Optimal Bayesian Design by Inhomogeneous Markov Chain Simulation

Peter Müller[1], **Bruno Sansó**[2] and **Maria De Iorio**[3]

Abstract

We consider decision problems defined by a utility function and an underlying probability model for all unknowns. The utility function quantifies the decision maker's preferences over consequences. The optimal decision maximizes the expected utility function where the expectation is taken with respect to all unknowns, i.e., future data and parameters. In many problems the solution is not analytically tractable. For example, the utility function might involve moments that can only be computed by numerical integration or simulation. Also, the nature of the decision space, i.e., the set of all possible actions, might have a shape or dimension that complicates the maximization. The motivating application for this discussion is the choice of a monitoring network when the optimization is performed over the high dimensional set of all possible locations of monitoring stations, possibly including choice of the number of locations.

We propose an approach to optimal Bayesian design based on inhomogeneous Markov chain simulation. We define a chain such that the limiting distribution identifies the optimal solution. The approach is closely related to simulated annealing. Standard simulated annealing algorithms assume that the target function can be evaluated for any given choice of the variable with respect to which we wish to optimize. For optimal design problems the target function, i.e., expected utility, is in general not available for efficient evaluation and might require numerical integration. We overcome the problem by defining an inhomogeneous Markov chain on an appropriately augmented space. The proposed inhomogeneous Markov chain Monte Carlo method addresses within one simulation both problems, evaluation of the expected utility as well as maximization.

KEY WORDS: Decision theory; Expected utility maximization; Markov chain Monte Carlo; Optimal design.

[1]Department of Biostatistics, The University of Texas M. D. Anderson Cancer Center, Houston, TX.

[2]AMS, University of California, Santa Cruz, CA, and CESMa, Universidad Simón Bolívar, Caracas, Venezuela

[3]Imperial College, London, U.K.

Research suppported by NIH/NCI grant 2 R01 CA75981-04A1 and NSF/DMS grant DMS-9704934.

1 INTRODUCTION

We propose a simulation based approach to expected utility optimization. Consider a decision problem described by a utility function $u(d, \theta, y)$ and a probability model $p_d(\theta, y)$ on all unknowns. Here d is the decision, θ are unknown parameters and y are the observable data. The utility function models preferences over consequences and $p_d(\theta, y)$ models beliefs, possibly influenced by actions. It can be argued (Raiffa and Schlaifer, 1961; DeGroot, 1970) that a rational decision maker chooses the action that maximizes expected utility $U(d) = \int u(d, \theta, y)\, \mathrm{d}p_d(\theta, y)$. The expectation is taken with respect to all unknowns at the time of decision making. Many decision problems consider utility functions that involve inference loss. See, for example, Chaloner and Verdinelli (1995) for a review of Bayesian approaches to such decision problems traditionally known as optimal design problems. Here we focus instead on problems with more general utility functions. We propose an approach which is suitable when the expected utility is an analytically intractable integral, possibly due to a complex probability model, the nature of the utility function or the nature of the action space. For example, the decision about shrinking a network of monitoring stations requires maximization over all possible patterns of included and excluded candidate sites, typically allowing neither analytical maximization nor full enumeration. Such problems are discussed, for example, in Clayton et al. (1999) or Nychka and Saltzman (1998). Decision problems that require simulation based solutions also frequently arise in biomedical research problems. Berry (1993), Spiegelhalter et al. (1994) and Berry and Stangl (1996) review related problems.

Simulation based approaches to decision problems have been successfully used in many contexts. For example, Carlin et al. (1998) and Brockwell and Kadane (2001) propose simulation based solutions to sequential decision problems. In these papers simulation is used to evaluate expected utility integrals. Müller and Parmigiani (1996) propose a similar strategy for general non-sequential problems, exploiting the continuity of $U(d)$ to reduce computational effort. Extensive literature in operations research deals with simulation based approaches to solve decision problems represented in so called influence diagrams. See, for example, Bielza et al. (1999) for a discussion and references.

Expected utility integration is only one of the computational challenges in optimal design. The other computational challenge is the maximization over the design space. One approach to implement simulation based maximization are simulated annealing methods. See van Laarhoven and Aarts (1987) for a review of simulated annealing. Geman and Ge-

man (1984) use simulated annealing for maximum *a posteriori* estimation. Fei and Berliner (1991) and Laud et al. (1992) discuss convergence issues and alternative probing distributions. The paradigm of simulated annealing is non-homogenuous Markov chain simulation with a sequence of stationary distributions which increasingly concentrate around the point of maximum (or minimum). The approach proposed in this paper builds on simulated annealing algorithms by blending simulated annealing and standard Markov chain Monte Carlo integration to develop an algorithm that simultaneously addresses maximization and integration.

A computational problem which is similar to that of expected utility maximization arises in the evaluation of marginal *a posteriori* modes (MMAP). In this case the expected utility is replaced by the marginal posterior distribution. Both, expected utility and marginal posterior distributions, require the calculation of possibly high dimensional integrals. Expected utility maximization and MMAP share the same formal description as maximizing analytically intractable integrals. In independent work Doucet et al. (2002) and Robert et al. (1999) discuss an MMAP strategy that is formally similar to the simulation based optimal design approach proposed in this paper.

In Section 2 we introduce the basic strategy. Section 3 gives the specific algorithm. Section 4 illustrates the proposed approach with examples. Section 5 concludes with a final discussion.

2 AUGMENTED PROBABILITY MODEL

Consider the generic decision problem defined by a utility function $u(d, \theta, y)$ and a probability model $p_d(\theta, y)$. Typically the probability model factors as $p_d(\theta, y \mid d) = p(\theta)\, p_d(y \mid \theta)$ into a prior $p(\theta)$ which is invariant under the decision d and a sampling distribution $p_d(y \mid \theta)$. The optimal decision problem is formally described as

$$d^* = \arg\max_{d \in \mathcal{D}} U(d) \text{ with } U(d) = \int u(d, \theta, y)\, \mathrm{d}p_d(y \mid \theta)\, \mathrm{d}p(\theta). \tag{1}$$

The assumptions about the probability model are not critical. Little changes in the following discussion if $p_d(\theta, y)$ does not factor as assumed in (1), or if the prior probability model $p_d(\theta)$ depends on d. When the target function $U(d)$ is not available in closed form standard optimization approaches are difficult to implement. However, $U(d)$ can easily be approximated by appropriate Monte Carlo integration. This is generally possible by independent Monte Carlo simulation since both, prior $p(\theta)$ and sampling model $p_d(y \mid \theta)$, are

usuallly available for efficient computer simulation. We can generate a Monte Carlo sample $(\theta_j, y_j) \sim p(\theta_j)\, p_d(y_j|\theta_j)$, $j = 1, \ldots, M$, to obtain an approximation

$$\widehat{U}(d) = 1/M \sum u(d, \theta_j, y_j). \tag{2}$$

The use of such approximations is a common technique in many simulation based optimal design approaches, including, for example, Sun et al. (1996), Carlin et al. (1998), or Müller and Parmigiani (1996).

We propose instead to solve the decision problem (1) by recasting it as a simulation from a sequence of augmented probability models. The central idea is to define an auxiliary distribution $h_J(d, \cdot)$ such that the marginal distribution in d is proportional to $U(d)^J$ for a positive integer J. Assume that $u(d, \theta, y)$ is non-negative and bounded. We define an artificial distribution h_J as

$$h_J(d, y_1, \theta_1, \ldots, y_J, \theta_J) \propto \prod_{j=1}^{J} u(d, y_j)\; p_d(y_j, \theta_j) \quad . \tag{3}$$

Marginalizing over $(y_1, \ldots, \theta_J)$ we find that

$$h_J(d) \propto \left[\int u(d, y)\; \mathrm{d}p_d(y, \theta) \right]^J = U(d)^J,$$

as desired. In (3) we have augmented the original probability model on parameters and data to include the design parameter d as a random variable. Treating d as a random variable we can from now on simplify notation and write $p(y \mid \theta, d)$ instead of $p_d(y \mid \theta)$, etc.

Fixing $J = 1$ the described procedure defines a probability model that randomly generates designs d with probability proportional to the expected utility $U(d)$. This is exploited in Bielza et al. (1999) by defining a Markov chain Monte Carlo (MCMC) simulation to generate from $h(d, \theta, y) \equiv h_1(d, \theta, y)$ using MCMC schemes as reviewed, for example, in Tierney (1994). A summary of the algorithm proposed in Bielza et al. (1999), as well as the discussion in Müller and Parmigiani (1996) appears in Müller (1999).

The fact that the target distribution is $h(\theta, y, \theta)$, instead of a posterior distribution $p(\theta|y)$ as in traditional posterior MCMC simulation, does not hinder application of the same simulation strategies. Appropriate summaries of the simulation output in d allow inference about the optimal design d^*. For example, for univariate d one can find the optimal design as the mode of the histogram of simulated d values. However, such strategies are limited to low dimensional design vectors d, say up to $\dim(d) = 4$. For higher dimensional design spaces it becomes impracticable to estimate the mode of $h(d)$ from the simulation output. An

exception is the special case when $U(d)$ is very peaked and gives only negligible probability to designs d significantly away from d^*. The optimum can then be inferred as the sample average of simulated d values.

This special case motivates the following computer simulation. We start by simulating from h. Suitable MCMC algorithms are discussed in Section 3, below. In subsequent iterations of the MCMC algorithm, $n = 1, 2, \ldots$, we change the target distribution to h_J, incrementing $J = J_n$ as we proceed. For sufficiently large J the marginal distribution $h_J(d)$ is tightly concentrated on the set of optimal designs d^* (d^* need not be unique).

The outlined algorithm is related to simulated annealing for a target function $f(d) = \log U(d)$. Central to simulated annealing is a sequence of equilibrium distributions $\pi_n(d) \propto \exp\{f(d)/T_n\}$. Defining $T_n = 1/J_n$ we find $\pi_n(d) = h_{J_n}(d)$, highlighting the similarity to simulated annealing. However, standard simulated annealing requires that the target function $\log U(d)$ be available for direct evaluation. But it is exactly this evaluation of the expected utility integral which complicates the optimal design problem (1) in the first place.

3 INHOMOGENEOUS MCMC CHAINS

We use inhomogeneous MCMC simulation to implement simulation from $h_J(d, \cdot)$, with $J = J_n$ increasing across iterations n. The chain is set up in such a way that the stationary distribution for fixed J is h_J. To describe the chain we use the alternative state vector (d, v) where v is the average log observed utility, $v = 1/J \sum \log u(d, \theta_j, y_j)$.

Algorithm 1.

0. Assume that the current state of the chain is $(J, d, y_1, \theta_1, \ldots, y_J, \theta_J)$, or (d, v).

1. Propose a new design $\tilde{d} \sim g(\tilde{d}|d)$. We discuss the choice of g below. For the moment assume $g(\tilde{d}|d)$ is symmetric, $g(\tilde{d}|d) = g(d|\tilde{d})$. For example, g might be a normal random walk $g(\tilde{d}|d) = N(d, S)$.

2. Propose simulated experiments $(\tilde{\theta}_j, \tilde{y}_j) \sim p(\theta, y \mid \tilde{d})$, $j = 1, \ldots, J$. Evaluate the corresponding v value and record $(\tilde{d}, \tilde{v})$ as a proposed new state of the Markov chain.

3. Evaluate the acceptance probability

$$\alpha_J = \min\left[1, \exp(J\tilde{v} - Jv)\right].$$

With probability α_J accept the proposal and set $v \equiv \tilde{v}$. Otherwise leave v unchanged.

4. Let J_n denote the current value of J. Increase J to $J_{n+1} \geq J_n$, following a chosen cooling schedule $(J_n, n = 1, 2, \ldots)$, such that $J_n \to \infty$.

5. Repeat steps 1 through 4 until the chain has practically converged.

Steps 1 through 5 describe a Markov chain (d_n, v_n) in the state vector (d, v). For fixed J, the chain is also Markovian in the extended state vector $(d, y_1, \theta_1, \ldots, y_J, \theta_J)$. The acceptance probability α_J is a Metropolis-Hastings acceptance probability for a target distribution $h_J(d, y_1, \theta_1, \ldots, y_J, \theta_J)$ and a proposal distribution $q(\tilde{v}, \tilde{y}_1, \ldots, \tilde{\theta}_J \mid d, y_1, \ldots, \theta_J) = g(\tilde{d} \mid d) \prod p(\tilde{\theta}_j, \tilde{y}_j \mid \tilde{d})$. The choice of $g(\tilde{d}|d)$ is essentially arbitrary, subject only to some technical conditions to ensure that the homogeneous chain which is obtained by fixing $J = J_n$ converges to the limiting distribution h_J. See Tierney (1994) for a discussion of appropriate conditions. In the context of the proposed algorithm, the only critical condition is irreducibility of the chain. Essentially, for any current state d and any candidate d_1, there must be a positive probability of reaching state d_1 starting from d in finitely many transitions. If the decision space is continuous, the event of reaching d_1 is replaced by the event of reaching a neighborhood of d_1. See Tierney (1994) for details.

The conditions in Tierney (1994) assure convergence to h_J for fixed J. Convergence of the inhomogeneous Markov chain (d_n, v_n) requires additional consideration. We first introduce some notation for the involved distributions. Discussion of Algorithm 1 includes three related, but different, probability models. We have already introduced the sampling model $p(\theta, y|d)$, and the homogeneous equilibrium distribution $h_J(d, y_1, \ldots, \theta_J)$. A third probability model is given by the transition probabilities $p(d_{n+1}, v_{n+1} \mid d_n, v_n)$ the chain. Also, $p(\theta, y|d)$ and $h_J(d, y_1, \ldots, \theta_J)$ imply corresponding probability models for (d, v). Let $p_J(v|d)$ denote the probability model on v implied by $(\theta_j, y_j) \sim p(\theta, y|d)$, $j = 1, \ldots, J$. Let $h_J(d, v)$ denote the distribution on (d, v) that is implicitely defined by $h_J(d, y_1, \ldots, \theta_{J_n})$. We will use $h(d, v)$ and $p(v \mid d)$, without subindex J, as short notation for h_1 and p_1. Note that the conditional distribution $h_J(v|d)$ differs from $p_J(v|d)$. For example, for $J = 1$, the distribution $h(\theta, y \mid d)$ includes an additional factor $u(\cdot)$ compared to $p(\theta, y \mid d)$. Let $\mu(d) = \int v \, dh(v|d)$ denote the expected log utility under $h(d, v)$, and let

$$S^* = \{(d^*, v^*),\ U(d^*) \geq U(d)\, \forall d \neq d^*,\ v^* = \mu(d^*)\}$$

denote the set of optimal decisions d^* and corresponding expectations. Finally, for $i = (d', v')$ and $j = (d, v)$ let

$$P_{ij}(m, n) = p(d_n = d, v_n = v \mid d_m = d', v_m = v')$$

denote the multi-step transition probabilities for the inhomogeneous chain (d_n, v_n) and let $P_{ij}(n) = P_{ij}(n, n+1)$.

The following result gives sufficient conditions for the convergence of the inhomogeneous chain under the following assumptions. We assume a finite state space S for $(d, v) \in S$. In practice this might require to round v to a finite grid. We assume that u is bounded, $0 < u_0 < u(d, \theta, y) < u_1$. The positive lower bound can be assumed without loss of generality. If u_0 were negative we add an appropriate offset without changing the optimal decision. Also, we assume $p(v|d) > 0$ for all (d, v).

Theorem 1 *Let (d_n, v_n) denote the states of the Markov chain defined in Algorithm 1, and let $T_n = 1/J_n$. If the annealing schedule $\{T_n,\ n = 1, 2, \ldots\}$ satisfies*

$$T_n \geq \gamma / \log(n + c)$$

with $\gamma \geq b$ where b is defined in the appendix and $c > 0$, then the inhomogeneous Markov chain (d_n, v_n) is strongly ergodic. That is, there exists a probability distribution π^* defined on S, such that for all $m \geq 1$, $i, j \in S$

$$\lim_{n \to \infty} P_{ij}(m, n) = \pi^\star(j) \tag{4}$$

and π^* is uniform over S^*. Further, $\pi^* = \lim h_{J_n}$ as $n \to \infty$.

Proof: see the Appendix.

For practical implementation we recommend to increase J only to the point where $h_J(d)$ is sufficiently peaked to identify the desired optimal design within meaningful accuracy. Let Δd be the minimum difference in d such that any two designs with difference Δd are still practically distinguishable. Stop cooling (incrementing J) when $J = \bar{J}$ such that $V_J(d) \leq \Delta d$ where $V_J(d)$ is some measure of dispersion for $h_J(d)$, for example, the sample standard deviation of the designs simulated under h_J.

When J is capped at $\bar{J}$, then the algorithm essentially reduces to a homogeneous chain. The early iterations, $J_n < \bar{J}$, define a finite burn-in only. The purpose of the early transient is to avoid simulations from getting trapped in local modes. In the rare case that $U(d)$ was known to be unimodal, the burn-in could be dropped without harm.

In either case, capping J allows an important generalization. Algorithm 1 requires i.i.d. sampling from $p(\theta, y|d)$ in step 2. In many decision problems this is easily possible by sampling from the prior $p(\theta)$ and the sampling model $p(y|\theta, d)$. Both, prior and sampling

distribution, are typically chosen as some well known distributions which allow efficient random variate generation. However, complications arise if some data y^o is already observed at the time of decision making, replacing $p(\theta, y \mid d)$ by $p(\theta, y \mid y^o, d)$. Such posterior and posterior predictive simulation conditional on observed data y^o might require MCMC simulation and hinder i.i.d. simulation as required in step 2. But this generalization is easily accomodated by replacing i.i.d. simulation by an appropriate MCMC proposal $q_y(\tilde{y}_j, \tilde{\theta}_j | y_1, \theta_1, \tilde{d})$, $j = 1, \ldots, J$, and changing the acceptance probability α_J accordingly. The exact nature of the MCMC proposal q_y is problem specific.

Algorithm 2. Proceed with steps 1 through 5, with the following modifications:

2'. Propose simulated experiments $(\tilde{\theta}_j, \tilde{y}_j) \sim q_y(\tilde{y}_j, \tilde{\theta}_j | y_j, \theta_j, \tilde{d})$, $j = 1, \ldots, J$. Evaluate $\tilde{v}$ as the corresponding v value.

3'. Evaluate the acceptance probability

$$\alpha_J = \min \left[1, exp(J\tilde{v} - Jv) \prod_{j=1}^{J} \frac{p(\tilde{\theta}_j, \tilde{y}_j | \tilde{d})}{q_y(\tilde{y}_j, \tilde{\theta}_j | y_1, \theta_1, \tilde{d})} \frac{q_y(y_j, \theta_j | \tilde{y}_1, \tilde{\theta}_1, d)}{p(\theta_j, y_j | d)} \right]$$

With probability α_J accept the proposal and set $v = \tilde{v}$, $d = \tilde{d}$, $y_1 = \tilde{y}_1$, etc.

4'. Let J_n denote the current value of J. Increase J to J_{n+1}, $\bar{J} \geq J_{n+1} \geq J_n$. If $J_{n+1} > J_n$, generate new values (y_j, θ_j), $j = J_n + 1, \ldots, J_{n+1}$ by resampling the currently imputed experiments (y_j, θ_j), $j = 1, \ldots, J_n$,

Algorithm 2 is no longer a Markov chain in (d, v). The Markov property holds for the state vector $(J, d, y_1, \ldots, \theta_J)$ only. But the homgeneous chain obtained by fixing $J = J_n$ still has the desired asymptotic distribution h_J, allowing use of Algorithm 2 with J increasing up to some fixed upper bound $\bar{J}$.

Another variation of Algorithm 1 occurs when $U(d)$ can be evaluated exactly. In this case step 2 is removed and v is replaced by $\log U(d)$. Without the uncertainty in the evaluation of $U(d)$ the chain reduces to a standard simulated annealing algorithm. A variation of such a simulated annealing algorithm is described in Müller (1999). Suppose the decision parameter can be written as a vector $d = (d_1, \ldots, d_p)$. To simulate a sample from $h_J(d)$ a Gibbs sampling scheme is defined by iteratively generating from $h_J(d_J \mid d_1, \ldots, d_{j-1}, d_{j+1}, \ldots, d_p)$, scanning over all coordinates $j = 1, \ldots, p$. The complete conditional draws from $h_J(d_J \mid \ldots)$ are implemented as nested Metropolis-Hastings chains. The resulting Metropolis-within-Gibbs

scheme provides an approximate draw from $h_J(d)$. The need for practical convergence for each run of the nested Metropolis-Hastings makes the scheme very computation intensive. Algorithm 1 avoids this complication by defining one Markov chain that achieves both, the integration and the maximization in the same Markov chain simulation.

4 NETWORK DESIGN AND OTHER APPLICATIONS

The proposed algorithm for optimal design is very general. It is applicable for any problem that allows pointwise evaluation of the utility function for assumed future outcomes and assumed parameter values. The probability model on parameters and future data must allow efficient computer simulation, either by independent simulation or by appropriate posterior MCMC simulation. However, we do not recommend to use Algorithms 1 and 2 as default approach for generic decision problems. In particular, in many problems the structure of the decision space and the nature of the utility function imply appropriate regularity conditions for the expected utility function $U(d)$ that allow to proceed with smoothing based methods. See Müller and Parmigiani (1996) for a discussion.

The proposed algorithm is most suitable for problems with complicated decision space. A typical application arises in the design of monitoring networks. Assume a current network of monitoring stations, $i = 1, \ldots, n$, and a desired decision about shrinking the network. The actions are described by binary variables d_i, with $d_i = 0$ ($d_i = 1$) indicating that station i is dropped from (retained in) the network. A typical utility function combines sampling cost with some notion of parameter learning, prediction and interpolation. There is no convenient formalization of continuity or regularity for expected utility $U(d)$ as a function of the n-dimensional binary vector. In sections 4.1 and 4.2 we describe two applications to network design. The first example uses a kriging model and a utility function that awards accurate prediction. The application is taken from Sansó and Müller (1999) where we use simulation with $J = 1$ to explore the expected utility. The second example uses a process convolution model (Higdon, 2002) and a utility function that combines criteria related to violation of a given maximum ozone level, prediction, and inference about the mean ozone level acoss the entire region of interest.

Another class of interesting applications arises in variable selection. Usually the problem of selecting variables in a regression problem is cast as a problem of posterior inference, considering the indicators for variable inclusion as part of the unknown parameter vector. See, for example, Chipman et al. (2001) and Clyde and George (2003) for a review. Alternatively

the variable selection problem can be considered as a decision problem, adding a loss function. This approach is chosen in Berger and Barbieri (2003), using a utility function that reduces expected utility maximization to finding the median probability model. See Berger and Barbieri (2003) for details. Brown et al. (1999) propose a setup for variable selection that allows computation efficient evaluation of expected utility. The expected utility maximization is then implemented as a simulated annealing algorithm. When variable selection is desired for a planned future experiment, and when the nature of the utility function or the sampling model prohibit an analytic evaluation or good approximation of expected utility, the variable selection problem can be addressed by algorithms proposed in Section 3. We show an example in Section 4.3. We set up an optimal design problem with decisions related to variable selection in a probit model, and a utility function combining classification and sampling cost.

4.1 Optimal Shrinkage of a Network of Rainfall Stations

We consider the problem of redesigning a network of 80 rainfall stations. Due to funding constraints the network has to be reduced in size. The stations are located in the State of Guárico, in the plains of central Venezuela, north of the Orinoco and Apure rivers and are managed by the Venezuelan authority of the environment (MARNR). The goal is to reduce the number of stations to approximately half. This should be done in a way such that the ability to interpolate and predict local rainfall over the state is mantained as much as possible whilst, at the same time, the cost of running the network is minimized. In Figure 1 and Figure 2 we show the locations of the $n = 80$ stations together with a plot of the annual rainfall for three of the stations during $T = 16$ years from 1968 to 1983.

Several authors have studied related problems of designing monitoring networks for athmospheric data: Caselton et al. (1992) and Guttorp et al. (1994) consider an approach based on a multivariate normal and inverse Wishart model and minimize entropy. Nychka and Saltzman (1998) use kriging and show how the problem of designing the network is equivalent to a variable selection problem in a linear model. Bras and Rodríguez-Iturbe (1985) consider the design problem over a grid and minimize a cost function that is similar to the one we consider in this paper, although they only deal with a small number of stations. The problem of redesigning the Guárico network was originally studied in Sansó and Müller (1999) who propose a stochastic algorithm that explores possible designs, but without a formal optimization step.

In this example we consider an isotropic Gaussian spatial model as in Handcock and Stein

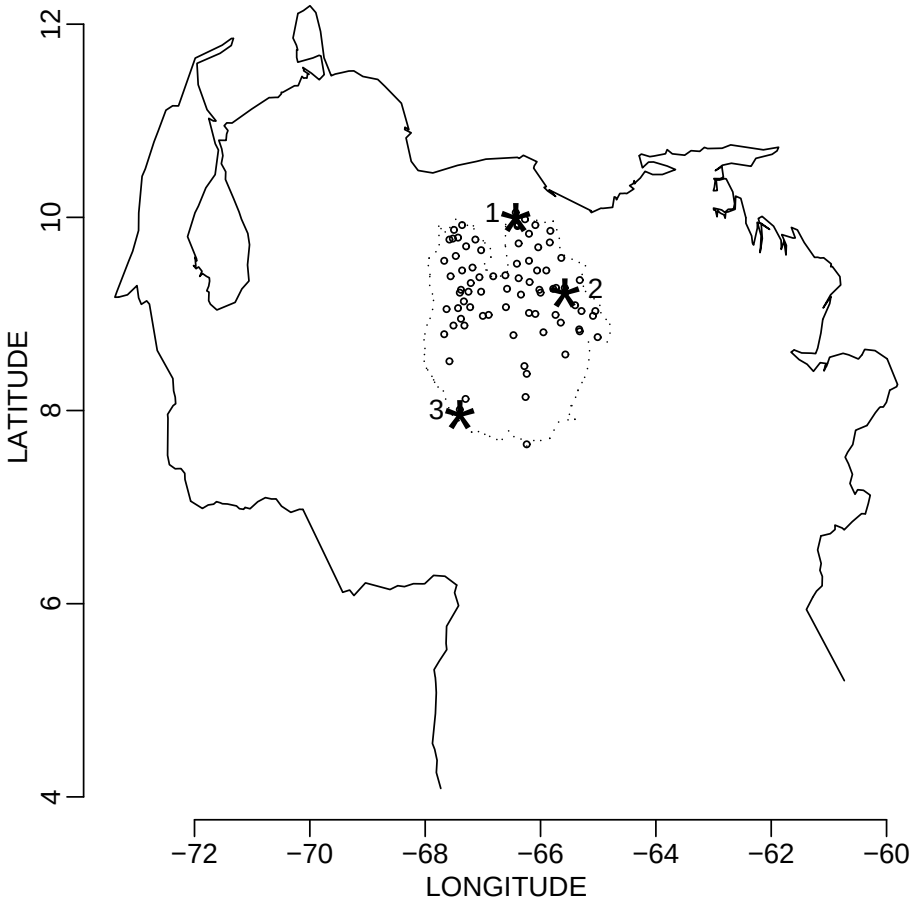


Figure 1: Locations of the 80 stations in Guárico, in the plains of central Venezuela, north of the Orinoco and Apure rivers. The dotted line shows the borders of Guárico. The locations labeled 1, 2 and 3 are the stations for which time series are shown in Figure 2.

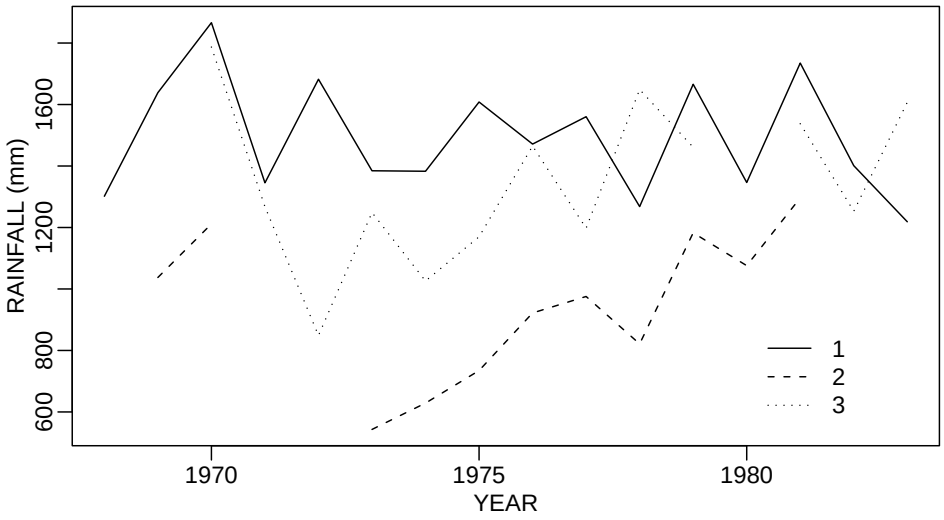


Figure 2: Annual rainfall at three locations, marked as 1 (solid line), 2 (dashed line) and 3 (dotted line) in Figure 1. The gaps are missing values.

(1993). We avoid the need of temporal modeling by considering annual rainfall and assuming that annual observations are exchangeable. To highlight the decision problem involved in the network shrinkage we keep the assumptions of the geostatistical model simple by using a standard model. A review of related, more complex models appears in Sansó and Guenni (1999).

4.1.1 The Probability Model

Let $y \in \mathbb{R}^n$ be the data that consist of rainfall observed at n stations scattered over the state of Guárico. Latitude (x_{i1}), longitude (x_{i2}), and elevation (x_{i3}) of the stations are known. We propose a Gaussian random field with a linear drift and an isotropic covariance structure, that is, we assume that correlations depend only on the relative distances between locations. Analysis of the historical data supports the model. We assume

$$y \sim N_n[x\beta, \sigma^2 V(\lambda)] \tag{5}$$

where $N_n(a, B)$ denotes a n-dimensional normal distribution with moments a and B. The design matrix x is an $(n \times p)$ matrix of p basis functions evaluated at the n station locations, β is an $(p \times 1)$ vector of linear regression parameters, $\sigma^2 > 0$ is a scale parameter and $V(\lambda)$ is a correlation matrix parametrised by $\lambda > 0$. We use $V_{ij}(\lambda) = \exp(-\lambda d_{ij})$ where d_{ij} is the distance between stations i and j. Let x_i denote the i-th row of the design matrix x. For the

mean function we choose a quadratic polynomial, i.e., $x_i\beta$ is a quadratic form in x_{i1} and x_{i2}. To simplify the model we do not use the elevations in the mean regression, a simplification that is sustained by the analysis of the historical data.

We complete the model with a prior distribution for the parameters $\theta = (\beta, \lambda, \sigma^2)$. We base the prior distribution on available historical data. Rather than formally defining a hierarchical probability model for joint inference on historical and future data, we choose the simpler mechanism of analyzing a separate probability model for the historical data, and using the posterior moments from that analysis to define the prior probability model for (5). Let z_{ij} be the log annual rainfall corresponding to station $i = 1, \ldots, n$ in year $j = 1, \ldots, T$. Recall that $n = 80$ and $T = 16$. Let z_j be the $(n \times 1)$ data vector for year j. We assume that the z_j are exchangeable across years and

$$z_j \sim N_n[x\gamma, \tau^2 V(\kappa)], \quad j = 1, \ldots, T, \tag{6}$$

as in (5), with (γ, κ, τ) replacing (β, λ, σ). In particular V is the same exponentially decaying correlation matrix. We assume *a priori* $p(\gamma, \kappa, \tau^2) \propto 1/\tau^2 p(\kappa)$ with $p(\kappa)$ being a gamma distribution with large variance and a mean chosen to match the empirical estimation of κ based on the covariogram of the data z.

Inference on (γ, κ, τ^2) was implemented by straightforward MCMC simulation. We now use posterior inference on (γ, κ, τ) to motivate prior choices for model (5). The marginal posterior distribution on κ is very peaked, leading us to fix λ at the posterior mean $E(\kappa|z)$. For (β, σ) we use a normal/inverse-Gamma prior, $1/\sigma^2 \sim Gamma(a, b)$ and $\beta|\sigma^2 \sim N(m, \sigma^2 S)$. The hyperparameters m, S, a and b are fixed to match the posterior means $E(\gamma, \tau^2 \mid z)$ and the ten-fold inflated marginal posterior variances, $10\,Var(\gamma \mid z)$ and $10\,Var(1/\tau^2 \mid z)$. Using inflated posterior variances to define prior distributions based on historical data is a standard procedure to account for lack of exchangeability across historical and future data.

4.1.2 Utility function

Let y_i denote log rainfall in a given year at station $i = 1, \ldots, n$ and let $d = (d_1, \ldots, d_n)$ be a vector that indicates a specific network design with $d_i = 0$ if station i is removed from the network and $d_i = 1$ if station i remains in the network. We need to specify a utility $u(d, y)$ associated with the decision d when the outcome y is observed. Let y_d be the subvector of y that corresponds to stations in the network under design d. Let $D = \{i : d_i = 1\}$ be the

set of stations in the network and $D^c = \{1, \ldots, n\} \setminus D$. We consider the utility function

$$u(d, y) = C \sum_{i \in D^c} \mathbf{1}\{\underbrace{y_i \in \hat{y}_i(y_d) \pm \delta}_{S_i}\} - \sum_{i \in D} c_i + C_0, \tag{7}$$

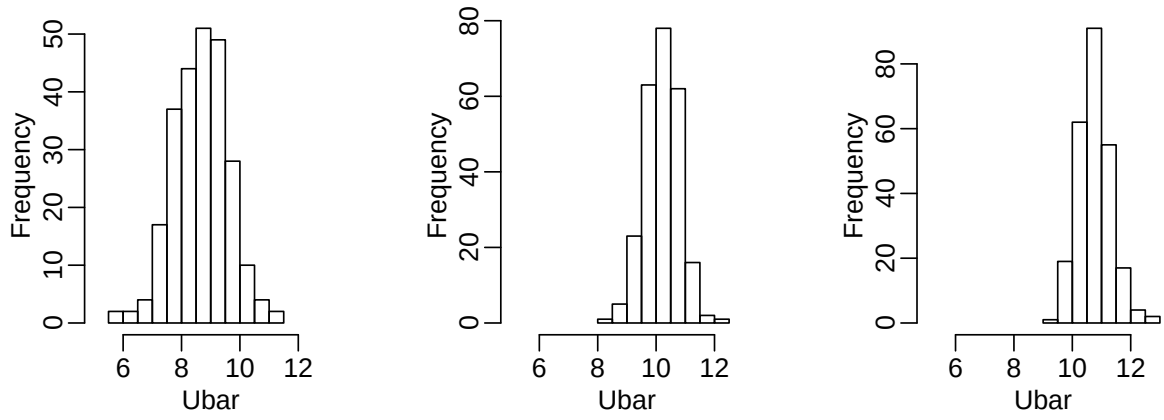
where $\mathbf{1}\{A\}$ denotes the indicator function of the event A and $\widehat{y}_i(y_d)$ is the prediction for station i based on the observed stations y_d, and C, C_0, δ and $c_i, i = 1, \ldots, n$ are positive constants. In words, we include a payoff C for predicting a station i that is dropped from the network, and we deduct a cost c_i for every station in the network. Prediction within $\pm\delta$ suffices. A constant C_0 is added to ensure a non-negative utility function, as needed in the conditions of Theorem 1. The idea behind this utility function is that a design d has a high utility if the total cost of running the network is low but the stations in the complement D^c can be reasonably well predicted by stations in D. Notice that $u(d, y)$ is bounded.

A key issue in the specification of the utility is the constant C. This involves a trade-off between sampling cost and payoff which is seldom easy to do. We interpret C as the cost of one individual measurement and assume that $(C - c_i) > 0$ since a station would not have been built if the cost of running it were greater than the utility of a measurement. The choice of δ is naturally related to the variability in the y. In our case we used the historical data and fixed δ as twice the residual standard deviation in fitting model (6).

We implemented Algorithm 1, changing J_n by increments of one every 25 iterations, running over 2500 iterations. The proposal distribution $g(\tilde{d} \mid d)$ is defined by randomly selecting a station $i \in \{1, \ldots, n\}$ and proposing to flip the corresponding indicator, i.e., $\tilde{d}_i = (1 - d_i)$. Other station indicators remain unchanged, $\tilde{d}_j = d_j$, $j \neq i$. Figure 3 reports the expected utilities of imputed network designs d against iterations. Since expected utilities $U(d)$ are not computed as part of the algorithm, we used separate large scale Monte Carlo simulation to compute $U(d)$ for the plotted designs. Figures 4 and 5 summarize some features of the selected designs.

4.2 An Ozone Monitoring Network

We consider optimal design for a network of ground level ozone monitoring stations in the eastern United States. The stations report daily maxima for eight hours running average ozone concentrations. Let y_{ti} denote the measurement on day t at station i. As probability model we use a process convolution model as described in Higdon (2002). Let $y_t = (y_{t1}, \ldots, y_{tn})$ denote the vector of all observations in period t, and let s_i denote the coordinates of the i-th station. To construct a probability model for y_t the process convolution model defines a set of knots, $\Omega = \{\omega_1, \ldots, \omega_m\}$, and kernels $k(s_i - \omega_j)$ centered at these



(a) $t = 1, \ldots, 250$ (b) $t = 1000, \ldots, 1250$ (c) $t = 2250, \ldots, 2500$

Figure 3: Estimated $U(d^{(t)})$ for simulated designs $d^{(t)}$ over iterations in the indicated range. Note how the level of observed $U(d)$ shifts to higher expected utility and the dispersion tightens. Evaluating $U(d^{(t)})$ is not part of the algorithm and was seperately computed for this plot, using a Monte Carlo approximation (2) with $M = 100$. The numerical standard errors for evaluating $U(d)$ are approximately 0.5.

knots. In our implementation we used a regular grid of $m = 27$ knots over the study region. A linear combination of these kernels is used to construct a spatial mean function. Adding independent normal residuals completes the model:

$$y_t = K\theta_t + \epsilon \text{ with } K_{ij} = k(s_i - \omega_j) \text{ and } \epsilon \sim N(0, \sigma^2 I_n), \tag{8}$$

where I_n denotes the $(n \times n)$ identity matrix. The $(m \times 1)$ vector θ_t parametrizes the mean surface. We assume a conjugate normal prior,

$$\theta_t \sim N(0, \tau^2 I_m). \tag{9}$$

The available data are for $T = 30$ days in summer 1999, collected by the U.S. Environmental Protection Agency (EPA). We extend the model to a spatio-temporal design model by assuming θ_t, $t = 1, \ldots, T$ *a priori* independent, and defining a conditional prior $p(\theta_{T+1} \mid \theta_1, \ldots, \theta_T)$ as $Pr(\theta_{T+1} = \theta_t) = 1/T$, $t = 1, \ldots, T$. In words, we simply assume that future mean ozone levels are generated by resampling one of the latent $\{\theta_1, \ldots, \theta_T\}$. This is a suitable design model since the resampling enforces the imputed future mean surface to be realistic realizations of ozone levels. However, the model would not be suitable as an analysis model. It

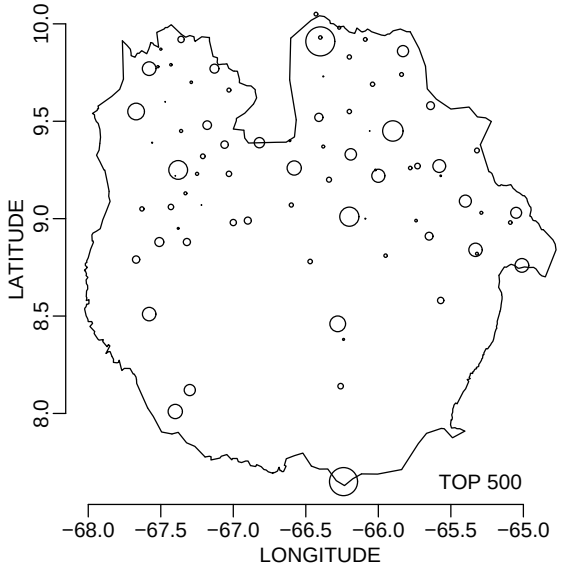


Figure 4: Frequency of occurence of individual stations in the top 500 designs generated during the simulation. Each circle corresponds to one station, centered at a location corresponding to longitude and latitude of the respective station. The sizes of the circles vary from small to large according to the frequency of the stations. The figure has to be interpreted with care since it provides only a marginal summary of the multivariate problem.

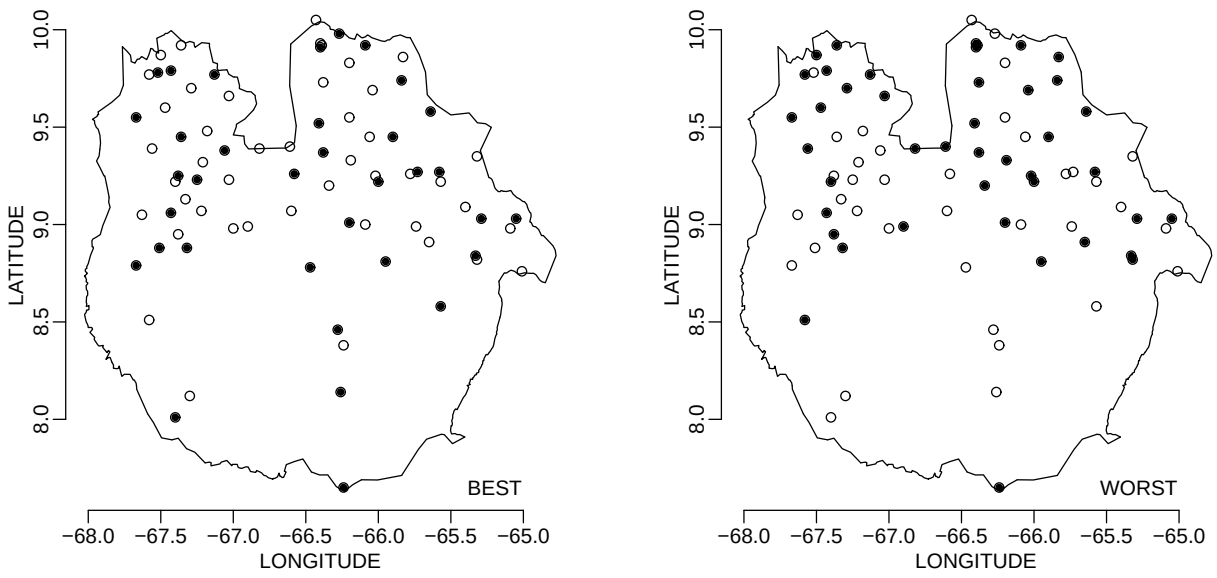


Figure 5: Best (left panel) and worst (right panel) design generated. These are the designs that correspond to the maximum and minimum in the plot of Figure 3. Selected stations are marked with solid dots, dropped stations are marked with empty circles.

fails to specify temporal dependence of θ_t, $t \leq T$. In particular we assume a different model for $p(\theta_{T+1} \mid \theta_1, \ldots, \theta_T)$ than for $p(\theta_{s+1} \mid \theta_1, \ldots, s)$ for any $s < T$. The use of different models for design and analysis is a common procedure (Etzioni and Kadane, 1993). It is perhaps most commonly seen in inference for clinical trials (Spiegelhalter et al., 1994), but equally appropriate in other areas.

The construction of a utility function formalizes several competing aims. First, we want to identify locations with mean ozone level beyond the air quality standard for ozone of 85 ppb (parts per billion). Second, we wish to estimate ozone levels at stations that are dropped from the network. Third, we wish to make inference about the response surface of mean ozone level across the eastern United States. Finally, we want to minimize cost. Let $y^o = (y_1, \ldots, y_t)$ denote the observed data. Let $D = \{i : \ d_i = 1\}$ be the set of chosen stations and $D^c = \{i : \ d_i = 0\}$ the set of stations removed from the network. Let $y_d = (y_{T+1,i},\, i \in D)$ denote the future data at selected stations, and let $y = y_{T+1}$ denote all future data, including the latent responses at stations dropped from the network. Let $\theta = \theta_{T+1}$ denote the parameters for the mean response surface at $T+1$ and let $\widehat{y} = E(y \mid y^o, y_d) = K\, E(\theta \mid y^o, y_d)$ denote the vector of estimated mean future responses. Finally, let

$K(s)$ denote the $(p \times 1)$ vector of kernels $k(s - \omega_j)$ evaluated at a generic location s, let $f(s;\theta) = K(s)\,\theta$ be the mean response surface evaluated at s and $\hat{f}(s) = E(f(s;\theta) \mid y^o, y_d)$. We use

$$u(d,\theta,y) = R \sum_{i \in D^c} \mathbf{1}\{(y_{T+1,i} > y^*) \text{ and } (\widehat{y}_i > y^*)\} + \\ + C_1 / \sum_{i \in D^c} (y_{T+1,i} - \widehat{y}_i)^2 + C_2 / \int_S (\hat{f}(s) - f(s;\theta))^2 \, \mathrm{d}s - \sum_{i \in D} c_i + C_0 \quad (10)$$

The terms correspond to the described competing aims plus a shift C_0 to ensure $u \geq 0$. The coefficients R, C_1, C_2 and c_i define the relative importance of the competing goals and the cost of including station i, respectively. The integral in the third term extends over the entire study region S. In the implementation we approximate the integral by a sum over a 10 by 10 grid.

Based on the EPA data we find an optimal network of ozone measurement stations by maximizing $U(d) = E(u \mid y^o, d)$. We use Algorithm 1., replacing $p(y, \theta \mid d)$ by $p(y, \theta \mid y^o, d)$. We fix the tradeoff coefficients as $R = 1, C_1 = C_2 = 10$ and $c_i = 5$. We simulated 30,000 iterations, discarding the first 10,000 as initial burn-in. Starting with iteration 10,000 we increment J by 1 every 1000 iterations, leading to a maximum power of $J = 20$. Figure 6 shows a summary of the networks d simulated in the course of the simulation. Figure 7a shows the proposed optimal design. For comparison Figure 7b shows the design proposed by the space filling algorithm discussed in Johnson et al. (1990), and as implemented in the software FIELDS (`http://www.cgd.ucar.edu/stats/Software/Fields/index.shtml`).

4.3 Variable Selection

Death status, defined as indicator for high probability of death, is a key input variable in *input-output* quality assessment of hospitals. A traditional approach is to fit a logistic regression to predict death status from sickness variables, using variable selection approaches to identify a subset of all potentially available sickness variables to include in the logistic regression. See, for example, Fouskakis and Draper (2002) for more discussion. This approach does not take into account the fact that measuring sickness variables takes time and costs money. A more cost effective approach is to consider the cheapest subset of sickness variables that have the highest predictive capability. The statistical setting of the decision problem is the following.

Consider data $y^o = \{y_i, s_{i1}, \ldots, s_{ip},\ i = 1, \ldots, n\}$ where y_i is an indicator for patient i dying within 30 days of admission, and $s_{ij}, j = 1, \ldots, p$ are all sickness variables recorded

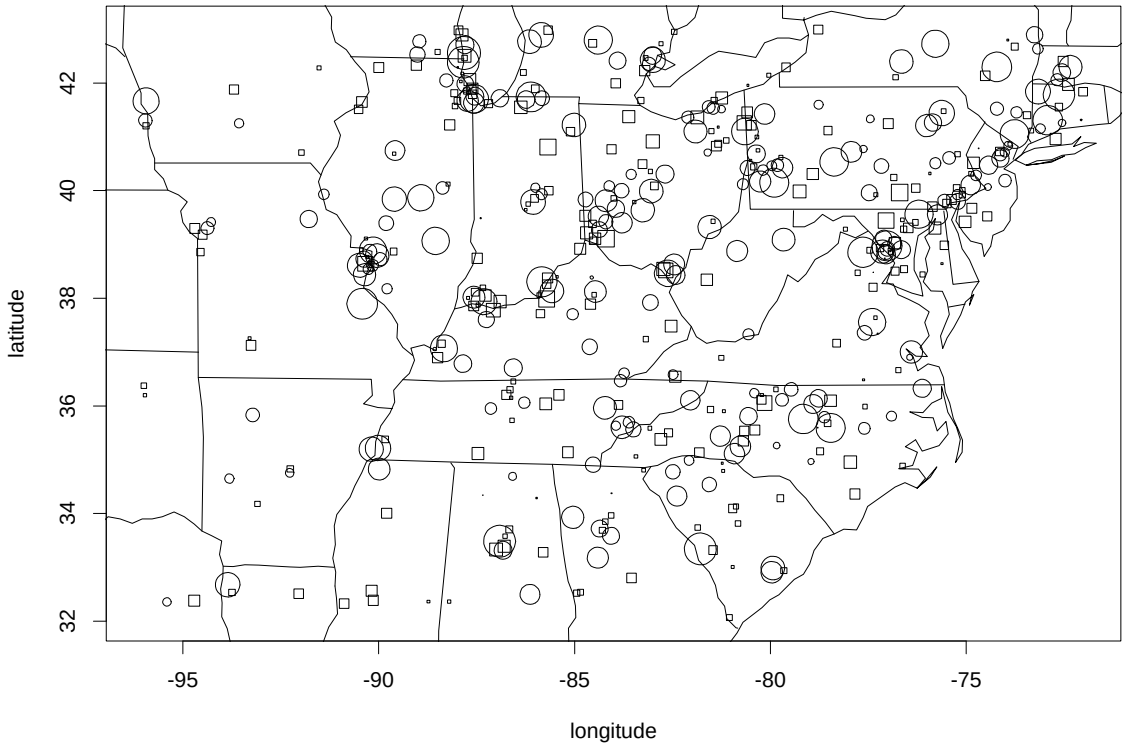
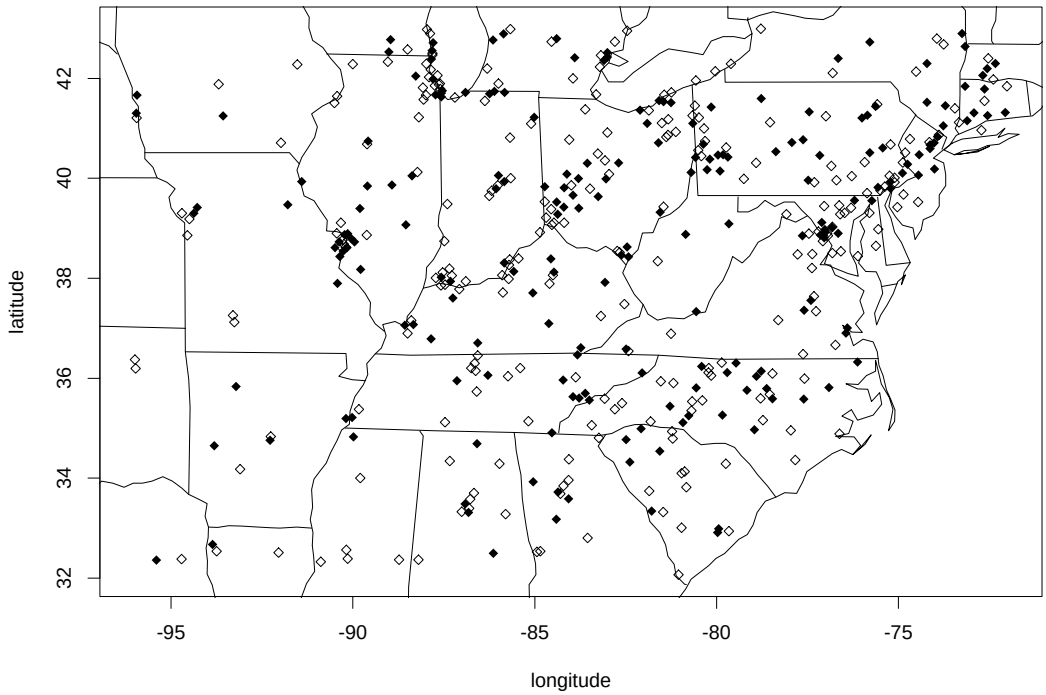
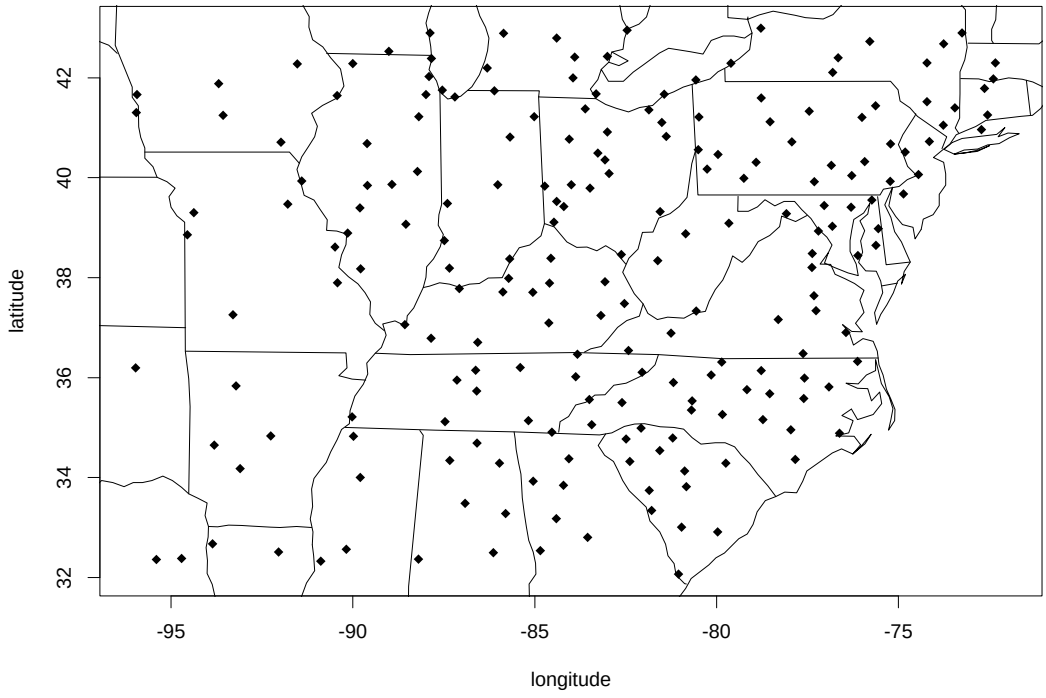


Figure 6: Frequency of occurence of individual stations during the simulation. Selected stations are marked with circles and dropped stations with squares. The area of the symbol is proportional to the frequency of the station.



(a) Optimal design d^*. Selected stations (solid diamond) and dropped stations (empty diamonds).



(b) Space filling design d^{**}, constrained to use the same number of selected stations as d^*.

Figure 7: Optimal network design.

for patient i. We assume a logistic regression

$$y_i \sim Ber(\pi_i), \quad \log\left(\pi_i/(1-\pi_i)\right) = \beta_0 + \beta_1 s_{i1} + \ldots + \beta_p s_{ip}. \tag{11}$$

The utility function has two components. One term is related to the cost of including a sickness variable. A second term measures the predictive capability of the model. Let d_j be an indicator for variable j being in the model. Then, for a given variable selection $d = (d_1, \ldots, d_p)$ we find total sampling cost $\sum_{j=1}^{p} c_j d_j$, where c_j is the cost of including variable j.

To measure the predictive capability of the model consider a threshold p^*. Let $s_i^d = (s_{ij},\ j \in D)$. We say the model predicts death for a future patient $n+1$ if $p(y_{n+1} = 1 \mid y^o, s_{n+1}^d) \geq p^*$. Let $p_{n+1}^d = p(y_{n+1} = 1 \mid s_{n+1}^d, y^o)$. We define predictive utility of the model as $K\,\mathbf{1}\{(y_{n+1} = 0,\ p_{n+1}^d < p^*)$ or $(y_{n+1} = 1,\ p_{n+1}^d \geq p^*)\}$. Here K is a constant that allows to monetize the utility of the precision. The total utility is equal

$$u(d, y_{n+1}) = K\,\mathbf{1}\{(y_{n+1} = 0,\ p_{n+1}^d < p^*) \text{ or } (y_{n+1} = 1, p_{n+1}^d \geq p^*)\} - \sum_{j=1}^{p} c_j d_j + C_0.$$

The shift C_0 ensures non-negativity. An additional complication arises from budget constraints, bounding total cost $\sum c_j\, d_j \leq C$ by a available funds C.

Based on available data y^o we could now proceed to find the best model d^* under utility $u(d, y)$ and the probility model (11) completed with a prior on the logistic regression parameters and a probability model for s_{n+1}. For the latter we recommend using the empirical distribution of the first n patients, i.e., simple resampling of $\{s_1, \ldots, s_n\}$. The nature of the design space as the set of $p-$dimensional indicator vectors makes the variable selection problem notoriously difficult. The inhomogeneous Markov chain simulation as proposed in algorithm 1 provides a useful alternative.

5 DISCUSSION

We have described the use of inhomogeneous MCMC simulation to solve expected utility maximization problems. The main feature of the proposed approach is its generality. The method only requires that under a given decision and assuming hypothesized future data y and parameter values θ the decision maker be able to evaluate utilities $u(d, \theta, y)$. Similarly the probability model can be quite general. As in posterior MCMC simulation, essentially any probability model that allows pointwise evaluation of posterior distributions for given parameter values can be used.

As in any complex MCMC simulation the main limitation of the proposed method is the computation intensive implementation. The approach strictly assumes a decision theoretic setup, and thus inherits any limitations related to this paradigm. In particular, we assume that there is one single utility function, with known tradeoffs between possibly competing goals. Also, the need for pointwise evaluation of utilities limits applicability of the method for problems with traditional inference loss involving often analytically intractable posterior variances.

In summary, the proposed algorithm is likely to be useful for decision problems with complicated design spaces, like in the network design example, and problems with utility functions that involve complicated functions of future data and parameters, like life histories of future patients undergoing screening for chronic disease.

For the sake of being specific the discussion was restricted to design problems, focusing on the motivating problem of network design. We defined the generic design problem as expected utility maximization, marginalizing with respect to parameters and future data to be collected in a planned experiment. One of the reasons why we chose this setting was that it often leads to the kind of difficult expected utility maximizations where the proposed algorithms are useful. However, nothing in the proposed inhomogeneous Markov chain Monte Carlo simulation limits the use to such settings. In particular, one inference question that leads to a similar problem structure is maximum *a posteriori* variable selection in a regression model. Let $d = (d_1, \ldots, d_p)$ denote a vector of binary indicators with $d_i = 1$ corresponding to the inclusion of the i-th potential covariate and let y denote the observed data. Finding the maximum *a posteriori* model becomes the problem of maximizing marginal posterior probability $\max_d p(d \mid y)$. The general setup is described in George and McCulloch (1993). A typical application appears, for example, in Sha et al. (2003) who describe a multivariate probit model for gene expression in large-scale microarray experiments.

References

Berger, J. O. and Barbieri, M. M. (2003), "Optimal predictive model selection," *Annals of Statistics*, 31, to appear.

Berry, D. (1993), "A case for Bayesianism in clincial trials (with discussion)," *Statistics in Medicine*, 12, 1377–1404.

Berry, D. A. and Stangl, D. K., eds. (1996), *Bayesian Biostatistics*, volume 151 of *Statistics: Textbooks and Monographs*, New York: Marcel Dekker.

Bielza, C., Müller, P., and Insua, D. R. (1999), "Monte Carlo Methods for Decision Analysis with Applications to Influence Diagrams," *Management Science*, 45, 995–1007.

Bras, R. L. and Rodríguez-Iturbe, I. (1985), *Random Functions and Hydrology*, Reading: Addison-Wesley.

Brockwell, A. E. and Kadane, J. B. (2001), "Sequential analysis by gridding sufficient statistics," Technical report, Carnegie Mellon University.

Brown, P. J., Fearn, T., and Vannucci, M. (1999), "The choice of variables in multivariate regression: A non-conjugate Bayesian decision theory approach," *Biometrika*, 86, 635–648.

Carlin, B., Kadane, J., and Gelfand, A. (1998), "Approaches for optimal sequential decision analysis in clinical trials," *Biometrics*, 54, 964–975.

Caselton, W. F., Kan, L., and Zidek, J. (1992), "Quality data networks that minimize entropy," in *Statistcs in the Environment and Earth Sciences*, eds. P. Guttorp and A. Walden, London: Griffin.

Chaloner, K. and Verdinelli, I. (1995), "Bayesian experimental design: a review," *Statistical Science*, 10, 273–304.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2001), *Model Selection*, volume 38 of *IMS Lecture Notes*, chapter The Practical Implementation of Bayesian Model Selection, IMS.

Clayton, M. K., et al. (1999), "Bayesian sequential design of a network of sensors," Technical report, University of Wisconsin.

Clyde, M. and George, E. I. (2003), "Model Uncertainty," Technical report, Duke University ISDS, USA.

DeGroot, M. (1970), *Optimal Statistical Decisions*, McGraw Hill.

Doucet, A., Godsill, S., and Robert, C. (2002), "Marginal Maximum A posteriori Estimation using Markov Chain Monte Carlo," *Statistics and Computing*, to appear.

Etzioni, R. and Kadane, J. (1993), "Optimal experimental design for another's analysis," *Journal of the American Statistical Association*, 88, 1401–1411.

Fei, L. and Berliner, L. M. (1991), "Asymptotic properties of stochastic probing for global optimization: The finite case," Technical report, Ohio State University.

Fouskakis, D. and Draper, D. (2002), "Stochastic optimization: a review," *International Statistical Review*, 70, 315–350.

Geman, S. and Geman, A. (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. Patt. Ana. Mac. Int.*, 6, 721–741.

George, E. I. and McCulloch, R. E. (1993), "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

Guttorp, P., Le, N., and Zidek, J. (1994), "Using entropy in the redesign of an environmental monitor network," in *Multivariate Environmental Statistics*, eds. G. Patil and C. Rao, 175–202, Elsevier Science Publisher.

Handcock, M. and Stein, M. (1993), "A Bayesian analysis of kriging," *Technometrics*, 35, 403–410.

Higdon, D. M. (2002), "Space and Space-Time Modeling Using Process Convolutions," in *Quantitative Methods for Current Environmental Issues*, ed. C. A. et al., 37–54, Springer-Verlag.

Isaacson, D. L. and Madsen, R. W. (1976), *Markov Chains: Theory and Applications*, Wiley.

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990), "Minimax and maximin distance designs," *Journal of Statistical Planning and Inference*, 26, 131–148.

Laud, P. W., Berliner, L. M., and Goel, P. K. (1992), "A stochastic probing algorithm for global optimization," *Journal of Global Optimization*, 2, 209–224.

Müller, A. (1999), "Simulation based optimal design," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, P. Dawid, and A. F. M. Smith, 459–474, Oxford University Press.

Müller, P. and Parmigiani, G. (1996), "Optimal design via curve fitting of Monte Carlo experiments," *Journal of the American Statistical Association*, 90, 1322–1330.

Nychka, D. and Saltzman, N. (1998), "Design of air-quality monitoring networks," in *Case Studies in Environmental Statistics*, eds. D. Nychka, W. Piegorsch, and L. Cox, New York.

Raiffa, H. and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Boston: Harvard University Press.

Robert, C. P., Doucet, A., and Godsill, S. J. (1999), "Marginal MAP Estimation using Markov Chain Monte Carlo," Technical report, CREST, INSEE.

Sansó, B. and Guenni, L. (1999), "Venezuelan rainfall data analysed using a Bayesian space-time model," *Applied Statistics*, 48, 345–362.

Sansó, B. and Müller, P. (1999), "Redesigning a network of rainfall stations," in *Case Studies in Bayesian Statistics, vol. IV*, eds. C. Gatsonis, R. Kass, B. Carlin, A. Carriquiry, A. Gelman, I. Verdinelli, and M. West, 497–510, Springer Verlag.

Sha, N., et al. (2003), "Gene selection in arthritis classification with large-scale microarray expression profiles," *Comparative and Functional Genomics*, 4, 171–181.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994), "Bayesian Approaches to Randomized Trials," *Journal of the Royal Statistical Society, Series A, General*, 157, 357–387.

Sun, D., Tsutakawa, R. K., and Lu, W.-S. (1996), "Bayesian Design of Experiment for Quantal Responses: What Is Promised Versus What Is Delivered," *Journal of Statistical Planning and Inference*, 52, 289–306.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *Annals of Statistics*, 22, 1701–1728.

van Laarhoven, P. and Aarts, E. (1987), *Simulated Annealing: Theory and Applications*, D. Reidel.

APPENDIX: PROOF OF THEOREM 1

The proof for Theorem 1 follows standard arguments for convergence of inhomogeneous Markov chains. See, for example, Isaacson and Madsen (1976) for related results. In the proof we will frequently need to refer to the equilibrium distribution h_J under $J = J_n$. To avoid double subscripts we write π_n for $\pi_n(d, v) = h_{J_n}(d, v)$.

The outline of the proof is as follows. We first establish that the chain is weakly ergodic, i.e., the chain forgets the starting condition in the limit. We show weak ergodicity by considering the series $s_n = \sum_{k=0}^{n}[1 - \tau_1(P(k))]$, where $\tau_1(P)$ is the coefficient of ergodicity for the transition matrix $P(n) = [P_{ij}(n)]$. Divergence of s_n implies weak ergodicity. To assess strong ergodicity and convergence to the desired stationary distribution π^* we verify that $\sum_{n=0}^{\infty} \parallel \pi_n - \pi_{n+1} \parallel < \infty$ for the sequence of stationary distributions π_n. This conditition together with weak ergodicity implies strong ergodicity.

We start with some observations about the time-homogenuous chains obtained by fixing $J = J_n$. The transition probabilities $P_{ij}(n)$ for the homogenuous chain at $J = J_n$ define a Metropolis-Hastings chain with the target distribution π_n and proposals defined by steps 1 and 2 in Algorithm 1. Let $i = (d_i, v_i)$ and $j = (d_j, v_j)$. Then

$$P_{ij}(n) = \begin{cases} g(d_j|d_i)\, p_J(v_j|d_j)\, \alpha_J(v_i, v_j) & \text{if } j \neq i \\ 1 - \sum_{(d,v)\neq j} g(d\ |d_i)\, p_J(v\ |d_i)\, \alpha_J(v_i, v\ ) & \text{if } j = i, \end{cases}$$

with $\alpha_J(v, w) = \min\{1, \exp(Jw - Jv)\}$.

The sequence of invariant distributions π_n converges to the uniform distribution π^* over the optimal set S^*. Marginally, for $\pi_n(d)$, this follows from $\pi_n(d) \propto U^{J_n}(d)$. Convergence of the conditional distributions $\pi_n(v|d)$ follows from the fact that v is the sample mean of J_n i.i.d. samples from $h(v|d)$. The law of large numbers implies convergence of $\pi_n(v|d)$ to a point mass at $\mu(d)$.

We establish a bound on $P_{ij}(n)$, which will later allow us to make an argument about the sequence of ergodic coefficients.

Lemma 2 *There are constants* $0 < a < \infty$ *and* $0 \leq b < \infty$ *such that*

$$P_{ij}(n) \geq a \exp\{-b/T_n\}$$

Proof. Recall the definition $v = 1/J \sum \log u(d, y_j, \theta_j)$. By assumption $u(\cdot)$ is bounded, $u_0 \leq u(d, \theta, y) \leq u_1$, and thus

$$\alpha_J(i, j) = \min\{1, \exp[J(\tilde{v} - v))]\} \geq (u_0/u_1)^J.$$

Also, we assumed $p(v|d) \geq \epsilon > 0$ for all (d, v), and thus $p_J(v|d) \geq \epsilon^J$. Additionally we asume $g(\tilde{d}|d) \geq a > 0$. Set $b = \log[u_1/(u_0\,\epsilon)]$ and let $T_n = 1/J_n$. Then

$$P_{ij}(n) \geq a\ (\epsilon\, u_0/u_1)^{J_n} = a \exp(-b/T_n).$$

The assumption $g(\tilde{d}|d) > 0$ is not critical. For the proof of Lemma 3 it suffices if there is an m such that $g_m(\tilde{d}|d) > 0$, $\forall (d, \tilde{d})$ for the $m-$step transition probability g_m. ■

We are now ready to show weak ergodicity

Lemma 3 *Under the cooling schedule*

$$T_n \geq \gamma/\log(n + c), \quad n = 0, 1, 2, \ldots,$$

$\gamma \geq b$ and $c > 0$, the Markov chain (d_n, v_n) is weakly ergodic.

Proof. Recall that $P(m, n)$ denotes the transition probabilities between steps m and n and $P(n) = P(n, n + 1)$. Let $\tau_1(P)$ denote the coefficient of ergodicity for the transition probabilties P, defined as

$$\tau_1(P) = 1 - \min_{i,j} \sum_{l=1}^{m} \min(P_{il}, P_{jl}).$$

An inhomogeneous Markov chain is weakly ergodic if and only if there is a strictly increasing sequence of positive numbers $\{n_l\}$, $l = 0, 1, 2, \ldots$ such that

$$\sum_{l=0}^{\infty} \{1 - \tau_1[P(n_l, n_{l+1})]\} = \infty. \tag{12}$$

Using the bound on $P_{ij}(n)$ we find $\tau_1(P(n)) \leq 1 - a \exp(b/T_n)$. Substituting $T_n = \gamma/\log(n+c)$ in (12) gives

$$\sum_{n=0}^{+\infty} [1 - \tau_1(P(n))] \geq a \sum_{n=0}^{+\infty} \exp(-b/T_n) \geq a \sum_{n=0}^{+\infty} 1/(n + c)^{b/\gamma}.$$

For $\gamma \geq b$ it follows that $a \sum_{n=n_0}^{+\infty} 1/(n + c)^{b/\gamma} = \infty$, implying weak ergodicity. ■

Theorem 4 *(Isaacson and Madsen, 1976). If there exists a sequence of probability vectors π_n such that $\pi_n = \pi_n P(n)$,*

$$\sum_{n=0}^{\infty} \parallel \pi_n - \pi_{n+1} \parallel < \infty, \tag{13}$$

and the time inhomogeneous Markov chain is weakly ergodic, then it is also strongly ergodic. If $\pi^\star = \lim_{n\to\infty} \pi_n$, then for all $m \geq 1$, $i, j \in S$,

$$\lim_{n\to\infty} P_{ij}(m, n) = \pi^\star(j)$$

We use Theorem 4 to prove strong ergodicity of (d_n, v_n). We have already established $\pi_n = \pi_n P(n)$ and $\lim \pi_n = \pi^*$. Only (13) is left to show. Without loss of generality assume the optimal set is a single point $S^* = \{(d^*, v^*)\}$. Let $S^o = \{(d, v),\ d \neq d^*\}$. Without loss of generality we assume

$$U(d^*) \geq 1 \text{ and } \sum_{d \in S^o} U(d) = \epsilon < 1$$

(if not, multiply $u(\cdot)$ and replace $U(d)$ by $U(d)^{J_0}$ to achieve the condition without changing the optimal solution). It follows that

$$h_J(d) = U^J(d)/[U^J(d^*) + \sum_{S^o} U^J(d)] \leq U^J(d)$$

and $\sum_{S^o} h_J(d) \leq |S^o|\, \delta^J$, where $|S^o|$ indicates the cardinality of S^o.

Finally, assume a finite fourth moment $E_h(v^4|d^*) = \int v^4\ dh_1(v|d^*) < \infty$ and let $v^* = \mu(d^*)$. We will use the following identity for the fourth moment of a sample mean $\bar{X}_n$ in a random sample $X_i \sim p(X)$, $i = 1, \ldots, n$

$$E(\bar{X}_n - EX)^4 = (1/n^3)E\{(X - EX)^4\} + [3n(n-1)/n^4]\left[E\{(X - EX)^2\}\right]^2.$$

Using Tschebychev's inequality and substituting for the fourth moment we get

$$\Pr\{|v - v^*| \geq \delta \mid d^*\} \leq \frac{1}{\delta^4} E\left\{(v - v^*)^4 \mid d^*\right\} \leq \frac{1}{\delta^4}(c_1/J^3 + c_2/J^2)$$

for some positive constants c_1 and c_2. Let $\delta = \min_v |v - v^*|$. Then

$$\begin{aligned}
\sum_n \parallel \pi_n - \pi_{n+1} \parallel &= \sum_J \parallel h_J - h_{J+1} \parallel \\
&= \sum_J \sum_{d \neq d^*} \sum_v |h_J(d, v) - h_{J+1}(d, v)| + \sum_J \sum_v |h_J(d^*, v) - h_{J+1}(d^*, v)| \\
&\leq \sum_J 2c\epsilon^J + 2 \sum_J 2/\delta^4 \left(c_1/J^3 + c_2/J^2\right) < \infty.
\end{aligned}$$

The equality in the first line holds since $\parallel \pi_n - \pi_{n+1} \parallel = 0$ for $J_n = J_{n+1}$. In the last line, the extra factor 2 in front of the second sum appears because we use $2/\delta^4\,(c_1/J^3 + c_2/J^2)$ as a bound for $\sum_v |h_J(d^*, v) - h_{J+1}(d^*, v)|$, as well as for $|h_J(d^*, v^*) - h_{J+1}(d^*, v^*)|$.