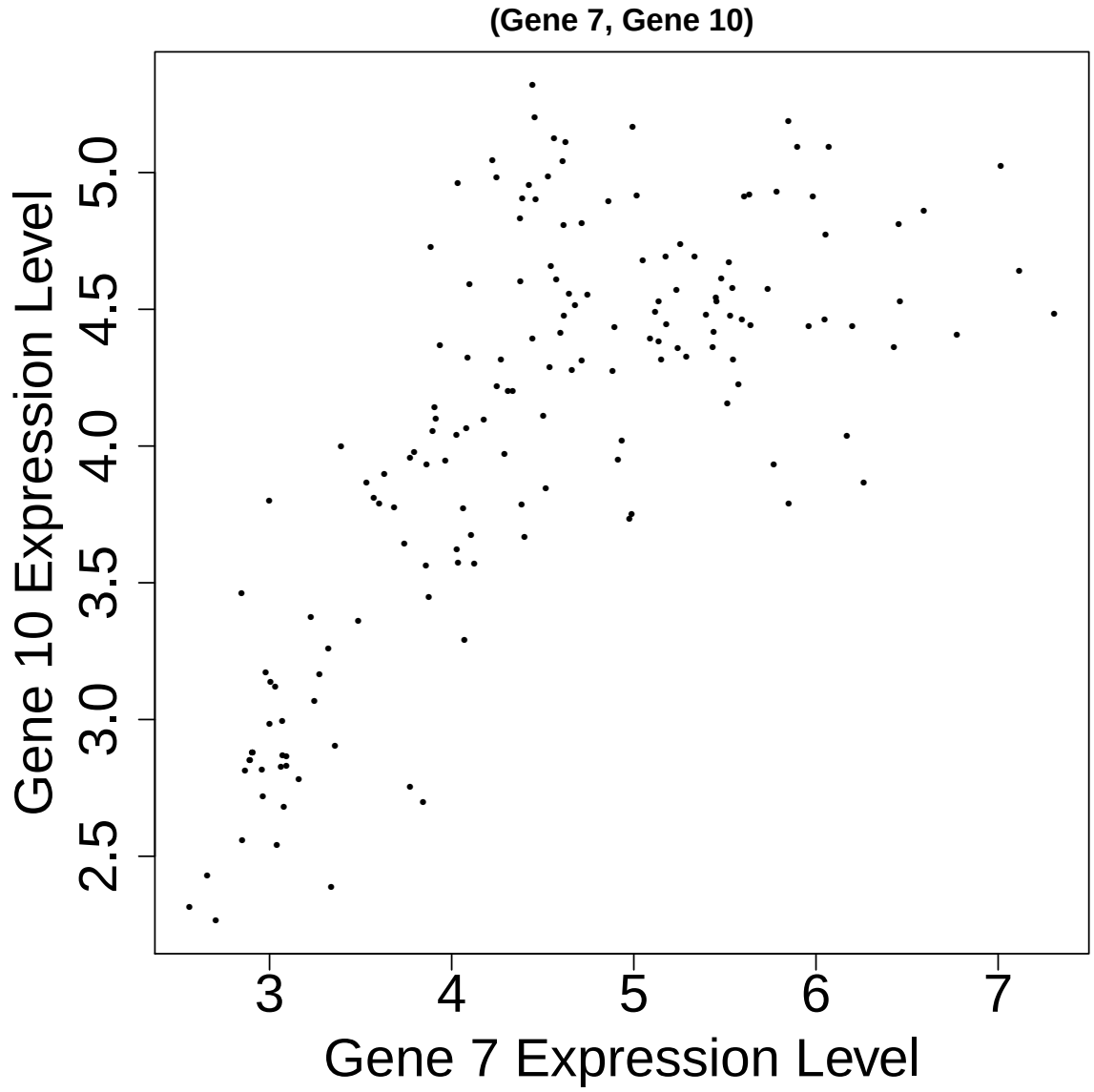

Figure: Pairwise expression levels for two genes from *The DREAM5 Network Inference Challenge*.

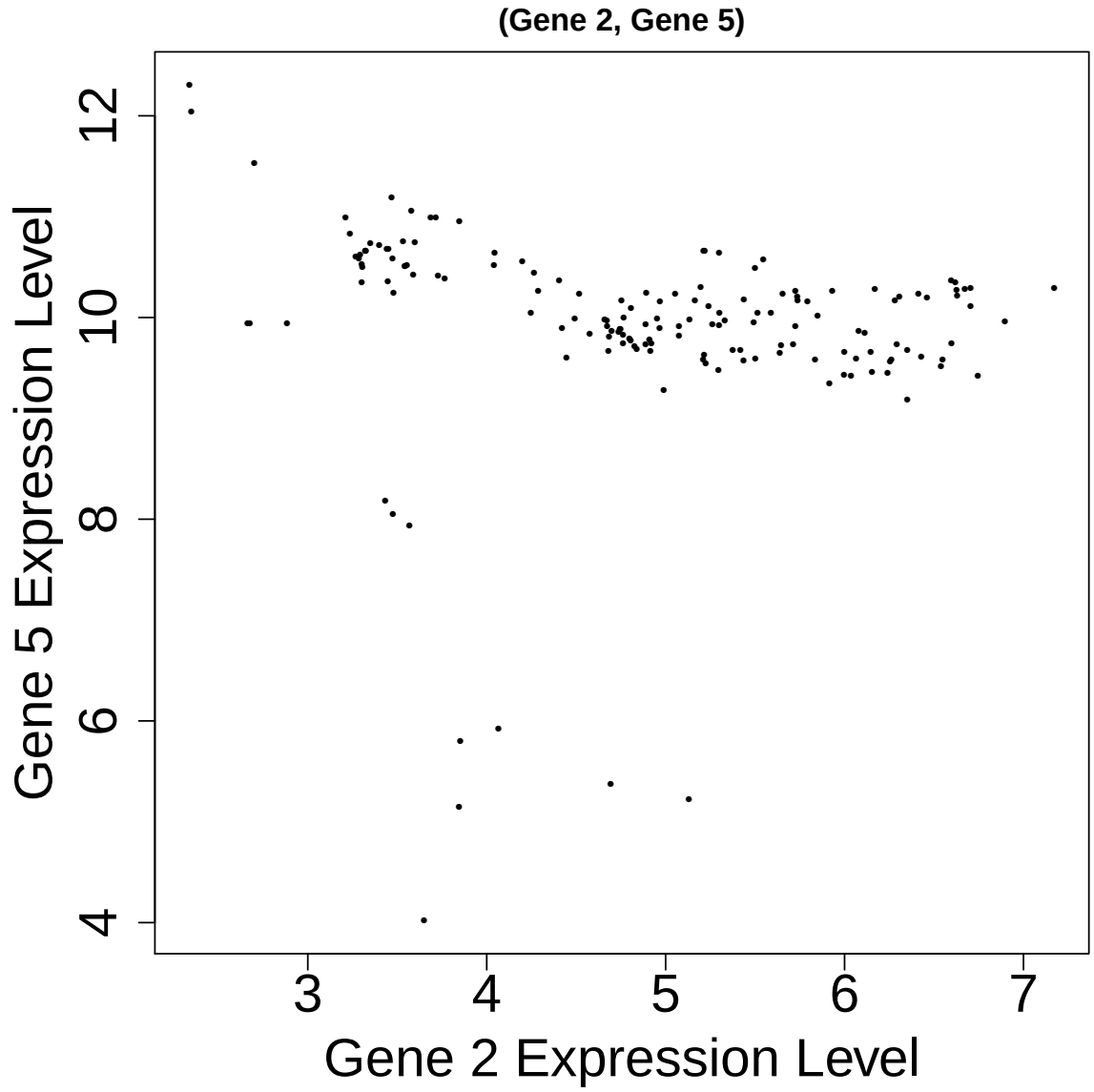

Figure: Pairwise expression levels for two genes from *The DREAM5 Network Inference Challenge*.

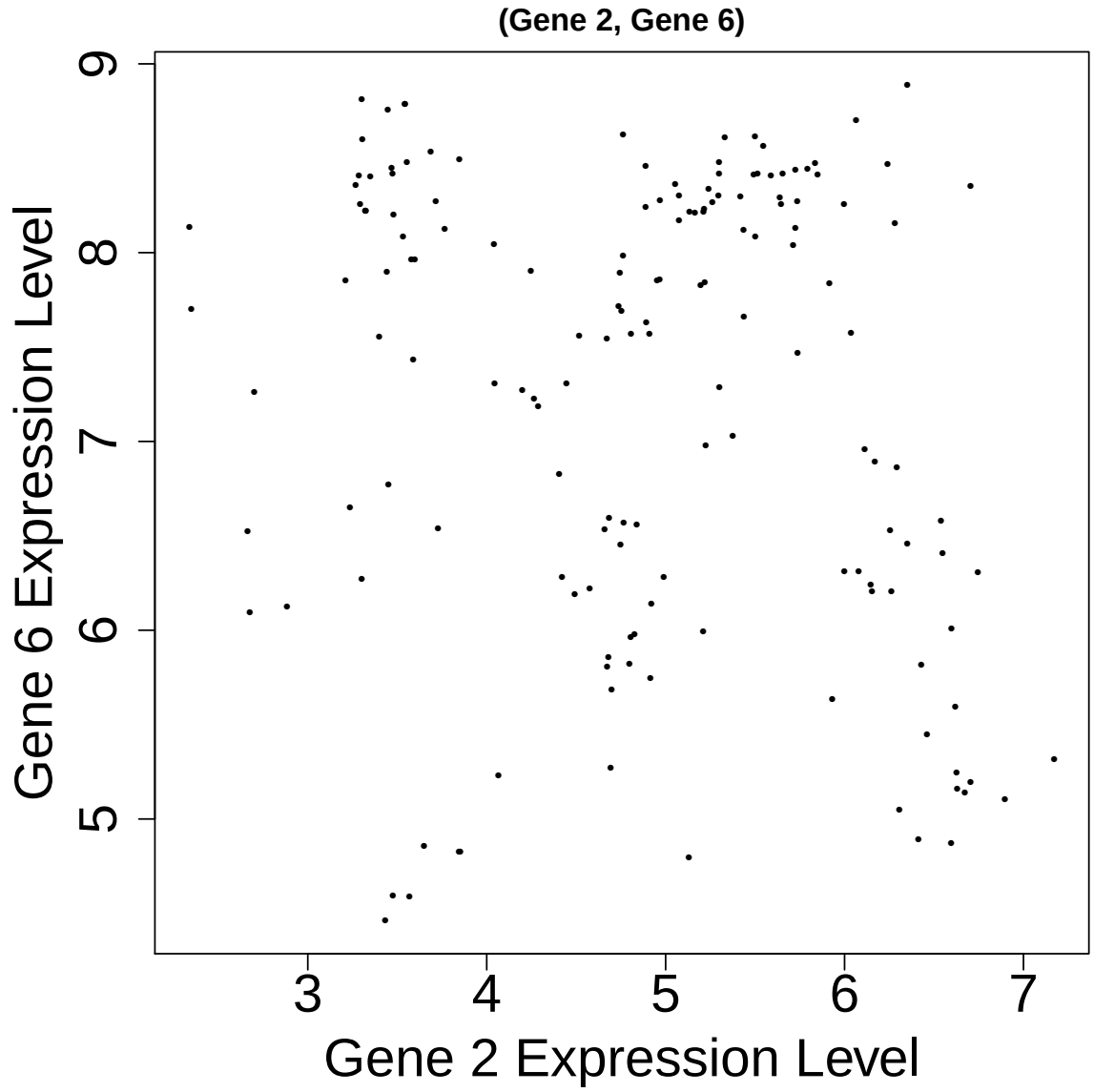

Figure: Pairwise expression levels for two genes from *The DREAM5 Network Inference Challenge*.

Measuring Dependence

The Setup – Takeaways

- It's hard to eyeball dependencies.
  - The mark-one human eyeball works well, but doesn't provide confidence intervals or P-values.
- Besides, we need to keep our friends in the AMSC department employed, so...
- Bring on the statisticians!

Measuring Dependence

The STAT100 Version of Dependence

- Let X and Y be two random variables. Then the (Pearson) correlation between X and Y is

$$\begin{aligned}\text{Corr}(X,Y) &= \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\ &= \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sqrt{E[(X-\mu_X)^2]E[(Y-\mu_Y)^2]}},\end{aligned}$$

  i.e. the normalized covariance between X and Y.

- **Theorem:** If X and Y are independent, then $\text{Corr}(X,Y) = 0$.
- The converse **need not be true**!
  - If $\text{Corr}(X,Y) = 0$, then X and Y need not be independent.

Measuring Dependence

What is Independence, Anyway?

- **Definition:** Two random variables X and Y with joint cumulative distribution $F_{X,Y}(x, y)$ and marginal cumulative distributions $F_X(x)$ and $F_Y(y)$ are *independent* if

$$F_{Y|X}(y|x) = F_Y(y), \text{ for all } x, y$$

or equivalently

$$F_{X|Y}(x|y) = F_X(x), \text{ for all } x, y$$

- i.e. Knowing something about X tells you **nothing** about Y, and vice versa.

Measuring Dependence

What is Independence, Anyway?

- **Theorem:** Two random variables X and Y with joint cumulative distribution $F_{X,Y}(x, y)$ and marginal cumulative distributions $F_X(x)$ and $F_Y(y)$ are *independent* if

$$F_{X,Y}(x, y) = F_X(x)F_Y(y), \text{ for all } x, y.$$

- 'The joint distribution factors as the product of the marginals.'
- If probability mass / density functions exist, then we also have that the mass / density functions factor, i.e.

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) \text{ for all } x, y.$$

Measuring Dependence

Information Theory – Claude Shannon Comes to the Party

- Information theory is an extension of probability theory that asks:
  - What happens when we take expectations of *distribution functions*?
- **Answer:** Lots of interesting things!

To the Board

Measuring Dependence

Discrete and Differential Entropies

The Discrete Case:

$$H[X] = -\sum_{x \in \mathcal{X}} p_X(x) \log p_X(x)$$
$$= -E_{p_X}[\log p_X(X)]$$

The Continuous Case:

$$h[X] = -\int_{x \in \mathcal{X}} f_X(x) \log f_X(x)\, dx$$
$$= -E_{f_X}[\log f_X(X)]$$

Measuring Dependence

Mutual Information

$$I[X \wedge Y] = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}$$
$$= E_{p_{X,Y}}\left[\log \frac{p_{X,Y}(X,Y)}{p_X(X)p_Y(Y)}\right]$$

$$I[X \wedge Y] = \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f_{X,Y}(x,y) \log \frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\, dA$$
$$= E_{f_{X,Y}}\left[\log \frac{f_{X,Y}(X,Y)}{f_X(X)f_Y(Y)}\right]$$

Measuring Dependence

Why Care About Mutual Information?

- **Theorem:** Two random variables X and Y are independent **if and only if** $I[X \wedge Y] = 0$.
- Thus, a *single number* tells us whether two random variables are independent.
  - This is frequently what people pretend the Pearson correlation does.
- 'Proof':

$$I[X \wedge Y] = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)}$$

Measuring Dependence

Okay, So Now What?

- We've been pretending that we know the joint distribution of X and Y.
- We *don't*, but we'll continue to pretend that such an object exists.
- How do we guess at $I[X \wedge Y]$ without knowing $f_{X,Y}(x, y)$?

$$I[X \wedge Y] = \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} f_{X,Y}(x, y) \log \frac{f_{X,Y}(x, y)}{f_X(x) f_Y(y)}\, dA$$

- We first need to make a guess at $f_{X,Y}(x, y)$.

Measuring Dependence

I Know: Histograms!

- Many methods exist for computing an estimator $\hat{f}_{X,Y}(x,y)$ for $f_{X,Y}(x,y)$:
  - Parametric fitting.
  - Histograms.
  - Kernel Density Estimators.
- Histograms tend to be the most popular choice:
  - Bin your data, and construct an empirical mass function $\hat{f}_{X^{\Delta_x},Y^{\Delta_y}}(x,y)$ by counting.
  - Now compute the mutual information between X^{Δ_x} and Y^{Δ_y}.
  - **Caveat**: This approach *does not* work for entropy estimation.

To the Board

Measuring Dependence

A Better Way

- Many methods exist for computing an estimator $\hat{f}_{X,Y}(x, y)$ for $f_{X,Y}(x, y)$:
  - Parametric fitting.
  - Histograms.
  - **Kernel Density Estimators.**
- **Paper:** *Exponential Concentration for Mutual Information Estimation with Application to Forests* by Han Liu, John Lafferty and Larry Wasserman:
  - Provide a kernel density estimator-based mutual information estimator, along with:
    - A convergence rate for the mean-squared error of the estimator.
    - Some idea about the bias and variance of the estimator.
  - In other words, the paper gives both an estimator *and* an idea of how well that estimator works.

Comparison of co-expression measures: mutual information, correlation, and model based indices

Lin Song, Peter Langfelder, Steve Horvath

Measuring Dependence

SLH

- The paper (hereafter referred to as *SLH*) investigates the appropriate dependence measure for constructing co-expression networks, with the ultimate goal of detecting modules in the co-expression network.
- The authors argue that Tukey's biweight midcorrelation outperforms correlation and mutual information for determining significantly enriched modules in the co-expression network.
  - Maybe.
- The authors also argue that mutual information can 'safely be replaced by correlation [...] when it comes to measuring co-expression relationships in stationary data.'
  - Not based on their argument.

Measuring Dependence

SLH

- The paper (hereafter referred to as *SLH*) investigates the appropriate dependence measure for constructing co-expression networks, with the ultimate goal of detecting modules in the co-expression network.
- The authors argue that Tukey's biweight midcorrelation outperforms correlation and mutual information for determining significantly enriched modules in the co-expression network.
  - Maybe.
- **The authors also argue that mutual information can 'safely be replaced by correlation [...] when it comes to measuring co-expression relationships in stationary data.'**
  - **Not based on their argument.**

Measuring Dependence

SLH – Experimentum Crucis

- The authors perform a simulation experiment where they assume the generative model

$$(X, Y) \sim \text{Bivariate Gaussian}(\rho).$$

- They simulate realizations $\{(X_i, Y_i)\}_{i=1}^n$ from this model with varying correlations ρ, and compute the (sample!) correlation, biweight midcorrelation, and mutual information between X and Y.
  - They take their sample size to be $n = 1000$ (!).

Measuring Dependence

SLH – Experimentum Crucis

- Their main result? When $(X, Y) \sim$ Bivariate Gaussian(ρ), an (empirical) simple relationship holds (approximately) between the (sample!) correlations and mutual informations.
- We can do one better. When $(X, Y) \sim$ Bivariate Gaussian(ρ), we have that (exactly!)

$$I[X \wedge Y] = -\frac{1}{2} \log_e \left(1 - \rho^2\right).$$

- In other words, if we estimate ρ^2 and $I[X \wedge Y]$, their estimated quantities should *also* lie somewhere near this curve.

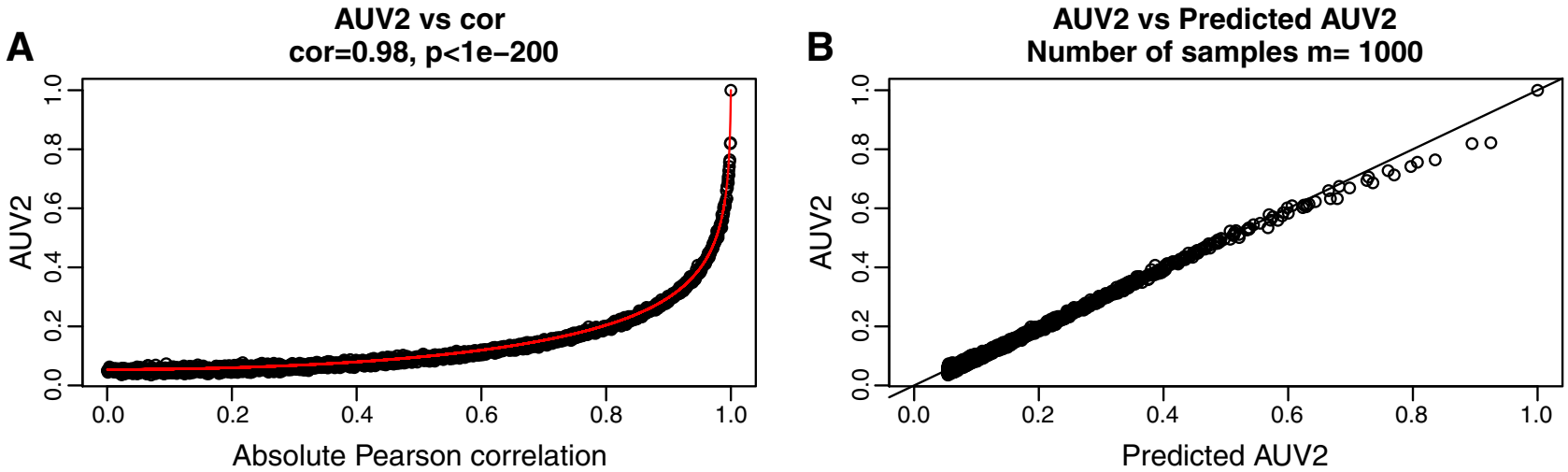

Figure: Using an estimator of correlation to predict an estimator for mutual information with *synthetic data*. From *SLH*.

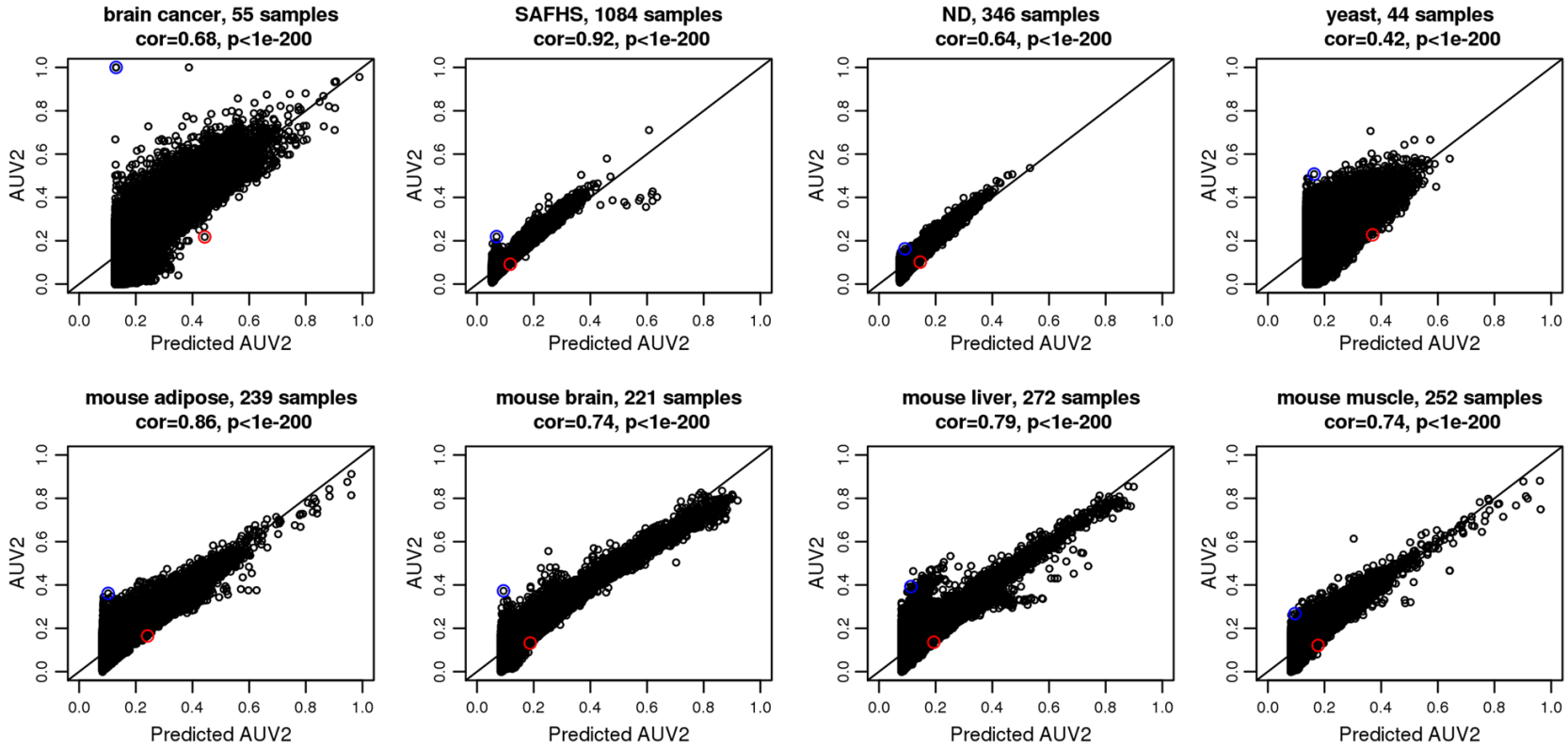

Figure: Using an estimator of correlation to predict an estimator for mutual information with *real data*. From *SLH*.

Measuring Dependence

SLH – Empirical Investigation

- What have the authors really found?
  - Most pairs of genes **don't look bivariate normal!**

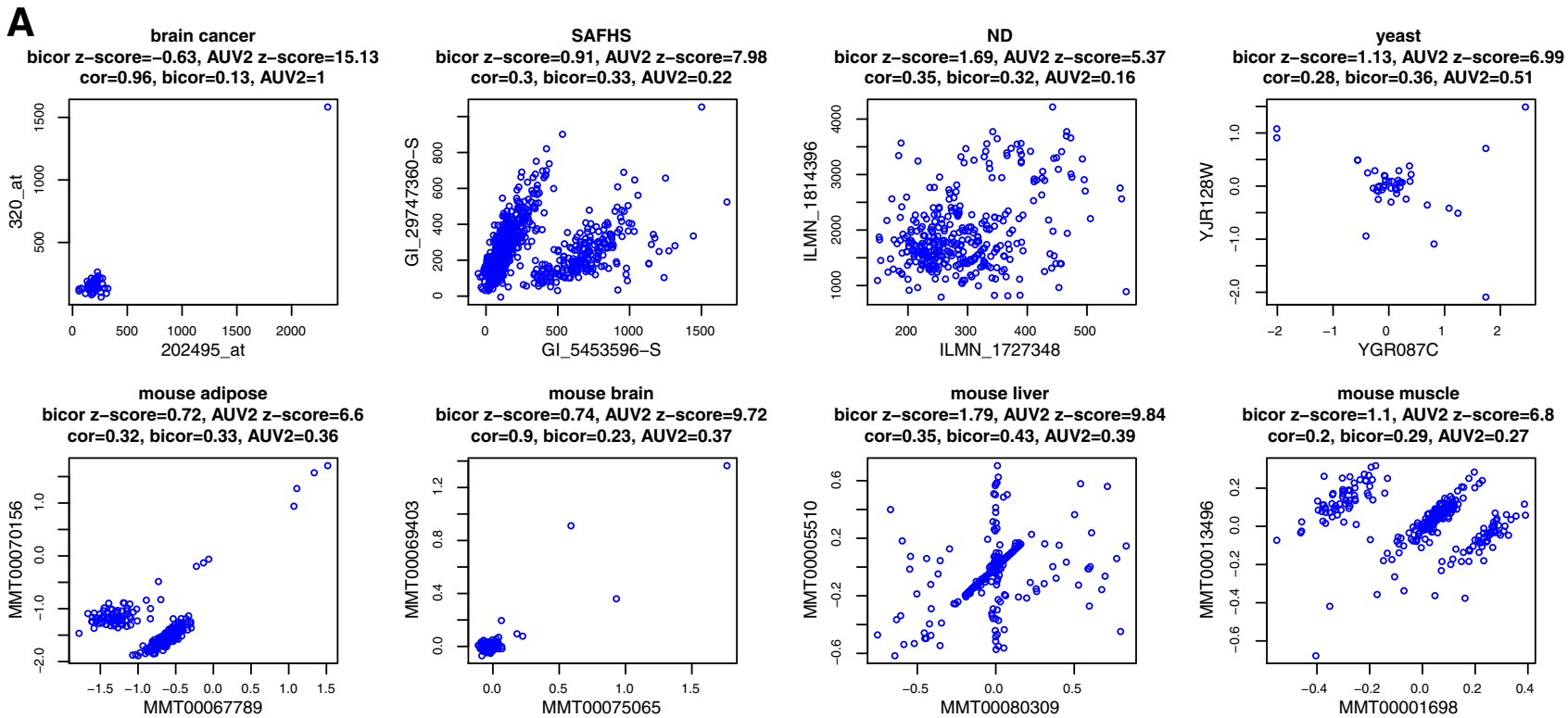

Figure: Gene expression levels for various gene pairs from *SLH*.

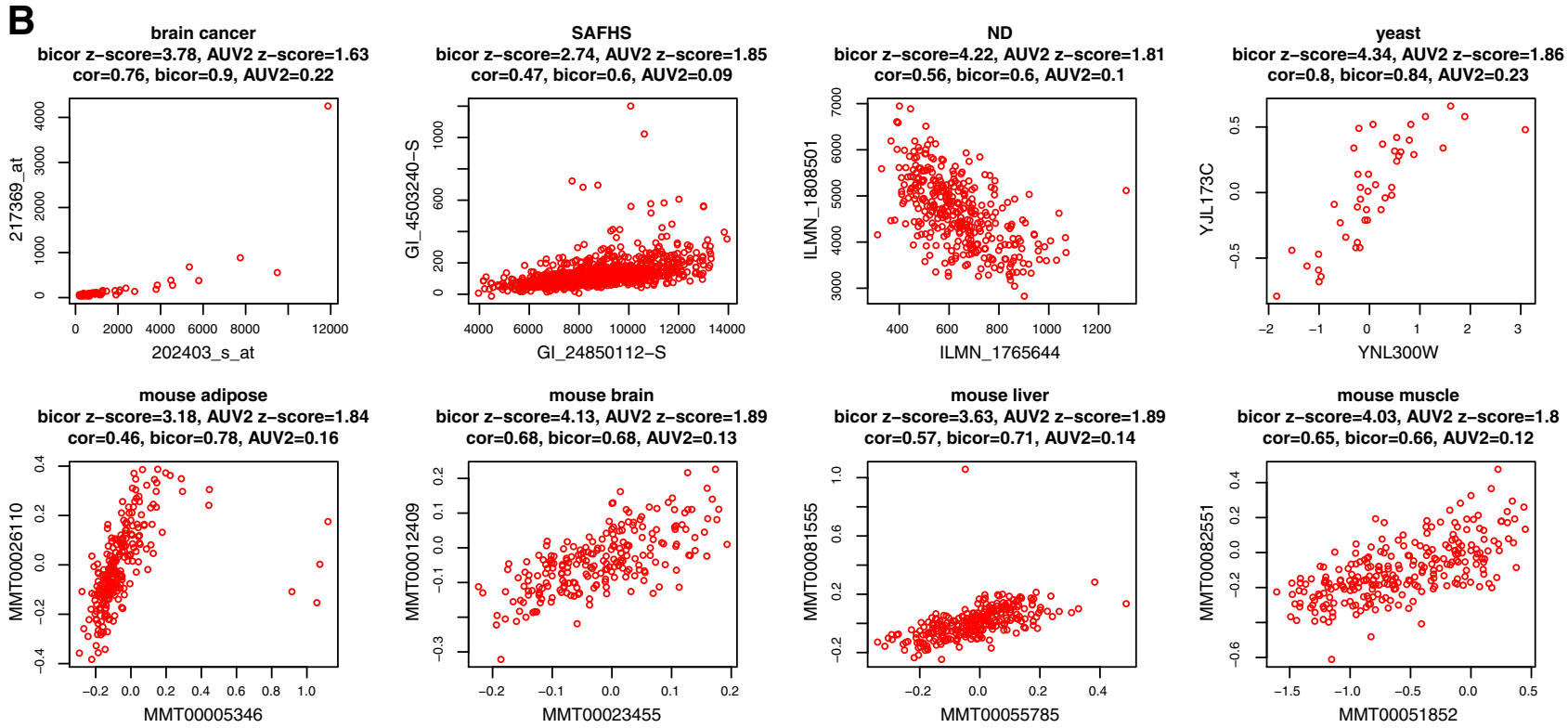

Figure: Gene expression levels for various gene pairs from *SLH*.

Measuring Dependence

SLH – Empirical Investigation

- What have the authors really found?
  - Most pairs of genes **don't look bivariate normal!**
- A favorite quote from the paper:
  - "The mouse liver data set displays a pairwise pattern that is neither commonly seen nor easily explained."
- What the authors really mean:
  - 'Mutual information [...] can safely be replaced by correlation [...] when it comes to measuring co-expression relationships in stationary data [**if you're willing to assume joint normality**].'