Centroid estimation in discrete high-dimensional spaces with applications in biology

Luis E. Carvalho and Charles E. Lawrence*

Division of Applied Mathematics, Brown University, 182 George Street, Providence, RI 02912

Communicated by David Mumford, Brown University, Providence, RI, December 28, 2007 (received for review May 24, 2007)

Maximum likelihood estimators and other direct optimization-based estimators dominated statistical estimation and prediction for decades. Yet, the principled foundations supporting their dominance do not apply to the discrete high-dimensional inference problems of the 21st century. As it is well known, statistical decision theory shows that maximum likelihood and related estimators use data only to identify the single most probable solution. Accordingly, unless this one solution so dominates the immense ensemble of all solutions that its probability is near one, there is no principled reason to expect such an estimator to be representative of the posterior-weighted ensemble of solutions, and thus represent inferences drawn from the data. We employ statistical decision theory to find more representative estimators, centroid estimators, in a general high-dimensional discrete setting by using a family of loss functions with penalties that increase with the number of differences in components. We show that centroid estimates are obtained by maximizing the marginal probabilities of the solution components for unconstrained ensembles and for an important class of problems, including sequence alignment and the prediction of RNA secondary structure, whose ensembles contain exclusivity constraints. Three genomics examples are described that show that these estimators substantially improve predictions of ground-truth reference sets.

prediction | statistical inference | computational biology | discrete decoding

In the past decade, high-throughput data-acquisition technologies have rendered datasets with sizes unimaginable to our predecessors, including the sequence of the human genome (1) and the products of numerous high-throughput technologies of the post-genome era (2), data warehouses of commercial and internet transactions (3), and surveys of the objects of the universe (4). Although the emergence of such large datasets seems to imply more precise parameter estimation, paradoxically just the opposite is becoming increasingly common. This paradox emerged because these technologies simultaneously opened opportunities to draw inferences on previously unanswerable high-dimensional questions.

Estimation and prediction have long been dominated by procedures that identify the most probable point, including maximum likelihood estimation (5), maximum *a posteriori* (MAP) estimates such as Viterbi decoding of hidden Markov models, and minimum "free-energy" structure predictions (6, 7). These types of estimators are referred as ML estimators (maximum likelihood-family estimators) in the remainder of this article. In addition, many algorithms that optimize scoring functions to produce estimates or predictions correspond to equivalent maximum likelihood estimation procedures (8, 9), and thus also yield ML estimators.

Historically, there have been good reasons for this dominance. ML estimates are intuitively appealing because they identify the point in the space of the unknowns for which the data have highest probability. In the prediction of molecular structures, if the energy of one structure is sufficiently lower than all of the others, its probability will be near one and thus it will dominate the ensemble. More importantly, the long dominance of ML estimators rests on a principled foundation showing that they possess a number of very desirable properties, at least in the historic setting in which they were developed and have been very successfully applied: low-dimensional continuous spaces. Specifically, ML estimators have three key advantageous properties, as reported by Wald and Cramér (10, 11): consistency, asymptotic normality, and asymptotic efficiency. However, these properties only hold asymptotically as the data increase relative to the number of unknowns, and only properly for continuous variables. Such conditions are not attained when interest is focused on the inference of high-dimensional (high-D) discrete unknowns. Thus, the principled foundation supporting ML estimators is absent in this new setting.

Furthermore, evidence has emerged indicating that, in practice, estimators that gather information from the full ensemble of solutions predict ground-truth reference sets better than ML estimators. Specifically, Miyazawa (9) described reliable alignments that outperformed maximum similarity alignment procedures in the prediction of protein structure. More recently, Ding *et al.* (12) derived centroid estimators for predicting RNA secondary structure, and showed that they outperform well established ML estimators. Thus, there is now evidence in principle and in practice suggesting the need for alternative estimation procedures.

Bayesian inference provides a very useful alternative that enjoys a number of advantages in continuous settings. However, mean values of these estimators are not applicable here because, in general, they will not provide discrete solutions. Also, when interest is not on the overall solution, but on individual components, maximization of the marginal probabilities has been proposed (13). In sequence alignment, Miyazawa developed reliable alignments that maximize marginal probabilities and showed that these estimates meet the problem's main constraints (9). For the special case of predicting RNA secondary structure, one of us and colleagues developed centroid predictions and showed that these meet this problem's constraints (12).

Here, we use statistical decision theory to broadly generalize and extend the results of Ding *et al.* (12), to formally develop an alternative class of centroid estimators, and to prove some related theorems.

Background

A common high-dimensional inference problem concerns the estimation of n correlated binary variables, θ, living on a subset of $\{0,1\}^n$. For example, in network identification one seeks to predict if pairs of nodes are either connected or not. In these applications binary variables, θ, are not observed directly, rather

Author contributions: L.E.C. and C.E.L. designed research; L.E.C. and C.E.L. performed research; L.E.C. analyzed data; and L.E.C. and C.E.L. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

*To whom correspondence should be addressed. E-mail: Charles_Lawrence@brown.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0712329105/DC1.

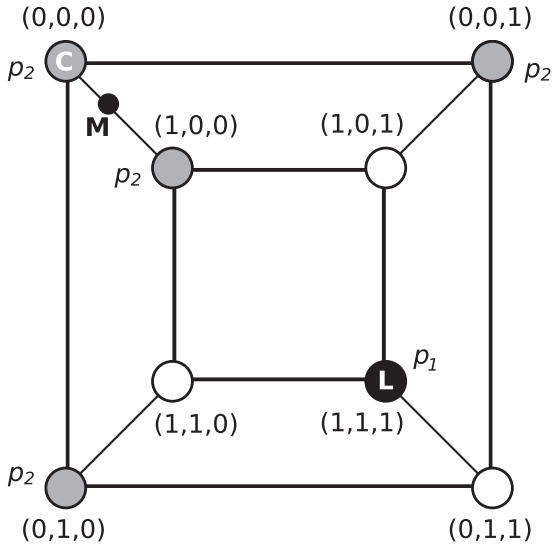

Fig. 1. Simple probability space and two different estimators: black point $L(1, 1, 1)$ has probability p_1 and is the ML estimate; gray points have probability p_2 as labeled, with $p_2 < p_1 < 2p_2$, including $C(0, 0, 0)$, and white, unshaded points have zero probability.

they have to be inferred from available data, S. For example, to predict gene networks, binary variables predict interactions between a pair of genes based on intensity data from microarray assays of expression levels of thousands of individual genes (14).

In maximum likelihood estimation, one finds the values of the unknowns that maximize the probability of the observed data, $\text{argmax}_\theta P(S|\theta)$. In Bayesian statistics, the other major statistical inference paradigm, Bayes's theorem gives the probability of allowable realizations of the set of unknown random variables θ after seeing the data S, and thus is called posterior probability:

$$P(\theta|S, \Lambda) = \frac{P(S|\theta, \Lambda)P(\theta|\Lambda)}{\sum_\theta P(S|\theta, \Lambda)P(\theta|\Lambda)} \quad [1]$$

where $P(\theta|\Lambda)$ is the prior probability of θ conditional on a set of parameters Λ and $P(S|\theta,\Lambda)$ is the likelihood of the data conditional on θ and Λ. The denominator in Eq. **1**, which is analogous to the partition function in statistical physics, is often called the marginal likelihood of the data. The common Bayesian ML estimator, the posterior probability maximizer, maximum *a posteriori* estimator is

$$\hat{\theta}_{\text{MAP}} = \text{argmax}_{\theta\in\Theta} P(\theta|S), \quad [2]$$

where, from now on, we denote by Θ the set of all possible realizations of θ.

Consider now the simple case of three correlated binary variables. Suppose that the eight combinations of these variables were equally likely *a priori* and that, after observing the data, the probabilities of the eight alternative combinations of these variables are those shown in Fig. 1, and $p_2 < p_1 < 2p_2$.

As Fig. 1 shows, although the point $L(1, 1, 1)$ is the most likely, it does not seem to represent the data well, because the only other points with positive probabilities are the four points that are about as different from this point as possible and together they are between two and four times more probable than point L. Point M is the mean with coordinates $(p_1 + p_2, p_1 + p_2, p_1 + p_2)$ and because, as discussed more formally below, $p_1 + p_2 < 1/2$, it is always closer to $C(0, 0, 0)$.

For the general case of n binary variables, where $p_2 < p_1 < (n - 1)p_2$, we could still have a cluster of $n + 1$ points that differ from some point C by no more than one component and such that each has probability p_2. The point at the opposite end of the hypercube L has probability p_1, all its neighbors have zero probability, and it differs from all of the rest with positive probability in at least $n - 1$ of its n components. Since $p_1 + (n + 1)p_2 = 1$, which implies that $1/(2n) < p_2 < 1/(n + 2)$, if we choose $p_2 = 1/(n + 3)$, then we would have $p_1 = 2/(n + 3)$. The ratio of the posterior mass for the cluster around point C to that around L is $(n + 1)p_2/p_1 = (n + 1)/2$. Thus, as the dimensionality of the problem increases, the proportion of posterior mass around point C becomes arbitrarily large when compared with L, and so does the distance between C and L. Nevertheless, since $p_2 < p_1$, point L is the ML estimate.

Although by design this is an extreme case, there are many other examples that present similar scenarios that raise concerns about the utility of ML estimators in high-dimensional discrete spaces. To examine these issues in less extreme and more realistic cases, we consider four questions: (*i*) Is there a principled basis that prevents ML estimators from being isolated from the region of space strongly recommended by the data? (*ii*) What alternative estimators better represent the data? (*iii*) Do these alternative estimators offer improved representation of the data in practice? (*iv*) Do these alternative estimators predict known references cases better?

Centroid Estimation

Is There a Principled Basis Preventing ML Estimators from Being Isolated from the Region of Space Strongly Recommended by the Data? In most high-dimensional circumstances, ML estimators are likely to have a very small probability because the single terms in the numerator of Eq. **1** will soon become swamped by the large number of terms in the denominator. We are not the first to recognize this fact, but many who recognize this often hoped that the ML estimator would be surrounded by a large number of similar solutions that together would contain a high proportion of the posterior probability mass (15). To find an estimator that will capture this hope, we employ statistical decision theory.

Statistical decision theory provides a principled means to optimally identify estimators with desirable properties through the use of loss functions. These functions assign losses to differences between the unknown value one seeks to infer and its estimate. Specifically, in the theory's jargon, we consider the loss $L(\hat{\theta}, \theta)$ associated with making the prediction $\hat{\theta}$ when the actual value of the unknown solution is θ, under loss function $L(\cdot)$ (11). To address the uncertainty inherent in estimation and prediction we seek an estimator that minimizes the expected loss, the *risk*. For example, to find estimators with minimum variance, the expected squared error loss is minimized. Specifically, the *posterior* risk is defined as the expected loss of choosing some estimator from S (16, 17),

$$\rho_L(\hat{\theta}(S)) = E_{\theta|S}[L(\theta,\hat{\theta}(S))] = \sum_{\theta\in\Theta} L(\theta,\hat{\theta}(S))P(\theta|S). \quad [3]$$

It is well known that ML estimators in this discrete setting are guaranteed to minimize posterior risk only under a zero–one loss function (13). Thus, any other point in the space that is different from the estimator incurs in a unitary penalty to the risk, regardless of how different they are.

Also, because these estimators ignore all other configurations but the most probable one, there is no principled reason for them not to differ greatly from all other solutions in the space, regardless of how strongly other points are supported by the data. Thus, because ML estimators in this setting only represent themselves and frequently have small posterior probabilities, these estimators are unlikely to be good representatives of the information contained in the observed data.

What Alternative Estimators Better Represent the Data? Although the hope of finding a high proportion of the posterior-weighted mass concentrated around an ML estimator has not been realized, nevertheless the concept of finding an estimator that does concentrate mass bares further consideration. Thus, to answer this question, we employ loss functions that incur higher losses as the difference in the number of components increases. The motivation is to better capture the character of the distribution of the posterior probability mass in the ensemble by seeking an estimator that represents collections of similar solutions, like point C in Fig. 1.

First, consider the *Hamming* loss,

$$H(z, y) = \sum_{i=1}^{n} I(z_i \neq y_i) = n - \sum_{i=1}^{n} I(z_i = y_i), \qquad [4]$$

where I is the indicator function, that is, $I(a) = 1$ if a is true and $I(a) = 0$ otherwise, and n is the dimension of both z and y. The Hamming loss function simply measures how many components differ between two members of a discrete solution space. Its posterior risk for some estimator $\hat{\theta}$ is

$$\begin{aligned}\rho_H(\hat{\theta}(S)) &= E_{\theta|S}[H(\theta, \hat{\theta}(S))] \\ &= \sum_{\theta\in\Theta} H(\theta, \hat{\theta}(S))P(\theta|S) \\ &= \sum_{\theta\in\Theta}\sum_{i=1}^{n} I(\theta_i \neq \hat{\theta}_i(S))P(\theta|S) \\ &= n - \sum_{\theta\in\Theta}\sum_{i=1}^{n} I(\theta_i = \hat{\theta}_i(S))P(\theta|S) \\ &= n - \sum_{i=1}^{n} P(\theta_i = \hat{\theta}_i(S)|S) \qquad [5]\end{aligned}$$

and so, it is immediate that, to minimize the risk, we can simply choose

$$\hat{\theta}_c(S) = \text{argmax}_{\hat{\theta}\in\Theta} \sum_{i=1}^{n} P(\theta_i = \hat{\theta}_i(S)|S), \qquad [6]$$

that is, the posterior *marginal* sum maximizer. We call $\hat{\theta}_C$ the *centroid estimator*, for which we have just presented the proof of the following equivalent definition:

Theorem 1. *$\hat{\theta}_C$ is the posterior Hamming loss risk minimizer.*

In the special case when $\Theta = A_1 \times A_2 \times \cdots \times A_n$, where A_i is the set of possible values for the ith entry in θ, we can take the *marginal* posterior maximizers for each position. That is, we choose an estimate by choosing the value of each component of the solution that is most probable, *consensus estimator*:

$$(\hat{\theta}^*_c)_i(S) = \text{argmax}_{a\in A_i} P(\theta_i = a|S), \quad i = 1, \ldots, n. \qquad [7]$$

Constrained Centroid Estimation

Centroid estimators also have their drawbacks. For instance, it might not be straightforward to derive a centroid estimator if the solution space is shaped by complex constraints. A naïve approach would be to employ the consensus estimator, but, since the inference is driven by the marginals, it is possible to find an estimate that is not feasible, for example, that does not belong to the solution space of the original problem. A simple example occurs when the space comprises three points (1, 0, 0), (0, 1, 0), and (0, 0, 1) with probabilities p_1, p_2, and $p_3 < 0.5$, respectively; the consensus estimator would be (0, 0, 0), which does not belong to the space. In general, taking the maximizers of the marginals is not a feasible solution in such a constrained problem, but under appropriate conditions it can be.

For many important applications, like RNA secondary structure or sequence alignment, the discrete unknowns are binary and the constraints have the characteristic form $\sum_{i\in J} \theta_i \leq 1$, where $J \subset \{1, \ldots, n\}$. Since, at most, one position in J can be matched, we say that J is restricted by an *exclusivity* constraint. This implies that, if we marginalize on J, then no two marginal sets of positions can have 1 at the same position. Therefore, we can reduce the problem to either selecting the alternative that has a probability greater than a half or, if none exists, assigning zero to all alternatives of this constraint. This would always yield a feasible centroid estimator. Formally, a more general result is available:

Theorem 2. *If $\Theta \subset \{0,1\}^n$ is such that $\theta \in \Theta$ satisfies a set of conditions $\{C_k\}_{k=1}^K$ of the form C_k: $\sum_{i\in J_k} \theta_i \leq 1$, where $\{J_k\}_{k=1}^K$ is a collection of index sets ($J_k \subset \{1,\ldots,n\}$, $1 \leq k \leq K$), then $\hat{\theta}^*_C$ also satisfies each condition C_k, $1 \leq k \leq K$, that is, $\hat{\theta}^*_C = \hat{\theta}_C$.* [*See supporting information (SI) Appendix for the proof.*]

Theorem 2 shows that, for problems in this class, consensus estimates will satisfy the original problem's constraint set even if the constraints overlap, and thus are centroid estimators. For example, in the sequence alignment problem there are two essential sets of constraints. First, if we view each solution as an array of binary variables θ_{ij}, $1 \leq i \leq n$ and $1 \leq j \leq m$ for sequences of size n and m, then each character in the first sequence should match with at most one other character in the second sequence, and vice versa: $\sum_{j=1}^{m} \theta_{ij} \leq 1$ for $1 \leq i \leq n$ and $\sum_{i=1}^{n} \theta_{ij} \leq 1$ for $1 \leq j \leq m$. The second set of constraints are collinearity constraints that prohibit the crossing of aligned character pairs: $\theta_{ij} + \theta_{kl} \leq 1$, $1 \leq i < k$ and $1 < j \leq n$. Because these are all exclusivity constraints, Theorem 2 applies.

Consider next pth power loss functions. These loss functions cover the broad class of loss functions that minimize pth order centered moments, including the important special case of the expected second centered moment. For categorical variables we can adopt a suitable binary representation to obtain the following result:

Theorem 3. *$\hat{\theta}_C$ is the posterior pth power loss risk minimizer.* (*See SI Appendix for the proof.*)

The estimator $\hat{\theta}_C$ minimizes the expected second moment centered around itself and so it is analogous to a multidimensional mean. As a matter of fact, under the same representation as before, it is the closest point to the mean.

Theorem 4. *$\hat{\theta}_C$ minimizes the squared distance to the posterior mean.* (*See SI Appendix for the proof.*)

Because this estimator is nearest to the center of mass of the posterior space, we call $\hat{\theta}_C$ the *centroid estimator.* Moreover, because $\hat{\theta}_C$ depends on the distance to other points and their probabilities, its behavior is quite different from $\hat{\theta}_{\text{MAP}}$: it seeks to find a point that minimizes the posterior-weighted distance to all points in the ensemble instead of choosing the single highest peak in the space. Thus, it pools the data's evidence from all points in the solution space.

Do These Alternative Estimators Offer Improved Representation of the Data in Practice?

Ding *et al.* (18) also applied centroid estimators to the characterization of messenger RNAs. In this study they showed that the variance about the ML estimate, minimum free-energy (MFE) structure, was on average 66% greater than the variance around

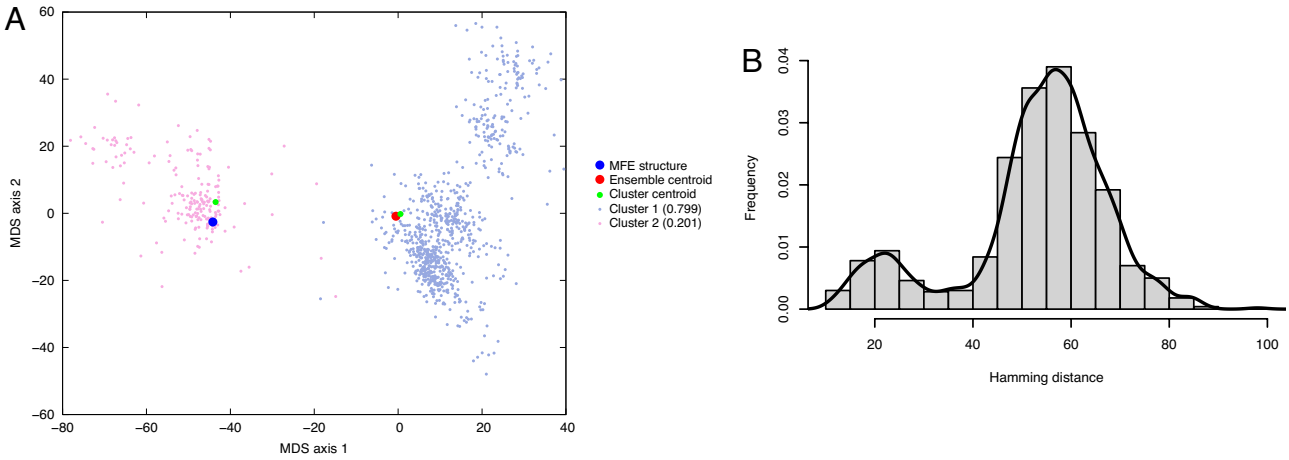

Fig. 2. Multidimensional scaled distribution (*A*) and histogram of distances to cluster 2 centroid (*B*) derived from 1,000 representative samples from Sfold for the secondary structure of *Dermocarpa* sp. ribonuclease P RNA.

the centroid estimate, indicating that posterior space is often asymmetric and that the most likely structure was often far from the center of mass of the posterior space. Although not as extreme as the illustrative example in Fig. 1, Fig. 2 shows an example in which the most likely structure, the MFE, lies in the periphery of the posterior space. Ding *et al.* (18) also showed that the posterior space often contained multiple clusters, and that the most likely structure was not in the largest cluster for 55 of the 100 mRNAs in their study.

Do These Alternative Estimators Predict Known References Cases Better?

We know of three applications of centroid estimators that have compared their predictions of known references to those of ML estimators. Ding *et al.* (12) made such a comparison by using an energy model that was identical to the model used by the most popular RNA structural prediction web server mfold (6) that predicts the most likely structure. Thus, their comparisons contrast directly only the two estimators. They found that, on average, the predicted base pairs of centroids have 30% fewer prediction errors (positive predictive value improvement) than those in the most likely structure, while also correctly predicting 3.5% more base pairs (sensitivity improvement). By using a different set of free-energy parameters (19), Mathews also showed that consensus estimators, thus centroid estimators, of RNA secondary structure improve positive predictive values by, on average, eight percentage points compared with the MFE structure (20).

In a article on the reliable alignment of protein sequences, Miyazawa (9) used a probabilistic model to identify the marginal probabilities of pairs of aligned protein residues and used a consensus, centroid estimator to estimate an alignment. He compared the ability of these alignments and optimal alignments that correspond to most probable alignments in their ability to predict the x-ray crystal structures of 1 of 109 pairs of proteins from that of the other. He found that the most probable alignments predicted reference's gold standard crystal structures better than centroid alignments by at least 0.25 Å root mean square deviation (rmsd) in only 4 of the 109 protein pairs. For these four, the most probable alignment was, on average, 0.41 Å rmsd closer to the reference structure than the centroid alignment. However, he found 29 pairs for which the centroid alignment predicted the reference structure better than the most probable alignment by at least 0.25 Å with an average improvement of 0.81 Å rmsd, thus demonstrating the centroid estimator's ability to improve the prediction of protein structure.

The computational identification of the locations of the regulatory sites of genes is another important area of study in genomics. Algorithms for this purpose are commonly known as motif-finding algorithms. Recently, Newberg *at al.* (21) developed a Gibbs sampling algorithm that seeks to identify regulatory sites by using the sequences from multiple related species. In this study they showed that centroid solutions consistently outperformed ML estimators. For example, in their simulation study of 1,000-bp-long sequences from five yeast species, they found that the centroid estimator made from 11% to 35% fewer prediction errors than the ML estimators with equal or better sensitivity. The larger differences occur when sites are more difficult to identify.

Conclusions

ML estimators have dominated prediction and estimation for years. Our results indicate that this paradigm has serious theoretical and practical limitations, and that there are better alternatives. Specifically, by using statistical decision theory with loss functions that incur increased penalties with increasing difference in their components, we develop centroid estimators. These estimators center themselves in the posterior-weighted ensemble by balancing the forces of the members based on their posterior probabilities and their component-wise distances from the centroid estimator. Given the findings in three computational biology applications that centroid estimators substantially improve prediction of ground-truth reference sets without modification of the underlying probabilistic model and perhaps more importantly their principled foundation, centroid estimators offer a promising avenue for improved estimation and prediction in discrete high-D inference problems that are becoming increasingly common in the twenty-first century.

Additional reports also show the utility of exploring the full ensemble of solutions. For example, Bradley *et al.* (7) in studies of protein structure prediction show that it is useful to sample the ensemble of solutions to identify probable energy wells. In CONTRAfold (22), conditional log-linear models are used to specify a probability distribution for RNA secondary structures conditional on RNA sequence. RNA structure estimation is then defined by the maximization of the expected accuracy, where accuracy weights correctly paired positions by a sensitivity/specificity trade-off parameter γ; when $\gamma = 1$ the estimator is the centroid estimator. Hartemink *et al.* (14) found posterior model

averaging useful after the application of simulated annealing to visit high-scoring regions in inference of genetic regulatory networks, although our experience differs somewhat from theirs. In our experience any use of preliminary optimization such as simulated annealing is detrimental (21). Also, Zhang and Liu (23) found that the incorporation of a side-chain entropy term in a simple free-energy function significantly improved the discrimination of native protein structures from decoys.

Centroid estimation has many other potential applications outside computational molecular biology. A common area of application of high-D discrete inference is variable selection in which discrete choices are made for inclusion of variables in a model. For example, Casella and Moreno (24) treat variable selection in normal regression models through the use of intrinsic priors and select the model with highest posterior probability. Smith and Fahrmeir (25) consider model selection in functional magnetic resonance imaging analysis with indicator variables for inclusion of regressors defined on a lattice, and use marginal probabilities for variable selection. Tadesse *et al.* (26) formulate a clustering problem by using a multivariate normal mixture model. Observations are allocated to classes according to the mode of a marginal posterior distribution, and variables are selected if their marginal posterior probability of above a threshold a; if $a = 1/2$ they have a centroid estimator.

Several caveats are appropriate. Since at this stage only a few applications have been examined, the assessment of how well these estimators will predict reference results in other settings awaits further study. The estimators developed here are appropriate in the important set of problems involving categorical variables. However, when discrete spaces involve ordinal or interval variables, estimators based on other loss functions that still achieve the goal of centering estimates in the posterior space may be more appropriate.

Ding *el al.* (18) showed evidence of multimodal posterior ensembles. When the clusters within these spaces are well separated, no single solution, including the centroid, is likely to represent the posterior space well. In such cases, multiple centroids, one for each cluster, may be required (12, 18). An alternate explanation for the results of Ding *el al.* (12) and Mathews (19) showing that there are better predictors of RNA structure than the minimum free-energy structure is that the energy model of these two works is incomplete. Thus, if evolution selects sequences that will adopt the minimum-energy states as the native states, the minimum of the incomplete (secondary structure) energy models of these two studies may not correspond with this overall energy minimum, and thus yields an incorrect prediction.

Although these estimators represent the posterior space in a defined manner, left unaddressed is the question of how representative a proposed estimator is of the specific ensemble from which it is drawn, thus leaving for future development the need to report how far the "true" state of nature may be from the proposed estimate. Whereas our findings on feasibility cover several important cases, for other cases further steps may be required to ensure feasibility. Dynamic programming offers a promising avenue to obtain these estimates while satisfying the constraints of the underlying problem (15, 21). Moreover, whereas for most constraints maintaining feasibility is important, it may not always be desirable to maintain all constraints. For example, it may be desirable to relax constraints imposed to achieve computability such as the no-pseudo-knot constraints of the RNA secondary structure algorithms. Also, centroid estimators focus on making reliable predictions. In the extreme case, posterior space is so widely dispersed the centroid estimator is empty, for example, with no margins >0.5, reflecting the fact that the data provide no support for any prediction. An estimation procedure that forces a result in this circumstance is available (15).

As we have shown in several important cases, centroid estimates can be derived from the marginal probabilities of solution components. However, obtaining marginal distributions is often a hard problem. In such cases, the approximation methods like variational Bayes (27, 28) and Markov chain Monte Carlo (MCMC) (16) can be applied. However, many problems present a sufficiently complicated joint probability structure to render ML estimation intractable by direct optimization. In such instances, it is common to apply sampling methods, such as MCMC. Given such a sample, estimating the marginal distributions can usually be completed with linear complexity in the number of solution components, whereas obtaining global maximum for ML estimation usually requires a more computationally intensive sampling approach, such as simulated annealing (29).

Centroid and ML estimators diverge more as the complexity of the probability space grows and becomes more multimodal, structured, and correlated. When the consensus estimator is the centroid estimator, the converse is easily seen, because, for probability spaces where each dimension is independent of the others, to maximize the joint distribution is equivalent to maximizing the marginals and so the two estimators always coincide.

Rapid improvements in data-acquisition technologies promise to continue to dramatically increase the pool of data in many fields. Although these data will be of great benefit, they also have opened a new universe of high-D inference and prediction problems that will likely provide major data analytic challenges in the coming decades. Among these is the development of point estimators in discrete spaces that are the focus of the centroid estimators developed here. But the more general point estimation challenge is to find one or a small number of feasible solutions among the many in the ensemble that is by some appropriate measure representative of the full ensemble and suitable for the data structural features of the solution space. These new high-D data and unknowns will also almost certainly force a reexamination of extant approaches to interval estimation, hypothesis tests, and predictive inference. In important ways these new challenges hark back to the early days of statistical physics in the age of Newtonian mechanics. For here again we are confronted with large ensembles and entropic effects arising from their shear size. But here it is often insufficient to deliver only distributions and averages for low-dimensional features, but rather specific high-D results are often demanded. Thus, centroid point estimates are only a small step into the challenges being driven by the rapid advances of data-acquisition technology.

ACKNOWLEDGMENTS. We thank Profs. Don McClure, Dan Weinreich, Richard Stratt, Ben Raphael, and Stuart Geman from Brown University and Drs. Lee Newberg, Ye Ding, and Clarence Chan of the Wadsworth Center (Albany, NY) for useful discussions and suggestions. This work was supported by Department of Energy Grant DE-FG02-04ER63942 and National Institutes of Health Grant R01-HG01257 (to C.E.L.) and by the Center for Computational Molecular Biology at Brown University.

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2006) GenBank. *Nucleic Acids Res* 34(Database issue):D16–D20.
2. ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) project. *Science* 5696:636–640.
3. Metwally A, Agrawal D, Abbadi AE (2005) *Using Association Rules for Fraud Detection in Web Advertising Networks* (Trondheim, Norway), pp 169–180.
4. York DG, *et al.* (2000) The SLOAN digital sky survey: technical summary. *Astron J* 120:1579–1587.
5. Fisher RA (1921) On the mathematical foundations of theoretical statistics. *Philos Trans R Soc Lond Ser A*, 222:309–368.
6. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31:3406–3415.
7. Bradley P, Misura KMS, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868–1871.
8. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Nat Acad Sci USA* 87:2264–2268.

9. Miyazawa S (1994) A reliable sequence alignment method based on probabilities of residue correspondences. *Protein Eng* 8:999–1009.
10. Wald A (1950) *Statistical Decision Functions* (Wiley, New York).
11. Lehmann EL, Casella G (2003) *Theory of Point Estimation* (Springer, New York).
12. Ding Y, Chan CY, Lawrence CE (2005) RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* 11:1157–1166.
13. Besag J (1986) On the statistical analysis of dirty pictures. *J R Stat Soc* 48:259–302.
14. Hartemink AJ, Gifford DK, Jaakkola TS, Young RA (2002) *Combining locations and expression data for principled discovery of genetics network models.* in Pacific Symposium on Biocomputing, Vol 7, pp 437–449.
15. Durbin R, Eddy SR, Krogh A, Mitchison G (1999) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge Univ Press, Cambridge, UK).
16. Gelman A, Carlin JB, Stern HS, Rubin DB (2003) *Bayesian Data Analysis* (Chapman and Hall/CRC, New York), 2nd Ed.
17. Carlin BP, Louis TA (2000) *Bayes and Empirical Bayes Methods for Data Analysis* (Chapman and Hall/CRC, New York), 2nd Ed.
18. Ding Y, Chan CY, Lawrence CE (2006) Clustering of RNA secondary structures with application to messenger RNAs. *J Mol Biol* 359:554–571.
19. Mathews DH, *et al.* (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci USA* 101:7287–7292.
20. Mathews DH (2004) Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* 10:1178–1190.
21. Newberg LA, *et al.* (2007) A phylogenetic Gibbs sampler that yields centroid solutions for cis regulatory site prediction. *Bioinformatics*, 10.1093/bioinformatics/btm241.
22. Do CB, Woods DA, Batzoglou S (2006) CONTRAfold: RNA secondary structure prediction without energy-based models. *Bioinformatics* 22:e90–e98.
23. Zhang J, Liu JS (2006) On side-chain conformational entropy of proteins. *PLoS Comput Biol* 2:e168.
24. Casella G, Moreno E (2006) Objective bayesian variable selection. *J Am Stat Assoc* 101:157–167.
25. Smith M, Fahrmeir L (2007) Spatial bayesian variable selection with application to functional magnetic resonance imaging. *J Am Stat Assoc* 102:417–431.
26. Tadesse MG, Sha N, Vannucci M (2005) Bayesian variable selection in clustering high-dimensional data. *J Am Stat Assoc* 100:602–617.
27. Attias H (2000) A variational Bayesian framework for graphical models. *Adv Neural Inf Process Syst* 12:209–215.
28. Beal MJ, Ghahramani Z (2003) The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures. *Bayesian Stat* **7**, pp 453–464.
29. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220:671–680.