Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks

Jean Hausser   JEAN.HAUSSER@UNIBAS.CH
Bioinformatics, Biozentrum
University of Basel
Klingelbergstr. 50/70
CH-4056 Basel, Switzerland

Korbinian Strimmer   STRIMMER@UNI-LEIPZIG.DE
Institute for Medical Informatics, Statistics and Epidemiology
University of Leipzig
Härtelstr. 16–18
D-04107 Leipzig, Germany

Abstract

We present a procedure for effective estimation of entropy and mutual information from small-sample data, and apply it to the problem of inferring high-dimensional gene association networks. Specifically, we develop a James-Stein-type shrinkage estimator, resulting in a procedure that is highly efficient statistically as well as computationally. Despite its simplicity, we show that it outperforms eight other entropy estimation procedures across a diverse range of sampling scenarios and data-generating models, even in cases of severe undersampling. We illustrate the approach by analyzing *E. coli* gene expression data and computing an entropy-based gene-association network from gene expression data. A computer program is available that implements the proposed shrinkage estimator.

Keywords: entropy, shrinkage estimation, James-Stein estimator, "small *n*, large *p*" setting, mutual information, gene association network

1. Introduction

Entropy is a fundamental quantity in statistics and machine learning. It has a large number of applications, for example in astronomy, cryptography, signal processing, statistics, physics, image analysis neuroscience, network theory, and bioinformatics—see, for example, Stinson (2006), Yeo and Burge (2004), MacKay (2003) and Strong et al. (1998). Here we focus on *estimating* entropy from small-sample data, with applications in genomics and gene network inference in mind (Margolin et al., 2006; Meyer et al., 2007).

To define the Shannon entropy, consider a categorical random variable with alphabet size p and associated cell probabilities $\theta_1, \ldots, \theta_p$ with $\theta_k > 0$ and $\sum_k \theta_k = 1$. Throughout the article, we assume

that p is fixed and known. In this setting, the Shannon entropy in natural units is given by[1]

$$H = -\sum_{k=1}^{p} \theta_k \log(\theta_k).$$ (1)

In practice, the underlying probability mass function are unknown, hence H and θ_k need to be *estimated* from observed cell counts $y_k \geq 0$.

A particularly simple and widely used estimator of entropy is the maximum likelihood (ML) estimator

$$\hat{H}^{\mathrm{ML}} = -\sum_{k=1}^{p} \hat{\theta}_k^{\mathrm{ML}} \log(\hat{\theta}_k^{\mathrm{ML}})$$

constructed by plugging the ML frequency estimates

$$\hat{\theta}_k^{\mathrm{ML}} = \frac{y_k}{n}$$ (2)

into Equation **1**, with $n = \sum_{k=1}^{p} y_k$ being the total number of counts.

In situations with $n \gg p$, that is, when the dimension is low and when there are many observation, it is easy to infer entropy reliably, and it is well-known that in this case the ML estimator is optimal. However, in high-dimensional problems with $n \ll p$ it becomes extremely challenging to estimate the entropy. Specifically, in the "small n, large p" regime the ML estimator performs very poorly and severely underestimates the true entropy.

While entropy estimation has a long history tracing back to more than 50 years ago, it is only recently that the specific issues arising in high-dimensional, undersampled data sets have attracted attention. This has lead to two recent innovations, namely the NSB algorithm (Nemenman et al., 2002) and the Chao-Shen estimator (Chao and Shen, 2003), both of which are now widely considered as benchmarks for the small-sample entropy estimation problem (Vu et al., 2007).

Here, we introduce a novel and highly efficient small-sample entropy estimator based on James-Stein shrinkage (Gruber, 1998). Our method is fully analytic and hence computationally inexpensive. Moreover, our procedure simultaneously provides estimates of the entropy *and* of the cell frequencies suitable for plugging into the Shannon entropy formula (Equation 1). Thus, in comparison the estimator we propose is simpler, very efficient, and at the same time more versatile than currently available entropy estimators.

2. Conventional Methods for Estimating Entropy

Entropy estimators can be divided into two groups: i) methods, that rely on estimates of cell frequencies, and ii) estimators, that directly infer entropy without estimating a compatible set of θ_k. Most methods discussed below fall into the first group, except for the Miller-Madow and NSB approaches.

---

1. In this paper we use the following conventions: log denotes the natural logarithm (*not* base 2 or base 10), and we define $0\log 0 = 0$.

2.1 Maximum Likelihood Estimate

The connection between observed counts y_k and frequencies θ_k is given by the multinomial distribution

$$\text{Prob}(y_1,\ldots,y_p;\theta_1,\ldots,\theta_p) = \frac{n!}{\prod_{k=1}^{p} y_k!} \prod_{k=1}^{p} \theta_k^{y_k}. \tag{3}$$

Note that $\theta_k > 0$ because otherwise the distribution is singular. In contrast, there may be (and often are) zero counts y_k. The ML estimator of θ_k maximizes the right hand side of Equation **3** for fixed y_k, leading to the observed frequencies $\hat{\theta}_k^{\text{ML}} = \frac{y_k}{n}$ with variances $\text{Var}(\hat{\theta}_k^{\text{ML}}) = \frac{1}{n}\theta_k(1-\theta_k)$ and $\text{Bias}(\hat{\theta}_k^{\text{ML}}) = 0$ as $E(\hat{\theta}_k^{\text{ML}}) = \theta_k$.

2.2 Miller-Madow Estimator

While $\hat{\theta}_k^{\text{ML}}$ is unbiased, the corresponding plugin entropy estimator $\hat{H}^{\text{ML}}$ is not. First order bias correction leads to

$$\hat{H}^{\text{MM}} = \hat{H}^{\text{ML}} + \frac{m_{>0}-1}{2n},$$

where $m_{>0}$ is the number of cells with $y_k > 0$. This is known as the Miller-Madow estimator (Miller, 1955).

2.3 Bayesian Estimators

Bayesian regularization of cell counts *may* lead to vast improvements over the ML estimator (Agresti and Hitchcock, 2005). Using the Dirichlet distribution with parameters $a_1, a_2, \ldots, a_p$ as prior, the resulting posterior distribution is also Dirichlet with mean

$$\hat{\theta}_k^{\text{Bayes}} = \frac{y_k + a_k}{n + A},$$

where $A = \sum_{k=1}^{p} a_k$. The flattening constants a_k play the role of pseudo-counts (compare with Equation 2), so that A may be interpreted as the *a priori* sample size.

Some common choices for a_k are listed in Table **1**, along with references to the corresponding plugin entropy estimators,

$$\hat{H}^{\text{Bayes}} = -\sum_{k=1}^{p} \hat{\theta}_k^{\text{Bayes}} \log(\hat{\theta}_k^{\text{Bayes}}).$$

a_k	Cell frequency prior	Entropy estimator
0	no prior	maximum likelihood
1/2	Jeffreys prior (Jeffreys, 1946)	Krichevsky and Trofimov (1981)
1	Bayes-Laplace uniform prior	Holste et al. (1998)
$1/p$	Perks prior (Perks, 1947)	Schürmann and Grassberger (1996)
$\sqrt{n}/p$	minimax prior (Trybula, 1958)	

Table 1: Common choices for the parameters of the Dirichlet prior in the Bayesian estimators of cell frequencies, and corresponding entropy estimators.

While the multinomial model with Dirichlet prior is standard Bayesian folklore (Gelman et al., 2004), there is no general agreement regarding which assignment of a_k is best as noninformative prior—see for instance the discussion in Tuyl et al. (2008) and Geisser (1984). But, as shown later in this article, choosing inappropriate a_k can easily cause the resulting estimator to perform *worse* than the ML estimator, thereby defeating the originally intended purpose.

2.4 NSB Estimator

The NSB approach (Nemenman et al., 2002) avoids overrelying on a particular choice of a_k in the Bayes estimator by using a more refined prior. Specifically, a Dirichlet mixture prior with infinite number of components is employed, constructed such that the resulting prior over the entropy is uniform. While the NSB estimator is one of the best entropy estimators available at present in terms of statistical properties, using the Dirichlet mixture prior is computationally expensive and somewhat slow for practical applications.

2.5 Chao-Shen Estimator

Another recently proposed estimator is due to Chao and Shen (2003). This approach applies the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) in combination with the Good-Turing correction (Good, 1953; Orlitsky et al., 2003) of the empirical cell probabilities to the problem of entropy estimation. The Good-Turing-corrected frequency estimates are

$$\hat{\theta}_k^{GT} = (1 - \frac{m_1}{n})\hat{\theta}_k^{ML},$$

where m_1 is the number of singletons, that is, cells with $y_k = 1$. Used jointly with the Horvitz-Thompson estimator this results in

$$\hat{H}^{CS} = -\sum_{k=1}^{p} \frac{\hat{\theta}_k^{GT} \log \hat{\theta}_k^{GT}}{(1 - (1 - \hat{\theta}_k^{GT})^n)},$$

an estimator with remarkably good statistical properties (Vu et al., 2007).

3. A James-Stein Shrinkage Estimator

The contribution of this paper is to introduce an entropy estimator that employs James-Stein-type shrinkage at the level of cell frequencies. As we will show below, this leads to an entropy estimator that is highly effective, both in terms of statistical accuracy and computational complexity.

James-Stein-type shrinkage is a simple analytic device to perform regularized high-dimensional inference. It is ideally suited for small-sample settings - the original estimator (James and Stein, 1961) considered sample size $n = 1$. A general recipe for constructing shrinkage estimators is given in *Appendix A*. In this section, we describe how this approach can be applied to the specific problem of estimating cell frequencies.

James-Stein shrinkage is based on averaging two very different models: a high-dimensional model with low bias and high variance, and a lower dimensional model with larger bias but smaller variance. The intensity of the regularization is determined by the relative weighting of the two models. Here we consider the convex combination

$$\hat{\theta}_k^{Shrink} = \lambda t_k + (1 - \lambda)\hat{\theta}_k^{ML}, \tag{4}$$

where $\lambda \in [0,1]$ is the shrinkage intensity that takes on a value between 0 (no shrinkage) and 1 (full shrinkage), and t_k is the shrinkage target. A convenient choice of t_k is the uniform distribution $t_k = \frac{1}{p}$. This is also the maximum entropy target. Considering that $\text{Bias}(\hat{\theta}_k^{\text{ML}}) = 0$ and using the unbiased estimator $\widehat{\text{Var}}(\hat{\theta}_k^{\text{ML}}) = \frac{\hat{\theta}_k^{\text{ML}}(1-\hat{\theta}_k^{\text{ML}})}{n-1}$ we obtain (cf. *Appendix A*) for the shrinkage intensity

$$\hat{\lambda}^{\star} = \frac{\sum_{k=1}^{p} \widehat{\text{Var}}(\hat{\theta}_k^{\text{ML}})}{\sum_{k=1}^{p} (t_k - \hat{\theta}_k^{\text{ML}})^2} = \frac{1 - \sum_{k=1}^{p} (\hat{\theta}_k^{\text{ML}})^2}{(n-1)\sum_{k=1}^{p} (t_k - \hat{\theta}_k^{\text{ML}})^2}. \tag{5}$$

Note that this also assumes a non-stochastic target t_k. The resulting plugin shrinkage entropy estimate is

$$\hat{H}^{\text{Shrink}} = -\sum_{k=1}^{p} \hat{\theta}_k^{\text{Shrink}} \log(\hat{\theta}_k^{\text{Shrink}}). \tag{6}$$

Remark 1 *There is a one to one correspondence between the shrinkage and the Bayes estimator. If we write $t_k = \frac{a_k}{A}$ and $\lambda = \frac{A}{n+A}$, then $\hat{\theta}_k^{Shrink} = \hat{\theta}_k^{Bayes}$. This implies that the shrinkage estimator is an empirical Bayes estimator with a data-driven choice of the flattening constants—see also Efron and Morris (1973). For every choice of A there exists an equivalent shrinkage intensity λ. Conversely, for every λ there exist an equivalent $A = n\frac{\lambda}{1-\lambda}$.*

Remark 2 *Developing $A = n\frac{\lambda}{1-\lambda} = n(\lambda + \lambda^2 + \ldots)$ we obtain the approximate estimate $\hat{A} = n\hat{\lambda}$, which in turn recovers the "pseudo-Bayes" estimator described in Fienberg and Holland (1973).*

Remark 3 *The shrinkage estimator assumes a fixed and known p. In many practical applications this will indeed be the case, for example, if the observed counts are due to discretization (see also the data example). In addition, the shrinkage estimator appears to be robust against assuming a larger p than necessary (see scenario 3 in the simulations).*

Remark 4 *The shrinkage approach can easily be modified to allow multiple targets with different shrinkage intensities. For instance, using the Good-Turing estimator (Good, 1953; Orlitsky et al., 2003), one could setup a different uniform target for the non-zero and the zero counts, respectively.*

4. Comparative Evaluation of Statistical Properties

In order to elucidate the relative strengths and weaknesses of the entropy estimators reviewed in the previous section, we set to benchmark them in a simulation study covering different data generation processes and sampling regimes.

4.1 Simulation Setup

We compared the statistical performance of all nine described estimators (maximum likelihood, Miller-Madow, four Bayesian estimators, the proposed shrinkage estimator (Equations 4–6), NSB und Chao-Shen) under various sampling and data generating scenarios:

- The dimension was fixed at $p = 1000$.

- Samples size n varied from 10, 30, 100, 300, 1000, 3000, to 10000. That is, we investigate cases of dramatic undersampling ("small n, large p") as well as situations with a larger number of observed counts.

The true cell probabilities $\theta_1, \ldots, \theta_{1000}$ were assigned in four different fashions, corresponding to rows 1-4 in Figure **1**:

1. Sparse and heterogeneous, following a Dirichlet distribution with parameter $a = 0.0007$,

2. Random and homogeneous, following a Dirichlet distribution with parameter $a = 1$,

3. As in scenario 2, but with half of the cells containing structural zeros, and

4. Following a Zipf-type power law.

For each sampling scenario and sample size, we conducted 1000 simulation runs. In each run, we generated a new set of true cell frequencies and subsequently sampled observed counts y_k from the corresponding multinomial distribution. The resulting counts y_k were then supplied to the various entropy and cell frequencies estimators and the squared error $\sum_{i=k}^{1000} (\theta_k - \hat{\theta}_k)^2$ was computed. From the 1000 repetitions we estimated the mean squared error (MSE) of the cell frequencies by averaging over the individual squared errors (except for the NSB, Miller-Madow, and Chao-Shen estimators). Similarly, we computed estimates of MSE and bias of the inferred entropies.

4.2 Summary of Results from Simulations

Figure **1** displays the results of the simulation study, which can be summarized as follows:

- Unsurprisingly, all estimators perform well when the sample size is large.

- The maximum likelihood and Miller-Madow estimators perform worst, except for scenario 1. Note that these estimators are inappropriate even for moderately large sample sizes. Furthermore, the bias correction of the Miller-Madow estimator is not particularly effective.

- The minimax and $1/p$ Bayesian estimators tend to perform slightly better than maximum likelihood, but not by much.

- The Bayesian estimators with pseudocounts $1/2$ and 1 perform very well even for small sample sizes in the scenarios 2 and 3. However, they are less efficient in scenario 4, and completely fail in scenario 1.

- Hence, the Bayesian estimators can perform better or worse than the ML estimator, depending on the choice of the prior and on the sampling scenario.

- The NSB, the Chao-Shen and the shrinkage estimator all are statistically very efficient with small MSEs in all four scenarios, regardless of sample size.

- The NSB and Chao-Shen estimators are nearly unbiased in scenario 3.

The three top-performing estimators are the NSB, the Chao-Shen and the prosed shrinkage estimator. When it comes to estimating the entropy, these estimators can be considered identical for practical purposes. However, the shrinkage estimator is the only one that simultaneously estimates cell frequencies suitable for use with the Shannon entropy formula (Equation 1), and it does so with high accuracy even for small samples. In comparison, the NSB estimator is by far the slowest method: in our simulations, the shrinkage estimator was faster by a factor of 1000.
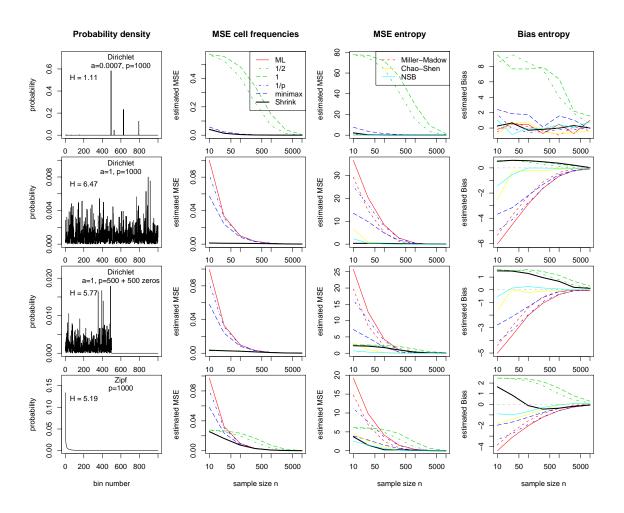
Figure 1: Comparing the performance of nine different entropy estimators (maximum likelihood, Miller-Madow, four Bayesian estimators, the proposed shrinkage estimator, NSB und Chao-Shen) in four different sampling scenarios (rows 1 to 4). The estimators are compared in terms of MSE of the underlying cell frequencies (except for Miller-Madow, NSB, Chao-Shen) and according to MSE and Bias of the estimated entropies. The dimension is fixed at $p = 1000$ while the sample size n varies from 10 to 10000.

5. Application to Statistical Learning of Nonlinear Gene Association Networks

In this section we illustrate how the shrinkage entropy estimator can be applied to the problem of inferring regulatory interactions between genes through estimating the nonlinear association network.

5.1 From Linear to Nonlinear Gene Association Networks

One of the aims of systems biology is to understand the interactions among genes and their products underlying the molecular mechanisms of cellular function as well as how disrupting these interactions may lead to different pathologies. To this end, an extensive literature on the problem of gene regulatory network "reverse engineering" has developed in the past decade (Friedman, 2004). Starting from gene expression or proteomics data, different statistical learning procedures have been proposed to infer associations and dependencies among genes. Among many others, methods have been proposed to enable the inference of large-scale correlation networks (Butte et al., 2000) and of high-dimensional partial correlation graphs (Dobra et al., 2004; Schäfer and Strimmer, 2005a; Meinshausen and Bühlmann, 2006), for learning vector-autoregressive (Opgen-Rhein and Strimmer, 2007a) and state space models (Rangel et al., 2004; Lähdesmäki and Shmulevich, 2008), and to reconstruct directed "causal" interaction graphs (Kalisch and Bühlmann, 2007; Opgen-Rhein and Strimmer, 2007b).

The restriction to linear models in most of the literature is owed at least in part to the already substantial challenges involved in estimating linear high-dimensional dependency structures. However, cell biology offers numerous examples of threshold and saturation effects, suggesting that linear models may not be sufficient to model gene regulation and gene-gene interactions. In order to relax the linearity assumption and to capture nonlinear associations among genes, entropy-based network modeling was recently proposed in the form of the ARACNE (Margolin et al., 2006) and MRNET (Meyer et al., 2007) algorithms.

The starting point of these two methods is to compute the mutual information $\mathrm{MI}(X,Y)$ for all pairs of genes X and Y, where X and Y represent the expression levels of the two genes for instance. The mutual information is the Kullback-Leibler distance from the joint probability density to the product of the marginal probability densities:

$$\mathrm{MI}(X,Y) = E_{f(x,y)} \left\{ \log \frac{f(x,y)}{f(x)f(y)} \right\}. \tag{7}$$

The mutual information (MI) is always non-negative, symmetric, and equals zero only if X and Y are independent. For normally distributed variables the mutual information is closely related to the usual Pearson correlation,

$$\mathrm{MI}(X,Y) = -\frac{1}{2}\log(1-\rho^2).$$

Therefore, mutual information is a natural measure of the association between genes, regardless whether linear or nonlinear in nature.

5.2 Estimation of Mutual Information

To construct an entropy network, we first need to *estimate* mutual information for all pairs of genes. The entropy representation

$$\mathrm{MI}(X,Y) = H(X) + H(Y) - H(X,Y), \tag{8}$$

1476

shows that MI can be computed from the joint and marginal entropies of the two genes X and Y. Note that this definition is equivalent to the one given in Equation **7** which is based on the Kullback-Leibler divergence. From Equation **8** it is also evident that $MI(X,Y)$ is the information shared between the two variables.

For gene expression data the estimation of MI and the underlying entropies is challenging due to the small sample size, which requires the use of a regularized entropy estimator such as the shrinkage approach we propose here. Specifically, we proceed as follows:

- As a prerequisite the data must be discrete, with each measurement assuming one of K levels. If the data are not already discretized, we propose employing the simple algorithm of Freedman and Diaconis (1981), considering the measurements of all genes simultaneously.

- Next, we estimate the $p = K^2$ cell frequencies of the $K \times K$ contingency table for each pair X and Y using the shrinkage approach (Eqs. 4 and 5). Note that typically the sample size n is much smaller than K^2, thus simple approaches such as ML are not valid.

- Finally, from the estimated cell frequencies we calculate $H(X)$, $H(Y)$, $H(X,Y)$ and the desired $MI(X,Y)$.

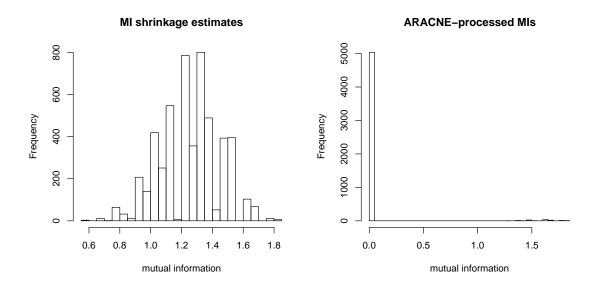### 5.3 Mutual Information Network for *E. Coli* Stress Response Data

Figure 2: Left: Distribution of estimated mutual information values for all 5151 gene pairs of the *E. coli* data set. Right: Mutual information values after applying the ARACNE gene pair selection procedure. Note that the most MIs have been set to zero by the ARACNE algorithm.

For illustration, we now analyze data from Schmidt-Heck et al. (2004) who conducted an experiment to observe the stress response in *E. Coli* during expression of a recombinant protein. This

data set was also used in previous linear network analyzes, for example, in Schäfer and Strimmer (2005b). The raw data consist of 4289 protein coding genes, on which measurements were taken at 0, 8, 15, 22, 45, 68, 90, 150, and 180 minutes. We focus on a subset of $G = 102$ differentially expressed genes as given in Schmidt-Heck et al. (2004).

Discretization of the data according to Freedman and Diaconis (1981) yielded $K = 16$ distinct gene expression levels. From the $G = 102$ genes, we estimated MIs for 5151 pairs of genes. For each pair, the mutual information was based on an estimated 16×16 contingency table, hence $p = 256$. As the number of time points is $n = 9$, this is a strongly undersampled situation which requires the use of a regularized estimate of entropy and mutual information.

The distribution of the shrinkage estimates of mutual information for all 5151 gene pairs is shown in the left side of Figure **2**. The right hand side depicts the distribution of mutual information values after applying the ARACNE procedure, which yields 112 gene pairs with nonzero MIs.

The model selection provided by ARACNE is based on applying the information processing inequality to all gene triplets. For each triplet, the gene pair corresponding to the smallest MI is discarded, which has the effect to remove gene-gene links that correspond to indirect rather than direct interactions. This is similar to a procedure used in graphical Gaussian models where correlations are transformed into partial correlations. Thus, both the ARACNE and the MRNET algorithms can be considered as devices to approximate the conditional mutual information (Meyer et al., 2007). As a result, the 112 nonzero MIs recovered by the ARACNE algorithm correspond to statistically detectable direct associations.

The corresponding gene association network is depicted in Figure **3**. The most striking feature of the graph are the "hubs" belonging to genes hupB, sucA and nuoL. hupB is a well known DNA-binding transcriptional regulator, whereas both nuoL and sucA are key components of the *E. coli* metabolism. Note that a Lasso-type procedure (that implicitly limits the number of edges that can connect to each node) such as that of Meinshausen and Bühlmann (2006) cannot recover these hubs.

6. Discussion

We proposed a James-Stein-type shrinkage estimator for inferring entropy and mutual information from small samples. While this is a challenging problem, we showed that our approach is highly efficient both statistically and computationally despite its simplicity.

In terms of versatility, our estimator has two distinct advantages over the NSB and Chao-Shen estimators. First, in addition to estimating the entropy, it also provides the underlying multinomial frequencies for use with the Shannon formula (Equation 1). This is useful in the context of using mutual information to quantify non-linear pairwise dependencies for instance. Second, unlike NSB, it is a fully analytic estimator.

Hence, our estimator suggests itself for applications in large scale estimation problems. To demonstrate its application in the context of genomics and systems biology, we have estimated an entropy-based gene dependency network from expression data in *E. coli*. This type of approach may prove helpful to overcome the limitations of linear models currently used in network analysis.

In short, we believe the proposed small-sample entropy estimator will be a valuable contribution to the growing toolbox of machine learning and statistics procedures for high-dimensional data analysis.
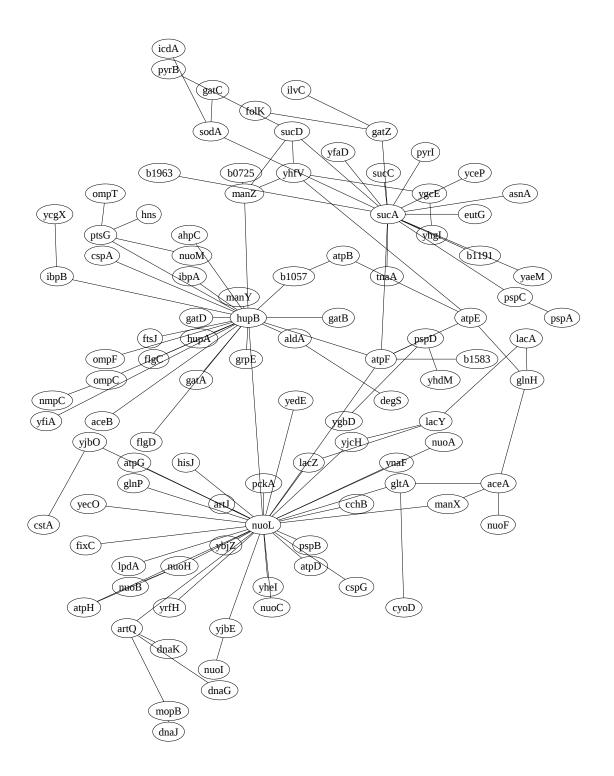
Figure 3: Mutual information network for the *E. coli* data inferred by the ARACNE algorithm based on shrinkage estimates of entropy and mutual information.

Acknowledgments

Appendix A. Recipe For Constructing James-Stein-type Shrinkage Estimators

The original James-Stein estimator (James and Stein, 1961) was proposed to estimate the mean of a multivariate normal distribution from a single ($n = 1$!) vector observation. Specifically, if x is a sample from $N_p(\mu, I)$ then James-Stein estimator is given by

$$\hat{\mu}_k^{\text{JS}} = (1 - \frac{p-2}{\sum_{k=1}^p x_k^2}) x_k.$$

Intriguingly, this estimator outperforms the maximum likelihood estimator $\hat{\mu}_k^{\text{ML}} = x_k$ in terms of mean squared error if the dimension is $p \geq 3$. Hence, the James-Stein estimator dominates the maximum likelihood estimator.

The above estimator can be slightly generalized by shrinking towards the component average $\bar{x} = \sum_{k=1}^p x_k$ rather than to zero, resulting in

$$\hat{\mu}_k^{\text{Shrink}} = \hat{\lambda}^\star \bar{x} + (1 - \hat{\lambda}^\star) x_k$$

with estimated shrinkage intensity

$$\hat{\lambda}^\star = \frac{p-3}{\sum_{k=1}^p (x_k - \bar{x})^2}.$$

The James-Stein shrinkage principle is very general and can be put to to use in many other high-dimensional settings. In the following we summarize a simple recipe for constructing James-Stein-type shrinkage estimators along the lines of Schäfer and Strimmer (2005b) and Opgen-Rhein and Strimmer (2007a).

In short, there are two key ideas at work in James-Stein shrinkage:

i) regularization of a high-dimensional estimator $\hat{\theta}$ by linear combination with a lower-dimensional target estimate $\hat{\theta}^{\text{Target}}$, and

ii) adaptive estimation of the shrinkage parameter λ from the data by quadratic risk minimization.

A general form of a James-Stein-type shrinkage estimator is given by

$$\hat{\theta}^{\text{Shrink}} = \lambda \hat{\theta}^{\text{Target}} + (1 - \lambda) \hat{\theta}. \tag{9}$$

Note that $\hat{\theta}$ and $\hat{\theta}^{\text{Target}}$ are two very different estimators (for the same underlying model!). $\hat{\theta}$ as a high-dimensional estimate with many independent components has low bias but for small samples a potentially large variance. In contrast, the target estimate $\hat{\theta}^{\text{Target}}$ is low-dimensional and therefore is generally less variable than $\hat{\theta}$ but at the same time is also more biased. The James-Stein estimate

is a weighted average of these two estimators, where the weight is chosen in a data-driven fashion such that $\hat{\theta}^{\text{Shrink}}$ is improved in terms of mean squared error relative to *both* $\hat{\theta}$ and $\hat{\theta}^{\text{Target}}$.

A key advantage of James-Stein-type shrinkage is that the optimal shrinkage intensity $\lambda^\star$ can be calculated analytically and without knowing the true value θ, via

$$\lambda^\star = \frac{\sum_{k=1}^{p} \text{Var}(\hat{\theta}_k) - \text{Cov}(\hat{\theta}_k, \hat{\theta}_k^{\text{Target}}) + \text{Bias}(\hat{\theta}_k)\,E(\hat{\theta}_k - \hat{\theta}_k^{\text{Target}})}{\sum_{k=1}^{p} E[(\hat{\theta}_k - \hat{\theta}_k^{\text{Target}})^2]}. \tag{10}$$

A simple estimate of $\lambda^\star$ is obtained by replacing all variances and covariances in Equation **10** with their empirical counterparts, followed by truncation of $\hat{\lambda}^\star$ at 1 (so that $\hat{\lambda}^\star \leq 1$ always holds).

Equation **10** is discussed in detail in Schäfer and Strimmer (2005b) and Opgen-Rhein and Strimmer (2007a). More specialized versions of it are treated, for example, in Ledoit and Wolf (2003) for unbiased $\hat{\theta}$ and in Thompson (1968) (unbiased, univariate case with deterministic target). A very early version (univariate with zero target) even predates the estimator of James and Stein, see Goodman (1953). For the multinormal setting of James and Stein (1961), Equation **9** and Equation **10** reduce to the shrinkage estimator described in Stigler (1990).

James-Stein shrinkage has an empirical Bayes interpretation (Efron and Morris, 1973). Note, however, that only the first two moments of the distributions of $\hat{\theta}^{\text{Target}}$ and $\hat{\theta}$ need to be specified in Equation **10**. Hence, James-Stein estimation may be viewed as a *quasi-empirical Bayes* approach (in the same sense as in quasi-likelihood, which also requires only the first two moments).

Appendix B. Computer Implementation

The proposed shrinkage estimators of entropy and mutual information, as well as all other investigated entropy estimators, have been implemented in R (R Development Core Team, 2008). A corresponding R package "entropy" was deposited in the R archive CRAN and is accessible at the URL `http://cran.r-project.org/web/packages/entropy/` under the GNU General Public License.

References

A. Agresti and D. B. Hitchcock. Bayesian inference for categorical data analysis. *Statist. Meth. Appl.*, 14:297–330, 2005.

A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA*, 97:12182–12186, 2000.

A. Chao and T.-J. Shen. Nonparametric estimation of Shannon's index of diversity when there are unseen species. *Environ. Ecol. Stat.*, 10:429–443, 2003.

A. Dobra, C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West. Sparse graphical models for exploring gene expression data. *J. Multiv. Anal.*, 90:196–212, 2004.

B. Efron and C. N. Morris. Stein's estimation rule and its competitors–an empirical Bayes approach. *J. Amer. Statist. Assoc.*, 68:117–130, 1973.

S. E. Fienberg and P. W. Holland. Simultaneous estimation of multinomial cell probabilities. *J. Amer. Statist. Assoc.*, 68:683–691, 1973.

D. Freedman and P. Diaconis. On the histogram as a density estimator: L2 theory. *Z. Wahrschein-lichkeitstheorie verw. Gebiete*, 57:453–476, 1981.

N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303:799–805, 2004.

S. Geisser. On prior distributions for binary trials. *The American Statistician*, 38:244–251, 1984.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, 2nd edition, 2004.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264, 1953.

L. A. Goodman. A simple method for improving some estimators. *Ann. Math. Statist.*, 24:114–117, 1953.

M. H. J. Gruber. *Improving Efficiency By Shrinkage*. Marcel Dekker, Inc., New York, 1998.

D. Holste, I. Große, and H. Herzel. Bayes' estimators of generalized entropies. *J. Phys. A: Math. Gen.*, 31:2551–2566, 1998.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, 47:663–685, 1952.

W. James and C. Stein. Estimation with quadratic loss. In *Proc. Fourth Berkeley Symp. Math. Statist. Probab.*, volume 1, pages 361–379, Berkeley, 1961. Univ. California Press.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. Roc. Soc. (Lond.) A*, 186:453–461, 1946.

M. Kalisch and P. Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Machine Learn. Res.*, 8:613–636, 2007.

R. E. Krichevsky and V. K. Trofimov. The performance of universal encoding. *IEEE Trans. Inf. Theory*, 27:199–207, 1981.

H. Lähdesmäki and I. Shmulevich. Learning the structure of dynamic Bayesian networks from time series and steady state measurements. *Mach. Learn.*, 71:185–217, 2008.

O. Ledoit and M. Wolf. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. *J. Empir. Finance*, 10:603–621, 2003.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.

A.A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Cali-fano. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 (Suppl. 1):S7, 2006.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, 34:1436–1462, 2006.

P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J. Bioinf. Sys. Biol.*, page doi:10.1155/2007/79879, 2007.

G. A. Miller. Note on the bias of information estimates. In H. Quastler, editor, *Information Theory in Psychology II-B*, pages 95–100. Free Press, Glencoe, IL, 1955.

I. Nemenman, F. Shafee, and W. Bialek. Entropy and inference, revisited. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, pages 471–478, Cambridge, MA, 2002. MIT Press.

R. Opgen-Rhein and K. Strimmer. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statist. Appl. Genet. Mol. Biol.*, 6:9, 2007a.

R. Opgen-Rhein and K. Strimmer. From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data. *BMC Systems Biology*, 1:37, 2007b.

A. Orlitsky, N. P. Santhanam, and J. Zhang. Always Good Turing: asymptotically optimal probability estimation. *Science*, 302:427–431, 2003.

W. Perks. Some observations on inverse probability including a new indifference rule. *J. Inst. Actuaries*, 73:285–334, 1947.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL http://www.R-project.org. ISBN 3-900051-07-0.

C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. L. Wild, and F. Falciani. Modeling T-cell activation using gene expression profiling and state space modeling. *Bioinformatics*, 20:1361–1372, 2004.

J. Schäfer and K. Strimmer. An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, 21:754–764, 2005a.

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.*, 4:32, 2005b.

W. Schmidt-Heck, R. Guthke, S. Toepfer, H. Reischer, K. Duerrschmid, and K. Bayer. Reverse engineering of the stress response during expression of a recombinant protein. In *Proceedings of the EUNITE symposium, 10-12 June 2004, Aachen, Germany*, pages 407–412, 2004. Verlag Mainz.

T. Schürmann and P. Grassberger. Entropy estimation of symbol sequences. *Chaos*, 6:414–427, 1996.

S. M. Stigler. A Galtonian perspective on shrinkage estimators. *Statistical Science*, 5:147–155, 1990.

D.R. Stinson. *Cryptography: Theory and Practice*. CRC Press, 2006.

S. P. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Letters*, 80:197–200, 1998.

J. R. Thompson. Some shrinkage techniques for estimating the mean. *J. Amer. Statist. Assoc.*, 63: 113–122, 1968.

S. Trybula. Some problems of simultaneous minimax estimation. *Ann. Math. Statist.*, 29:245–253, 1958.

F. Tuyl, R. Gerlach, and K. Mengersen. A comparison of Bayes-Laplace, Jeffreys, and other priors: the case of zero events. *The American Statistician*, 62:40–44, 2008.

V. Q. Vu, B. Yu, and R. E. Kass. Coverage-adjusted entropy estimation. *Stat. Med.*, 26:4039–4060, 2007.

G. Yeo and C. B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comp. Biol.*, 11:377–394, 2004.