Maximum entropy sampling

Jon Lee

Volume 3, pp 1229–1234

in

Encyclopedia of Environmetrics
(ISBN 0471 899976)

Edited by

Abdel H. El-Shaarawi and Walter W. Piegorsch

Maximum entropy sampling

The goal of maximum entropy sampling is to choose a most informative subset of s random variables from a set of n random variables, subject to side constraints. A typical side constraint might be a budget restriction, where we have a cost for observing each random variable. Other possibilities include logical constraints (e.g. multiple choice or precedence constraints). In many situations, we can assume that the random variables are Gaussian, or that they can be suitably transformed.

We briefly set our notation. We assume that we have n Gaussian random variables

$$Y_j, \quad j \in N = \{1, 2, \ldots, n\} \tag{1}$$

Our goal is to choose the 'most informative' subset S from N, having s elements, possibly subject to additional constraints

$$\sum_{j \in S} a_{ij} \le b_i, \quad i \in M = \{1, 2, \ldots, m\} \tag{2}$$

We let Y_S denote the set of random variables indexed by S, and we let f_S denote the joint density function of Y_S. Our measure of information, which we seek to maximize, is the *Boltzmann–Shannon entropy*

$$h_N(S) = -E(\ln f_S(Y_S)) \tag{3}$$

We assume that the random variables have a joint Gaussian distribution, and we let C denote the covariance matrix for Y_N. Then, letting $C[S, T]$ denote the submatrix of C with rows indexed by S and columns indexed by T, we have that $C[S, S]$ is the covariance matrix of Y_S. It turns out that in this Gaussian case, the entropy $h_N(S)$ is just an increasing linear function of

$$H_N(S) = \ln \det C[S, S] \tag{4}$$

where det denotes determinant. So, in what follows, we refer to $H_N(S)$ as the entropy associated with Y_S, and it is this quantity that we seek to maximize.

The term entropy was coined by R. Clausius. Boltzmann [4] developed the concept mathematically when he built the foundations of statistical mechanics. Shannon [25] popularized entropy in the field of information theory. Shewry and Wynn [26] introduced the sampling problem (without side constraints) presented above, in the context of the optimal design of spatial sampling networks (see [5]–[9], [24] and the entry **Spatial design, optimal**).

Environmental Monitoring

The maximum entropy sampling problem takes on considerable importance in experimental situations in which we seek to gain information concerning potential observations at a large number of points while observing only a few. One such situation involves configuring or reconfiguring a network of spatially disbursed environmental monitoring stations. For example, the US National Atmospheric Deposition Program/National Trends Network (NADP/NTN) is a nationwide network of precipitation monitoring stations (see `http://nadp.sws.uiuc.edu`). Currently there are approximately 200 stations, mostly in the continental US, with data from some of these dating back to 1978. Data is collected on the chemistry of precipitation for monitoring geographical and temporal trends (*see* **Trend, detecting**). Precipitation at each station is collected weekly and analyzed for pH, hydrogen, sulfate, nitrate, ammonium, chloride, calcium, magnesium, potassium, and sodium. As a means of evaluating existing networks and assessing their configuration and possible reconfiguration, we can ask which set of stations, having some prespecified number, provides the most information. We can formalize this by focusing on a single chemical and deriving a covariance matrix. The covariance matrix is estimated in part from historical observations, and in part by interpolation using available data with an appropriate model (see [12], for example). Aggregating weekly data over months and applying a logarithmic transformation has been found to be valuable toward meeting the assumption that the underlying random variables have a joint Gaussian distribution. Wu and Zidek [27] carried out a detailed analysis of part of the NADP/NTN network using the maximum entropy framework.

In this setting, budget constraints are quite natural. We can also incorporate other types of logical constraints to meet potential historical or political concerns.

Computational Complexity

Ko et al. [16] established that the maximum entropy sampling problem is 'NP-Hard' even when there are no side constraints. Problems that are NP-Hard are those that are at least as hard as the class of decision (i.e. Yes/No) problems that can be efficiently solved (i.e. solved in polynomial time) by a nondeterministic Turing machine. Such a machine is an extremely powerful abstraction that can simultaneously explore an ever-increasing set of potential solutions. This places our problem beyond those for which theoretically efficient algorithms are likely to exist (see [11] for more details). This theoretical result is of a worst-case and asymptotic nature, so we should not give up hope for developing practical methods for solving actual instances of problems with moderate size. However, the result suggests that we should restrict our attention in looking for practical procedures to local search heuristics and semi-enumerative methods – note that pure enumeration of the $\binom{n}{s}$ possible solutions is already impractical for $n = 30$, $s = 15$ where we have 155 117 520 possible solutions.

Heuristics

Heuristics, although suffering from some degree of myopia, can be used to find reasonably good solutions. Below, we discuss some methods which work reasonably well, when there are no side constraints.

Mitchell [21, 22] implemented some exchange procedures for finding a local optimum. Guttorp et al. [12] used a greedy constructive procedure: Start with $S = \emptyset$; then, for $j = 1, 2, \ldots, s$, choose $k \in N - S$ so as to maximize $H_N(S + k)$, and then adjoin k to S. Ko et al. [16] experimented with this greedy approach and the following dual greedy variant: Start with $S = N$; then, for $j = 1, 2, \ldots, n - s$, choose $l \in S$ so as to maximize $H_N(S - l)$, and then remove l to S. They also used a simple exchange method, beginning from the output set S of the greedy methods described above. Namely, while possible, choose $k \in N - S$ and $l \in S$ so that $H_N(S + k - l) > H_N(S)$, and replace S with $S + k - l$. Mitchell's method is a variation on this idea, exploring a wider region of neighbors of S.

All of these methods can be implemented rather efficiently by employing various matrix identities (see [14] and [15] for a wealth of such identities). For the data sets with which we experimented, these various heuristic methods worked remarkably well, often finding an optimal solution when n is small (say less than 30).

Branch-and-bound

Exact algorithms for the maximum entropy sampling problem are based on the branch-and-bound framework. The **branch-and-bound algorithm** is a standard technique for solving integer *linear* programs (see [23], for example). Its application to maximum entropy sampling is not straightforward, because the entropy criterion is not a linear function in 0/1 indicator variables for the n stations. Nonetheless, Ko et al. [16] proposed the first branch-and-bound algorithm for the maximum entropy sampling problem.

Branching

Branch-and-bound works with subproblems of the original problem. Subproblems are determined by deleting some stations U from consideration and forcing some stations F to be chosen. It remains then to choose $s - |F|$ stations from the $n - u - f$ stations $N - U - F$, so as to maximize the conditional entropy

$$H_{N-U-F}(S|F) = \ln\det C_F[S, S] \tag{5}$$

where

$$C_F = C[N - F, N - F] - C[N - F, F](C[F, F])^{-1}C[F, N - F] \tag{6}$$

is the conditional covariance matrix for Y_{N-F} given Y_F. Adding $H_N(F)$ to an upper bound on $H_{N-U-F}(S|F)$ yields an upper bound for $H_N(S)$. If we have a good solution S^*, generated by a heuristic, and if this upper bound is less than $H_N(S^*)$, then we conclude that $F \not\subset S \not\subset N - U$ for every optimal solution S. A branching strategy selectively tests various sets U and F, with the hope of eventually discovering an optimum. In practice, we maintain a list of 'active subproblems', each of which is determined by its sets U and F. Initially, the only active subproblem is the original problem (for which $U = F = \emptyset$). At each stage of the algorithm, an active subproblem is selected and an upper bound is calculated for it. If the upper bound on $H_{N-U-F}(S|F)$ plus $H_N(F)$ is less than $H_N(S^*)$, then the subproblem is discarded. Otherwise, a station $j \in N - U - F$ is selected, and the subproblem is replaced with up to two new subproblems: one with j adjoined to U and the other with j adjoined to F. Then, for each of these two potential subproblems, we check if there is only one S' having s elements with $F \subset S' \subset N - U$. If there is, and this S' satisfies the side constraints and has $H_N(S') > H_N(S^*)$, then we replace S^* with S' and discard the subproblem.

Spectral Bounds

For the problem without side constraints, Ko et al. [16] used a spectral bounding method to calculate upper bounds on the optimum value. Specifically, they established the spectral upper bound

$$\sum_{l=1}^{s} \ln \lambda_l(C[N,N]) \tag{7}$$

where λ_l denotes the lth greatest eigenvalue. Using a Lagrangian methodology, Lee [17] demonstrated how this bound can be improved when there are side constraints. Specifically, Lee established and described a method for calculating a Lagrangian spectral upper bound as the solution π of the convex program:

$$\min_{\pi \in \mathbb{R}^m_+} \left\{ \sum_{l=1}^{s} \ln \lambda_l(D^\pi C[N,N] D^\pi) + \sum_{i \in M} \pi_i b_i \right\} \tag{8}$$

where D^π is the diagonal matrix having diagonal elements

$$D^\pi_{jj} = \exp\left\{ -\frac{1}{2} \sum_{i \in M} \pi_i a_{ij} \right\} \tag{9}$$

Continuous Relaxation

Anstreicher et al. [1, 2] introduced a continuous convex relaxation of the problem. They introduced the continuous nonlinear programming upper bound

$$\begin{aligned} & \max f(x) \\ & \text{such that } \sum_{j \in N} a_{ij} x_j \le b_i, \quad \forall i \in M \\ & \sum_{j \in N} x_j = s \\ & 0 \le x_j \le 1, \quad \forall j \in M \end{aligned} \tag{10}$$

where

$$\begin{aligned} f(x) = \ln\det \Big(& \operatorname{diag}\big(x_j^{p_j/2}\big) C[N,N] \operatorname{diag}\big(x_j^{p_j/2}\big) \\ & + \operatorname{diag}\big(d_j^{x_j} - d_j x_j^{p_j}\big)\Big) \end{aligned} \tag{11}$$

the constants $d_j > 0$ and $p_j \ge 1$ satisfy $d_j \le \exp(p_j - \sqrt{p_j})$, and $\operatorname{diag}(d_j) - C[N,N] \succeq 0$. Anstreicher et al. report success in solving environmental monitoring problems with n as large as 63.

Semidefinite Programming

Helmberg (personal communication) pointed out how the problem can be relaxed as a semidefinite program. Specifically, we have the semidefinite programming upper bound

$$\begin{aligned} & \max \ln\det\ ((C[N,N] - I) \circ Y + I) \\ & \text{such that } Y - \operatorname{diag}(Y)\operatorname{diag}(Y)^{\mathrm{T}} \succeq 0 \\ & \mathcal{A}(Y) \le \beta \end{aligned}$$

where Y is a symmetric matrix of variables which replaces xx^{T}, '$\circ$' denotes the Hadamard (i.e. element-wise) product, and the linear constraints on x are incorporated in the linear constraints $\mathcal{A}(Y) \le \beta$ (see [19] for more details). It does not seem that there have been any computational experiments with this bound.

More Bounds

By exploiting the identity

$$\begin{aligned} \ln\det C[S,S] = & \ln\det C \\ & + \ln\det C^{-1}[N-S, N-S] \end{aligned} \tag{12}$$

Anstreicher et al. [1, 2] demonstrated how *any* bound for the complementary problem of choosing a maximum entropy set of $n - s$ points with respect to the covariance matrix C^{-1} translates to a bound for the original problem (note that the side constraints must be appropriately complemented).

Hoffman et al. [13] used an inequality of [10] to demonstrate how *any* bound for a problem with 'extra independence' yields a bound for the original problem. Specifically, they showed that for any partition of N into nonempty sets $N_1, N_2, \ldots, N_p$ (for some p between 1 and n), we can reset c_{ij} to 0 when i and j are in different blocks of the partition. Although the entropy of any subset of N cannot decrease with this extra independence, the upper bounds may decrease. Hoffman et al. [13] also discussed local search methods for finding an appropriate choice of extra independence. Lee and Williams [20] developed additional ideas that use this idea of partitioning.

Remote Sampling

Anstreicher et al. [3] considered a related problem (see also [18]). In remote sampling, the goal is not to gain as much information about Y_N as possible, but to gain as much information about a disjoint set of unobservable 'target' random covariates Y_T. They adapted many of the techniques of Anstreicher et al. [1, 2] and Lee [17] to this maximum entropy remote sampling problem.

Concluding Remarks

Although most of the methods surveyed above have been tested to some extent, there is a dearth of publically available software. Computational experiments suggest that a combination of some of the methods surveyed are capable of routinely finding exact solutions to problems with perhaps $n = 75$. Besides a need for reliable software implementing the methods discussed above, improved bounds are needed to routinely find optimal solutions to larger problems.

References

[1] Anstreicher, K.M., Fampa, M., Lee, J. & Williams, J. (1996). Continuous relaxations for constrained maximum entropy sampling, in *Integer Programming and Combinatorial Optimization*, Vancouver, BC, 1996, Springer-Verlag, Berlin, pp. 234–248.

[2] Anstreicher, K.M., Fampa, M., Lee, J. & Williams, J. (1999). Using continuous nonlinear relaxations to solve constrained maximum entropy sampling problems, *Mathematical Programming* **85**, 221–240.

[3] Anstreicher, K.M., Fampa, M., Lee, J. & Williams, J. (2001). Maximum entropy remote sampling, *Discrete Applied Mathematics* **108**, 259–274.

[4] Boltzmann, L. (1877). Beziehung zwischen dem zweiten haupstatz der wärmetheorie und der wahrscheinlichkeitsrechnung resp. den sätzen über das wärmegleichgewicht (complexionen-theorie), *Sitzungsberichte der Kaiserlichen Akademie der Wissenschaften, Wien; Mathematisch-Naturwissenschaftliche Klasse* **76^{2}**, 373, 1877.

[5] Caselton, W.F. & Zidek, J. (1984). Optimal monitoring network designs, *Statistics and Probability Letters* **2**, 223–227.

[6] Caselton, W.F., Kan, L. & Zidek, J.V. (1991). Quality data network designs based on entropy, in *Statistics in the Environmental and Earth Science*, P. Guttorp & A. Walden, eds, Griffin, London.

[7] Fedorov, V. & Müller, W. (1989). Comparison of two approaches in the optimal design of an observation network, *Statistics* **20**, 339–351.

[8] Fedorov, V., Leonov, S., Antonovsky, M. & Pitovranov, S. (1987). The experimental design of an observation network: software and examples. Technical Report WP-87-05, International Institute for Applied Systems Analysis, Laxenburg, Austria.

[9] Fedorov, V. & Hackl, P. (1996). Optimal experimental design: spatial sampling, *Nova Journal of Mathematics Game Theory and Algebra* **4**, 55–78.

[10] Fischer, E.S. (1908). Über den Hadamardschen determinantensatz, *Archiv für Mathematik und Physik* **13**, 32–40.

[11] Garey, M.R. & Johnson, D.S. (1979). A guide to the theory of NP-completeness, *Computers and Intractability*, W.H. Freeman, San Francisco.

[12] Guttorp, P., Le, N.D., Sampson, P.D. & Zidek, J.V. (1992). Using entropy in the redesign of an environmental monitoring network. Technical Report 116, Department of Statistics, University of British Columbia.

[13] Hoffman, A., Lee, J. & Williams, J. (2001). New upper bounds for maximum entropy sampling. Technical Report RC21679, IBM, 2000. To appear in the proceedings of MODA 6.

[14] Horn, R.A. & Johnson, C.R. (1985). *Matrix Analysis*, Cambridge University Press, Cambridge.

[15] Horn, R.A. & Johnson, C.R. (1991). *Topics in Matrix Analysis*, Cambridge University Press, Cambridge.

[16] Ko, C.W., Lee, J. & Queyranne, M. (1995). An exact algorithm for maximum entropy sampling, *Operational Research* **43**, 684–691.

[17] Lee, J. (1998). Constrained maximum entropy sampling, *Operational Research* **46**, 655–664.

[18] Lee, J. (1998). Discussion on: a state-space-model approach to optimal spatial sampling design based on entropy, *Environmental and Ecological Statistics* **5**, 45–46.

[19] Lee, J. (2000). Semidefinite programming in experimental design, in *Handbook of Semidefinite Programming: Theory, Algorithms and Applications*, H. Wolkowicz,

R. Saigal & L. Vandenberghe, eds, Kluwer, Dordrecht, pp. 528–532.

[20] Lee, J. & Williams, J. (2000). A linear integer programming bound for maximum entropy sampling. Technical Report RC21845, IBM.

[21] Mitchell, T.J. (1974). An algorithm for the construction of ‘*D*-optimal’ experimental designs, *Technometrics* **16**, 203–210.

[22] Mitchell, T.J. (1974). Computer construction of ‘*D*-optimal’ first-order designs, *Technometrics* **16**, 211–220.

[23] Nemhauser, G.L. & Wolsey, L.A. (1988). *Integer and Combinatorial Optimization*, Wiley, New York.

[24] Sebastiani, P. & Wynn, H.P. (2000). Maximum entropy sampling and optimal Bayesian experimental design, *Journal of the Royal Statistical Society, Series B* **62**, 145–157.

[25] Shannon, C.E. (1948). A mathematical theory of communication, *The Bell System Technical Journal* **27**, 379–423, 623–656.

[26] Shewry, M.C. & Wynn, H.P. (1987). Maximum entropy sampling, *Journal of Applied Statistics* **46**, 165–170.

[27] Wu, S. & Zidek, J.V. (1992). An entropy based review of selected NADP/NTN network sites for 1983–86, *Atmospheric Environment* **26A**, 2089–2103.

(*See also* **Optimization**)

JON LEE