Learning the Sparsity Parameter in a Generalized Fast Subset Sums Framework for Bayesian Event Detection

Kan Shao
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

We present the Generalized Fast Subset Sums (GFSS) method, an extension of the recently proposed Multivariate Bayesian Scan Statistic (MBSS) and Fast Subset Sums (FSS) approaches for detecting irregularly shaped spatial clusters efficiently and effectively. The MBSS framework (Neill and Cooper, 2010) can integrate multiple data streams for detection of emerging events, but its detection power is primarily limited by computational considerations, which limit it to searching over circular spatial regions. The FSS method (Neill, 2011) enables more accurate and timely detection by defining a hierarchical prior over all subsets of the N locations. The GFSS method generalizes the FSS framework by introducing a sparsity parameter p to describe how likely each location in the neighborhood (defined by the center location and the neighborhood size) is to be affected. The sparsity parameter allows us to consider all possible subsets of locations (including irregularly-shaped spatial regions) but also puts higher weight on more compact regions. In this study, we first learn the distribution of the sparsity parameter p from a fully labeled dataset, given the spatial extent of each labeled event. The detection power of GFSS, using the learned distribution of p, was compared to GFSS with uniformly distributed p, as well as the original FSS and MBSS approaches. Our evaluation results (on synthetic disease outbreaks injected into real-world hospital Emergency Department data) show that the distribution of p can be learned reasonably well based on 100 simulated outbreaks, and that the GFSS method with learned sparsity parameter has higher detection power and spatial accuracy than MBSS and FSS, particularly when the outbreak region is irregular or elongated. We also demonstrate that the learned models can be used for event characterization, accurately distinguishing between two otherwise identical outbreak types based solely on the sparsity of the affected spatial region.

1. INTRODUCTION

The multivariate Bayesian scan statistic (MBSS), proposed by Neill and Cooper [1], is a general framework that detects and characterizes emerging events (such as an outbreak of disease) using multivariate spatial time series data. MBSS has multiple advantages over previously proposed, frequentist spatial scan methods: it is computationally efficient, can accurately differentiate between multiple event types, and its results (the posterior probability distributions of each event type over time and space) can be easily visualized and used for decision-making [1,2]. However, as is the case for many spatial scan methods, MBSS is primarily limited by computational considerations, which only allow circular spatial regions to be searched.

The Fast Subset Sums (FSS) method [2] is an extension of the previous MBSS framework. By introducing a hierarchical prior distribution over regions that assigns non-zero prior probability to each of the 2^N subsets of locations, FSS can compute the total posterior probability of an event and its spatial distribution by computing the *sum* of the exponentially many region posterior probabilities efficiently and effectively.

In present study, we propose a generalized version of the FSS framework which improves the detection power of the original FSS method, especially for irregularly shaped outbreaks. A new parameter p, representing the *sparsity* of the outbreak region, is introduced into the framework. This parameter can be viewed as the expected proportion of locations affected in the local neighborhood consisting of a center location and its k – 1 nearest neighbors. We show that two specific values of p, p = 1 and p = 0.5, reduce to the previously proposed MBSS and FSS methods respectively, but detection performance can often be improved by considering a range of possible p values from 0 to 1. We show that the distribution of the p parameter can be accurately learned from labeled training data, and that the resulting learned distribution of p can be incorporated into the GFSS detection framework, resulting in substantially improved detection power and spatial accuracy.

2. METHODOLOGY

In this section, we first briefly review the recently proposed Multivariate Bayesian Scan Statistics (MBSS) and Fast Subset Sums (FSS) methods, present the new "Generalized Fast Subset Sums" (GFSS) framework, and finally consider how the distribution of the sparsity parameter can be learned from labeled training data within this framework.

2.1 Multivariate Bayesian Scan Statistics

The MBSS methodology aims at detecting emerging events (such as disease outbreaks), identifying the type of event and pinpointing the affected locations. MBSS compares a set of alternative hypotheses $H_1(S, E)$ with the null hypothesis H_0, where each hypothesis $H_1(S, E)$ represents the occurrence of some event type E in some subset of locations S, and the null hypothesis H_0 assumes that no events have occurred. All of

these hypotheses are mutually exclusive. Therefore, according to Bayes' Theorem, the posterior probability of each hypothesis (given the observed dataset D) can be expressed as:

$$\Pr(H_1(S,E)|D) = \frac{\Pr\left(D\middle|H_1(S,E)\right)\Pr\left(H_1(S,E)\right)}{\Pr(D)}$$

$$\Pr(H_0|D) = \frac{\Pr(D|H_0)\Pr(H_0)}{\Pr(D)}$$

where the total probability of the data can be written as: $\Pr(D) = \Pr(D|H_0)\Pr(H_0) + \sum_{S,E}\Pr(D|H_1(S,E))\Pr(H_1(S,E))$. The standard MBSS method introduced in [1] assumes that the prior $\Pr(H_1(S,E))$ is uniformly distributed over all event types and all possible circular spatial regions S. Only circular regions are considered because this simplification reduces the computation time from exponential to quadratic in the number of locations N; however, this assumption reduces the power of MBSS to detect non-circular clusters, especially if the affected region is highly elongated or irregular.

The dataset D in the MBSS framework consists of multiple data streams D_m, m=1…M, and each stream contains spatial time series data collected from a set of locations s_i, for i=1…N. For each location s_i and data stream D_m, we have a time series of observed counts $c_{i,m}^t$ and the corresponding expected counts (or baselines) $b_{i,m}^t$, where the baselines are estimated from time series analysis of the historical data for the given location and data stream. The subscript t=0 represents the current time step, and t=1…T represent from 1 to T time steps ago respectively. For instance, a given count $c_{i,m}^t$ may represent for a given zip code on a given day the total number of Emergency Department (ED) visits for fever symptoms, and the corresponding baseline $b_{i,m}^t$ would represent the expected number of fever ED visits for that zip code on that day [1, 2].

Besides the prior $\Pr(H_1(S,E))$, another important quantity to consider is the likelihood function $\Pr(D|H)$, as shown in Figure 1. We assume that the observed count $c_{i,m}^t$ is modeled using the Poisson distribution: $c_{i,m}^t \sim \text{Poisson}(q_{i,m}^t b_{i,m}^t)$, where $q_{i,m}^t$ is the relative risk. Further, the relative risk is modeled as $q_{i,m}^t \sim \text{Gamma}(\alpha_m, \beta_m)$ under the null hypothesis, and as $q_{i,m}^t \sim \text{Gamma}(x_{i,m}^t \alpha_m, \beta_m)$ under the alternative hypothesis, where α_m and β_m are parameter priors calculated from historical data, and $x_{i,m}^t$ is the impact of the outbreak for the given data stream D_m, location s_i, and time step t. The distribution of $x_{i,m}^t$ is conditioned on the outbreak region S, the event type E, the temporal window W and the severity parameter θ, which is assumed to be drawn from a discrete uniform distribution Θ. The temporal window W is drawn uniformly at random between 1 and W_{max}, the maximum temporal window size. In present study, W_{max} is set to 3 for better detecting more quickly emerging outbreaks, as suggested by Neill [3].

The total likelihood of the data given the alternative hypothesis $H_1(S,E)$ can be expressed as: $\Pr\left(D\middle|H_1(S,E)\right) = \frac{1}{W_{max}|\Theta|}\sum_{\theta\in\Theta}\sum_{W\in 1\ldots W_{max}}\Pr(D|H_1(S,E),\theta,W)$. The

likelihood ratio for each location in a given region S can be expressed as $LR_i = \prod_{m=1\ldots M} \prod_{t=0\ldots W-1} \frac{\Pr(c_{i,m}^t | b_{i,m}^t, x_m \alpha_m, \beta_m)}{\Pr(c_{i,m}^t | b_{i,m}^t, \alpha_m, \beta_m)}$ as described in [1, 2]. The likelihood ratio for each spatial region S, conditioned on W, θ, and E, is obtained by multiplying the likelihood ratios of all locations in that region. We then marginalize over these parameters to obtain the total likelihood of region S, and combine the prior with the likelihood (as given above) to obtain the posterior probability of each event type E in each region S.

2.2 Fast Subset Sums

As discussed in the previous section, the MBSS method is primarily restricted by its exhaustive computation over spatial regions S, which requires us to limit our search space to a small fraction of all the $O(2^N)$ possible subsets of locations. More precisely, all non-circular regions are assumed to have zero prior probability, thus reducing the method's computation time but also its detection power for irregular clusters. However, two important insights allow us to circumvent this limitation: the total posterior probability of an outbreak is the sum of the region probabilities $\Pr(H_1(S,E)|D)$, and the posterior probability that each spatial location s_i has been affected is a sum over all spatial regions which contain s_i. The FSS framework defines a non-uniform, hierarchical prior distribution $\Pr(H_1(S,E)|E)$ such that all 2^N subsets have non-zero prior probability, but more compact regions have a larger prior. The spatial region S affected by the outbreak is assumed to be drawn from a generative distribution conditioned on two latent variables: the "center" location s_c and the "neighborhood size" k. Here we assume that the center s_c is drawn uniformly at random from the set of N locations, and the neighborhood size k is also drawn uniformly at random between 1 and some constant maximum neighborhood size k_{max}. The center and neighborhood size together define the local neighborhood S_{ck}, consisting of the center location s_c and its k – 1 nearest neighbors. Finally, we assume that the affected subset of locations S is drawn uniformly at random from the neighborhood S_{ck}, i.e. all 2^k subsets of S_{ck} are equally likely. This hierarchical prior, and particularly the assumption that S is drawn uniformly at random given the center and neighborhood size, enables us to calculate the posterior probabilities much more efficiently. Conditioning on the outbreak type E, severity θ, temporal window W, center location s_c, and neighborhood size k, we can compute the average likelihood ratio over all 2^k subsets of S_{ck}: $\sum_{S \subseteq S_{ck}} \Pr(S|D) \propto \frac{1}{2^k} \prod_{s_i \in S_{ck}} (1 + LR_i) = \prod_{s_i \in S_{ck}} \left(\frac{1+LR_i}{2}\right)$, as shown by Neill [2]. We must then marginalize over the five variables E, θ, W, s_c, and k to compute the total posterior probability of an outbreak.

2.3 Generalized Fast Subset Sums

The fundamental improvement of the newly proposed method in the present study is that a sparsity parameter is introduced into the FSS framework. For a given local neighborhood S_{ck}, we do not simply choose the subset of affected locations S uniformly

at random, but instead we independently choose whether to include or exclude each location s_i in S_{ck}. Each location is included with probability *p* or excluded with probability 1 – *p*, where *p* is a constant (0 < p ≤ 1) which we term the sparsity parameter. The sparsity parameter *p* can also be viewed as the expected proportion of locations affected within a given (circular) local neighborhood. Hence, the previous MBSS and FSS approaches are two special cases in the GFSS framework: when *p* = 1, all of the locations in a circular region S_{ck} are affected, and thus GFSS becomes equivalent to the MBSS framework. When p = 0.5, each location is equally likely to be affected or unaffected: thus each of the 2^k subsets of S_{ck} is equally probable, and GFSS becomes equivalent to FSS. Higher values of the sparsity parameter p result in improved detection of more compact clusters, while lower values of p enhance detection of more elongated or irregular clusters. We now explain the GFSS detection framework, and then present algorithm for learning p from labeled training data.

The detection method of GFSS is developed based on the original FSS method, and still enables us to efficiently compute the average likelihood ratio for the 2^k subsets in a given local neighborhood S_{ck}. The only difference between GFSS and FSS is that the sparsity parameter is now applied when we calculate the likelihood ratio for each location in the region S_{ck}. Thus, the sum of the 2^k products of the locations' likelihood ratios can again be expressed as a product of k sums: $\sum_{S \subseteq S_{ck}} \Pr(S|D) = \sum_{S \subseteq S_{ck}} \prod_{s_i \in S} LR_i = \prod_{s_i \in S_{ck}} \big((1-p) + p * LR_i\big)$. Furthermore, the total posterior probability of an outbreak is now marginalized over the distribution of the sparsity parameter p as well as the event type E, temporal window W, severity θ, center location s_c, and neighborhood size k.

Based on fully labeled data, we can also *learn* the distribution of the sparsity parameter p to improve the detection ability of the GFSS method. The assumption of fully labeled data means that we are given the affected subset of locations S for each labeled training example; however, we are not given the values of the three latent variables (center s_c, neighborhood size k, and sparsity p). Let $\vec{S}$ = $S_1 \ldots S_J$ represent a set of J labeled training examples. We can then apply Bayes' Theorem to learn the posterior distribution of p:

$$\Pr\big(p|\vec{S}\big) \propto \Pr(p)\Pr\big(\vec{S}|p\big) \propto \Pr(p)\prod_{j=1}^{J}\sum_{s_c}\sum_{k}\Pr(S_j|p,s_c,k)\Pr(s_c)\Pr(k)$$

where $\vec{S}$ is the fully labeled training dataset, p is the sparsity parameter, and *J* is the total number of outbreak regions in the training dataset $\vec{S}$. The likelihood of outbreak region S_j given the center s_c, the neighborhood size k, and the sparsity parameter p can be further expressed as $\Pr\big(S_j|p,s_c,k\big) = p^{|S_j|}(1-p)^{|S_{ck}|-|S_j|}$ if all of the affected locations in S_j are fully contained in the local neighborhood S_{ck}, and $\Pr\big(S_j|p,s_c,k\big) = 0$ otherwise, where $|S_j|$ is the total number of locations in outbreak S_j and $|S_{ck}|$ = k is the total number of locations in region S_{ck}. We consider a discrete uniform prior distribution of p with ten distinct values (p = 0.1, 0.2, …, 1) and obtain the posterior probability for each value of p.

2.4 Related work

The present study proposes the Generalized Fast Subset Sums (GFSS) method to extend the current FSS [2] and MBSS [1] methods for multivariate Bayesian event detection. The Bayesian spatial scan framework is a variant of the traditional frequentist, hypothesis test-based spatial scan methods [4, 5]. Two recently proposed frequentist spatial scan methods, the Kulldorff's parametric scan [6] and the nonparametric scan [7], also allow integration of multiple data streams for detection; however, unlike MBSS and FSS, these methods cannot differentiate between multiple event types. The recently proposed "linear-time subset scanning" (LTSS) method [8], enables an efficient search over the 2^N subsets of locations while only evaluating O(N) subsets. However, the LTSS method simply finds the most anomalous (highest scoring) subset, and cannot be used to compute the total posterior probability of an outbreak or its posterior distribution in space and time, which require summing over all subsets of locations. Previously, incorporating learning approaches to improve detection power by using non-uniform priors on each search region were explored in [1, 9]. However, these methods are still constrained by the computational issues presented in the MBSS framework, preventing them from being used to define priors over all subsets of the data rather than just circular regions.

3. EVALUATION

In this section, we evaluate the learning performance, as well as compare the detection power and spatial accuracy of the GFSS method with the MBSS and FSS approaches. The dataset and simulation procedure are introduced first.

3.1 Description of training and testing data

The training and testing data are generated based the same dataset used in (Neill, 2011). The original dataset contains de-identified Emergency Department visit records collected from ten hospitals in Allegheny County from January 1, 2004 to December 31, 2005. The records are further classified into various data streams according to the different chief complaint types, such as "cough" and "nausea" which are the two streams used in the present study. For each data stream, we have the count of ED visits of that type (cough or nausea) on each day for each Allegheny County zip code.

We injected simulated outbreaks into the two original data streams to generate simulated training and testing datasets. The simulated outbreaks were generated using the same hierarchical generative model as assumed in the GFSS framework; evaluation for the case where the model is misspecified, and the simulated outbreaks are not generated according to this framework, will be addressed in future work. We considered six different outbreak types: outbreaks generated using five different values of the sparsity parameter p (p = 0.2, 0.4, …, 1.0), and a sixth outbreak type which consisted of an equal mixture of p = 0.2 and p = 0.8. For each combination of the value

of sparsity parameter p and data stream, 100 outbreaks were injected to form a set of training data and another 100 outbreaks for forming the testing data. This gives us a total of six datasets for training and another six corresponding datasets for testing, with each pair of datasets assuming a different value or mixture of the sparsity parameter p.

3.2 Learning performance

In present study, we set the number of components of p equal to 10, which means that a multinomial distribution with ten components was used to describe the probability of each of the ten possible values of p (from 0.1 to 1.0). For each of the 12 training datasets (six for each data stream), the posterior distribution of the sparsity parameter was learned and shown in Figures 2 and 3. For all 12 training datasets, we observe that the learned distribution of p peaks either at the true value of p or at a value that is 0.1 less than the true value. This is occurring probably because of the sparsity of the data: occasionally, a zip code does not get any cases injected, and therefore it is not counted as part of the "correct" inject region. So even though we attempt to inject into proportion p of the zip codes, we are actually injecting into a slightly lower proportion of zip codes, and this discrepancy is correctly captured by the learned distribution of p.

3.3 Detection power and spatial accuracy

In this section, we compare the detection power and spatial accuracy of four different methods: (1) original MBSS; (2) original FSS; (3) GFSS with uniform distribution of sparsity parameter p (each value of p from p = 0.1 to p = 1.0 has an equal probability of 0.1); and (4) GFSS with a learned distribution of p.

The comparison of detection times for each of the two data streams is shown in Figures 4 and 5 respectively. The average detection time of each method, assuming a fixed false positive rate of one false positive per month, is displayed on the graphs. When the value of p is small, corresponding to an elongated or irregular outbreak region, GFSS with learned p is able to detect the outbreaks substantially earlier than the other methods. The FSS method (equivalent to putting of all of the probability mass at p = 0.5) performs well for values of p near 0.5, and the MBSS method (equivalent to putting all of the probability mass at p = 1.0) performs well for values of p near 1.0, as expected, but both methods lose detection power when the assumed value of p is incorrect. Next we evaluated the spatial accuracy of each method by computing the average overlap coefficient between the true and detected clusters. The spatial accuracy results shown in Figures 6 and 7 also demonstrate that the GFSS method with learned distribution of p has higher spatial accuracy when the outbreak region is elongated or irregular, while achieving similar performance to the other methods for more compact outbreak regions.

3.4 Detection ability for mixture outbreak type

In this section, we examine the detection time and spatial accuracy of these different methods for the mixed outbreak type (half of outbreaks generated with p = 0.2 and half

of outbreaks generated with p = 0.8). We first consider a single distribution of p learned from the mixed outbreak type, as compared to MBSS, FSS, and GFSS with a uniform distribution of p. The last graphs of Figure 2 and 3 demonstrate that the single model can accurately capture the bimodal distribution of p.

The results of detection time and spatial accuracy for the mixed outbreaks by using four different methods are listed in Table 1. The GFSS with learned p slightly outperforms FSS and GFSS with uniform p for detecting mixture outbreaks, with all three methods outperforming MBSS by a large margin.

Table 1. Comparison of detection time and spatial accuracy for the mixed outbreak type

		MBSS	FSS	GFSS with uniform p	GFSS with learned p
Cough Data	Days to detect (1 fp/month)	6.05	**5.37**	5.52	5.51
	Spatial overlap coefficient	0.495	0.589	0.611	**0.617**
Nausea Data	Days to detect (1 fp/month)	4.11	3.91	**3.88**	**3.88**
	Spatial overlap coefficient	0.516	0.671	0.691	**0.695**

Next we assumed that the two values of p in the mixed outbreak type corresponded to two different outbreaks, and evaluated the ability of the GFSS framework to distinguish between these two outbreak types. Figure 8 is the result of learning the mixed type of outbreaks by using two GFSS models. We note that each model can capture each outbreak type quite well for both data streams. Additionally, using two models to learn the mixed outbreak type can also help us improve the ability to discriminate between the two different outbreak types. Figures 9 and 10 show the posterior conditional probability of the correct outbreak type, Pr(correct type | Data) / (Pr(correct type | Data) + Pr(incorrect type | Data)), as a function of the outbreak day. As we can see, near the start of the outbreak, the posterior probability of an outbreak is divided nearly 50/50 between the correct and incorrect outbreak type, but by the end of the outbreak, posterior conditional probability of the correct outbreak type has risen to 72% for a cough outbreak or 75% for a nausea outbreak.

Finally, we note that, in addition to learning the distribution of the sparsity parameter, we can also learn the distribution of each outbreak type's relative effects on the two data streams from the same labeled training data, as in [1]. We considered two outbreak types which had both different values of p for the injected outbreaks (p = 0.2 and p = 0.8, as above) and also different relative effects on the two data streams (one outbreak type affected the cough stream twice as much as the nausea stream, and one type affected nausea twice as much as cough). As can be seen from Figure 11, either learning the sparsity or learning the effect alone enables the correct outbreak type to be identified with approximately 80% probability, but learning both the sparsity and effect

simultaneously allows the probability of identifying the correct outbreak type to increase to approximately 90%.

3.5 Learning performance vs. training sample size

The last aspect we evaluate is the relationship between the learning performance and the training sample size. For each of the six outbreak types described above, we form four different training sets consisting of 25, 50, 100, and 200 injects respectively. And then, a separate probability distribution of p was learned from each training dataset. Figures 12 to 17 illustrate that the distribution of the sparsity parameter p can be learned reasonably well even if the size of training outbreaks is as small as 25.

4. Conclusion

The Generalized Fast Subset Sums (GFSS) method is an extension of the previously proposed Fast Subset Sums framework. The new sparsity parameter p in the GFSS framework describes the expected proportion of locations affected within a given circular neighborhood, and thus can be varied to emphasize detection of more compact or more dispersed clusters. We demonstrate that the posterior distribution of the sparsity parameter can be learned accurately based on fully labeled training data even when the size of the training sample is small. With the learned sparsity parameter, the GFSS method has higher detection power and higher spatial accuracy than the previously proposed FSS and MBSS method, especially for elongated or irregular outbreaks. Additionally, learning two different models for outbreak types which have different sparsities (but are otherwise identical) allows us to precisely distinguish between the two outbreak types. Finally, we demonstrate that the relative effects of each outbreak type on the two monitored data streams can also be learned from the training data, and that learning the effects and the sparsity together further improves performance as compared to learning either parameter alone.

In future work, we will evaluate the impact of using a larger or smaller number of discrete components for the learned distribution of the sparsity parameter p. We will extend the GFSS framework by also learning the distributions of the center location s_c and the neighborhood size k from labeled training data. We will also consider the case of partially labeled data, when only a subset of the affected locations is identified. Finally, we will also evaluate the performance of the GFSS method in the case of a misspecified model, when the outbreaks are not generated under the same framework that was used for detection.

References:

[1] Neill DB, Cooper GF. (2010) A multivariate Bayesian scan statistic for early event detection and characterization. Machine Learning 79(3): 261-282.

[2] Neill DB. (2011). Fast Bayesian scan statistics for multivariate event detection and visualization. Statistics in Medicine 30(5): 455-469.

[3] Neill DB. (2009). Expectation-based scan statistics for monitoring spatial time series data. International Journal of Forecasting 25:498–517.

[4] Kulldorff M. (1997). A spatial scan statistic. Communications in Statistics: Theory and Methods: 26(6):1481–1496.

[5] Kulldorff M, Nagarwalla N. (1995). Spatial disease clusters: detection and inference. Statistics in Medicine 14:799–810

[6] Kulldorff M, Mostashari F, Duczmal L, Yih WK, Kleinman K, Platt R. (2007). Multivariate scan statistics for disease surveillance. Statistics in Medicine 26:1824–1833.

[7] Neill DB, Lingwall J. (2007) A nonparametric scan statistic for multivariate disease surveillance. Advances in Disease Surveillance 4:106.

[8] Neill DB. (2008). Fast and flexible outbreak detection by linear-time subset scanning. Advances in Disease Surveillance 5: 48.

[9] Makatchev M, Neill DB. (2008). Learning outbreak regions in Bayesian spatial scan statistics. Proceedings of the ICML/UAI/COLT Workshop on Machine Learning for Health Care Applications, Helsinki, Finland.

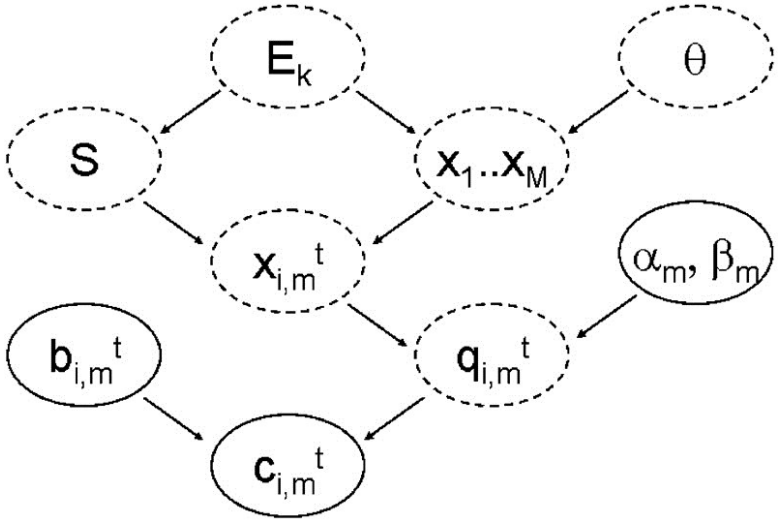

Figure 1. The structure of the multivariate Bayesian scan statistic framework, from Neill and Cooper [1].

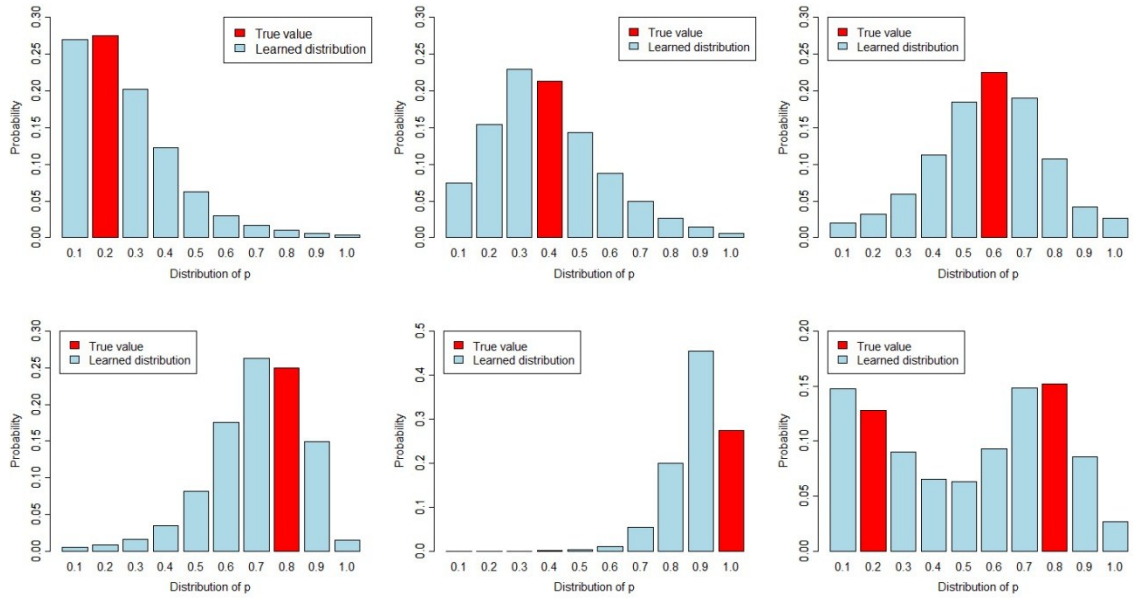

Figure 2. True value and learned distribution of sparsity parameter p, for six different simulated outbreak types injected into cough data from Allegheny County, PA.

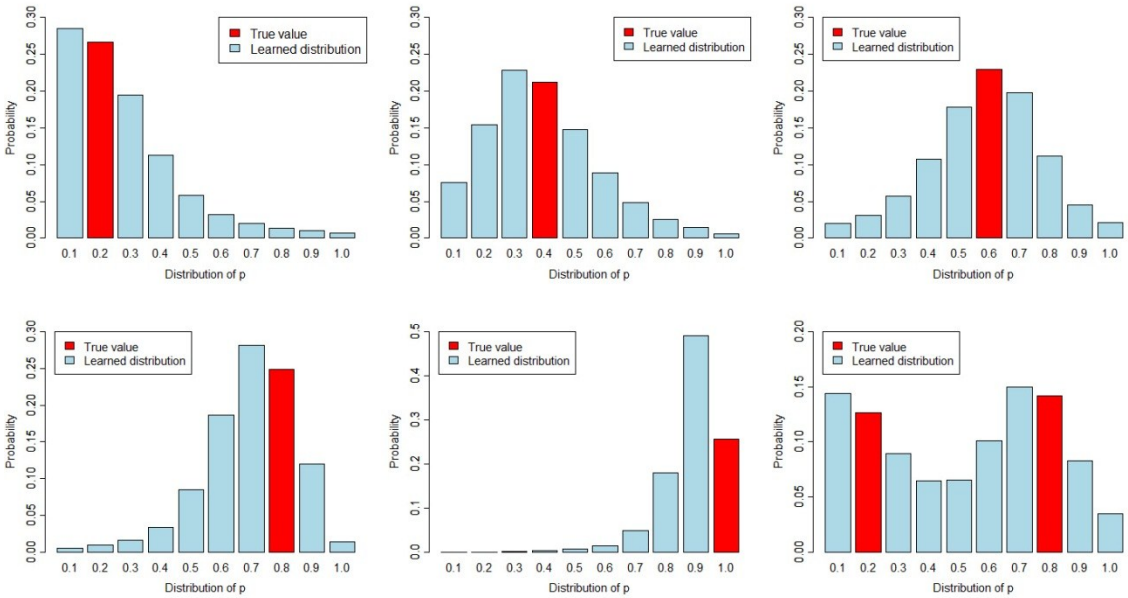

Figure 3. True value and learned distribution of sparsity parameter p, for six different simulated outbreak types injected into nausea data from Allegheny County, PA.

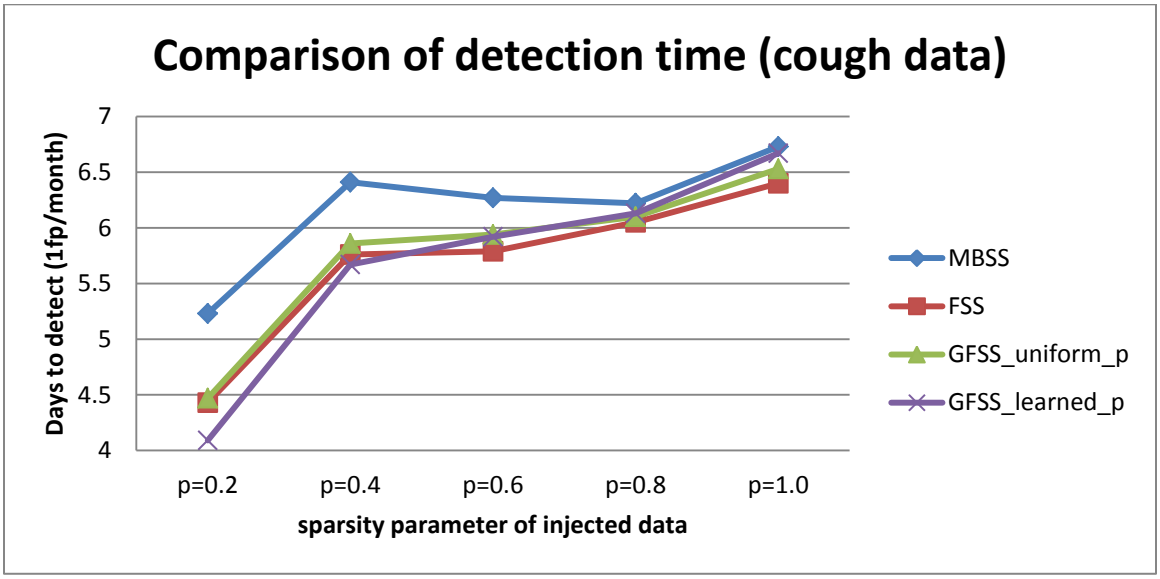

Figure 4. The detection time of four competing methods, for five different simulated outbreak types injected into cough data from Allegheny County, PA.

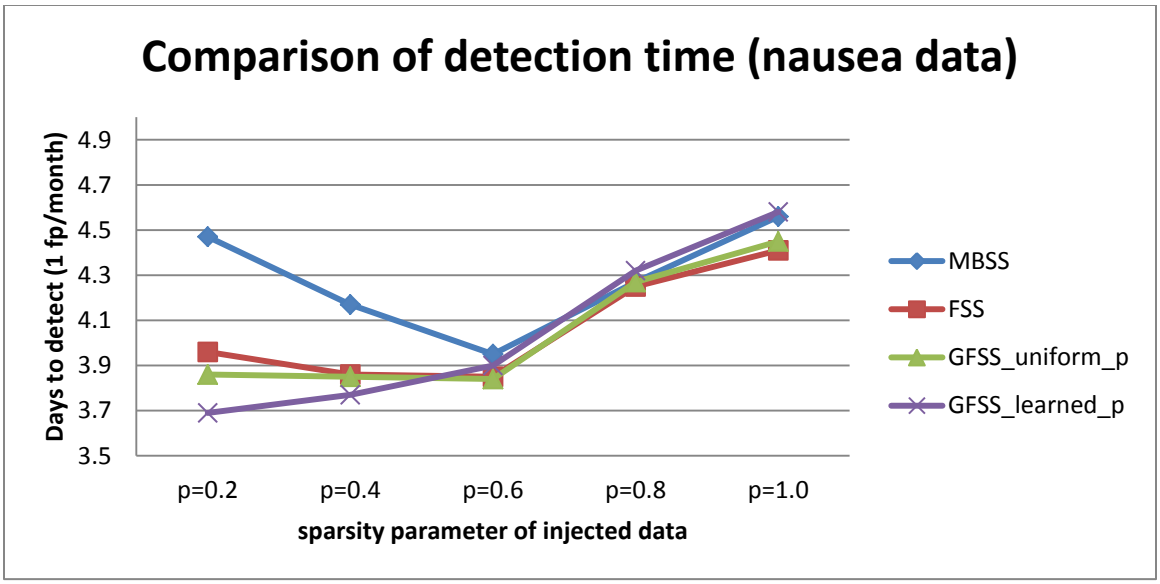

Figure 5. The detection time of four competing methods, for five different simulated outbreak types injected into nausea data from Allegheny County, PA.

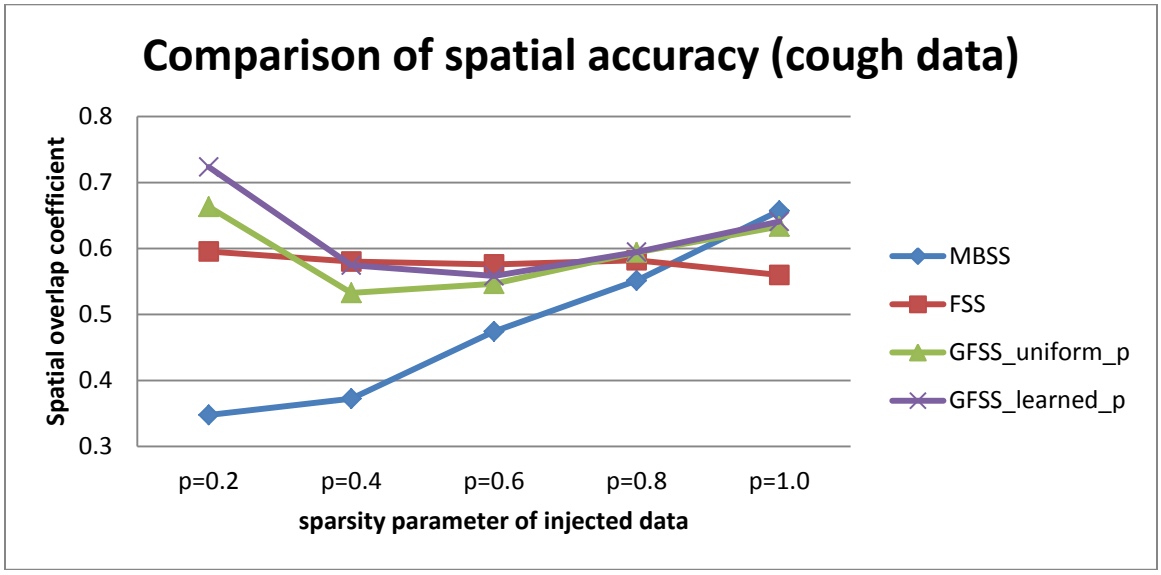

Figure 6. The spatial accuracy (overlap coefficient) of four competing methods, for five different simulated outbreak types injected into cough data from Allegheny County, PA.

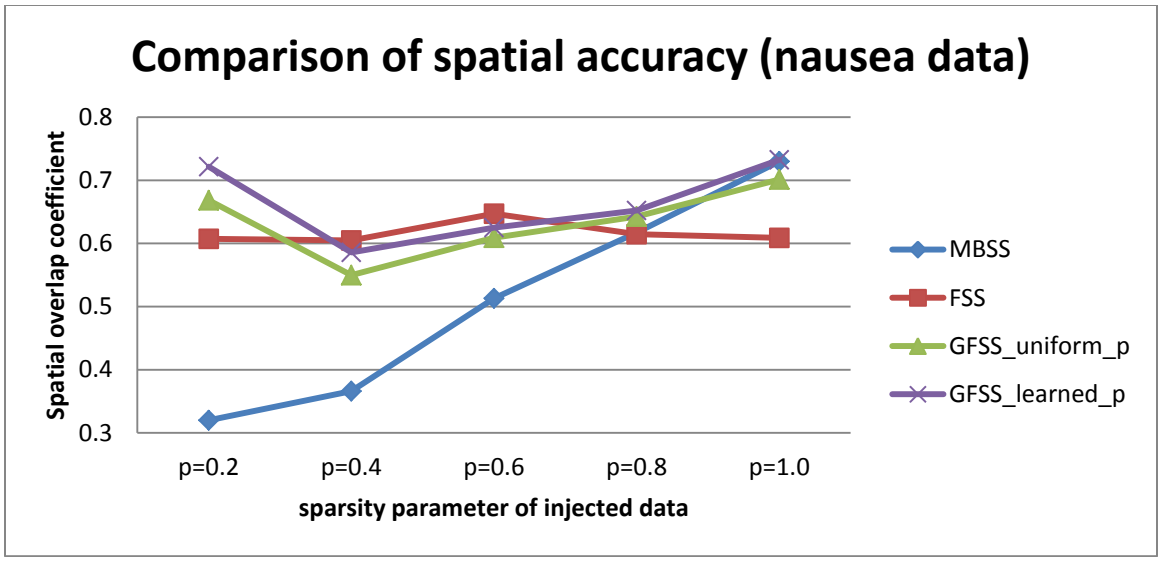

Figure 7. The spatial accuracy (overlap coefficient) of four competing methods, for five different simulated outbreak types injected into nausea data from Allegheny County, PA.

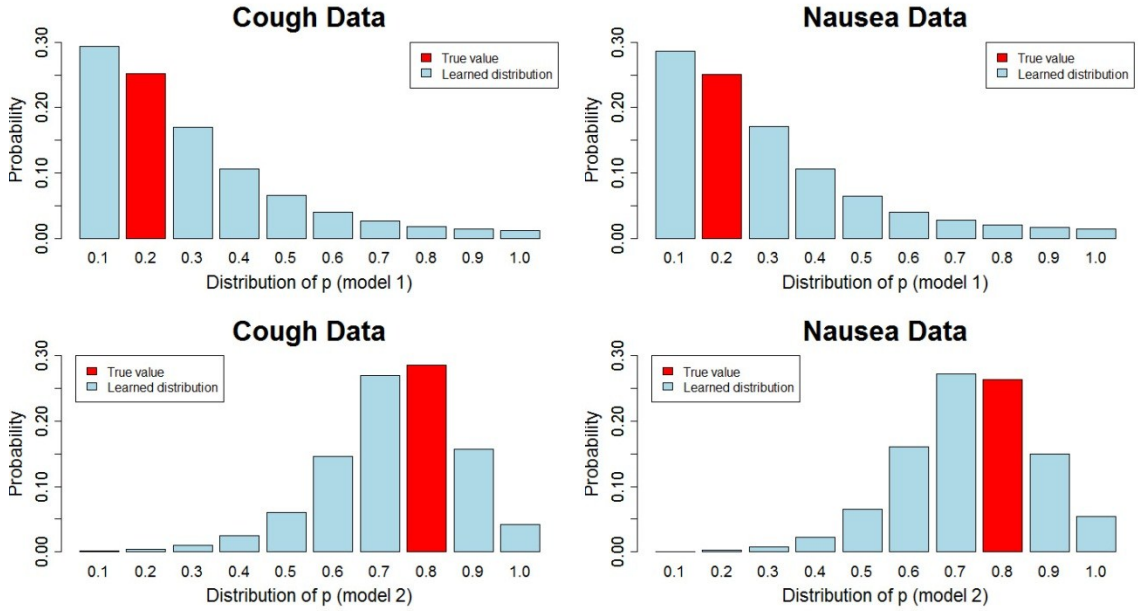

Figure 8. True value and learned distribution of the sparsity parameter p for the mixed outbreak type (equal mixture of p = 0.2 and p = 0.8) assuming two different outbreak models.

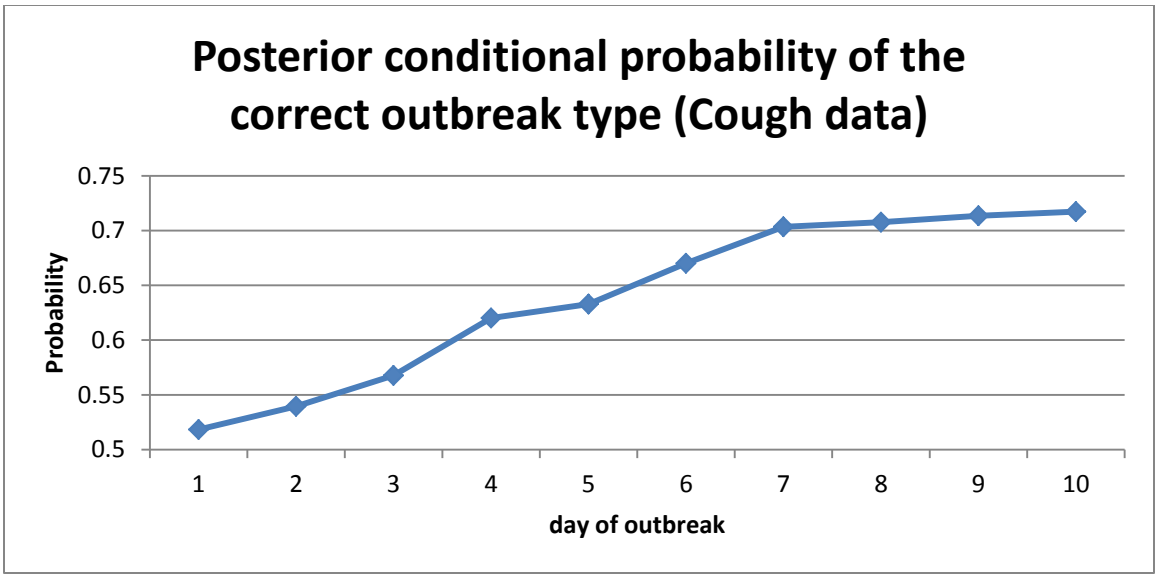

Figure 9. Posterior probability of the correct outbreak type as a function of day of outbreak, for cough data.

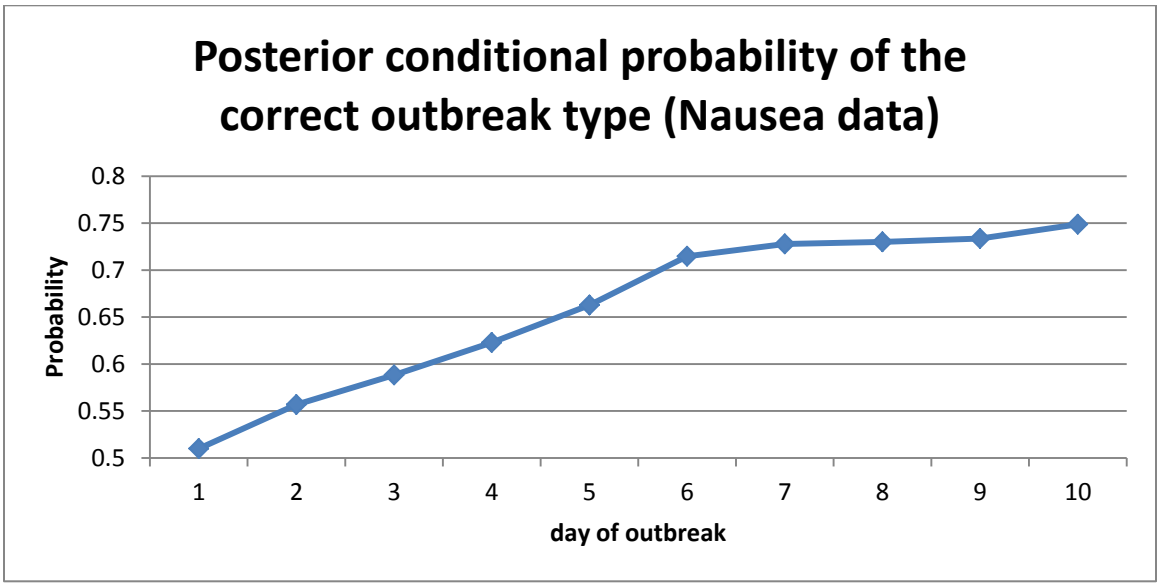

Figure 10. Posterior probability of the correct outbreak type as a function of day of outbreak, for nausea data.

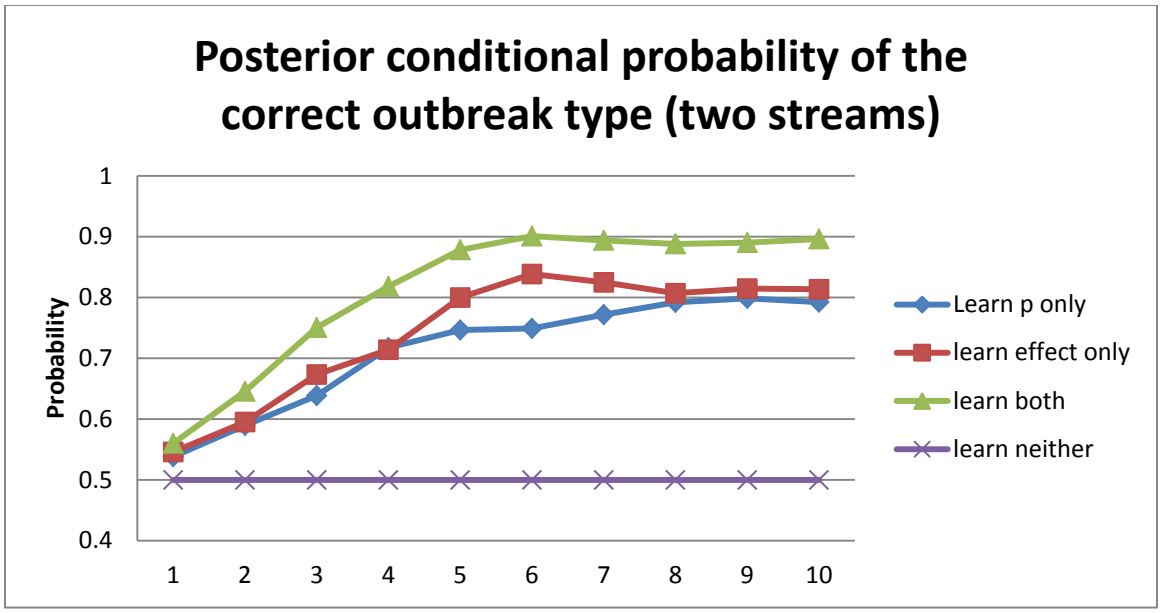

Figure 11. Posterior probability of the correct outbreak type as a function of day of outbreak, assuming two outbreak models and monitoring two data streams.

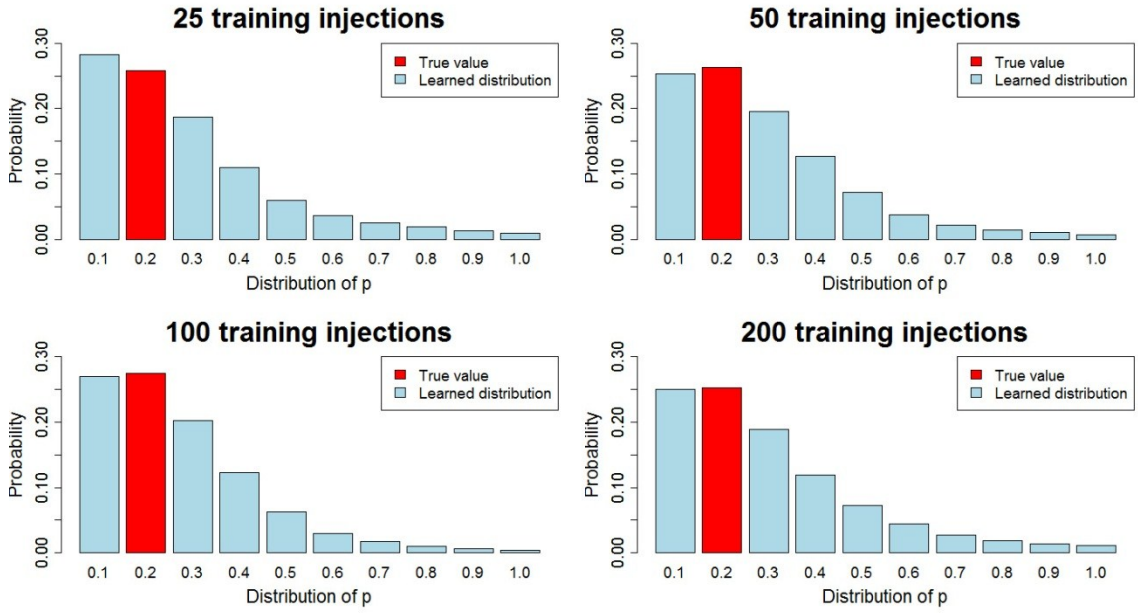

Figure 12. Learned distributions of sparsity parameter p given different training sample sizes (true p = 0.2).

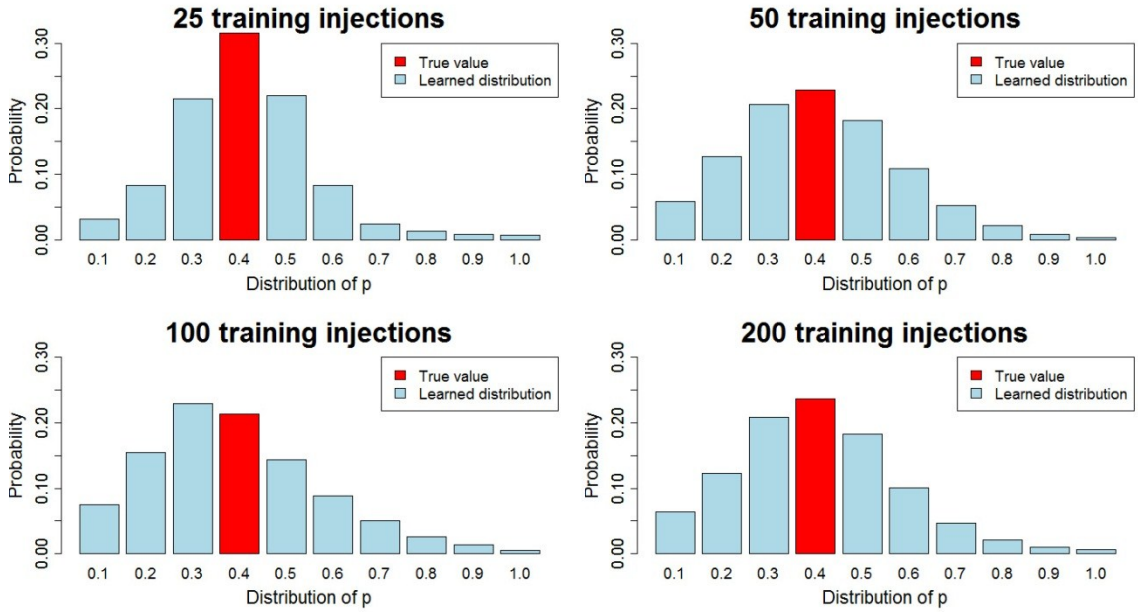

Figure 13. Learned distributions of sparsity parameter p given different training sample sizes (true p = 0.4).

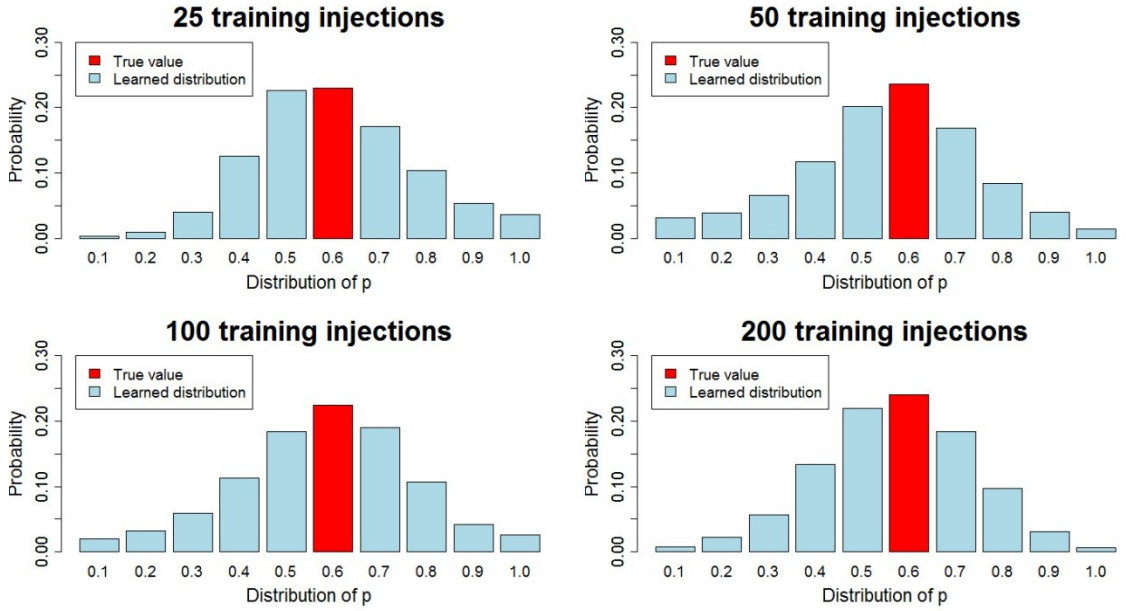

Figure 14. Learned distributions of sparsity parameter p given different training sample sizes (true p = 0.6).

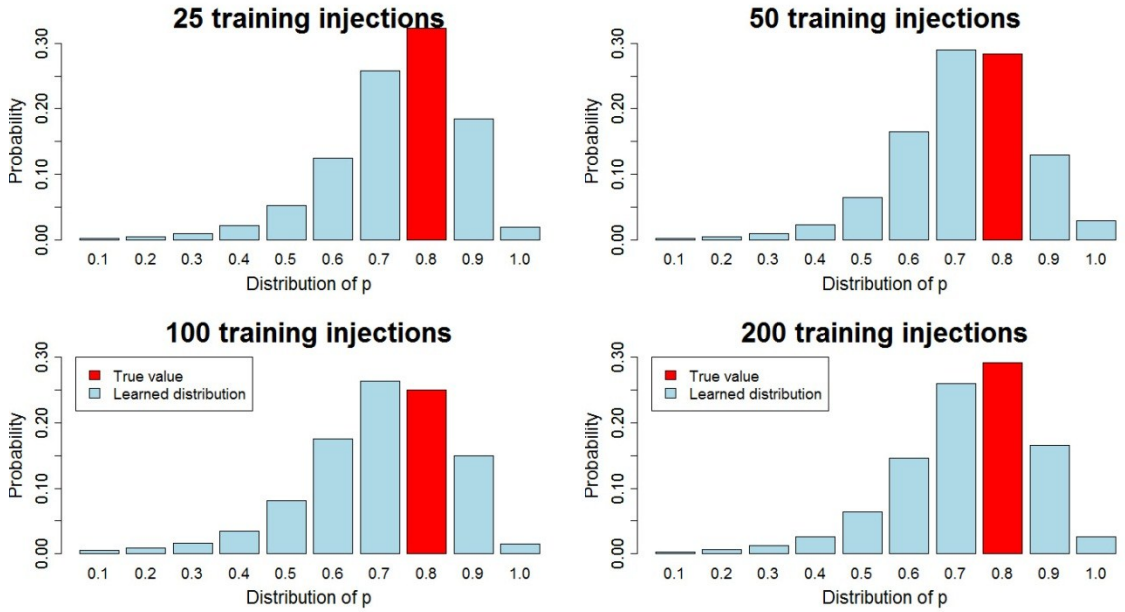

Figure 15. Learned distributions of sparsity parameter p given different training sample sizes (true p = 0.8).

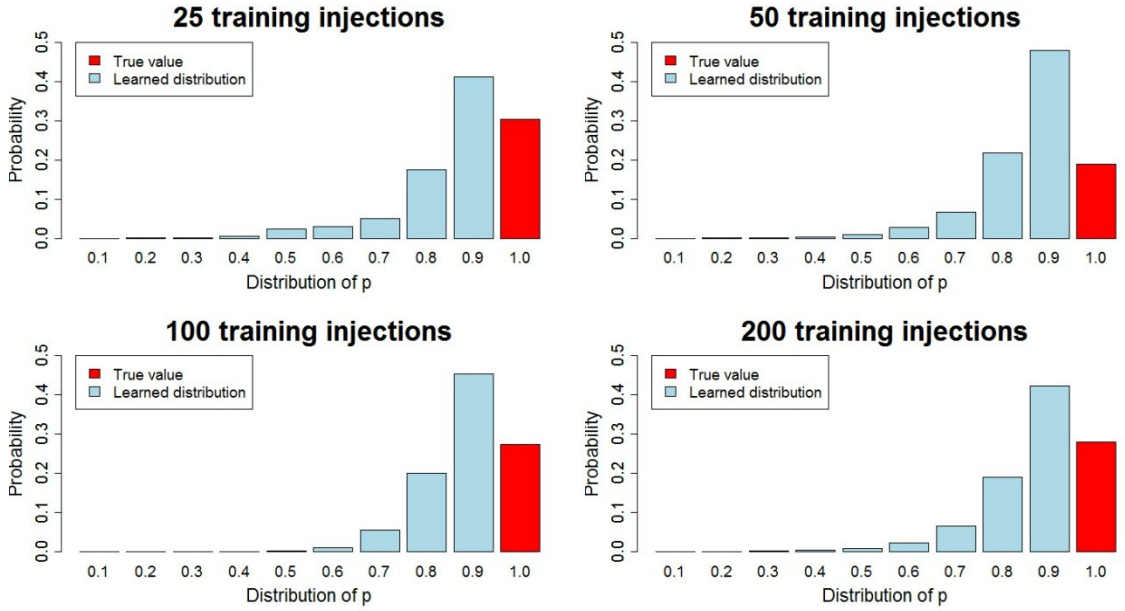

Figure 16. Learned distributions of sparsity parameter p given different training sample sizes (true p = 1.0).

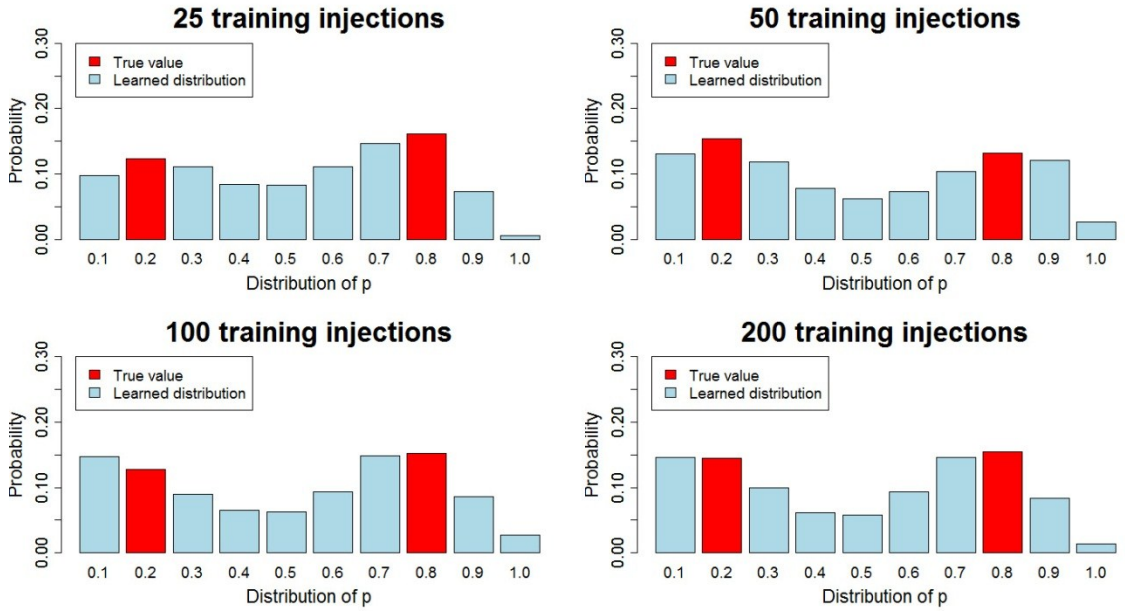

Figure 17. Learned distributions of sparsity parameter p given different training sample sizes (true p is a mixture of p = 0.2 and p = 0.8 with equal probabilities).