Active Surveillance via Group Sparse Bayesian Learning

Journal:	*Transactions on Pattern Analysis and Machine Intelligence*
Manuscript ID	TPAMI-2019-08-0686.R1
Manuscript Type:	Regular
Keywords:	Epidemic Dynamics, Diffusion, Sensor Deployment, Dynamical System, Automatic Relevance Determination

Our Responses to Comments

Submission ID: TPAMI-2019-08-0686

Active Surveillance via Group Sparse Bayesian Learning

Dear Editor-in-Chief, Handling Associate Editor, and Reviewers,

First of all, we would like to sincerely thank the handling associate editor and all the reviewers for your detailed comments and constructive suggestions.

We have carefully studied them to make sure that they are adequately addressed and incorporated in our revisions. In the following, we provide our detailed responses to the comments and indicate the corresponding changes made in the revisions. To help you to check our responses, we reproduce the original comments in italic, followed by the responses.

Sincerely,

Hongbin Pei, Bo Yang, Jiming Liu, Kevin Chen-Chuan Chang

Contents

1 Responses to Handling Associate Editor ..... Page 2
2 Responses to Reviewer #1 ..... Page 7
3 Responses to Reviewer #2 ..... Page 10
4 Responses to Reviewer #3 ..... Page 15

April 2, 2020

Handling Associate Editor

Comment 1

"Contribution of this submission with respect to the authors's AAAI'18 paper given that the theoretical contribution claimed by the paper can be found in other papers on group sparse Bayesian method."

Response:

Thanks so much for your constructive comments. A preliminary version of this work was presented as a regular paper at the 32nd AAAI Conference on Artificial Intelligence (AAAI-18). The current submission is a "more than 70% (more than 13 out of 18 pages have now been newly added or updated)" substantial revision of our preliminary conference publication. Please kindly refer to the enclosed document, *"Major Differences between our Preliminary AAAI Paper and the New TPAMI Submission"*, for details.

In what follows, we will combine our responses to your Comment 1 and Comment 2, in view that both of them are about the novelty of theoretic results in this submission.

Comment 2

"Please explain the novelty of the theoretic results particularly regarding the paper below: Z. Zhang and B. D. Rao. "Extension of SBL Algorithms for the Recovery of Block Sparse Signals With Intra-Block Correlation." IEEE Transactions on Signal Processing, 61.8 (2013): 2009-2015"

Response:

In this submission, a new group sparse Bayesian learning method is proposed that is aimed to tackle the challenge of active surveillance, which is innovative in both theory and application. In the literature, Zhang's work is related to ours, where the group sparse Bayesian learning was used to recover block sparse signals [1][2][3] (see references on page 6 of this document). Since both Zhang's work and ours focus on modeling the group sparsity feature of Bayesian models, there will be a connection between them in form. However, the theoretical results as contributed by this submission, including both computational models and algorithms, are novel and fundamentally different from Zhang's work.

Now, let us elaborate on the primary distinctions between the two as follows.

1. **Different Gaussian prior.**
   Although both Zhang's work and ours adopt a zero-mean multivariate Gaussian distribution as the prior to induce group sparsity, the two Gaussian priors employ different intra-group correlation modelling. We show the two different covariance matrices of prior in Table 1.

Table 1: Covariance matrix of Gaussian prior

$\boldsymbol{\Sigma_0} = \begin{bmatrix} \gamma_1\mathbf{I}_N & & \\ & \ddots & \\ & & \gamma_N\mathbf{I}_N \end{bmatrix}$	$\boldsymbol{\Sigma_0} = \begin{bmatrix} \gamma_1\mathbf{B}_1 & & \\ & \ddots & \\ & & \gamma_N\mathbf{B}_N \end{bmatrix}$
In this submission	In Zhang's studies

   In Zhang's work, the correlation in each group is modelled by a learnable matrix $\mathbf{B}_i, i = [1, N]$ (see details in Section II in [1][2][3]). In our work, we ignore such intra-group correlation by using an identity matrix $\mathbf{I}_N$ (see Eq. 9 in this submission).

We uncover, for the first time, two important properties of the designed priors to enforce group sparsity by analyzing their sparsity effect on parameters in Section 3.2.1 of this submission. By introducing a non-information auxiliary hyper-prior, we show that the marginal prior is a spike-and-slab distribution and that it has two nice properties to enforce group sparsity: (1) it is a concave and monotonic function; (2) it is a heavy-tailed distribution (see the prior distribution in Fig. 2 of this submission). We do not find any related sparsity analysis on prior in Zhang's papers [1][2][3].

Besides, the biggest advantage brought by our prior modeling is that it can **significantly reduce the time complexity of Bayesian inference** from $O(N^5\alpha)$ to $O(N^2\alpha)$, where $\alpha = \max(T, N)$, by simplifying the posterior covariance matrix into a diagonal-block matrix and then coming up with two fast matrix operations applied to it (see Section 3.4 in this submission). It should be noted that the two fast matrix operations we propose are **exact and flexible**, in addition to group sparse Bayesian inference, they can also be applied to other kinds of tasks involving diagonal-block matrices.

In contrast, Zhang's models have high time complexity. For instance, their *T-SBL* model [2], most relevant to ours, has a time complexity $O(M^2L^2N)$, or $O(N^5\alpha)$ after we make their notations consistent with ours. As they discussed in [2], by using some approximation techniques such as the speed-up scheme proposed by [4], the time of *T-SBL* model can be further reduced to $O(MN^2)$, e.g. the *T-MSBL* model, being approximately on the same order as ours.

2. **Different likelihood function.**
We model continuous data and discrete data by Gaussian likelihood and Bernoulli likelihood, respectively (see Eqs. 9 and 11 in this submission). In contrast, Zhang's work only model continuous data via Gaussian distributions, since the signals they processed for recovery are continuous [1][2][3].

   To the best of our knowledge, this submission is the first work that proposes a group sparse Bayesian model for handling discrete data. In this submission, modelling discrete data is motivated by the fact that many diffusion processes can be naturally represented as discrete signals, e.g., whether a piece of news is posted on a blog in the information spread setting.

3. **Different posterior distribution.**
As the adopted priors and likelihoods are different, the posterior distributions in this submission and Zhang's work are accordingly different, as shown in Table 2.

Table 2: Summary of differences in terms of posterior distributions

	This submission		Zhang's work
Data type	Continuous data	Discrete data	Continuous data
Posterior	Gaussian	Approximate Gaussian	Gaussian
Covariance	$(\boldsymbol{\Sigma}_0^{-1} + \lambda^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}$	$(\boldsymbol{\Sigma}_0^{-1} + 2\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\boldsymbol{\xi})\boldsymbol{\Phi})^{-1}$	$(\boldsymbol{\Sigma}_0^{-1} + \lambda^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}$
Expectation	$\lambda^{-1}\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}$	$2^{-1}\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{\Phi}^{\mathrm{T}}(2\boldsymbol{y} - \mathbf{1})$	$\lambda^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\mathbf{D}^{\mathrm{T}}\boldsymbol{y}$

For discrete data, the posterior in this submission is an approximate Gaussian, which is obviously different from the posterior in Zhang's studies, a Gaussian. And $\boldsymbol{\xi}$ in approximate Gaussian is a learnable variational parameter vector.

As for continuous data, although the posteriors in this submission and in Zhang's work are Gaussian, however, it should be noted that the two posteriors are intrinsically different, due to the two different prior covariance $\boldsymbol{\Sigma}_0$, as shown in Table 1. The $\boldsymbol{\Sigma}_0$ in this submission simplifies the posterior covariance into a diagonal-block matrix, which enables us significantly reduce the time complexity of Bayesian inference.

4. **Different Bayesian inference method.**
   We employ the expectation–maximization (EM) method and the variational EM method to infer the posterior distribution and hyper-parameters in the cases of continuous data and discrete data, respectively.

   In contrast, Zhang's work only uses the EM method. Moreover, we introduce a Gaussian lower bound for the Bernoulli likelihood to address the challenge that original Q function in EM cannot be analytically solved in the case of discrete data (see Theorem 2, and its proof in this submission).

Action:
We have revised this submission by adding a new subsection, entitled "A connection to block sparse Bayesian learning", to incorporate the above response. Please refer to Section 5.1 on page 13 of this submission for details.

Comment 3

"Why is formula (2) claimed to be a new proposed measure?"

Response:
Formula (2) might be confused with formula (15), i.e., the importance measure for selecting sentinels, vs. the updating rule for the hyper-parameter, since they look a bit alike in form. Please allow us to make the following further clarifications.

Formula (15) indicates how to update the hyper-parameter γ in the optimization process. When the updating converges, our algorithm will use formula (2) to calculate the importance of each component, and thereafter select sentinels according to these values. From this perspective, formula (2) can be regarded as a measure of importance.

Now, let us explain the novelty of the importance measure given by formula (2). This measure is new in two key aspects: First, in the domain of sensor placement and active surveillance, it is a new measure for evaluating and hence determining important sentinels. Second, it is derived from two newly proposed group sparse Bayesian models, as given in this submission.

In the literature, the tasks of sensor placement and active surveillance have often been modeled as a linear inverse problem, a Gaussian process interpolation, or a matrix completion problem (see details in Section 2 of this submission). **To the best of our knowledge, it is the first time to model a sensor placement task as a group sparse Bayesian learning problem.** In this submission, we show that the data-dependent hyper-parameter γ (i.e., formula (2)) is capable of indicating the importance of each component for predicting the behavior of the whole system, and thus, it can serve as an effective and easy to calculate measure for selecting as a fewer sentinels as possible to be actively monitored. Our comprehensive theoretical analysis and empirical validations provide the rigorous support for this finding.

In this submission, we propose two new group sparse Bayesian models, and derive formula (2) by optimizing their hyper-parameters. More specifically, the two models are formulated for handling continuous data and discrete data, respectively (please refer to our responses to Comments 1 and 2 above), and the EM method and the variational EM method are employed to optimize their respective hyper-parameters. **Interestingly, through theoretical derivations, we find that, whether for the continuous model or the discrete model, their learning rule for hyper-parameter γ has exactly the same form (i.e., formula (15) and (2)), although the two models adopt entire different likelihood function.** This unified form of hyper-parameter is also a new finding in the study of group sparse Bayesian learning.

Action:
We have further improved our submission based on the above response. Please refer to the explanation for formula (2) in Section 3.2 on page 3 of this submission for details.

Comment 4

" What is the novelty about the hierarchical prior modelling to induce sparsity for modelling?"

Response:
Thanks very much for the question.

Hierarchical prior modelling to induce sparsity can be considered as a standard paradigm, which has been widely used in sparse Bayesian modelling. In this paradigm, the novelty comes from a new way of elaborating suitable priors and likelihoods, according to different domain problems.

In the literature of group sparse Bayesian learning, the existing work, such as Zhang's studies [1][2][3], adopt the hierarchical prior to model group sparsity as well. Compared with these methods, here we design different priors and likelihoods, and thus derive different posteriors. In addition to solving problems in different fields, the additional benefit is that the time complexity of posterior inference is greatly reduced.

Moreover, the existing hierarchical prior modeling for group sparsity only focuses on continuous data, and does not provide models to deal with discrete data, which is also common in the field of signal processing, especially for active surveillance. For this reason, we propose a new prior and a likelihood suitable for dealing with discrete data, as well as the corresponding posterior inference algorithm. The related discussions have been provided in our responses to Comments 1, 2 and 3 above. Here, what we would like to emphasize is that, in addition to the contribution of modelling design, in the submission, we also provide a new perspective to understand the theoretical properties of the designed Bayesian models.

More specifically, we uncover two important properties of the designed priors to enforce group sparsity by analyzing their sparsity effect on parameters. To obtain the "true" shape of the designed conditional sparse prior, we introduce a non-information auxiliary hyper-prior, an inverse gamma distribution, and thereafter integrate out hyper-parameters from joint prior, thereby obtaining an equivalent marginal prior distribution over parameters (please refer to Section 3.2.1 of this submission for details). We find that the marginal prior is a spike-and-slab distribution and that it has two nice properties to enforce group sparsity: (1) it is a concave and monotonic function, which can effectively restrict the probability on trivial groups; (2) it is a heavy-tailed distribution, which allows the important groups with large weights (see the prior distribution in Fig. 2 of this submission).

Comment 5

"The theoretical justifications in this submission can be found in the Supporting Information of the original AAAI-18 paper. Can the authors clarify this issue?"

Response:
Please note that the "Support Information" was on the web server of **arXiv** as an uncopyrighted/unpublished material[1], which is not a publication by itself, nor a part of the official AAAI-18 conference paper publication. Due to the page submission limit, we could not put very detailed theoretical justifications into the conference paper, but without such proofs it could be difficult for readers to fully understand those theorems, thus we put up them as an unofficial material on **arXiv**, and just provided its URL link in the AAAI-18 conference paper.

Action:
We have double-checked, and guarantee to strictly follow, the IEEE policy[2] about the preprint manuscripts on **arXiv** as follows.

- Can an author post his manuscript on a preprint server such as **arXiv**?
  "Yes. The IEEE recognizes that many authors share their unpublished manuscripts on public sites. Once manuscripts have been accepted for publication by IEEE, an author is required to post an IEEE copyright notice on his preprint. Upon publication, the author must replace the preprint with either 1) the full citation to the IEEE work with Digital Object Identifiers (DOI) or a link to the paper's abstract in IEEE Xplore, or 2) the accepted version only (not the IEEEpublished version), including the IEEE copyright notice and full citation, with a link to the final, published paper in IEEE Xplore. "

[1] https://arxiv.org/abs/1712.00328
[2] https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/author_faq.pdf

[1] Z. Zhang and B. D. Rao. "Extension of SBL Algorithms for the Recovery of Block Sparse Signals With Intra-Block Correlation." IEEE Transactions on Signal Processing, 2013, 61(8): 2009-2015.
[2] Z. Zhang and B. D. Rao. "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning." IEEE Journal of Selected Topics in Signal Processing, 2011, 5(5): 912-926.
[3] Z. Zhang and B. D. Rao. "Recovery of block sparse signals using the framework of block sparse Bayesian learning." ICASSP 2012, Japan, March, 2012.
[4] D. P. Wipf and B. D. Rao. "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem." IEEE Transactions on Signal Processing, 2007, 55(7): 3704-3716.

Reviewer #1

Comment 1

"On the spike-and-slab prior shown in Figure 2, some more clarifications on the figure can be helpful. The Eq. 7 is not a probability density function. How to plot its distribution shown in Figure 2?"

Response:

We appreciate very much your detailed comments.

The Eq. 7 is indeed not the probability density function (PDF) of the spike-and-slab prior, but is an auxiliary function being proportional to the prior. The prior actually cannot be calculated analytically because of the intractable limit operation in the Eq. 7 when $a \to 0$ and $b \to 0$. Instead of using PDF to represent the prior, we uncover and illustrate the shape of the prior by introducing an auxiliary function which is proportional to the PDF of prior and easy to calculate, as shown in the right side of Eq. 7. The auxiliary function has a same shape with the prior because it is proportional to the prior PDF. Therefore, the prior distribution can be illustrated by plotting the auxiliary function, as shown in Figure 2 in this submission.

To plot the auxiliary function in a 3-D space, there are two steps, sampling and interpolating. In sampling step, we random sample a large amount of points (i.e., (x, y)) in a 2-D feasible region of the auxiliary function, i.e., $\{(x, y)||x| < l, |y| < l, x \neq 0, y \neq 0\}$, which is a square region without the points on every axis. Then we calculate the function value (i.e., z) on every sampled point though feeding the point to the auxiliary function, i.e., the right side of Eq. 7. After standardizing the summation of all function values z to one, we obtain the final data samples (i.e., (x, y, z)) outlining the auxiliary function.

In step two, we employ a smoother interpolating method in MATLAB toolbox to form a continuous 2D surface from the discrete data samples. This surface is then plotted in Figure 2, from which we find two nice properties of the prior distribution to induce group sparsity of parameters.

Action:

We have revised this submission by adding a new section in its appendix to incorporate the above content. Please refer to Section D on page 3 of Appendix for more details.

Comment 2

"In Section 3.4, two efficient matrix operations for diagonal-block matrix are proposed to calculate the time-consuming multiplication $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ and matrix inverse $\mathbf{\Sigma}_s^{-1}$. However, it seems that the variational parameter $\boldsymbol{\xi}$ destroys the diagonal-block property of the covariance matrix $\mathbf{\Sigma}_s$ in the logistical discrete system (Theorem 2). Could you explain how the efficient operations work on this matrix?"

Response:

This inapplicable problem exists in the proposed accelerated algorithm, and we address it by modelling the variational parameter vector $\boldsymbol{\xi} \in \mathcal{R}^{TN}$ being with the following particular group structure,

$$\boldsymbol{\xi} = [\underbrace{\hat{\xi}_1, \ldots, \hat{\xi}_1}_{N}, \underbrace{\hat{\xi}_2, \ldots, \hat{\xi}_2}_{N}, \ldots, \underbrace{\hat{\xi}_T, \ldots, \hat{\xi}_T}_{N}]^{\mathrm{T}} \tag{1}$$

by which the diagonal-block property of $\mathbf{\Sigma}_s$ is kept and the two efficient matrix operations can apply to the logistical discrete system. Such modelling is to expand the variational lower bound functions on T points, $\hat{\xi}_i, i \in [1, \ldots, T]$, instead of TN points in the original algorithm. In this submission, we improperly omitted the details about the modelling because of the page limitation. We will provide a detailed description of the modelling to make the revision be completed and self-contained.

We first describe and analyze the inapplicable problem. The two efficient matrix operations is designed for matrix with diagonal-block structure because such matrix can be effectively compressed as block projective matrix without any information loss (see Figure 5 in this submission). However, the logistical discrete system introduce variational parameter vector $\boldsymbol{\xi}$, and the corresponding matrix $\boldsymbol{\pi}(\boldsymbol{\xi})$ is not a diagonal-block matrix because the diagonal entries in its square sub-matrices are determined by $\boldsymbol{\xi}$ and not the same. Since $\boldsymbol{\pi}(\boldsymbol{\xi})$ is a part of both $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\boldsymbol{\xi})\boldsymbol{\Phi}$ and the covariance matrix $\boldsymbol{\Sigma}_s$ (see Theorem 2 in this submission), $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\boldsymbol{\xi})\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}_s$ become not diagonal-block matrices. Thus, the efficient matrix operations no longer apply to calculating the multiplication $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\xi)\boldsymbol{\Phi}$ and the inverse $\boldsymbol{\Sigma}_s^{-1}$ directly.

Based on the above analysis, we can see the problem is caused by the introduced variational parameter $\boldsymbol{\xi}$, and the problem will be addressed if we can modify $\boldsymbol{\xi}$ so as to shape $\boldsymbol{\pi}(\boldsymbol{\xi})$ being a diagonal-block matrix. By doing so, both $\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\xi)\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}_s$ will keep diagonal-block structure and the efficient matrix operations can be applied. Toward this end, we model the variational parameter $\boldsymbol{\xi} \in \mathcal{R}^{TN}$ with particular group structure in accelerated algorithm, as shown in Eq. 1. Specially, we let the N variational parameters for the N dynamic data at time $t, t \in [1, \ldots, T]$ be the same and equal to $\hat{\xi}_t$, i.e., $\boldsymbol{\xi}_{(t-1)N+i} = \hat{\xi}_t, \forall i \in [1, \ldots, N]$. After such modelling, the corresponding matrix $\boldsymbol{\pi}(\boldsymbol{\xi})$ becomes a diagonal-block matrix.

From the Taylor expansion perspective, the variational parameter vector $\boldsymbol{\xi}$ consists of the contact points where the lower bound approximation functions are tangent to the original functions [1] (see references on page 9 of this document). Thus, the modelling for $\boldsymbol{\xi}$ in accelerated algorithm is actually to expand the variational lower bound functions (Eq. 2 in Appendix) on T points, instead of TN in original algorithm. Here, decreasing contact points may influence adversely the variational approximation because the lower bound functions are not expended on the most suitable points. However, experimental results of accelerated algorithm show the influence is limited.

In the case of new modelling for $\boldsymbol{\xi}$ in accelerated algorithm, the learning rule for variational parameter becomes

$$\hat{\xi}_t \leftarrow \frac{1}{N} \sum_{n=(t-1)N+1}^{tN} \sqrt{\boldsymbol{\Phi}_n(\Sigma_{\boldsymbol{s}} + \boldsymbol{\mu}_{\boldsymbol{s}}\boldsymbol{\mu}_{\boldsymbol{s}}^{\mathrm{T}})\boldsymbol{\Phi}_n^{\mathrm{T}}}, \ t \in [1, \ldots, T]. \tag{2}$$

There is no change in learning rules for other hyper-parameters.

Action:
We have further improved our submission by adding a new subsection in Appendix to incorporate the above content. Please refer to Section C.3 on page 3 of Appendix for more details.

Comment 3

"In the experiments, the proposed method is compared with the GP-MI method. The GPs-MI method is sub-modular with provable bounds. The nature of the sub-modularity is the monotonicity and diminishing gain. This is also shown in Figure 6, when the number of sentinels is small, GPs-MI is actually better. It's hard to understand for me why the error curve of GP-MI has fluctuation, which is against the monotonicity."

Response:
The function in GPs-MI method is indeed sub-modular and thus can guarantee the greed algorithm is monotone and have near-optimal solution with provable bounds. However, the sub-modularity is about mutual information criterion to select sensor locations, but not about the accuracy of dynamics prediction. In other words, GP-MI method can guarantee the mutual information criterion increase continuously during its greed selection, but cannot guarantee its prediction error curve is monotone.

The fluctuations in the prediction error curve of GP-MI is caused by ill-conditioned problem in Gaussian processing regression when the number of sentinels is small. It's inevitable. In fact, we can observe some similar fluctuations in the original paper proposed GP-MI method [2], e.g., Figure 12 (see references on page 9 of this document).

Comment 4

"Though the paper provides a good summary of related work, it will be nice to add some very recent works on sensor placement."

Response:

We do our best to survey recent papers on the topics related to active surveillance to further enhance the related work section of this submission. We find three recent relevant papers [3][4][5] (see references on page 9 of this document), which both focus on optimal sensor deployment.

Action:

We have further improved our submission by incorporating the three relevant papers into related work section. Please refer to Section 2 on page 2 of this submission for more details.

Comment 5

"The real-world applications in experiments are interesting. The authors are suggested to share the source code to the public."

Response:

We have released the code of our algorithm on GitHup. The implementation is based on MATLAB. Please access the code via link: https://github.com/hp17illinois/Active-Surveillance-via-Group-Sparse-Bayesian-Learning.

[1] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.

[2] Krause A, Singh A, Guestrin C. "Near-optimal sensor placements in Gaussian processes: Theory, efficient algorithms and empirical studies." Journal of Machine Learning Research, 2008, 9(Feb): 235-284.

[3] Jiang C, Chen Z, Su R, et al. "Group Greedy Method for Sensor Placement." IEEE Transactions on Signal Processing, 2019, 67(9): 2249-2262.

[4] Rusu C, Thompson J, Robertson N M. "Sensor scheduling with time, energy, and communication constraints." IEEE Transactions on Signal Processing, 2017, 66(2): 528-539.

[5] Rusu C, Thompson J. "On the use of tight frames for optimal sensor placement in time-difference of arrival localization." EUSIPCO 2017, Greece, September, 2017.

Reviewer #2

Comment 1

"The authors claimed that this new submission has improved their previous conference paper in both theoretical and experimental perspectives. In fact, the technique used in this paper is the well-known block sparse Bayesian learning algorithm, where theoretical results have already been available. In particular, the analysis in Section 3.2.1, and the Theorem 1 and Theorem 2 in Section 3.2.3 can be found in the existing paper on group SBL such as the one below.
Z. Zhang and B. D. Rao, "Extension of SBL Algorithms for the Recovery of Block Sparse Signals With Intra-Block Correlation," in IEEE Transactions on Signal Processing, 61.8 (2013): 2009-2015.
The authors should explain explicitly how their theoretic results are unique and cannot be found elsewhere."

Response:

Thanks very much for your helpful comments.

A preliminary version of this work was presented as a regular paper at the 32nd AAAI Conference on Artificial Intelligence (AAAI-18). The current submission is a "more than 70% (more than 13 out of 18 pages have now been newly added or updated)" substantial revision of our preliminary conference publication. Please kindly refer to the enclosed document, *"Major Differences between our Preliminary AAAI Paper and the New TPAMI Submission"*, for details.

In what follows, we will elaborate on the novelty of theoretic results in this submission compared to the existing block sparse Bayesian learning in the literature.

In this submission, a new group sparse Bayesian learning method is proposed that is aimed to tackle the challenge of active surveillance, which is innovative in both theory and application. In the literature, Zhang's work is related to ours, where the group sparse Bayesian learning was used to recover block sparse signals [1][2][3] (see references on page 14 of this document). Since both Zhang's work and ours focus on modeling the group sparsity feature of Bayesian models, there will be a connection between them in form. However, the theoretical results as contributed by this submission, including both computational models and algorithms, are novel and fundamentally different from Zhang's work. We will elaborate on the primary distinctions between them in the following, especially the parts mentioned in this comment, "the analysis in Section 3.2.1" and "the Theorem 1 and Theorem 2 in Section 3.2.3".

1. **Different Gaussian prior.**
   Although both Zhang's work and ours adopt a zero-mean multivariate Gaussian distribution as the prior to induce group sparsity, the two Gaussian priors employ different intra-group correlation modelling. We show the two different covariance matrices of prior in Table 3.

Table 3: Covariance matrix of Gaussian prior

$\boldsymbol{\Sigma_0} = \begin{bmatrix} \gamma_1 \mathbf{I}_N & & \\ & \ddots & \\ & & \gamma_N \mathbf{I}_N \end{bmatrix}$	$\boldsymbol{\Sigma_0} = \begin{bmatrix} \gamma_1 \mathbf{B}_1 & & \\ & \ddots & \\ & & \gamma_N \mathbf{B}_N \end{bmatrix}$
In this submission	In Zhang's studies

In Zhang's work, the correlation in each group is modelled by a learnable matrix $\mathbf{B}_i, i = [1, N]$ (see details in Section II in [1][2][3]). In our work, we ignore such intra-group correlation by using an identity matrix $\mathbf{I}_N$ (see Eq. 9 in this submission).

We uncover, for the first time, two important properties of the designed priors to enforce group sparsity by analyzing their sparsity effect on parameters in Section 3.2.1 of this submission. By introducing a non-information auxiliary hyper-prior, we show that the marginal prior is a spike-and-slab distribution and that it has two nice properties to enforce group sparsity: (1) it is a concave and monotonic function; (2) it is a heavy-tailed distribution (see the prior distribution in Fig. 2 of this submission). We do not find any related sparsity analysis on prior in Zhang's papers [1][2][3].

Besides, the biggest advantage brought by our prior modeling is that it can **significantly reduce the time complexity of Bayesian inference** from $O(N^5\alpha)$ to $O(N^2\alpha)$, where $\alpha = \max(T, N)$, by simplifying the posterior covariance matrix into a diagonal-block matrix and then coming up with two fast matrix operations applied to it (see Section 3.4 in this submission). It should be noted that the two fast matrix operations we propose are **exact and flexible**, in addition to group sparse Bayesian inference, they can also be applied to other kinds of tasks involving diagonal-block matrices.

In contrast, Zhang's models have high time complexity. For instance, their *T-SBL* model [2], most relevant to ours, has a time complexity $O(M^2L^2N)$, or $O(N^5\alpha)$ after we make their notations consistent with ours. As they discussed in [2], by using some approximation techniques such as the speed-up scheme proposed by [4], the time of *T-SBL* model can be further reduced to $O(MN^2)$, e.g. the *T-MSBL* model, being approximately on the same order as ours.

2. **Different likelihood function.**
We model continuous data and discrete data by Gaussian likelihood and Bernoulli likelihood, respectively (see Eqs. 9 and 11 in this submission). In contrast, Zhang's work only model continuous data via Gaussian distributions, since the signals they processed for recovery are continuous [1][2][3].

   To the best of our knowledge, this submission is the first work that proposes a group sparse Bayesian model for handling discrete data. In this submission, modelling discrete data is motivated by the fact that many diffusion processes can be naturally represented as discrete signals, e.g., whether a piece of news is posted on a blog in the information spread setting.

3. **Different posterior distribution.**
As the adopted priors and likelihoods are different, the posterior distributions in this submission and Zhang's work are accordingly different, i.e., Theorem 1 and Theorem 2 in Section 3.2.3 of this submission, as shown in Table 4.

Table 4: Summary of difference in terms of posterior distribution

	This submission		Zhang's work
Data type	Continuous data	Discrete data	Continuous data
Posterior	Gaussian	Approximate Gaussian	Gaussian
Covariance	$(\mathbf{\Sigma}_0^{-1} + \lambda^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}$	$(\mathbf{\Sigma}_0^{-1} + 2\mathbf{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\boldsymbol{\xi})\mathbf{\Phi})^{-1}$	$(\mathbf{\Sigma}_0^{-1} + \lambda^{-1}\mathbf{D}^{\mathrm{T}}\mathbf{D})^{-1}$
Expectation	$\lambda^{-1}\mathbf{\Sigma}_{\boldsymbol{s}}\mathbf{\Phi}^{\mathrm{T}}\boldsymbol{y}$	$2^{-1}\mathbf{\Sigma}_{\boldsymbol{s}}\mathbf{\Phi}^{\mathrm{T}}(2\boldsymbol{y} - \mathbf{1})$	$\lambda^{-1}\mathbf{\Sigma}_{\boldsymbol{x}}\mathbf{D}^{\mathrm{T}}\boldsymbol{y}$

For discrete data, the posterior in this submission is an approximate Gaussian, which is obviously different from the posterior in Zhang's studies, a Gaussian. And $\boldsymbol{\xi}$ in approximate Gaussian is a learnable variational parameter vector.

As for continuous data, although the posteriors in this submission and in Zhang's work are Gaussian, however, it should be noted that the two posteriors are intrinsically different, due to the two different

prior covariance $\mathbf{\Sigma}_0$, as shown in Table 1. The $\mathbf{\Sigma}_0$ in this submission simplifies the posterior covariance into a diagonal-block matrix, which enables us significantly reduce the time complexity of Bayesian inference.

4. **Different Bayesian inference method.**
We employ the expectation–maximization (EM) method and the variational EM method to infer the posterior distribution and hyper-parameters in the cases of continuous data and discrete data, respectively.

   In contrast, Zhang's work only uses the EM method. Moreover, we introduce a Gaussian lower bound for the Bernoulli likelihood to address the challenge that original Q function in EM cannot be analytically solved in the case of discrete data (see Theorem 2, and its proof in this submission).

Action:
We have revised this submission by adding a new subsection, "A connection to block sparse Bayesian learning", to incorporate the above response. Please refer to Section 5.1 on page 13 of this submission for details.

Comment 2

"In the formulated model, the authors employ hierarchical prior modelling to induce sparsity, which is a standard approach in group sparse Bayesian modelling. I am confused on what is the actual novelty of the paper from modelling perspective. The author should explicitly elaborate their novelty of the proposed model in the manuscript."

Response:
Thanks very much for the question.

Hierarchical prior modelling to induce sparsity can be considered as a standard paradigm, which has been widely used in sparse Bayesian modelling. In this paradigm, the novelty comes from a new way of elaborating suitable priors and likelihoods, according to different domain problems.

In the literature of group sparse Bayesian learning, the existing work, such as Zhang's studies [1][2][3], adopt the hierarchical prior to model group sparsity as well. Compared with these methods, here we design different priors and likelihoods, and thus derive different posteriors. In addition to solving problems in different fields, the additional benefit is that the time complexity of posterior inference is greatly reduced.

Moreover, the existing hierarchical prior modeling for group sparsity only focuses on continuous data, and does not provide models to deal with discrete data, which is also common in the field of signal processing, especially for active surveillance. For this reason, we propose a new prior and a likelihood suitable for dealing with discrete data, as well as the corresponding posterior inference algorithm. The related discussions have been provided in our responses to Comments 1, 2 and 3 above. Here, what we would like to emphasize is that, in addition to the contribution of modelling design, in the submission, we also provide a new perspective to understand the theoretical properties of the designed Bayesian models.

More specifically, we uncover two important properties of the designed priors to enforce group sparsity by analyzing their sparsity effect on parameters. To obtain the "true" shape of the designed conditional sparse prior, we introduce a non-information auxiliary hyper-prior, an inverse gamma distribution, and thereafter integrate out hyper-parameters from joint prior, thereby obtaining an equivalent marginal prior distribution over parameters (please refer to Section 3.2.1 of this submission for details). We find that the marginal prior is a spike-and-slab distribution and that it has two nice properties to enforce group sparsity: (1) it is a concave and monotonic function, which can effectively restrict the probability on trivial groups; (2) it is a heavy-tailed distribution, which allows the important groups with large weights (see the prior distribution in Fig. 2 of this submission).

Comment 3

"The author claimed a new measure is proposed as in (2). However, the measure in (2) is the same as the

updating formula for γ in (15). If I understand correctly, the measure (2) is not a proposed formula, but it is from the updating formula of γ obtained from the EM derivation of group SBL. Thus, I am confused why the formula (2) is claimed to be a new proposed measure."

Response:
Formula (2) might be confused with formula (15), i.e., the importance measure for selecting sentinels, vs. the updating rule for the hyper-parameter, since they look a bit alike in form. Please allow us to make the following further clarifications.

Formula (15) indicates how to update the hyper-parameter γ in the optimization process. When the updating converges, our algorithm will use formula (2) to calculate the importance of each component, and thereafter select sentinels according to these values. From this perspective, formula (2) can be regarded as a measure of importance.

Now, let us explain the novelty of the importance measure given by formula (2). This measure is new in two key aspects: First, in the domain of sensor placement and active surveillance, it is a new measure for evaluating and hence determining important sentinels. Second, it is derived from two newly proposed group sparse Bayesian models, as given in this submission.

In the literature, the tasks of sensor placement and active surveillance have often been modeled as a linear inverse problem, a Gaussian process interpolation, or a matrix completion problem (see details in Section 2 of this submission). **To the best of our knowledge, it is the first time to model a sensor placement task as a group sparse Bayesian learning problem.** In this submission, we show that the data-dependent hyper-parameter $\boldsymbol{\gamma}$ (i.e., formula (2)) is capable of indicating the importance of each component for predicting the behavior of the whole system, and thus, it can serve as an effective and easy to calculate measure for selecting as a fewer sentinels as possible to be actively monitored. Our comprehensive theoretical analysis and empirical validations provide the rigorous support for this finding.

In this submission, we propose two new group sparse Bayesian models, and derive formula (2) by optimizing their hyper-parameters. More specifically, the two models are formulated for handling continuous data and discrete data, respectively (please refer to our responses to Comments 1 and 2 above), and the EM method and the variational EM method are employed to optimize their respective hyper-parameters. **Interestingly, through theoretical derivations, we find that, whether for the continuous model or the discrete model, their learning rule for hyper-parameter γ has exactly the same form (i.e., formula (15) and (2)), although the two models adopt entire different likelihood function.** This unified form of hyper-parameter is also a new finding in the study of group sparse Bayesian learning.

Action:
We have further improved our submission based on the above response. Please refer to the explanation for formula (2) in Section 3.2 on page 3 of this submission for details.

Comment 4

"Is the number of sentinel components known in practise? I would also assume that number of sentinel components are much less than the number of components of interest N? The authors should add description of these in the paper to justify the use of sparsity in this context."

Response:
Yes. The number of sentinel components k is known in practise, which actually depends on the amount of budget. A trade-off between prediction accuracy and budget is practically necessary: the more sentinels are selected, the more predictive accuracy is expected, while more cost is needed.

No. We do not assume that number of sentinel components k is much less than the number of components of interest N. Our proposed algorithm is much flexible, which can output an appropriate sentinel network according to the pre-given number of sentinel components k. With the discovered sentinel network, one can further test whether the specified k is enough in term of prediction accuracy by conducting validations on test set, as what we did in the experiment section of this submission.

Empirically, our experiments indeed show that a small number of sentinels k always can achieve pretty good predictions for all N components of interest. It's because that a few hub components actually dominate those dynamical systems. For instance, only 7 sentinel towns are selected for malaria surveillance in Tengchong city of 18 towns. In contract, in a dynamical system where every components are almost independent, the number of sentinels k should be large to achieve a good enough prediction.

Action:
We have further improved our submission based on the above response. Please refer to the description in Section 3.2.5 on page 7 of the new submission for details.

Comment 5

"For the computational complexity comparison, I am supervised that group LASSO is the slowest as there are various fast implementations available. The author should specify how the group LASSO was implemented in the comparison to avoid confusion."

Response:
We implement a standard group LASSO algorithm for linear/logistic regression [5] based on MATLAB by ourself. It's well known that the objective of group LASSO regression has two terms to optimize, reconstruction error term and regularization term. In this submission, the reconstruction error term brings the most of time cost of group LASSO regression, rather than the regularization term. Thus, fast implementations of group LASSO may has limited help to reduce the time cost.

In the computational complexity comparison, the group LASSO regression is the slowest because it's based on a linear/logistic regression which involves in calculating multiplication of feature matrix $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$. Considering the large size of $\mathbf{\Phi} \in \mathcal{R}^{TN \times N^2}$, the time complexity of one iteration of group LASSO will be $O(N^5 \times \max(T, N))$, where N and T denote the number of component of interest and the length of historical dynamics, respectively. In contrast, the time complexity of our proposed algorithm with efficient matrix operations is $O(N^2 \times \max(T, N))$ (see Section 3.4 on page 8 of this submission). If one equips the standard group LASSO with our proposed efficient matrix operations, the time complexity of group LASSO will decrease to $O(N^2 \times \max(T, N))$, the same order as our algorithm.

Besides, we have released the code of our experiments to public on GitHup. Please access the code via link: https://github.com/hp17illinois/Active-Surveillance-via-Group-Sparse-Bayesian-Learning.

[1] Z. Zhang and B. D. Rao. "Extension of SBL Algorithms for the Recovery of Block Sparse Signals With Intra-Block Correlation." IEEE Transactions on Signal Processing, 2013, 61(8): 2009-2015.
[2] Z. Zhang and B. D. Rao. "Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning." IEEE Journal of Selected Topics in Signal Processing, 2011, 5(5): 912-926.
[3] Z. Zhang and B. D. Rao. "Recovery of block sparse signals using the framework of block sparse Bayesian learning." ICASSP 2012, Japan, March, 2012.
[4] D. P. Wipf and B. D. Rao. "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem." IEEE Transactions on Signal Processing, 2007, 55(7): 3704-3716.
[5] Meier L, Van De Geer S, Buhlmann P." The group lasso for logistic regression." Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2008, 70(1): 53-71.

Reviewer #3

Comment 1

"Many paragraphs of this manuscript are too similar to the AAAI-18 paper. I think the authors should rephrase and rewrite them."

Response:

Thanks so much for your detailed comments.

We have rewritten the entire sections of Abstract, Introduction, and Conclusion, which are now entirely different from those of the AAAI-18 paper.

Moreover, we have sufficiently met the requirement[3] of IEEE about the previous conference paper, i.e., "When a TPAMI submission is based on a previous conference paper, IEEE requires that the journal paper be a 'substantial revision' of the previous publication (30% is generally considered "substantial"). " A preliminary version of this work was presented as a regular paper at the 32nd AAAI Conference on Artificial Intelligence (AAAI-18). The current submission is a "more than 70% (more than 13 out of 18 pages have been newly added or updated)" substantial revision of our preliminary conference publication. The main improvements include providing in-depth state-of-the-art analysis, theoretical supporting for the proposed method, a comprehensive survey on related works, as well as systematic empirical validations and comparisons. Please kindly refer to the enclosed document, *"Major Differences between our Preliminary AAAI Paper and the New TPAMI Submission"*, for details.

Comment 2

"In the AAAI-18 paper, all the theoretical results (Theorem 1, Theorem 2, Theorem 3) are supported by the "Supporting Information". So it is unclear for me about the authors' claim on their new contribution of providing theoretical supports for the proposed method in the AAAI-18 paper. Could the authors clarify the differences between the Supporting Information in the AAAI-18 paper and the new supplementary in this manuscript?"

Response:

Thanks so much for your detailed comments. Please note that the "Support Information" was on the web server of **arXiv** as an uncopyrighted/unpublished material[4], which is not a publication by itself, nor a part of the official AAAI-18 conference paper publication. Due to the page submission limit, we could not put very detailed theoretical justifications into the conference paper, but without such proofs it could be difficult for readers to fully understand those theorems, thus we put up them as an unofficial material on **arXiv**, and just provided its URL link in the AAAI-18 conference paper.

Action:

In addition to the proofs of Theorems 1 to 4 and the derivations of hyper-parameter estimation, we also added new sections to the appendix during the major revision according to reviewers' constructive suggestions. For instance, we added more clarifications on the spike-and-slab prior shown in Figure 2 to explain how to plot this Figure 2, and please refer to the Section D in Appendix. Moreover, in order to make Theorem 2 more readable, we added more details about the proposed matrix operations for algorithm acceleration, and please refer to the Section C.3 in Appendix.

Moreover, we have double-checked, and guarantee to strictly follow, the IEEE policy[5] about the preprint manuscripts on **arXiv** as follows.

[3] Information for authors in https://ieeexplore.ieee.org/xpl/aboutJournal.jsp?punumber=34

[4] https://arxiv.org/abs/1712.00328

[5] https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/author_faq.pdf

- Can an author post his manuscript on a preprint server such as **arXiv**?
  "Yes. The IEEE recognizes that many authors share their unpublished manuscripts on public sites. Once manuscripts have been accepted for publication by IEEE, an author is required to post an IEEE copyright notice on his preprint. Upon publication, the author must replace the preprint with either 1) the full citation to the IEEE work with Digital Object Identifiers (DOI) or a link to the paper's abstract in IEEE Xplore, or 2) the accepted version only (not the IEEEpublished version), including the IEEE copyright notice and full citation, with a link to the final, published paper in IEEE Xplore. "

Comment 3

"Some key references are missing in this manuscript, e.g., for the expectation maximization (EM) algorithm; variational EM; ARD."

Response:

We have further improved our submission by adding the key references. Specifically, the references are expectation maximization (EM) algorithm [1], variational EM algorithm [2], automatic relevance determination (ARD) mechanism[3][4], respectively (please see references on page 17 of this document).

Comment 4

"Further clarifications are needed, in particular:
4.1. Definition of the variable x in Fig. 1.
4.2. Definition of the "importance" of a sentinel for prior and posterior.
4.3. Definitions of μ_s^i, Σ_s^i in (2). As I understood, the authors assumed that γ_i is a Gaussian with the mean and covariance matrix given above.
4.4. An explicit estimate for "a small approximation error" in (19)–I personally think that the paper should be self-contained here.
4.5. A more detailed explanation about the "concave and monotonic" and "heavy-tailed" properties of the prior in (7)."

Response:

Comment 4.1 We have added the definition of the variable x, i.e., x_t^i denotes the state of component i at time t. Besides, we further clarified the linear equation system in the caption of Fig. 1. Please refer to Figure 1 on page 4 of this submission for details.

Comment 4.2 In this submission we define the importance of a component as its capacity for predicting the future dynamics of the whole system (see the Basic idea in Section 3.2 on page 3 of this submission). We discover the importance can be reflected by the value of hyper-parameter $\boldsymbol{\gamma}$ (see Eq. 2 on page 4 of this submission). And then we analyze and explain how the $\boldsymbol{\gamma}$ value measures the importance from prior perspective in Section 3.2.1 of this submission, and from posterior perspective in Section 3.2.3 of this submission.

From prior perspective, the $\boldsymbol{\gamma}$ value is able to measure the importance because it's a hyper-parameter with automatic relevance determination (ARD) mechanism. When $\boldsymbol{\gamma}_i$ is small, the group i in $\boldsymbol{s}$ (the vector representation of sentinel network) is sparse and vice versa; when the group i is sparse, the links sent from component i are very weak in $\mathbf{S}$ (sentinel network). That is, component i is unimportant and can be pruned out without losing much in prediction accuracy.

From posterior perspective, the $\boldsymbol{\gamma}$ value is able to measure the importance because of the structure of learning rule for $\boldsymbol{\gamma}$. There are two meaningful terms that contribute to the $\boldsymbol{\gamma}$ value according to its learning rule Eq. 15 in this submission. The first term is the inner product term $(\boldsymbol{\mu}_{\boldsymbol{s}}^i)^{\mathrm{T}}\boldsymbol{\mu}_{\boldsymbol{s}}^i$, which denotes the sum of the squares of the weight means of the overall links emitted from node i in the sentinel network. In other words, it characterizes the *influence strength* of component i for the dynamics prediction. The second term is the trace term $\mathrm{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{s}}^i]$, which denotes the variance of the posterior estimation on the links emitted from node i. That is to say, this term features the *influence uncertainty* of component i. In other words, a large $\boldsymbol{\gamma}$ value

corresponds to a node that has many links with large or uncertain influences on other nodes; a small γ value corresponds to a node that assuredly has only trivial influences on other nodes.

Comment 4.3 We have added the definitions of $\boldsymbol{\mu}_s^i$ and $\boldsymbol{\Sigma}_s^i$ in Eq. 2, i.e., $\boldsymbol{\mu}_{\boldsymbol{s}}^i$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}}^i$ denote the mean and variance of posterior Gaussian distribution with respect to group i of the sentinel network ($\boldsymbol{s}$ is the vector representation of sentinel network), which are both inferred from dynamics data $\mathbf{D}$. Please refer to Section 3.2 on page 3 of this submission for details.

Here, we assume the weights of edges in sentinel network obey a zero-mean multivariate Gaussian prior distribution, and γ_i is the hyper-parameter of the prior distribution (see Eq. 3 on page of 4 this submission).

Comment 4.4 We have presented the explicit upper bound for "a small approximation error" in Eq. 19, i.e., the supremum of approximation error in logistical approximation is 0.02. Please refer to Section 3.3 on page 8 of this submission for details.

Comment 4.5 We have added explanations about the "concave and monotonic" and "heavy-tailed" properties of the prior in Eq. 7. The prior is concave and monotonic because the prior is proportional to a power function which is concave and monotonic. For the same reason, the prior is a power-law distribution and it's well known that power-law distribution is heavy-tailed. Please refer to Section 3.2.1 on page 4 of this submission for details.

[1] Moon, Todd K. "The expectation-maximization algorithm." IEEE Signal processing magazine, 1996, 13(6): 47-60.
[2] Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.
[3] MacKay DJ. "Bayesian interpolation." Neural computation, 1992, 4(3):415-47.
[4] Tipping ME. "Sparse Bayesian learning and the relevance vector machine." Journal of machine learning research, 2001, 1(Jun):211-44.

Active Surveillance via Group Sparse Bayesian Learning

Hongbin Pei, Bo Yang, Jiming Liu, *Fellow, IEEE*, and Kevin Chen-Chuan Chang

Abstract—The key to the effective control of a diffusion system lies in how accurately we could predict its unfolding dynamics based on the observation of its current state. However, in the real-world applications, it is often infeasible to conduct a timely and yet comprehensive observation due to resource constraints. In view of such a practical challenge, the goal of this work is to develop a novel computational method for performing active observations, termed active surveillance, with limited resources. Specifically, we aim to predict the dynamics of a large spatio-temporal diffusion system based on the observations of some of its components. Towards this end, we introduce a novel measure, the γ value, that enables us to identify the key components by means of modeling a sentinel network with a row sparsity structure. Having obtained a theoretical understanding of the γ value, we design a backward-selection Sentinel Network Mining Algorithm (SNMA) for deriving the sentinel network via *group sparse Bayesian learning*. In order to be practically useful, we further address the issue of scalability in the computation of SNMA, and moreover, extend SNMA to the case of a non-linear dynamical system that could involve complex diffusion mechanisms. We show the effectiveness of SNMA by validating it using both synthetic datasets and five real-world datasets. The experimental results are appealing, which demonstrate that SNMA readily outperforms the state-of-the-art methods.

Index Terms—Epidemic dynamics, diffusion, sensor deployment, dynamical systems, automatic relevance determination.

1 Introduction

Diffusion systems are ubiquitous in the real world. A good understanding of the dynamics of such systems, such as how an infectious disease spread over time in different locations, would be essential in finding effective ways to cope with (e.g., control) the underlying diffusion processes. Here, if we adopt the notion of a dynamical system [1], the task of epidemic dynamics prediction can be stated as that of estimating the future states of epidemic dynamics based on the observation of its current states. In the case of infectious diseases surveillance, this observation entails timely monitoring and reporting of the current infection cases in all the locations of a geographical area in question. Practically speaking, such a comprehensive all-around surveillance would be rather unrealistic, if not impossible, simply due to the limitation of resources (e.g., public health personnel and equipment). In other words, it would be infeasible to observe the states of all the components of such a dynamical system given the resource constraints. This challenge would become even more acute if we are to deal with diffusion phenomena (e.g., outbreaks) across a very large spatio-temporal range, e.g., infectious disease spreading in a country [2], as well as air contaminant diffusion in a large city [3] or hot topics/meme forwarding on social media [4].

Consequently, due to a lack of systematic deployment of often limited surveillance resources, disease surveillance tends to suffer from low reporting rates, biased sampling, and lengthy reporting time-lags [5]. Here is one of the typical examples. Tengchong city is a malaria endemic region in China. It has 18 towns, which consist of 221 villages, 167,964 households, and 658,207 residents, spatially distributed over a mountainous area of 5,845 square kilometers. During the time period from 2005 to 2011, Tengchong had 7,835 confirmed malaria cases reported, while the Tengchong Centers for Disease Control (CDC) had only a handful of disease surveillance staff, making it impossible to conduct the time-consuming case surveys.

In response to such a practical need for comprehensive surveillance with limited resources, in this work, we introduce the notion of active surveillance that enables the prediction of epidemic dynamics by proactively deploying and surveying only a small number of selected locations, referred to as *sentinels*. The observation of the states from such sentinels is expected to help achieve reasonable predictions about the future states of all the locations (not only the observed ones), resulting to a good trade-off between prediction accuracy and surveillance cost [6]. Using the dynamical systems terminology, the goal of active surveillance can be stated as that of computing and hence identifying which components of the system are most useful for dynamics prediction. In order to achieve this goal, we need to overcome three fundamental challenges. First of all, the diffusion phenomena of a dynamical system come into play due to the underlying interactions of its components, which may be characterized as **hidden interaction structure**. In the case of disease diffusion, the social contact network plays such a role as the hidden interaction structure [7]. Second, what makes the matter even more intriguing is that the fact that diffusion phenomena often exhibit non-linear complex behaviors, resulting from the underlying **complex diffusion mechanism** of the system. Such behaviors are difficult to predict. Third, there is a further challenge of **divers dynamics data-type**. Active surveillance involves tasks of both regression and classification.

- *H. Pei and B. Yang are with the College of Computer Science and Technology, Jilin University, China; Department of Computer Science, University of Illinois at Urbana-Champaign, USA; Key Laboratory of Symbolic Computation and Knowledge Engineer (Jilin University), Ministry of Education, China. E-mail: peihb15@mails.jlu.edu.cn, ybo@jlu.edu.cn.*
- *J. Liu is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. E-mail: jiming@comp.hkbu.edu.hk.*
- *K. Chang is with Department of Computer Science, University of Illinois at Urbana-Champaign, USA. Email: kcchang@illinois.edu.*

Corresponding author: Bo Yang.
Manuscript received XX XX, XXXX; revised XX XX, XXXX

As a result, both real and categorical variables will be present. For instance, a real number may denote temperature measurements for heat diffusion, while a Boolean value may indicate whether a piece of news is posted in a blog in information spread case. Over the years, various related methods (e.g., sensor deployment [8]) have been proposed and could be useful, yet the above-mentioned three challenges remain open.

This work is aimed to address these challenges by introducing a novel notion of importance measure, the γ value, that enables us to decide to what extent a component should be chosen that would contribute to the task of epidemic dynamics prediction. We provide a theoretical analysis of the γ value from both prior and posterior perspectives. Based on this notion, we then design a backward-selection Sentinel Network Mining Algorithm (SNMA) for uncovering a sentinel network, a row sparse network that contains only the influential links emitted from the sentinels. With the discovered sentinel network, we will be able to predict the future states of the overall system dynamics based on the observation of the states of only those sentinels.

The present work represents a major departure from the existing work in several ways. First, we model the task of active surveillance as that of row sparse sentinel network mining for the first time, and accordingly develop an effective method of group sparse Bayesian learning for this task. The developed method incorporates both the expectation-maximization (EM) method and the variational approximation technique. Next, we examine the task of sentinel network discovery in two most widely used dynamical systems; namely, linear continuous systems and logistical discrete system. In doing so, we address the issue of scalability by utilizing more efficient multiplication and inverse operations on diagonal-block metrics. More importantly, we extend SNMA to non-linear dynamical systems that could allow for handling complex diffusion mechanisms by means of employing the technique of basic function embedding. To evaluate the effectivensss of the SNMA method, we systematically perform validations and comparisons using both synthetic and five real-world diffusion datasets, including disease spread (2009 Hong Kong H1N1 flu and 2005-2009 Tengchong malaria), as well as heat diffusion (temperature measurements in Intel Berkeley Lab and thermal diffusion on UltraSPARC T1 microprocessor) and information spread (hot words diffusion in Baidu Tieba). The experimental results demonstrate that the proposed method outperforms the existing methods in terms of a novel criterion, *cumulative mean square error*.

Our work substantially extends the previous preliminary work on the above idea [6] by providing in-depth state-of-the-art analysis, theoretical supporting for the proposed method, a comprehensive survey on related works, as well as systematic empirical validations and comparisons. The major contributions of the present work can be summarized as follows:

1) We formulate the active surveillance task as a row sparse sentinel network mining problem for the first time.
2) We propose and theoretically analyze a novel measure, the γ value, that can be used to identify sentinel components by means of discovering a sentinel network with a row sparsity structure.
3) Drawing on the γ value, we propose the SNMA algorithm for active surveillance that employs group sparse Bayesian learning to uncover a sentinel network.
4) We solve the scalability problem to a certain degree, and moreover, extend SNMA to the case of non-linear dynamical systems.
5) We provide a novel evaluation criterion for solving the active surveillance problem, and validate SNMA by performing extensive comparisons with the existing state-of-the-art methods on synthetic and five real-world datasets.

2 Related work

In this paper, we study active surveillance on epidemic dynamics, which is similar to traditional sensor deployment problem [8]. The foundation of sensor deployment methods lies in an interpolation prediction model, which is used to predict the unobserved locations based on the observation of certain sentinels. Various selection methods have been proposed in the past that aim to identify the sentinels capable of achieving accurate predictions. These methods generally fall into the following categories: greedy strategy [8], [9], [10], [11], [12], [13], [14], [15], heuristics method [16], [17], [18], and convex optimization [19], [20]. In what follows, we will focus on the interpolation prediction model, and provide a comprehensive survey of the related work with respect to: linear inverse problem [11], [12], [13], [17], [19], [20], [21], [22], [23], Gaussian process interpolation [8], [9], [10], [18], and matrix complement technique [14], [15].

2.1 Linear inverse problem

Linear inverse problem (LIP) originates from classical experimental design [24]. It defines a linear mapping from a measured physical field to a low-dimensional representation through a pre-given sensing matrix $\mathcal{T}$. Given a set of sentinels, LIP-based methods first estimate the low-dimensional representation from sentinels' observation, and then predict the unobserved physical field by solving the linear mapping inversely. In this linear framework, the sentinels are identified by selecting the rows of matrix $\mathcal{T}$ that can achieve the minimum prediction error $\mathrm{Tr}(\mathcal{T}^{\mathrm{T}}\mathcal{T})^{-1}$. Towards this end, many criteria about $\mathcal{T}$ have been proposed to guide the selection, such as condition number [11], frame potential [12], [22], minimum eigenspace [13], and multi-criteria [23]. FrameSense is able to provide a near-optimal solution with respect to the mean square error (MSE) under strict spectrum stability conditions [12]. By relaxation, this combination optimization problem can also be solved via convex optimization approximately [19], [20]. Group-greedy method integrates three criteria (mean squared error, the volume of the confidence ellipsoid, and the worst-case error variance) and optimizes them simultaneously, thereby increasing the searching space and improving model performance [23]. Besides, Kalman filtering has been used in information fusion to reduce inaccuracies from observation and prediction model [11].

Note that the performance of LIP-based methods relies heavily on the quality of a pre-given sensing matrix, whose column vectors compose a basis of the physical field. Such a proper sensing matrix is in practice extremely difficult to construct, especially when the interaction structure among components cannot be observed directly, such as the social contact network that accounts for infectious disease transmission among people [7]. Some existing methods construct the sensing matrix based on historical dynamics data via principal component analysis (PCA) technique [12], [25]. However, our experiments show that the constructed sensing matrix introduces a serious ill-conditioned problem, especially when the noise level is high.

2.2 Gaussian process interpolation

In spatial statistics, Gaussian process (GP) has been used as an effective non-parametric spatial interpolation method under the spatial correlation assumption. With a GP model, many classical information theoretic criteria can be employed to characterize the monitoring importance of locations, mainly including entropy [9], cross-entropy [18], and mutual information [8], [10]. According to the criteria, those methods perform a greedy selection of the sensors. Some work has claimed that an ϵ near-optimal sensor set can be found due to the submodularity of the adopted criterion [8], [10]. Here it is important to point out that such a near-optimal solution is about the criterion (e.g., entropy) rather than the prediction accuracy for the unobserved, which is the real goal of our active surveillance task. To handle complex dynamics, Krause and Guestrin have also studied kernel functions and non-stationary GP for sensor deployment [10].

A key limitation of GP-based methods is that it is difficult to incorporate the prior knowledge about dynamics (e.g., susceptible-infectious-recovered (SIR) model for disease spread [26] or independent cascading (IC) model for information spread [27]) because GP is model-free. Although one can use kernel functions to incorporate certain prior knowledge in theory, it is usually hard to design an equivalent kernel for a known physical model. If such available prior knowledge could be adequately incorporated, the performance of learning and prediction would be significantly improved, especially when the diffusion mechanism is very complex, as shown in our experiments.

2.3 Matrix completion method

Recently, a new active sensing strategy has been introduced that utilizes the matrix completion method, which could recover the whole matrix through active queries on only a few entries. The method has been used to build an effective urban traffic monitoring system that is composed of a small number of sentinel vehicles [14], [15]. However, this method could suffer two problems if we apply it to the active surveillance tasks; namely, 1) it can only recover historical data but not forecast future data, due to the limitation of matrix completion; 2) the sentinels identified are not stationary and have to switch from one entry to another, which is suitable for sentinel vehicles to monitor city traffic. While, in many applications the construction of sentinels is very expensive and thus these sentinels will be fixed there and running for a long time, e.g., sentinel hospitals and air quality monitor stations.

Although many related studies have been conducted, to the best of our knowledge, most of the existing methods focus on continuous dynamics data. For a dynamical system whose state is characterized by discrete data (e.g., information spread case in our experiment), those methods cannot be used. As compared with the existing work, our proposed method SNMA has the following advantages:

1) SNMA is parameter-free and does not need a pre-given sensing matrix.
2) SNMA can readily incorporate various prior knowledge.
3) SNMA can deal with discrete epidemic dynamics data.
4) SNMA outperforms the state-of-the-art methods.

3 Active Surveillance Framework

We first propose the framework of active surveillance on epidemic dynamics. It consists of three main steps.

S1: collect historical epidemic dynamics data in N components of interest.

S2: mine the sentinel network from the data. In the network, the number of sentinel components k is according to a budget.

S3: with the sentinel network, predict future epidemic dynamics of the overall N components based on the data collected from the k sentinel components.

The second and the last steps constitute the foundation of the framework, and we will elaborate them in the next section.

3.1 Problem formulation

Consider a diffusion among N components in a stable dynamical system. Let matrix $\mathbf{D} \in \mathbb{R}^{T\times N}$=$[\mathbf{D}_1, \cdots, \mathbf{D}_T]^{\mathrm{T}}$ be the epidemic dynamics during a time window $[1, T]$. Specifically for a t in $[1, T]$, row vector $\mathbf{D}_t$=$[\mathbf{D}_{t,1}, \cdots, \mathbf{D}_{t,N}]$, where each entry $\mathbf{D}_{t,i}$ denotes the state of component i at the time t. The entry $\mathbf{D}_{t,i}$ may be a real number (e.g., temperature measurement in the location i) or a Boolean value (e.g., whether a piece of news is posted in the blog i). Let $\mathbf{D}^{\boldsymbol{s}} \in \mathbb{R}^{T\times N}$denote the surveillance data collected by k sentinel components. Specifically, $\mathbf{D}^{\boldsymbol{s}}_{t,i}$ is equal to $\mathbf{D}_{t,i}$ when component i is a sentinel, and empty otherwise.

Let $f(\mathbf{D}^{\boldsymbol{s}}_t; \mathbf{S})$ be a dynamical system function achieving the epidemic dynamics prediction. Let matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$ denote a sentinel network, which is a set of key parameters to be estimated in the dynamical system function. The sentinel network $\mathbf{S}$ depicts the influential relationship from the sentinels to all components, where each link $\mathbf{S}_{i,j}$ encodes the effect of sentinel i on component j by its weight. Now, the active surveillance can be formulated as to predict the future components' states $\mathbf{D}_{t+1}$ based on the surveillance data $\mathbf{D}^{\boldsymbol{s}}_t$ and the sentinel network $\mathbf{S}$:

$$\mathbf{D}_{t+1} \approx \hat{\mathbf{D}}_{t+1} = f(\mathbf{D}^{\boldsymbol{s}}_t; \mathbf{S}). \quad (1)$$

From Eq. 1, two computational issues need to be addressed for the goal of active surveillance:

I) Sentinel identification: How to identify the sentinels from all components and mine the sentinel network $\mathbf{S}$ according to a given budget from the dynamics data $\mathbf{D}$?

II) Sentinel prediction: How to predict the future dynamics $\mathbf{D}_{t+1}$ from the current surveillance data $\mathbf{D}^{\boldsymbol{s}}_t$ based on the discovered $\mathbf{S}$?

3.2 Sentinel identification

Our basic idea to identify sentinels is intuitive: *In a dynamical system, the components having little influence on others are unimportant for predicting others' states, while those exerting a heavy influence on others dominate the system dynamics and should be selected as sentinels.*

From a graph perspective, one can determine whether a component is important or not by inferring the row sparsity of the sentinel network $\mathbf{S}$, where its link $\mathbf{S}_{i,j}$ encodes the effect of component i on component j by its weight. That is, unimportant components are associated with sparse rows, in which zeros are much more than non-zeros; on the other hand, important ones are associated with non-sparse rows. The important components emit much more links than the unimportant ones in $\mathbf{S}$. Figure1 shows an illustration by taking a linear dynamical system as an example.

Based on this idea, we propose a novel index, the $\boldsymbol{\gamma}$ value, to measure components' importance in predicting the epidemic dynamics: *a component is important if it is important in both prior and posterior structures of a sentinel network*. Specifically, the $\boldsymbol{\gamma}$

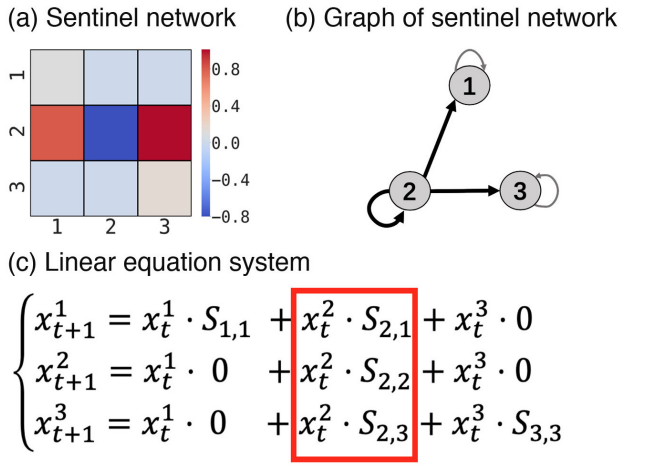


Fig. 1. Unimportant = row sparse. (a) Sentinel network $\mathbf{S}$; (b) the graph of $\mathbf{S}$; (c) the equations of a linear dynamical system, where x^i_t denotes the state of component i at time t and link S_{ij}, encodes the effect of component i on component j by its weight. The component 2 dominates the system. Unimportant component 3 is associated with a sparse row.

value is defined as the data-dependent hyper-parameter of the prior of the sentinel network, and also that reflecting the profiles of the posterior of the sentinel network:

$$\gamma_i \coloneqq ((\boldsymbol{\mu}^i_{\boldsymbol{s}})^{\mathrm{T}}\boldsymbol{\mu}^i_{\boldsymbol{s}} + \mathrm{Tr}[\boldsymbol{\Sigma}^i_{\boldsymbol{s}}])N^{-1}, \tag{2}$$

where γ_i is the $\boldsymbol{\gamma}$ value of component i. And $\boldsymbol{\mu}^i_{\boldsymbol{s}}$ and $\boldsymbol{\Sigma}^i_{\boldsymbol{s}}$ denote the mean and variance of the posterior Gaussian distribution with respect to group i of the sentinel network ($\boldsymbol{s}$ is the vector representation of sentinel network), which are both inferred from dynamics data $\mathbf{D}$.

We will elaborate on and theoretically analyze this importance index from the prior and posterior perspectives in the following, respectively. Specifically, we first give a group sparse prior over the sentinel network with key hyper-parameter $\boldsymbol{\gamma}$, and illustrate the prior can enforce group sparsity effectively in Section 3.2.1; then we model the sentinel network for two kinds of dynamical systems widely used to characterize diffusion in Section 3.2.2; finally, we present the learning rule for $\boldsymbol{\gamma}$, and show the $\boldsymbol{\gamma}$ value is able to indicate the importance of component for dynamics prediction in Section 3.2.3.

3.2.1 Prior perspective

From the basic idea, we expect to mine a sentinel network which has a row sparse structure. Thus, we adopt a zero-mean multivariate Gaussian prior for each row of sentinel network:

$$p(\mathbf{S}_{i,\cdot}|\boldsymbol{\gamma}_i) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\gamma}_i\mathbf{I}_N),\ i = 1, \ldots, N \tag{3}$$

where vector $\mathbf{S}_{i,\cdot} \in \mathbb{R}^N$ denotes the ith row of the sentinel network, and $\mathbf{I}_N \in \mathbb{R}^{N\times N}$ is an identity matrix. By this modeling, $\boldsymbol{\gamma}_i$ controls the diversity of row i from a zero vector.

For conciseness, we vectorize matrix $\mathbf{S}$, i.e., let $\boldsymbol{s} = \mathrm{vec}(\mathbf{S}^{\mathbf{T}})$, where operator $\mathrm{vec}(\cdot)$ denotes the vectorization of the input matrix by stacking its columns into a column vector. By doing so, the vector $\boldsymbol{s} \in \mathbb{R}^{N^2}$ consists of N groups of length N, where each group is associated with a row in $\mathbf{S}$. Now, the row sparse structure of $\mathbf{S}$ is equal to the group sparse structure of $\boldsymbol{s}$. In terms of the prior on $\mathbf{S}$ Eq. 3, the Gaussian sparse prior over $\boldsymbol{s}$ is

$$p(\boldsymbol{s}|\boldsymbol{\gamma}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0), \tag{4}$$

where vector $\boldsymbol{\gamma}= (\gamma_1, \cdots, \gamma_N)^{\mathrm{T}}$ and the covariance matrix $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{N^2\times N^2}$ is a diagonal matrix:

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} \gamma_1\mathbf{I}_N & & \\ & \ddots & \\ & & \gamma_N\mathbf{I}_N \end{bmatrix}. \tag{5}$$

As mentioned before, the links sent from a component reflect its effect on the system dynamics. Now, the links sent from i in $\mathbf{S}$ (i.e., the entries of group i in $\boldsymbol{s}$) are tied together and controlled by a common data-dependent hyper-parameter γ_i. This modeling is a type of automatic relevance determination (ARD) mechanism [28], [29], which can transfer the variable/model selection problem from a discrete space of variables/models to a continuous hyper-parameter space by applying parameterized sparse prior distributions. In the proposed prior Eq. 4, when γ_i is small, the group i in $\boldsymbol{s}$ is sparse and vice versa; when the group i is sparse, the links sent from i are very weak in $\mathbf{S}$. That is, component i is unimportant and can be pruned out without losing much in prediction accuracy.

To offer further insight on the prior and $\boldsymbol{\gamma}$, we try to uncover the true effect of the prior on the network $\boldsymbol{s}$. It is improper to directly analyze the prior Eq. 4 because it is a conditional prior and the data-dependent hyper-parameter $\boldsymbol{\gamma}$ is unknown. Thus, we need to integrate out $\boldsymbol{\gamma}$ to discover a 'true' prior over $\boldsymbol{s}$. To this end, we firstly introduce an auxiliary hyper-prior distribution over γ_i which is an inverse gamma (IG) distribution,

$$p(\gamma_i|a, b) = \Gamma(a)^{-1}b^a\gamma_i^{-a-1}e^{-b/\gamma_i}, \tag{6}$$

where $\Gamma(a)$ is a gamma function. By let $a \to 0$ and $b \to 0$, the gamma distribution becomes a flat uniform distribution, the so called Jeffrey's non-information prior.

By introducing this hyper-prior, we obtain a joint distribution of $\boldsymbol{s}_i$ and γ_i through $p(\boldsymbol{s}_i, \gamma_i|a, b)$=$p(\boldsymbol{s}_i|\gamma_i)p(\gamma_i|a, b)$. Then, we can calculate the 'true' marginal prior over $\boldsymbol{s}_i$ by integrate out γ_i from the joint distribution,

$$\begin{aligned} p(\boldsymbol{s}_i|a, b) &= \int p(\boldsymbol{s}_i, \gamma_i|a, b)d\gamma_i \\ &= (\frac{\pi}{2})^{\frac{N}{2}}\frac{\Gamma(a+\frac{N}{2})}{\Gamma(a)(2b)^{-a}}(2b + ||\boldsymbol{s}_i||_2^2)^{-a-\frac{N}{2}}. \end{aligned}$$

As we let $a \to 0$ and $b \to 0$, the marginal distribution becomes

$$p(\boldsymbol{s}_i) \propto (\frac{1}{||\boldsymbol{s}_i||_2})^N. \tag{7}$$

The marginal prior $p(\boldsymbol{s}_i)$, Eq. 7, has a same effect on $\boldsymbol{s}$ as the conditional prior $p(\boldsymbol{s}_i|\gamma_i)$, Eq. 4, because we set hyper-prior $p(\gamma_i|a, b)$, Eq. 6, as a uniform distribution via letting $a \to 0$ and $b \to 0$. Finally, we get a completed prior on $\boldsymbol{s}$, $p(\boldsymbol{s}) = \prod_i p(\boldsymbol{s}_i)$, since the prior of each group $p(\boldsymbol{s}_i)$ is independent, and we then visualize a profile of the 'true' prior $p(\boldsymbol{s})$ in Figure 2. In the figure, the 2-dimensional surface denotes $\log p(\boldsymbol{s})$, which is the log-distribution of a toy $\boldsymbol{s}$ contained only two groups $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$. Note that, the index of the two horizontal axes is the 2-norm of each group, i.e., $||\boldsymbol{s}_1||_2$ and $||\boldsymbol{s}_2||_2$, which indicate whether a group is sparse or not.

As shown Figure 2, prior $p(\boldsymbol{s})$ is a spike-and-slab distribution which is capable of enforcing group sparsity of $\boldsymbol{s}$ due to its two nice properties. Firstly, the most of probability mass is concentrated along two 'spines' ($||\boldsymbol{s}_1||_2 = 0$ and $||\boldsymbol{s}_2||_2 = 0$), namely, this prior hardly tolerate a $\boldsymbol{s}$ whose groups are both non-sparse. As the prior is proportional to a power function, the

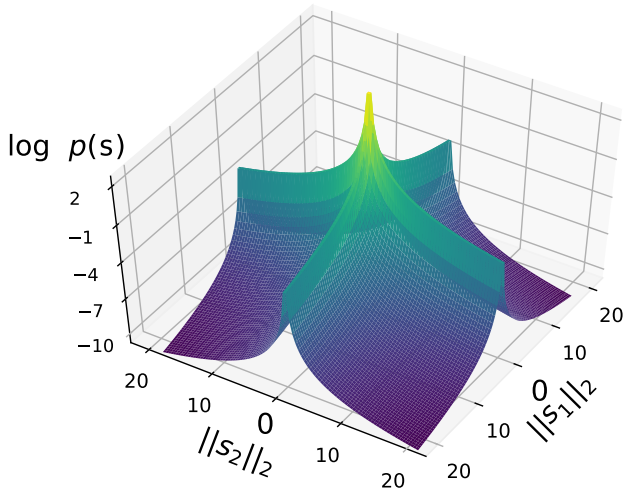


Fig. 2. The log-distributions of the 'ture' prior $\log p(\boldsymbol{s})$. Note that it is a heavy-tailed distribution and the most of probability mass is concentrated along two 'spines' ($||\boldsymbol{s}_1||_2 = 0$ and $||\boldsymbol{s}_2||_2 = 0$). The index of the two horizontal axises is the 2-norm of each group, which indicate whether a group is sparse or not.

shown is a concave and monotonic function which is well-known to encourage sparse representations [30]. Secondly, the prior is obviously a heavy-tailed prior (power-law distribution), which is generally considered to be desirable for variable selection because large weights are possible to occur, such as a group i with a large 2-norm $||\boldsymbol{s}_i||_2$ in our case. In short, this prior can prune out trivial groups by enforcing group sparse pattern, and allow of the important groups with a large 2-norm because of its heavy-tailed property.

3.2.2 Dynamical system modeling

With the sparse prior, we model sentinel network for two kinds of dynamical systems widely used to characterize diffusion phenomena in the real world: a linear continuous system and a logistical discrete system. The linear continuous system is suitable for characterizing continuous dynamics data (e.g., temperature measurements for heat diffusion), and logistical discrete system is used to model discrete dynamics data (e.g., whether a news is posted in a blog in information spread case).

Likelihood of linear continuous system. Starting from the linear continuous system, we first give the likelihood function and illustrate the pre-processing of dynamics data. The system function of a linear continuous system is

$$\mathbf{Y} = \mathbf{X}\mathbf{S} + \mathbf{V}, \tag{8}$$

where $\mathbf{Y} = \mathbf{D}_{2:T+1} \in \mathbb{R}^{T\times N}$ and $\mathbf{X} = \mathbf{D}_{1:T} \in \mathbb{R}^{T\times N}$ are both extracted from the dynamics data $\mathbf{D}$. The output $\mathbf{Y}$ is the epidemic dynamics later than the input $\mathbf{X}$ one time-unit. $\mathbf{V}$ is a Gaussian noise matrix.

For convenience, we further transform Eq. 8 into a vector form,

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{s} + \mathbf{v},$$

where vector $\boldsymbol{y} = \text{vec}(\mathbf{Y}^{\mathrm{T}}) \in \mathbb{R}^{TN\times 1}$, $\boldsymbol{s} = \text{vec}(\mathbf{S}^{\mathbf{T}}) \in \mathbb{R}^{N^2\times 1}$, and $\mathbf{v} = \text{vec}(\mathbf{V}^{\mathrm{T}}) \in \mathbb{R}^{TN\times 1}$. The matrix $\boldsymbol{\Phi} = \mathbf{X} \otimes \mathbf{I}_N \in \mathbb{R}^{TN\times N^2}$, where the operator $\otimes$ represents the Kronecker product. Now, based on the Gaussian noise assumption, the likelihood of the linear continuous system can be given as

$$p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{s}, \lambda) \sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{s}, \lambda\mathbf{I}_N), \tag{9}$$

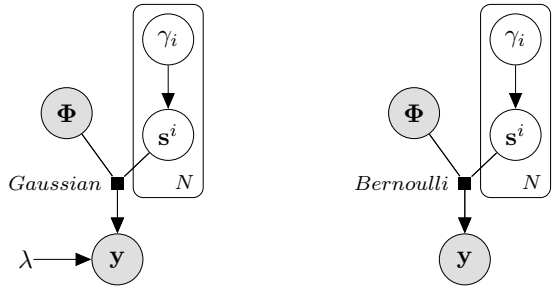
where λ denotes the noise level and $\mathbf{I}_N$ is an identity matrix. That is, we assume a same noise level for all data.

Likelihood of logistical discrete system. For the logistical discrete system, the entry in the dynamics data $\mathbf{D}_{t,i}$ is represented by a Boolean value, 0 or 1, indicating whether component i is 'infected' at time t. After the same data pre-processing as linear continuous system, we adopt a Bernoulli distribution over each entry of $\boldsymbol{y}$, i.e., $\boldsymbol{y}_n$:

$$p(\boldsymbol{y}_n = 1|\boldsymbol{\Phi}_n, \boldsymbol{s}) = \sigma(\boldsymbol{\Phi}_n\boldsymbol{s}),\ n = 1, \ldots, TN, \tag{10}$$

where $\sigma(\boldsymbol{\Phi}_n\boldsymbol{s}) = 1/(1 + e^{-\boldsymbol{\Phi}_n\boldsymbol{s}})$ denotes the sigmoid function, and $\boldsymbol{\Phi}_n$ is the nth row of $\boldsymbol{\Phi}$. Then, the likelihood of the dynamics can be written as

$$p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{s}) = \prod_{n=1}^{TN} \sigma(\boldsymbol{\Phi}_n\boldsymbol{s})^{\boldsymbol{y}_n}(1 - \sigma(\boldsymbol{\Phi}_n\boldsymbol{s}))^{1-\boldsymbol{y}_n}. \tag{11}$$



(a) Linear continuous system (b) Logistic discrete system

Fig. 3. Graphical models of two kinds of dynamical systems. $\boldsymbol{\gamma}_i$ is the $\boldsymbol{\gamma}$ value of component i. $\boldsymbol{s}^i$ denotes the ith group in $\boldsymbol{s}$.

3.2.3 Posterior perspective

The graphical models of the two kinds of dynamical systems are shown in Fig. 3. Now, based on the aforementioned prior and the likelihoods we have the following two theorems about the posterior of the sentinel network:

Theorem 1. *For the linear continuous system, the posteriors of the sentinel network is a Gaussian distribution,*

$$p(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\Phi}, \boldsymbol{\gamma}, \lambda) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{s}}, \boldsymbol{\Sigma}_{\boldsymbol{s}})$$
$$\boldsymbol{\mu}_{\boldsymbol{s}} = \lambda^{-1}\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y},\ \boldsymbol{\Sigma}_{\boldsymbol{s}}^{-1} = \boldsymbol{\Sigma}_0^{-1} + \lambda^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$$

Proof. See the Appendix A.1. □

Theorem 2. *For the logistical discrete system, the posteriors of the sentinel network is an approximate Gaussian distribution,*

$$p(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\Phi}, \boldsymbol{\gamma}, \boldsymbol{\xi}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{s}}, \boldsymbol{\Sigma}_{\boldsymbol{s}})$$
$$\boldsymbol{\mu}_{\boldsymbol{s}} = 2^{-1}\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{\Phi}^{\mathrm{T}}(2\boldsymbol{y} - \mathbf{1}),\ \boldsymbol{\Sigma}_{\boldsymbol{s}}^{-1} = \boldsymbol{\Sigma}_0^{-1} + 2\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\boldsymbol{\xi})\boldsymbol{\Phi}$$

where $\mathbf{1} \in \mathbb{R}^{TN}$ *is a vector whose entries are both one,* $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_{TN}]^{\mathrm{T}}$ *denotes variational parameters, and* $\boldsymbol{\pi}(\boldsymbol{\xi}) \in \mathbb{R}^{TN\times TN}$ *is a diagonal matrix. Specifically,* $\boldsymbol{\pi}(\boldsymbol{\xi})_{n,n} = -(\sigma(\xi_n) - 1/2)/2\xi_n$

Proof. See the Appendix A.2. □

Estimation of hyper-parameters. Theorem 1 and 2 give the posterior distributions of the sentinel network, which are a Gaussian and an approximate Gaussian. In the following, we elaborate that the $\boldsymbol{\gamma}$ value reflects the profile of the posterior of the sentinel network by giving and analyzing the learning rules of $\boldsymbol{\gamma}$.

For both the two systems, we have to iteratively estimate the hyper-parameters because they cannot be obtained in a closed

form. Thus, we use *evidence* maximization or Type-II Maximum Likelihood method to estimate the hyper-parameters [31]. For the linear continuous system, *evidence* maximization is to maximize the marginal likelihood,

$$p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{\gamma},\lambda)=\int p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s},\lambda)p(\boldsymbol{s}|\boldsymbol{\gamma})d\boldsymbol{s}. \tag{12}$$

We can get the *evidence* of the logistical discrete system by replacing λ with ξ in Eq. 12.

By treating the $\boldsymbol{s}$ as hidden variables, we employ Expectation-Maximization (EM) method [32] to maximize the *evidence*, thereby estimating the hyper-parameters. When the integral in the Q function can be analytically solved (in the linear system), EM obtains a well-formed solution to estimating the hyper-parameters; otherwise, we apply variational EM [33] to estimate the hyper-parameters by optimizing an approximate low bound of the marginal likelihood (in the logistical system). Existing work on ARD prior show that EM method can guarantee the convergence of the estimation globally except for fixed points where $\boldsymbol{\gamma}_i = 0$. Fortunately, the fixed points have no effects on our method because $\boldsymbol{\gamma}_i \to 0$ indicates the component i is trivial in our active surveillance framework, thus all component i related parameters will be pruned out.

Here we only present the final form of the learning rules for the hyper-parameters, and derivation details can be found in the Appendix B. In the linear continuous system, the learning rule for the noise parameter λ is

$$\lambda \leftarrow (TN)^{-1}(\parallel \boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\mu_s}\parallel_2^2 +\mathrm{Tr}[\boldsymbol{\Sigma_s}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}]). \tag{13}$$

In the logistical discrete system, we update the variational parameter ξ_n by

$$\xi_n \leftarrow \sqrt{\boldsymbol{\Phi}_n(\Sigma_{\boldsymbol{s}}+\boldsymbol{\mu_s}\boldsymbol{\mu_s}^{\mathrm{T}})\boldsymbol{\Phi}_n^{\mathrm{T}}},\ n=1,\ldots,TN. \tag{14}$$

It is extremely interesting that in both the linear and logistical systems the learning rule for the $\boldsymbol{\gamma}$ value is the same,

$$\boldsymbol{\gamma}_i \leftarrow ((\boldsymbol{\mu_s}^i)^{\mathrm{T}}\boldsymbol{\mu_s}^i+\mathrm{Tr}[\boldsymbol{\Sigma_s}^i])N^{-1},\ i=1,\ldots,N, \tag{15}$$

where the vector $\boldsymbol{\mu_s}^i \in \mathbb{R}^N$ and matrix $\boldsymbol{\Sigma_s}^i \in \mathbb{R}^{N\times N}$ denote the ith group of $\boldsymbol{\mu_s}$ and $\boldsymbol{\Sigma_s}$, respectively.

Intuitively, there are two meaningful terms that contribute to the $\boldsymbol{\gamma}$ value according to its learning rule Eq. 15. The first term is the inner product term $(\boldsymbol{\mu_s}^i)^{\mathrm{T}}\boldsymbol{\mu_s}^i$, which denotes the sum of the squares of the weight means of the overall links emitted from node i in the sentinel network. In other words, it characterizes the *influence strength* of component i for the dynamics prediction. The second term is the trace term $\mathrm{Tr}[\boldsymbol{\Sigma_s}^i]$, which denotes the variance of the posterior estimation on the links emitted from node i. That is to say, this term features the *influence uncertainty* of component i. In other words, a large $\boldsymbol{\gamma}$ value corresponds to a node that has many links with large or uncertain influences on other nodes; a small $\boldsymbol{\gamma}$ value corresponds to a node that assuredly has only trivial influences on other nodes.

Summarily, $\boldsymbol{\gamma}$ is a data-dependent hyper-parameter of a group sparse prior, and it integrates the profiles of both the prior and posterior of the sentinel network. The $\boldsymbol{\gamma}$ value is an index by which the importance of a component for predicting the epidemic dynamics of an entire system can be measured. For a trivial component in the system, its $\boldsymbol{\gamma}$ value will converge to zero during the Bayesian learning. For an important component, whose $\boldsymbol{\gamma}$ value larger than zero, its $\boldsymbol{\gamma}$ value could indicate its monitoring priority. The components with large $\boldsymbol{\gamma}$ values domain the entire dynamical systems, thus their state data should be collected for making the epidemic dynamics prediction.

3.2.4 Geometrical perspective

To further reveal the relationship between the $\boldsymbol{\gamma}$ value and dynamics prediction, we revisit the *evidence* maximization from a geometrical perspective. Our analysis shows that the $\boldsymbol{\gamma}$ value is capable of indicating monitoring priority of a component for dynamics prediction by giving two conclusions:

1) The $\boldsymbol{\gamma}$ value of a trivial component will converge to zero during the Bayesian learning.
2) The $\boldsymbol{\gamma}$ value of an important component could indicate its contribution for predicting dynamics.

Here we only analyze the $\boldsymbol{\gamma}$ value in the linear system, and the same conclusions still hold in the logistical system.

We first analytically integrate the marginal likelihood, Eq. 12, with respect to $\boldsymbol{s}$, the integrating details can be found in the proof of Theorem 1 (Appendix A.1). By doing so, we obtain an *evidence* of a Gaussian process regression,

$$p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{\gamma},\lambda)\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{C}),\ \boldsymbol{C}=\lambda\mathbf{I}_{TN}+\boldsymbol{\Phi}\boldsymbol{\Sigma_0}\boldsymbol{\Phi}^{\mathrm{T}}, \tag{16}$$

where $\mathbf{I}_{TN}\in\mathbb{R}^{TN\times TN}$ is an identity matrix. From a Gaussian process view, we can consider each element of output, $\boldsymbol{y}_i$, as an observation of a Gaussian process. The mean of this Gaussian process is zero, and its covariance matrix is $\boldsymbol{\Phi}\boldsymbol{\Sigma_0}\boldsymbol{\Phi}^{\mathrm{T}}$. $\lambda\mathbf{I}_{TN}$ characterize its observation noise. Now, the *evidence* maximization is to optimize $\boldsymbol{C}$ by updating the hyper-parameters, λ and $\boldsymbol{\gamma}$ (i.e., $\boldsymbol{\Sigma_0}$), so as to make the all observations $\boldsymbol{y}$ most probable.

We then decompose Eq. 16, the *evidence* of all observations, into N independent *evidences* of observations from each component j, according to the consistency property of Gaussian process [34],

$$\begin{gathered}p(\mathbf{Y}_{\cdot,j}|\mathbf{X},\boldsymbol{\gamma},\lambda)\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{c}),\ j=1,\ldots,N\\ \boldsymbol{c}=\lambda\mathbf{I}_T+\sum\nolimits_{i=1}^{N}\boldsymbol{\gamma_i}\mathbf{X}_{\cdot,i}\mathbf{X}_{\cdot,i}^{\mathrm{T}},\end{gathered} \tag{17}$$

where $\mathbf{Y}_{\cdot,j}$, the jth column of the matrix output $\mathbf{Y}$, denotes the observation from component j. $\mathbf{X}_{\cdot,i}$ is the ith column of $\mathbf{X}$, and it can be seen as a base whose outer product contribution to the covariance matrix $\boldsymbol{c}$ is modulated by the hyper-parameter $\boldsymbol{\gamma}_i$. These N *evidences* defined in Eq. 17 are equivalent to the one defined in Eq. 16 because each $\mathbf{Y}_{\cdot,j}$ is independent in term of the covariance $\boldsymbol{C}$ in Eq. 16.

It is quite interesting to note that every $\mathbf{Y}_{\cdot,j}$ obeys an identical Gaussian distribution in Eq. 17. In other words, each component j observe the same Gaussian process from a different perspective, and its observation is $\mathbf{Y}_{\cdot,j}$. After this decomposition, the *evidence* maximization becomes optimizing the common covariance matrix $\boldsymbol{c}$ to make the N observations $\mathbf{Y}_{\cdot,j}$ most probable, which is equivalent to maximize N log *evidences*,

$$\log p(\mathbf{Y}_{\cdot,j}|\mathbf{X},\boldsymbol{\gamma},\lambda)\propto-\log(\det(\boldsymbol{c}))-\mathbf{Y}_{\cdot,j}^{\mathrm{T}}\boldsymbol{c}^{-1}\mathbf{Y}_{\cdot,j},$$

where $j=1,\ldots,N$, and $\det(\cdot)$ denotes the determinant of the input matrix. Now, we can see that the *evidence* maximization is actually to optimize $\boldsymbol{c}$, thereby minimizing $\mathbf{Y}_{\cdot,j}^{\mathrm{T}}\boldsymbol{c}^{-1}\mathbf{Y}_{\cdot,j}$ term, the Mahalanobis distance between $\mathbf{Y}_{\cdot,j}$ and the zero-mean, as well as $\log(\det(\boldsymbol{c}))$ term, the normalisation term of Gaussian distribution.

We illustrate geometrically the *evidences* maximization by projecting the original observation space $\mathbb{R}^T$ into a 2D space, as shown in Figure 4. Red points denote the N observations

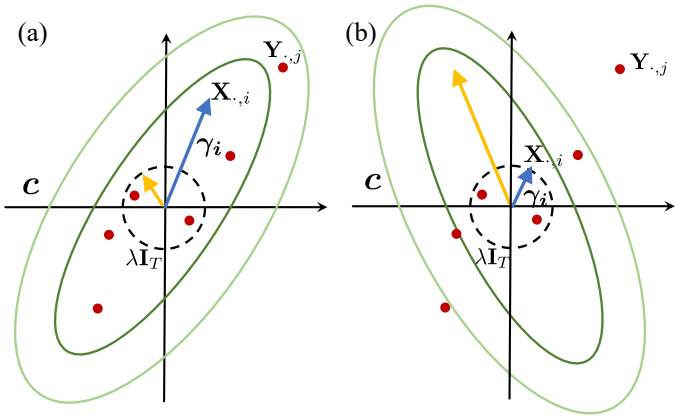


Fig. 4. A geometrical perspective of *evidences* maximization, showing in 2D projection of the T-dimension original observation space. Each red point denotes an observation $\mathbf{Y}_{\cdot,j}$, and green confidence ellipses illustrate the covariance $\boldsymbol{c}$. Although the observations are same, their Mahalanobis distances to the origin in panel (a) are much less than the distances in panel (b) because their $\boldsymbol{c}$ are different. Each arrow shows the weighted direction of a base $\mathbf{X}_{\cdot,i}$. Blue arrows denote an important base, and yellow arrows denote a trivial base. A dashed circle denotes the noise variance $\lambda\mathbf{I}_T$.

$\mathbf{Y}_{\cdot,j}$, $j = 1,\ldots,N$, and green confidence ellipses illustrate the covariance $\boldsymbol{c}$. Deeper green indicates higher confidence. All points on the same confidence ellipse have the same Mahalanobis distance to the origin (zero-mean), and the higher the confidence, the smaller the Mahalanobis distance. Thus, to minimize the Mahalanobis distances is to adjust the orientation, shape, and size of the confidence ellipses (controlled by the parameters in $\boldsymbol{c}$), so that the observations $\mathbf{Y}_{\cdot,j}$ can be distributed in high confidence ellipses. On the other hand, the normalisation term $\log(\det(\boldsymbol{c}))$ is proportional to the size of the ellipses, i.e., for a given confidence level, the larger the corresponding ellipse, the larger the $\log(\det(\boldsymbol{c}))$, and thus to minimize the normalisation term requires restricting the size of the ellipses, which prohibits arbitrarily enlarging their size to make more observations included in higher confidence ellipses. In other words, to maximize the *evidence* (Eq. 17) is to obtain an ideal covariance matrix $\boldsymbol{c}$ which can optimally tradeoff the minimization of Mahalanobis distances and normalisation term. For example, in Figure 4, for the same observations, the $\boldsymbol{c}$ in (a) is better than the $\boldsymbol{c}$ in (b) although they have the same $\log(\det(\boldsymbol{c}))$, because most observations in (a) are distributed in higher confidence ellipses, thus having much smaller Mahalanobis distances and greater *evidence*.

Given the input $\mathbf{X}_{\cdot,i}$, the covariance matrix $\boldsymbol{c}$ is determined by $\boldsymbol{\gamma}$ and λ in Eq. 17. In Figure 4, each base $\mathbf{X}_{\cdot,i}$ specifies a direction in the space, and the weight $\boldsymbol{\gamma}_i$ characterizes the scale of the base. Thus, adjusting $\boldsymbol{\gamma}_i$ can change the orientations, shapes, and sizes of the confidence ellipses. While the adjustment of the noise variance parameter λ changes the confidence ellipses equally in all directions.

From this geometrical perspective, the importance of component i for dynamics prediction implies a contribution of its base $\mathbf{X}_{\cdot,i}$ to minimizing the Mahalanobis distances under the regularization term $\log(\det(\boldsymbol{c}))$. Based on the understanding, a trivial component i is the one whose base $\mathbf{X}_{\cdot,i}$ is substantially orthogonal with most observations $\mathbf{Y}_{\cdot,j}$, or the base can be represented by a combination of other bases. Bayesian learning will optimize the $\boldsymbol{\gamma}_i$ of such component i to zero because otherwise, this will only increase the regularization term, and will hardly contribute to minimizing the Mahalanobis distances. Compared with the bases of such trivial components, the Mahalanobis distances can be better minimized by increasing the noise variance $\lambda\mathbf{I}_T$, which will enlarge the confidence ellipse in all directions without increasing the regularization item too much. For instance, in Figure 4 (b) the ellipses extend along a trivial base (yellow arrow), which is even worse than only enlarging the dash circle (noise variance $\lambda\mathbf{I}_T$) to minimize the Mahalanobis distances.

On the contrary, an important component i is the one whose base $\mathbf{X}_{\cdot,i}$ is well 'aligned' with most observations $\mathbf{Y}_{\cdot,j}$. If the confidence ellipses extend along the base of an important component, the Mahalanobis distances significantly decrease as shown in Figure 4 (a). As $\boldsymbol{\gamma}_i$ characterizes the scale of the base, it actually measures the influence of the base $\mathbf{X}_{\cdot,i}$ on changing the ellipses to make every observation $\mathbf{Y}_{\cdot,j}$ more probable. The greater $\boldsymbol{\gamma}_i$, the greater the contribution to maximizing the *evidence*. Thus, we can treat $\boldsymbol{\gamma}_i$ as the monitoring priority of component i for making accurate dynamics prediction.

3.2.5 Sentinel network mining algorithm

Based on the $\boldsymbol{\gamma}$ value, we propose a backward-selection algorithm called the SNMA (**S**entinel **N**etwork **M**ining **A**lgorithm) to infer the posterior sentinel network, as shown in Algorithm 1. SNMA in turn performs a parameter optimization step and a sentinel selection step. It starts with all N components of interest and removes one component at a time until only k components are left (k is the pre-given sentinel amount according to a budget). The component that is removed should be chosen as the one with the minimum $\boldsymbol{\gamma}$ value.

A trade-off between prediction accuracy and budget is practically necessary: the more sentinels are selected, the more predictive accuracy is expected, while more cost is needed. In dynamical systems dominated by a few hub components, a small number of sentinels can achieve pretty good predictions for dynamics of the overall system; in contract, in a dynamical system where every components are almost independent, the number of sentinels should be large to obtain a good enough prediction.

Similar to the sensor placement problem, to construct an optimal subset of components set for active surveillance is an NP-hard combination optimization problem. Fortunately, the form of backward-selection algorithm is theoretically guaranteed to pick an optimal sentinel set if we only remove the trivial components whose $\boldsymbol{\gamma}$ value is zero and the system perturbation is small enough [35].

3.3 Sentinel prediction

Once we have obtained the posterior structure of the sentinel network, the epidemic dynamics of the overall system, $\mathbf{D}$, can be predicted based on the surveillance data $\mathbf{D}^{\boldsymbol{s}}$ collected by k sentinel components. Let $\mathbf{D}^{\boldsymbol{s}}_{+}$ be a new set of surveillance data, where only the values on k sentinels' locations are kept and the rest are empty. As mentioned above, we obtain $\boldsymbol{\Phi}^{\boldsymbol{s}}_{+}$ through the data pre-processing. Then, a predictive distribution over the following system states $\boldsymbol{y}_{+}$ is given by

$$p(\boldsymbol{y}_{+}|\boldsymbol{\Phi}^{\boldsymbol{s}}_{+},\boldsymbol{y},\boldsymbol{\Phi}) = \int p(\boldsymbol{y}_{+}|\boldsymbol{\Phi}^{\boldsymbol{s}}_{+},\boldsymbol{s})p(\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi})d\boldsymbol{s}. \qquad (18)$$

Algorithm 1: **S**entinel **N**etwork **M**ining **A**lgorithm

Input: epidemic dynamics $\mathbf{D}$, number of components of interest N, number of sentinels k;
Output: posterior distribution of sentinel network, i.e., mean vector $\boldsymbol{\mu_s}$, covariance matrix $\boldsymbol{\Sigma_s}$;

```
Pre-processing: extract y and Φ from D;
Randomly initialize γ, λ (the linear) or ξ (the logistical);
L ← N;
while L > k do
    // Optimization step
    while γ is not converged do
        update μ_s and Σ_s via Theorem 1 or Theorem 2;
        update γ, λ or ξ via Eq. 15, 13 or 14;
    // Selection step
    for i = 1, ..., L do
        if γ_i is the minimum entry in the γ then
            update Φ, μ_s, Σ_s, γ through pruning out the
              ith group in them;
            L ← L − 1;
return μ_s, Σ_s
```

Linear continuous system. In this case, the integral in Eq. 18 is a Gaussian convolution (refer to the proof of Theorem 1), whose analytical solution is still a Gaussian. Then, we have

$$p(\boldsymbol{y}_+|\boldsymbol{\Phi}^{\boldsymbol{s}}_+, \boldsymbol{y}, \boldsymbol{\Phi}) \sim \mathcal{N}(\mu_{y_+}, v^2_{y_+})$$

with parameters $\mu_{y_+} = \boldsymbol{\Phi}^{\boldsymbol{s}}_+\boldsymbol{\mu_s}$, $v^2_{\boldsymbol{y}_+} = \lambda + \boldsymbol{\Phi}^{\boldsymbol{s}}_+\boldsymbol{\Sigma_s}(\boldsymbol{\Phi}^{\boldsymbol{s}}_+)^{\mathrm{T}}$.

Logistical discrete system. The predictive distribution of discrete data is a Bernoulli distribution. By substituting the variational approximation posterior given in Theorem 2 for the term of $\boldsymbol{s}$ posterior in Eq. 18, we have the following predictive distribution:

$$p(\boldsymbol{y}_{+n} = 1|\boldsymbol{\Phi}^{\boldsymbol{s}}_+, \boldsymbol{y}, \boldsymbol{\Phi}) \approx \int p(\boldsymbol{y}_{+n} = 1|\boldsymbol{\Phi}^{\boldsymbol{s}}_+, \boldsymbol{s}) q(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\Phi}) d\boldsymbol{s},$$

where $\boldsymbol{y}_{+n}$ is the nth entry of $\boldsymbol{y}_+$, and $n = 1, \ldots, N$.

However, the integral in this approximation is a convolution of a logistical function with a Gaussian distribution, which cannot be analytically integrated. By introducing an error function, it can be approximated as a re-parameterized logistical function [36]:

$$\int p(\boldsymbol{y}_+{=}1|\boldsymbol{\Phi}^{\boldsymbol{s}}_+, \boldsymbol{s}) q(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\Phi}) d\boldsymbol{s} \approx (1 + e^{-\tau \boldsymbol{\Phi}^{\boldsymbol{s}}_+ \boldsymbol{\mu_s}})^{-1}, \quad (19)$$

where $\tau = (1 + \frac{\pi}{8}\boldsymbol{\Phi}^{\boldsymbol{s}}_+\boldsymbol{\Sigma_s}(\boldsymbol{\Phi}^{\boldsymbol{s}}_+)^{\mathrm{T}})^{-\frac{1}{2}}$. A very small approximation error is guaranteed in theory, which is always less than 0.02 [37].

3.4 Scaling up

As mentioned before, diffusion phenomena often span over very large spatio-temporal ranges, thus algorithmic scalability is very important for active surveillance to be practically deployable. In the SNMA, the most expensive operations are the matrix multiplication ($\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$) and the matrix inverse ($\boldsymbol{\Sigma}_{\boldsymbol{s}}^{-1}$) for calculating the posterior parameters in Theorem 1 and 2. Considering the large size of $\boldsymbol{\Phi} \in \mathbb{R}^{TN \times N^2}$ and $\boldsymbol{\Sigma_s} \in \mathbb{R}^{N^2 \times N^2}$, the time complexity of one iteration will be $\mathcal{O}(N^5 \times \max(T, N))$, making it infeasible to handle many real-world problems. We now propose two fast matrix operations to solve the scalability problem. The time complexity after scaling up is reduced to $\mathcal{O}(N^2 \times \max(T, N))$, which is in the same level with the state-of-the-art methods.

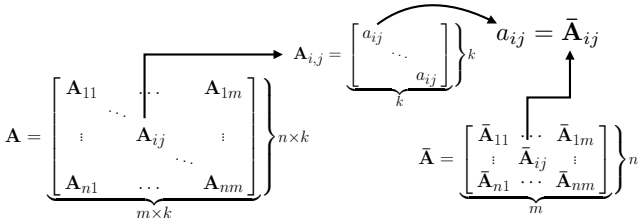


Fig. 5. Illustration of the relationship between a diagonal-block matrix and block projective matrix.

Definition 1. *(Diagonal-block matrix). Matrix $\mathbf{A} \in \mathbb{R}^{nk \times mk}$ is called a diagonal-block matrix if it consists of n by m square sub-matrices, and each sub-matrix $\mathbf{A}_{i,j} \in \mathbb{R}^{k \times k}$ is a diagonal matrix with all diagonal entries having the same value a_{ij}, i=1 : n, j=1 : m.*

Definition 2. *(Block projective matrix). Let $\mathbf{A} \in \mathbb{R}^{nk \times mk}$ be a diagonal-block matrix. Matrix $\bar{\mathbf{A}} \in \mathbb{R}^{n \times m}$ is called the block projective matrix of $\mathbf{A}$ if each entry $\bar{\mathbf{A}}_{i,j} = a_{ij}$, i=1 : n, j=1 : m.*

The structure of the two matrices and their relationship is shown in Fig. 5. The block projective matrix $\bar{\mathbf{A}}$ is much smaller than its original diagonal-block matrix $\mathbf{A}$. In the meantime, it contains the overall distinct elements in $\mathbf{A}$. Based on the definitions, we have the following theorems, which can significantly reduce the computational cost of the two kinds of operations.

Theorem 3. *Let $\mathbf{A}, \mathbf{B}$ be two diagonal-block matrices with the same size of sub-matrix. Their product $\mathbf{C} = \mathbf{AB}$ is also a diagonal-block matrix, and those block projective matrices satisfy $\bar{\mathbf{C}} = \bar{\mathbf{A}}\bar{\mathbf{B}}$.*

Proof. See the Appendix A.3. □

Theorem 4. *Let $\mathbf{A}$ be a square diagonal-block matrix. Its inverse matrix $\mathbf{A}^{-1}$ is also a diagonal-block matrix and satisfies $\overline{\mathbf{A}^{-1}} = (\bar{\mathbf{A}})^{-1}$.*

Proof. See the Appendix A.4. □

In the SNMA, $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma_s}$ are two diagonal-block matrices. When we alternatively calculate the multiplication ($\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$) and the matrix inverse ($\boldsymbol{\Sigma}_{\boldsymbol{s}}^{-1}$) on the block projective matrices $\bar{\boldsymbol{\Phi}} \in \mathbb{R}^{T \times N}$ and $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{s}} \in \mathbb{R}^{N \times N}$ via the two theorems, the time complexity of one iteration can be reduced by 3 orders of magnitude (i.e., N^3). Note that this scaling-up technique not only solves the difficulty faced by our algorithm, but can also address other tasks involving diagonal-block matrices, such as the computational obstacle of a multiple measurement vector (MMV) model in compressed sensing [38] and signal reconstruction [39]. Besides, SNMA can be readily paralleled in practice by adopting the group testing strategy, which has been used for parallel feature selection [40].

3.5 Embedding non-linear dynamical models

As mentioned before, complex diffusion mechanism is an important challenge for active surveillance. To address this challenge, we extend SNMA to non-linear dynamical systems by means of embedding basic functions, where non-linear dynamical system functions are represented as the combinations of basic functions [1]. And a basic function, $\psi_k(x), k \in [1, \cdots, m]$, is an arbitrary function, e.g., a polynomial function for polynomial dynamical systems [41].

We illustrate the technological details based on the linear continuous system Eq. 8 as an example, where $\mathbf{X}_{t,i}$ denotes the state of component i at time t. Let vector $\mathbf{\Psi}(\mathbf{X}_{t,i}) = [\psi_1(\mathbf{X}_{t,i}), \cdots, \psi_m(\mathbf{X}_{t,i})] \in \mathbb{R}^m$ be the mapping of $\mathbf{X}_{t,i}$ through m basic functions. For all components, let vector $\mathbf{\Psi}(\mathbf{X}_t) = [\mathbf{\Psi}(\mathbf{X}_{t,1}), \cdots, \mathbf{\Psi}(\mathbf{X}_{t,N})] \in \mathbb{R}^{mN}$ denote the mapping of all components at time t, and let matrix $\mathbf{\Psi}(\mathbf{X}) = [\mathbf{\Psi}(\mathbf{X}_1); \cdots; \mathbf{\Psi}(\mathbf{X}_T)]^{\mathrm{T}} \in \mathbb{R}^{T \times mN}$ denote the mapping of all components during the overall time window. Then, the system function Eq.8 can be extended as:

$$\mathbf{Y} = \mathbf{\Psi}(\mathbf{X})\dot{\mathbf{S}} + \mathbf{V}, \tag{20}$$

where the augmented sentinel network $\dot{\mathbf{S}} \in \mathbb{R}^{mN \times N}$ characterizes the interactions from the mN mappings to the N components. We adopt the sparse Gaussian prior Eq. 3 over m rows related to component i in $\dot{\mathbf{S}}$. By doing so, $\boldsymbol{\gamma}_i$ indicates not only the influence from component i, but also the influence from basic functions related to component i. Eq. 20 is able to represent non-linear dynamical systems if non-linear basic functions are adopted. For instance, we can readily construct basic functions for the susceptible-infectious-recovered (SIR) model, a well-studied and widely adopted disease spread model, from its reproduction matrix form [26].

To integrate the extended system function Eq.20 into the current algorithm SNMA, just let $\mathbf{\Phi} = \mathbf{\Psi}(\mathbf{X}) \otimes \mathbf{I}_N \in \mathbb{R}^{TN \times mN^2}$ and $\boldsymbol{s} = \mathrm{vec}(\dot{\mathbf{S}})^{\mathrm{T}} \in \mathbb{R}^{mN^2}$. Meanwhile, the group size in $\boldsymbol{s}$ needs be changed from N to mN. The rest of the steps are the same as the SNMA shown in Algorithm 1. For the logistical discrete system, the extended process is analogous.

4 VALIDATIONS

In this section, we analyze the performance of SNMA and compare it with five related methods for both sentinel identification and sentinel prediction task. We propose a novel evaluation criterion, *cumulative mean square error* (CMSE in short), for sentinel prediction. The experiments are conducted on both synthetic and five real-world datasets. We released the code implementation of our algorithm and competitor methods on GitHup[1]. The implementation is based on MATLAB.

Comparative Study. We chose three kinds of state-of-the-art method as competitors: linear inverse problem (LIP) based method, Gaussian process (GP) based method, and group sparse learning method.

LIP-based method originates from classical experimental design, and it deploys sensors by solving linear inverse problem [19]. We select three very recent methods, FrameSense [12], MNEP [13], and MPME [13], as representatives for this kind of method. Notice that LIP-based methods relay on a pre-given sensing matrix, which however, is always unknown in real-world applications. To address this challenge, existing methods construct an approximate sensing matrix from historical data via principal component analysis (PCA) technique [25].

We propose a new method to construct the sensing matrix by using Gaussian process. We model a Gaussian process regression for sensor deployment as a linear inverse problem,

$$\mathbf{X}^{\mathrm{T}} = \Sigma_{\mathbf{XX}}\Sigma_{\mathbf{YX}}^{-1}\mathbf{Y}^{\mathrm{T}} + \mathbf{V}, \tag{21}$$

1. https://github.com/hp17illinois/Active-Surveillance-via-Group-Sparse-Bayesian-Learning

where $\Sigma_{\mathbf{XX}}$ denotes the covariance of $\mathbf{X}$, and $\Sigma_{\mathbf{YX}}$ denotes the covariance between $\mathbf{Y}$ and $\mathbf{X}$ (each column as a variable and each row as a data sample). $\mathbf{V}$ is a Gaussian noise matrix. Here the sensing matrix is $\mathcal{T} = \Sigma_{\mathbf{XX}}\Sigma_{\mathbf{YX}}^{-1}$. And the linear inverse problem is to select rows of matrix $\mathcal{T}$, which can achieve the most precise prediction of $\mathbf{Y}$, in the sense of minimizing error $\mathrm{Tr}(\mathcal{T}^{\mathrm{T}}\mathcal{T})^{-1}$.

We construct six LIP-based competitors by combining the two sensing matrices (by PCA and GP) and the three LIP-based methods. For convenience, we denote FrameSense method with sensing matrix constructed by PCA as FrameSense-PCA, and the one by GP as FrameSense-GP, respectively. The naming rule is same for MNEP and MPME. By studying these combinations, we analysis the effects from sensing matrix on LIP-based methods.

The second kind of competitor is GP-based method [8] [10]. In those methods, sensor locations are optimized according to information theory criteria (e.g., entropy or mutual information) to achieve predictions on unobserved locations by leveraging GP interpolation. Here we choose GP-based mutual information (GP-MI) as the representative for this kind of method in view of the fact that it is the state-of-the-art method [8].

For LIP-based and GP-based methods, a fundamental limitation is that they cannot deal with logistical discrete system because GP and linear regression are only able to work on continuous data. To extend those methods to discrete cases, we feed their continuous outputs to a simple threshold classification, in which a continuous variable is changed to be one/zero if it is larger/less than a threshold of 0.5. Although this extension is modest, it achieves a quite good performance in experiments.

The third kind of competitor is group sparse learning method, which is used to test the performance of the proposed group sparse Bayesian learning method. We choose group lasso [42] as the representative for group sparse learning method. Although group lasso cannot address active surveillance problem directly, but it has a similar capacity as the proposed group sparse Bayesian learning method. To achieve active surveillance, we use our proposed framework and replace the group sparse Bayesian learning with group lasso. Linear and logistical regression with a group lasso regularization are adopted for the linear continuous system and the logistical discrete system respectively.

4.1 Validations on synthetic dynamics data

Synthetic Data Generation. We generate synthetic datasets by imagining diffusion processes taking place in the linear continuous system or the logistical discrete system. The main steps are summarized as follows. (1) Random generation of a ground truth $\boldsymbol{\gamma}$ value vector with 500 entries, where 200 entries are sampled from $\mathcal{N}(0, 10)$ and the other 300 from $\mathcal{N}(0, 0.1)$, i.e., 200 sentinels and 300 trivial components; (2) Random sampling of a ground truth sentinel network $\mathbf{S}$ of 500 nodes via the prior Eq.3 based on the ground truth $\boldsymbol{\gamma}$ value; (3) Based on the $\mathbf{S}$, simulation of the epidemic dynamic (length of time window T = 500) via the dynamical system functions, i.e., Eq. 8 for the continuous data, or Eq. 10 for the discrete data. (4) Simulation of the non-linear epidemic dynamic by embedding a quadratic basic function $\phi(x) = x^2$ as shown in Eq. 20. Finally, we obtain four datasets as shown in Table 1 (two continuous datasets and two discrete datasets).

The comparison experiments are conducted under three noise levels. We adopt *Signal-to-noise ratio* (SRN) and *bit error rate* (BER) [43] to measure noise levels in the linear system and the

logistical system, respectively. We specify the noise level as high (SRN=10 or BER=0.15), medium (SRN=20 or BER=0.1), and low (SRN=30 or BER=0.05).

4.1.1 Sentinel identification experiments

We first evaluate the methods with regard to the sentinel identification task, which is to identify optimal sentinels from all components. *failure rate* is adopted as a criterion to measure whether a method can identify the correct sentinels. Failure rate is defined as the percentage of wrong sentinels given by a method,

$$\text{Failure rate} = 1 - |\mathbf{\Gamma} \cap \hat{\mathbf{\Gamma}}|/|\mathbf{\Gamma}|,$$

where $\mathbf{\Gamma}$ is the set of the ground truth sentinels, and $\hat{\mathbf{\Gamma}}$ is the set given by a method. $\mathbf{\Gamma}$ and $\hat{\mathbf{\Gamma}}$ have the same size $|\mathbf{\Gamma}| = |\hat{\mathbf{\Gamma}}|$. The smaller failure rate a method obtains, the better sentinel identification it achieves.

We adopt a 5-fold cross-validation strategy in experiments on synthetic datasets. Based on the training data, we identify 200 sentinels via one method, and then calculate the failure rate by comparing with the ground truth. For each method, the average failure rate of sentinel identification is shown in Table. 1. Obviously, the proposed SNMA achieve the best performance in every setting. For LIP-based methods, the results show that the sensing matrix constructed by GP (Eq. 21) is better than the one constructed by PCA in most cases.

4.1.2 Sentinel prediction experiments

We then evaluate the methods with regard to the sentinel prediction task, which is to predict future dynamics based on surveillance data collected from k discovered sentinels. With the dynamics predicted by k sentinels, *mean square error* (MSE) is an intuitive criterion to measure prediction error. However, we find that the winning method changes with k in term of MSE as shown in Fig. 6, where GP-MI achieves lower prediction error when $k < 4$, but SNMA outperforms GP-MI when $k \geq 4$. Obviously, it is not objective to evaluate the methods at a certain k because k depends on the given surveillance budget in practice. How to measure the sentinel prediction error in a comprehensive manner? This evaluation problem is common and has not been addressed in existing sensor deployment literatures [8], [12], [13].

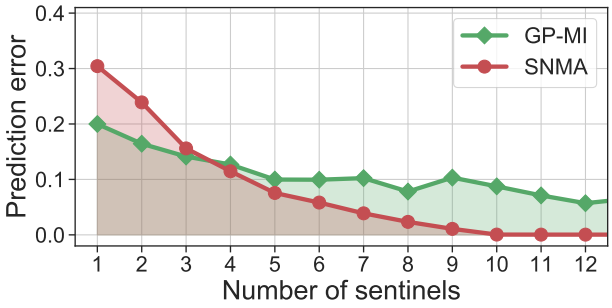


Fig. 6. An illustration of sentinel prediction error. x-axis denotes the number of sentinels k. Prediction error is measured by MSE. GP-MI achieves lower prediction error when $k < 4$, but SNMA outperforms GP-MI when $k \geq 4$

From Fig. 6, we find the area under the curve of prediction error can summary the prediction errors at all the feasible k. Based on this observation, we design a novel criterion, *cumulative mean square error* (CMSE in short), to measure the sentinel prediction error. Specifically,

$$\text{CMSE} = \sum\nolimits_{k \in \mathcal{K}} ||\mathbf{Y} - \hat{\mathbf{Y}}_k||_2^2/(TN), \qquad (22)$$

where $\mathbf{Y}$ is the real data on all N components during T time units, and $\hat{\mathbf{Y}}_k$ is the predicted dynamics based on the k discovered sentinels. $\mathcal{K}$ is a set which contains all feasible k. CMSE evaluates the methods at all feasible k, thus it is a comprehensive and objective criterion for sentinel prediction.

We also adopt a 5-fold cross-validation strategy. We firstly identify $k \in \mathcal{K}$ sentinels from the training data. Here, we specify $\mathcal{K} = [20, 40, 60, \ldots, 200]$ as the number of ground truth sentinel is 200. With the k discovered sentinels, we then make prediction by feeding the surveillance data from the sentinels to the corresponding prediction model of each method, such as Eq. 18 for the proposed SNMA. After experiments on all $k \in \mathcal{K}$, we calculate CMSE for each method. The average CMSE over 5-folds is shown in Table 2. The results show the proposed SNMA is superior to all the competitor methods once again. For LIP-based methods, the sensing matrix constructed by GP is better than the one constructed by PCA in most cases, especially when the noise level is high.

4.1.3 Running time comparisons

The time complexity of SNMA is $\mathcal{O}(N^2 \times \max(T, N))$ as we discussed in section 3.4.1. For GP-MI, FrameSense, MNEP, and MPME, their most time-consuming operation is calculating the covariance of $\mathbf{X} \in \mathbb{R}^{T \times N}$, whose time complexity is $\mathcal{O}(N^2 \times T)$. Thus, the competitors are faster when $T < N$, otherwise SNMA is faster. We further test the running time of each method on a PC with an 8-core 3.2GHz CPU and 16GB memory coded by Matlab 2018a.

Fig. 7 presents the average running time of one iteration of the five methods. Since GP-MI, MNEP, and MPME employ forward greedy strategy (fast when k is small), but SNMA and FrameSense use backward selection method (fast when k is large), it is fair to evaluate them by comparing the running time of one iteration of each algorithm. We can see that group lasso is the slowest, and the other methods are almost at the same level.

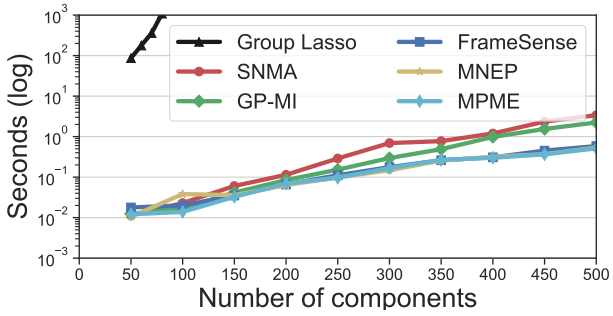


Fig. 7. Comparison on running time of one iteration. y-axis is the running time (seconds) on a log scale. x-axis denotes the size of a system.

4.2 Validations on real data

We validate SNMA and illustrate its applications on five real-world datasets, including disease spread, heat diffusion, and information spread. We only evaluate sentinel prediction because the ground truth sentinels are unknown in real cases.

4.2.1 2009 Hong Kong H1N1 flu pandemic

The epidemic dynamics data of 2009 Hong Kong H1N1 influenza [44], was provided by Centre for Health Protection (CHP), Department of Health, Government of the Hong Kong Special Administrative Region. During this pandemic, the first imported

TABLE 1
Sentinel Identification Results (Average Failure Rate %) on Synthetic Data

Data type	continuous data						discrete data					
Non-linear embedding	no			yes			no			yes		
Noise level	low	medium	high	low	medium	high	low	medium	high	low	medium	high
GP-MI	0.32	0.38	0.44	0.41	0.45	0.46	0.41	0.42	0.46	0.43	0.44	0.49
FrameSense-PCA	0.33	0.39	0.47	0.34	0.44	0.51	0.23	0.24	0.26	0.18	0.24	0.24
FrameSense-GP	0.20	0.31	0.43	0.09	0.20	0.33	0.25	0.30	0.37	0.26	0.27	0.31
MNEP-PCA	0.28	0.30	0.40	0.30	0.42	0.43	0.36	0.39	0.46	0.37	0.39	0.42
MNEP-GP	0.24	0.35	0.41	0.42	0.45	0.44	0.23	0.26	0.28	0.26	0.30	0.32
MPME-PCA	0.29	0.32	0.39	0.32	0.38	0.42	0.35	0.40	0.41	0.36	0.40	0.43
MPME-GP	0.05	0.13	0.27	0.25	0.30	0.37	0.22	0.24	0.28	0.24	0.27	0.29
Group Lasso	0.13	0.16	0.21	0.23	0.25	0.26	0.27	0.31	0.36	0.31	0.34	0.38
SNMA	**0.00**	**0.00**	**0.04**	**0.00**	**0.01**	**0.03**	**0.00**	**0.07**	**0.13**	**0.05**	**0.12**	**0.14**

TABLE 2
Sentinel Prediction Results (Average CMSE) on Synthetic Data

Data type	continuous data						discrete data					
Non-linear embedding	no			yes			no			yes		
Noise level	low	medium	high	low	medium	high	low	medium	high	low	medium	high
GP-MI	0.02	0.03	0.06	2.56	3.62	6.17	0.23	0.44	0.70	0.63	1.19	1.52
FrameSense-PCA	0.16	0.40	2.26	9.21	20.97	27.50	2.24	2.37	2.60	2.20	1.94	2.23
FrameSense-GP	0.17	0.31	0.37	7.75	10.30	11.97	0.29	0.48	0.66	0.85	1.27	1.52
MNEP-PCA	0.04	0.09	0.33	3.21	4.79	30.28	2.19	3.12	3.47	1.72	1.69	2.17
MNEP-GP	0.12	0.24	0.37	4.75	6.29	6.96	0.32	0.49	0.67	0.81	1.24	1.57
MPME-PCA	0.02	0.18	0.33	2.99	4.78	31.94	3.35	3.25	3.44	1.67	1.65	1.83
MPME-GP	0.03	0.05	0.25	2.75	3.64	7.95	0.33	0.54	0.69	0.73	1.23	1.55
Group Lasso	0.25	0.36	0.46	2.95	4.34	8.54	0.92	1.54	1.76	0.91	1.13	1.47
SNMA	**0.01**	**0.02**	**0.05**	**0.49**	**0.83**	**1.65**	**0.15**	**0.37**	**0.53**	**0.49**	**0.71**	**1.12**

TABLE 3
Sentinel Prediction Results (CMSE) on Real Data

Data type	continuous data				discrete data
Epidemic dynamics	disease spread		heat diffusion		information spread
Dataset	Hong Kong H1N1	Tengchong malaria	Berkeley Lab	UltraSPARC T1	Baidu Tieba
GP-MI	15.69	3.34	0.65	0.64	0.48
FrameSense-PCA	71.77	23.33	28.57	14.21	1.98
FrameSense-GP	29.28	3.69	0.43	1.81	2.38
MNEP-PCA	27.81	10.06	13.59	15.51	1.04
MNEP-GP	29.37	3.69	0.43	1.79	2.20
MPME-PCA	33.42	10.56	10.09	21.13	0.87
MPME-GP	29.22	3.70	0.43	1.80	2.27
Group Lasso	34.57	3.81	17.25	20.16	1.13
SNMA	**12.48**	**3.27**	**0.39**	0.22	0.40
SNMA-embedding	47.21	3.32	0.48	**0.13**	**0.29**

case of human swine influenza (HSI) was confirmed on May 1, 2009. As of September 2010, there were over 36,000 laboratory confirmed cases of HSI in Hong Kong, among which about 290 were severe cases and over 80 of them died [45]. We consider the number of disease infections for 105 days since the disease onset in June 1st 2009 based on the confirmed cases of H1N1 infection reported by CHP[2], which gives the dynamics of disease in terms of the number of newly reported infected cases (i.e., spatial position and infection time of each case are both provided by the report).

Hong Kong consists of 18 administrative districts, i.e., 18 components of interest. By merging the confirmed cases during 105 days at 3 days interval, the infectious period of this flu is 3.2 days, we obtain an epidemic dynamics of N=18 and T=35. We set the continuous dynamics data from Jun.1 to Aug.15 as training data and the one from Aug. 15 to Sep. 15 as test data. We specify $\mathcal{K} = [2, 4, \ldots, 16]$, which contains feasible sentinel number k to test. Then according to each $k \in \mathcal{K}$, sentinel districts of Hong Kong are identified via every method. Finally, based on the sentinels, we predict the newly cases on test data and calculate the sentinel prediction error (CMSE) according to Eq. 22.

As shown in Table 3, SNMA outperforms the competitors obviously. In real cases, we construct non-linear SNMA-embedding model by embedding a quadratic basic function $\phi(x) = x^2$ to SNMA via Eq. 20. In Hong Kong flu experiments, the poor performance of SNMA-embedding is caused by overfitting because embedding basic functions introduce new parameters. Fig. 8 (a) shows a case of spatial distribution of sentinel districts, where 8 sentinel districts are selected via SNMA. It is interesting to note that four sentinels gather in the center area (i.e., Kowloon and Hong Kong island) where most people live in Hong Kong.

4.2.2 2005-2009 Tengchong malaria outbreak

Tengchong City, Yunnan Province, China, has 18 towns, and 658,207 residents that are distributed in a wide area of 5,845 km^2 in 2011. Because of the suitable climate for mosquito habitats, Tengchong has a quite serious malaria outbreak. Five years' (2005-2009) monthly malaria cases data at the town level, were

2. http://www.chp.gov.hk

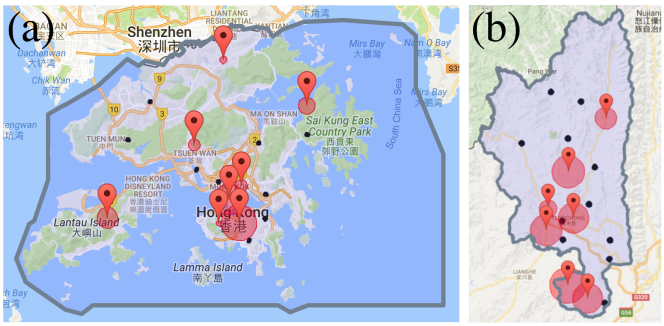


Fig. 8. The spatial distribution of sentinels in Hong Kong and Tengchong. The red bubble markers denote sentinel locations, and the radius of red circle depicts its importance for dynamics prediction (its γ value). The black points are unmonitored locations. (a) 8 sentinel districts in Hong Kong for H1N1 flu surveillance; (b) 7 sentinel towns in Tengchong city for malaria surveillance.

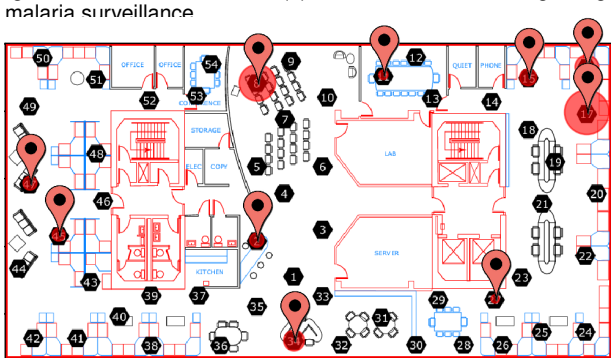


Fig. 9. The spatial distribution of 10 sentinels in Intel Berkeley Lab for temperature monitoring. The red bubble markers denote sentinel locations, and the radius of red circle depicts its importance (its γ value).

collected by Tengchong CDC and can be obtained from the annual reports of National Institute of Parasitic Disease, China CDC. By merging malaria cases during the 5 years at one-month interval and eliminating the missing data from 2005 Jun. to Dec., we get malaria dynamics, which contains $N = 18$ components and $T = 53$ months.

We set the malaria dynamics from 2005 to 2008 as training data and the one during 2009 as test data. Similar to the Hong Kong case, we specify $\mathcal{K} = [2, 4, \ldots, 16]$ and identify the sentinel towns for each $k \in \mathcal{K}$. Then we predict the malaria dynamics on the test data. Table 3 shows the sentinel prediction error, and SNMA achieves the best performance once again. Fig. 8 (b) shows a case of spatial distribution of sentinel districts, where 7 towns are selected as sentinels in Tengchong via SNMA.

4.2.3 Temperature measurements in Intel Berkeley Lab

Besides disease spread cases, we also validate SNMA on heat diffusion data. In Intel Berkeley Lab, temperature measurements was collected from $N = 54$ sensors deployed between Feb. 28th and Apr. 5th, 2004. The sensors were arranged according to the diagram shown in Fig. 9. This dataset is freely available[3].

We average the temperature measurements of each sensor at 30 minutes intervals on 5 consecutive days (starting Feb. 28th 2004, T=240). We set the temperature data on the first 3 day as training data, the one on the two following days as test data. We specify $\mathcal{K} = [5, 10, \ldots, 45]$, and identify the sentinels for each $k \in \mathcal{K}$. SNMA achieves the best sentinel prediction as shown in Table 3.

3. http://db.csail.mit.edu/labdata/labdata.html

Fig. 9. shows a case of spatial distribution of sentinels, where 10 sentinels are selected for temperature monitoring via SNMA.

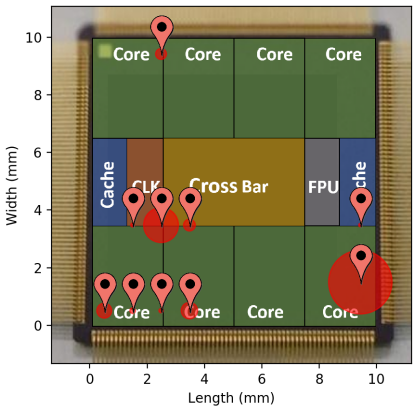


Fig. 10. The spatial distribution of 10 sentinels on UltraSPARC T1 microprocessor for thermal surveillance. The red bubble markers denote sentinel locations, and the radius of red circle depicts its importance.

4.2.4 Thermal diffusion on UltraSPARC T1 microprocessor

In microprocessor architecture, on-chip sensors are placed to measure local temperatures and prevent the uncontrolled thermal runaway situations [25]. However, one can not place too many on-chip sensors to achieve all-round monitoring due to the space limitation of microprocessor. Thus, how to deploy the limited sensors effectively is significant in microprocessor design.

Taking UltraSPARC T1 8-core microprocessor shown in Fig. 10 as a case study, we first generate dynamics data by simulating thermal diffusion on the microprocessor using 3D-ICE [46]. 3D-ICE is a compact transient thermal model for the thermal simulation of 3D integrated circuits. We simulate $T = 1000$ thermal maps at 0.1 second intervals during 100 seconds, which report the temperature of overall $N = 100$ cells of equal area (approximately 1 mm^2) on the microprocessor. We set the first 800 thermal maps as training data, and the rest 200 maps as test data. We specify $\mathcal{K} = [5, 10, \ldots, 45]$, and identify the sentinels for each $k \in \mathcal{K}$. Here SNMA-embedding achieves the best sentinel prediction because of sufficient training data as shown in Table 3. Fig. 10 shows a case of spatial distribution of sentinels on the chip, where 6 sentinels are placed in boundary areas. It indicates boundary area may have more information for dynamics prediction than the other areas on the microprocessor.

4.2.5 Hot words diffusion in Baidu Tieba

We finally validate SNMA on an information spread case where epidemic dynamics data is discrete. Baidu Tieba[4], one of the largest online community platforms in China, is a collection of thousands of topic-specific web forums. Tieba users can post any hot words (vocabularies that are widely used in Tieba during a short period) in any forums. Hot words often present diffusion phenomena in Baidu Tieba: a hot word first appears in only a few forums, and then is posted gradually in many other forums by the overlapping users who are active in multiple forums. Thus, Tieba can be regarded as a logistical discrete system, where forums are components, hot words are contagions, and the infected state of a component is 0 or 1.

We tracked the dynamics of 11 independent hot words cascading among the top-200 ($N = 200$) active forums in Baidu

4. http://tieba.baidu.com

Tieba between Apr. 2014 and Oct. 2015 (18 months). Only the dynamics during the words' bursting period is preserved, and the total bursting period of the words is 738 days, i.e., T=738 by using one day as a time-unit. We set the dynamics of 10 hot words as training data, and the rest one be test data. We specify $\mathcal{K} = [5, 10, \ldots, 45]$, and identify the sentinels for each $k \in \mathcal{K}$. As mentioned before, GP-based and LIP-based methods cannot deal with discrete data, thus we extend those methods by adding a simple threshold classification. SNMA-embedding outperforms the competitors as shown in Table 3. In this discrete case, the embedding basic function is $\phi(x_t) = x_t - x_{t-1}$, which could characterize the changing information of dynamics.

Summarily, the proposed SNMA and SNMA-embedding outperform the competitors on the real-world datasets in term of sentinel prediction error (CMSE). SNMA-embedding may fall into overfitting when data volume is insufficient because embedding basic functions introduce parameters. GP-MI is better than LIP-based methods in most cases. For LIP-based methods, experiment results show the sensing matrix constructed by GP (Eq. 21) is better than the one constructed by PCA which usually suffers from an ill-conditioned problem. Although in the same active surveillance framework, SNMA achieves better performance than Group Lasso because the proposed group sparse Bayesian learning can effectively handle the uncertainty from both data and model.

5 Discussions and Conclusions

5.1 A connection to block sparse Bayesian learning

In the literature, there exist works on block sparse Bayesian (BSB) modeling [38] [47] [48] related to our Bayesian modelling, which were used to recover block sparse signals. Since both these BSB models and ours focus on modeling the group sparsity feature of Bayesian models, there are a connection between them in form. However, the theoretical results as contributed by this paper, including both computational models and algorithms, are novel. We elaborate on the primary distinctions between them as follows.

Different Gaussian prior. Although the BSB models and ours adopt a zero-mean multivariate Gaussian distribution as the prior to induce group sparsity, the two Gaussian priors employ different intra-group correlation modelling. In the BSB models, the correlation in each group is modelled by a learnable matrix $\mathbf{B}_i, i \in [1, 2, \ldots, N]$. In this paper, we ignore such intra-group correlation by using an identity matrix. It's important to note that, we uncover, for the first time, two important properties of the designed priors to enforce group sparsity by analyzing their sparsity effect on parameters in Section 3.2.1.

Besides, the biggest advantage brought by our prior modeling is that it can significantly reduce the time complexity of Bayesian inference by simplifying the posterior covariance matrix into a diagonal-block matrix and then coming up with two fast matrix operations applied to it, as discussed in Section 3.4. In contrast, the BSB models have high time complexity. For instance, the *T-SBL* model [38], most relevant to ours, has a time complexity $O(M^2L^2N)$, or $O(N^5 \times \max(T, N))$ after we make the notations consistent with ours. As discussed in [38], by using some approximation techniques such as the speed-up scheme proposed by [49], the time of *T-SBL* model can be further reduced to $O(MN^2)$, e.g. the *T-MSBL* model, being approximately on the same order as ours.

Different likelihood function. We model continuous data and discrete data by Gaussian likelihood and Bernoulli likelihood in this paper, respectively. In contrast, the BSB models only model continuous data via Gaussian distributions, since the signals they processed for recovery are continuous. To the best of our knowledge, this paper is the first work that proposes a group sparse Bayesian model for handling discrete data. To model discrete data is motivated by the fact that many diffusion processes can be naturally represented as discrete signals, e.g., whether a piece of news is posted on a blog in the information spread setting.

Different posterior distribution. As the adopted priors and likelihoods are different, the posterior distributions in this paper and the BSB models are accordingly different. For discrete data, the posterior in this paper is an approximate Gaussian distribution with learnable variational parameters, which is obviously different from the posterior in the BSB models, a Gaussian distribution. As for continuous data, although the posteriors in this paper and in the BSB models are both Gaussian, however, it should be noted that the two posteriors are intrinsically different, due to the two different prior modelling.

Different Bayesian inference method. We employ the expectation–maximization (EM) method and the variational EM method to infer the posterior distribution and hyper-parameters in the cases of continuous data and discrete data, respectively. In contrast, the BSB models only adopt the EM method. Moreover, we introduce a Gaussian lower bound for the Bernoulli likelihood to address the challenge that original Q function in EM cannot be analytically solved in the case of discrete data in Theorem 2 and its proof.

5.2 Conclusions

In this work, we proposed and demonstrated a comprehensive active surveillance framework for solving the practical problem of predicting the entire dynamics of a diffusion system by means of monitoring only some of its sentinel components. Unlike the existing work, we solved the above problem by modeling the task of active surveillance as that of a row sparse sentinel network mining and addressing three of its related challenges. For our framework, we provided a novel importance measure, the $\boldsymbol{\gamma}$ value, for assessing the monitoring priority of a component for dynamics prediction, and we theoretically analyzed the $\boldsymbol{\gamma}$ from both prior and posterior perspectives. Based on the measure, we designed a backward-selection SNMA algorithm to mine the sentinel network in the cases of both linear dynamical system and logistical discrete system via group sparse Bayesian learning. We then solved the scalability problem and extended SNMA to non-linear dynamical systems by embedding basic functions for handling complex diffusion mechanism. The results of our experiments, involving both synthetic and five real-world datasets (e.g., disease transmission, heat diffusion, and information diffusion), showed that our developed sentinel network-based active surveillance method could achieve better performances than the state-of-the-art counterparts, thus is more promising for tackling the real-world, resource-constrained diffusion prediction challenges.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grants 61876069 and 61572226, the Jilin Province Key Scientific and Technological Research and Development Project under grants 20180201067GX and 20180201044GX, and Hong Kong Research Grants Council (RGC) General Research Fund (GRF) under grant RGC/HKBU12201318. A preliminary version of this work was published in AAAI'18 [6].

REFERENCES

[1] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[2] P. M. Polgreen, Z. Chen, A. M. Segre, M. L. Harris, M. A. Pentella, and G. Rushton, "Optimizing influenza sentinel surveillance at the state level," *American journal of epidemiology*, vol. 170, no. 10, pp. 1300–1306, 2009.

[3] Y. Zheng, F. Liu, and H.-P. Hsieh, "U-air: when urban air quality inference meets big data," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Chicago, USA: ACM, 2013, pp. 1436–1444.

[4] Y. Chen, H. Amiri, Z. Li, and T.-S. Chua, "Emerging topic detection for organizations from microblogs," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Ireland: ACM, 2013, pp. 43–52.

[5] L. Gerardo-Giorda, G. Puggioni, R. J. Rudd, L. A. Waller, and L. A. Real, "Structuring targeted surveillance for monitoring disease emergence by mapping observational data onto ecological process," *Journal of The Royal Society Interface*, vol. 10, no. 86, p. 20130418, 2013.

[6] H. Pei, B. Yang, J. Liu, and L. Dong, "Group sparse bayesian learning for active surveillance on epidemic dynamics," in *Thirty-Second AAAI Conference on Artificial Intelligence*. USA: AAAI, 2018, pp. 800–807.

[7] B. Yang, H. Pei, H. Chen, J. Liu, and X. Shang, "Characterizing and discovering spatiotemporal social contact patterns for healthcare," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 8, pp. 1532–1546, 2017.

[8] A. Krause, A. Singh, and C. Guestrin, "Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies," *Journal of Machine Learning Research*, vol. 9, no. Feb, pp. 235–284, 2008.

[9] N. Cressie, *Statistics for spatial data*. Wiley, 1991.

[10] A. Krause and C. Guestrin, "Nonmyopic active learning of gaussian processes: an exploration-exploitation approach," in *Proceedings of the 24th international conference on Machine learning*. USA: ACM, 2007, pp. 449–456.

[11] M. Shamaiah, S. Banerjee, and H. Vikalo, "Greedy sensor selection: Leveraging submodularity," in *49th IEEE conference on decision and control*. USA: IEEE, 2010, pp. 2572–2577.

[12] J. Ranieri, A. Chebira, and M. Vetterli, "Near-optimal sensor placement for linear inverse problems," *IEEE Transactions on signal processing*, vol. 62, no. 5, pp. 1135–1146, 2014.

[13] C. Jiang, Y. C. Soh, and H. Li, "Sensor placement by maximal projection on minimum eigenspace for linear inverse problems," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5595–5610, 2016.

[14] R. Du, C. Chen, B. Yang, and X. Guan, "Vanet based traffic estimation: A matrix completion approach," in *2013 IEEE Global Communications Conference*. USA: IEEE, 2013, pp. 30–35.

[15] R. Du, C. Chen, B. Yang, N. Lu, X. Guan, and X. Shen, "Effective urban traffic monitoring by vehicular sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 1, pp. 273–286, 2015.

[16] K. Cohen, S. Siegel, and T. McLaughlin, "A heuristic approach to effective sensor placement for modeling of a cylinder wake," *Computers & fluids*, vol. 35, no. 1, pp. 103–120, 2006.

[17] L. Yao, W. A. Sethares, and D. C. Kammer, "Sensor placement for on-orbit modal identification via a genetic algorithm," *AIAA journal*, vol. 31, no. 10, pp. 1922–1928, 1993.

[18] H. Wang, K. Yao, G. Pottie, and D. Estrin, "Entropy-based sensor selection heuristic for target localization," in *Proceedings of the 3rd international symposium on Information processing in sensor networks*. USA: ACM, 2004, pp. 36–45.

[19] S. Joshi and S. Boyd, "Sensor selection via convex optimization," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 451–462, 2009.

[20] S. P. Chepuri and G. Leus, "Continuous sensor placement," *IEEE Signal Processing Letters*, vol. 22, no. 5, pp. 544–548, 2014.

[21] C. Rusu, J. Thompson, and N. M. Robertson, "Sensor scheduling with time, energy, and communication constraints," *IEEE Transactions on Signal Processing*, vol. 66, no. 2, pp. 528–539, 2017.

[22] C. Rusu and J. Thompson, "On the use of tight frames for optimal sensor placement in time-difference of arrival localization," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 1415–1419.

[23] C. Jiang, Z. Chen, R. Su, and Y. C. Soh, "Group greedy method for sensor placement," *IEEE Transactions on Signal Processing*, vol. 67, no. 9, pp. 2249–2262, 2019.

[24] A. C. Atkinson, "Recent developments in the methods of optimum and related experimental designs," *International Statistical Review/Revue Internationale de Statistique*, vol. 56, no. 2, pp. 99–115, 1988.

[25] J. Ranieri, A. Vincenzi, A. Chebira, D. Atienza, and M. Vetterli, "Eigenmaps: Algorithms for optimal thermal maps extraction and sensor placement on multicore processors," in *Proceedings of the 49th Annual Design Automation Conference*. USA: ACM, 2012, pp. 636–641.

[26] J. Wallinga, M. van Boven, and M. Lipsitch, "Optimizing infectious disease interventions during an emerging epidemic," *Proceedings of the National Academy of Sciences*, vol. 107, no. 2, pp. 923–928, 2010.

[27] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Washington, USA: ACM, 2010, pp. 1019–1028.

[28] D. J. MacKay, "Bayesian interpolation," *Neural computation*, vol. 4, no. 3, pp. 415–447, 1992.

[29] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of machine learning research*, vol. 1, no. Jun, pp. 211–244, 2001.

[30] C.-H. Zhang, T. Zhang *et al.*, "A general theory of concave regularization for high-dimensional sparse estimation problems," *Statistical Science*, vol. 27, no. 4, pp. 576–593, 2012.

[31] S. Liu, Y. D. Zhang, T. Shan, and R. Tao, "Structure-aware bayesian compressive sensing for frequency-hopping spectrum estimation with missing observations," *IEEE Transactions on Signal Processing*, vol. 66, no. 8, pp. 2153–2166, 2018.

[32] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal processing magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[33] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin: Springer, 2006.

[34] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*. Cambridge, Massachusetts: the MIT Press, 2006.

[35] C. Couvreur and Y. Bresler, "On the optimality of the backward greedy algorithm for the subset selection problem," *SIAM Journal on Matrix Analysis and Applications*, vol. 21, no. 3, pp. 797–808, 2000.

[36] P. Maragakis, F. Ritort, C. Bustamante, M. Karplus, and G. E. Crooks, "Bayesian estimates of free energies from nonequilibrium work data in the presence of instrument noise," *The Journal of Chemical Physics*, vol. 129, no. 2, p. 024102, 2008.

[37] G. E. Crooks, "Logistic approximation to the logistic-normal integral," *Technical Report Lawrence Berkeley National Laboratory*, 2009.

[38] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 912–926, 2011.

[39] S. Liu, H. Wu, Y. Huang, Y. Yang, and J. Jia, "Accelerated structure-aware sparse bayesian learning for three-dimensional electrical impedance tomography," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 9, pp. 5033–5041, 2019.

[40] Y. Zhou, U. Porwal, C. Zhang, H. Q. Ngo, X. Nguyen, C. Ré, and V. Govindaraju, "Parallel feature selection inspired by group testing," in *Advances in Neural Information Processing Systems*, Montreal,Canada, 2014, pp. 3554–3562.

[41] A. S. Jarrah, R. Laubenbacher, B. Stigler, and M. Stillman, "Reverse-engineering of polynomial dynamical systems," *Advances in Applied Mathematics*, vol. 39, no. 4, pp. 477–489, 2007.

[42] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 1, pp. 53–71, 2008.

[43] L. E. Frenzel, *Handbook of Serial Communications Interfaces*. Elsevier, 2016.

[44] J. T. Wu, E. S. Ma, C. K. Lee, D. K. Chu, P.-L. Ho, A. L. Shen, A. Ho, I. F. Hung, S. Riley, L. M. Ho *et al.*, "The infection attack rate and severity of 2009 pandemic H1N1 influenza in Hong Kong," *Clinical Infectious Diseases*, vol. 51, no. 10, pp. 1184–1191, 2010.

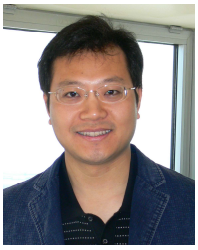
[45] "Summary report on the surveillance of adverse events following HSI immunisation and expert group's comment on the safety of hsi vaccine in Hong Kong," Centre for Health Protection, Hong Kong, Tech. Rep., September 2010.

[46] A. Sridhar, A. Vincenzi, M. Ruggiero, T. Brunschwiler, and D. Atienza, "3d-ice: Fast compact transient thermal modeling for 3d ics with inter-tier liquid cooling," in *2010 IEEE/ACM International Conference on Computer-Aided Design*. USA: IEEE, 2010, pp. 463–470.

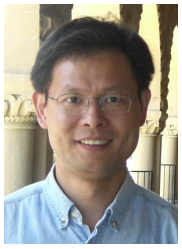
[47] Z. Zhang and B. D. Rao, "Recovery of block sparse signals using the framework of block sparse bayesian learning," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 3345–3348.

[48] ——, "Extension of sbl algorithms for the recovery of block sparse signals with intra-block correlation," *IEEE Transactions on Signal Processing*, vol. 61, no. 8, pp. 2009–2015, 2013.
[49] D. P. Wipf and B. D. Rao, "An empirical bayesian strategy for solving the simultaneous sparse approximation problem," *Signal Processing, IEEE Transactions on*, vol. 55, no. 7, pp. 3704–3716, 2007.

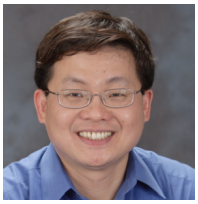
Hongbin Pei is currently a Ph.D. student in the College of Computer Science and Technology, Jilin University. He is also a visiting scholar in the Department of Computer Science, University of Illinois at Urbana-Champaign. He received the M.S. and B.S. degrees from Jilin University in 2015 and 2012, respectively. His research interests include network data analysis and spatio-temporal data mining, with applications to health informatics and intelligent transportation.



Bo Yang is currently a professor in the College of Computer Science and Technology, Jilin University. He is also the director of the Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, China. His current research interests are in the areas of data mining, complex network analysis, self-organized and self-adaptive multi-agent systems, with applications to knowledge engineering and intelligent health informatics.



Jiming Liu is currently the Chair Professor of Computer Science and Associate Vice President (Research) at Hong Kong Baptist University. He received his M.Eng. and Ph.D. degrees from McGill University. His research interests include Data Analytics, Data Mining and Machine Learning, Complex Network Analytics, Data-Driven Complex Systems Modeling, and Health Informatics. He is a Fellow of the IEEE. Prof. Liu has served as the Editor-in-Chief of Web Intelligence Journal (IOS), and the Associate Editor of Big Data and Information Analytics (AIMS), IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Cybernetics, and Computational Intelligence (Wiley), among others.



Kevin Chen-Chuan Chang is a Professor in University of Illinois at Urbana-Champaign. His research addresses large-scale information access, for search, mining, and integration across structured and unstructured big data including Web data and social media. He also cofounded Cazoodle for deepening vertical dataaware search over the Web.

Appendix to the paper "Active Surveillance via Group Sparse Bayesian Learning"

Hongbin Pei, Bo Yang, Jiming Liu, *Fellow, IEEE*, and Kevin Chen-Chuan Chang

APPENDIX A
DERIVATION OF POSTERIOR DISTRIBUTION

In this section, we present the derivation of posterior distribution of sentinel network, linear continuous system in Section A.1 and logistical discrete system in Section A.2.

A.1 Theorem 1

Proof. The posterior distribution over $\boldsymbol{s}$ of the linear continuous system is given by:

$$p(\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi},\boldsymbol{\gamma},\lambda)=\frac{p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s},\lambda)p(\boldsymbol{s}|\boldsymbol{\gamma})}{p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{\gamma},\lambda)}, \tag{1}$$

where the likelihood $p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s},\lambda)\sim\mathcal{N}(\boldsymbol{\Phi}\boldsymbol{s},\lambda\mathbf{I}_N)$ and the prior $p(\boldsymbol{s}|\boldsymbol{\gamma})\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_0)$ (refer to Eq. 9 and Eq. 4 in the main paper). Thus, the numerator of Eq. 1 is a product of two Gaussians, which is still a Gaussian.

The denominator is the marginal likelihood,

$$p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{\gamma},\lambda)=\int p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s},\lambda)p(\boldsymbol{s}|\boldsymbol{\gamma})d\boldsymbol{s},$$

which is a convolution of two Gaussians. This convolution can be analytically calculated and its result is also a Gaussian. Then the posterior of $\boldsymbol{s}$ can be obtained by division, which is the Gaussian distribution given by theorem 1 in the main paper. □

A.2 Theorem 2

Proof. The posterior distribution over $\boldsymbol{s}$ in the logistical discrete system is given by:

$$p(\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi},\boldsymbol{\gamma})=\frac{p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s})p(\boldsymbol{s}|\boldsymbol{\gamma})}{p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{\gamma})},$$

where the likelihood $p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s})$ is a Bernoulli distribution and the prior $p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{\gamma})$ is a Gaussian (refer to Eq. 11 and Eq. 4 in the main paper). We cannot directly calculate this posterior because its denominator, marginal likelihood $p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{\gamma})=\int p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s})p(\boldsymbol{s}|\boldsymbol{\gamma})d\boldsymbol{s}$, cannot be analytically integrated.

We calculate the posterior approximately by using a local variational method. We introduce a lower bound function having a Gaussian form for the sigmoid function (refer to Eq. 10 in the main paper),

$$\sigma(\boldsymbol{\Phi}_n\boldsymbol{s})\geq\sigma(\xi_n)e^{\frac{1}{2}(\boldsymbol{\Phi}_n\boldsymbol{s}-\xi_n)-\boldsymbol{\pi}(\xi_n)((\boldsymbol{\Phi}_n\boldsymbol{s})^2-\xi_n^2)},$$

where $n=1,\ldots,TN$, $\boldsymbol{\pi}(\xi_n)=-\frac{1}{2\xi_n}(\sigma(\xi_n)-\frac{1}{2})$ and ξ_n is an introduced variational parameter corresponding to each label $\boldsymbol{y}_n$.

By substituting the lower bound into the likelihood function of the logistical system (refer to Eq. 11 in the main paper), we get a lower bound of the likelihood,

$$\begin{aligned}&p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s})\geq h(\boldsymbol{s},\boldsymbol{\xi})\\&=\prod_{n=1}^{TN}\sigma(\xi_n)e^{\boldsymbol{\Phi}_n\boldsymbol{s}\boldsymbol{y}_n-\frac{1}{2}(\boldsymbol{\Phi}_n\boldsymbol{s}+\xi_n)-\boldsymbol{\pi}(\xi_n)((\boldsymbol{\Phi}_n\boldsymbol{s})^2-\xi_n^2)}.\end{aligned} \tag{2}$$

Based on the likelihood lower bound, we can calculate the lower bound function for the joint distribution of $\boldsymbol{s}$ and $\boldsymbol{y}$,

$$\begin{aligned}&p(\boldsymbol{y},\boldsymbol{s}|\boldsymbol{\Phi},\boldsymbol{\gamma})=p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s})p(\boldsymbol{s}|\boldsymbol{\gamma})\\&\geq h(\boldsymbol{s},\boldsymbol{\xi})p(\boldsymbol{s}|\boldsymbol{\gamma})\\&=e^{\frac{1}{2}\boldsymbol{s}^{\mathrm{T}}\Sigma_0^{-1}\boldsymbol{s}+\sum_{n=1}^{TN}(\boldsymbol{\Phi}_n\boldsymbol{s}(\boldsymbol{y}_n-\frac{1}{2})-\boldsymbol{\pi}(\xi_n)(\boldsymbol{\Phi}_n\boldsymbol{s})^2)+const}.\end{aligned} \tag{3}$$

The result of Eq. 3 cannot be interpreted as a probability distribution because it's not normalized. Note that, the exponent term of Eq. 3 is a quadratic function of $\boldsymbol{s}$. By identifying the linear and quadratic terms of $\boldsymbol{s}$ and then normalizing over $\boldsymbol{s}$, we obtain a Gaussian variational distribution over $\boldsymbol{s}$. This approximate variational posterior is given by theorem 2 in the main paper, where the entry of the diagonal matrix $\boldsymbol{\pi}(\boldsymbol{\xi})_{n,n}$ is equal to $\boldsymbol{\pi}(\xi_n)$ in the Eq. 3. □

APPENDIX B
DERIVATION OF HYPER-PARAMETERS ESTIMATION

We estimate the hyper-parameters (λ and $\boldsymbol{\gamma}$ for the linear continuous system, and $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$ for the logistical discrete system) by employing Expectation-Maximization (EM) and variational EM to maximize the marginal likelihood (refer to Eq. 12 in the main paper). Specifically, we first give the Q function of the hyper-parameters by treating $\boldsymbol{s}$ as hidden variables, and then deduce the update rule of the hyper-parameters via optimizing the Q functions.

- *H. Pei and B. Yang are with the College of Computer Science and Technology, Jilin University, China; Department of Computer Science, University of Illinois at Urbana-Champaign, USA; Key Laboratory of Symbolic Computation and Knowledge Engineer (Jilin University), Ministry of Education, China. E-mail: peihb15@mails.jlu.edu.cn, ybo@jlu.edu.cn.*
- *J. Liu is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. E-mail: jiming@comp.hkbu.edu.hk.*
- *K. Chang is with Department of Computer Science, University of Illinois at Urbana-Champaign, USA. Email: kcchang@illinois.edu.*

B.1 Linear continuous system

Based on the posterior distribution given in Theorem 1, the conditional expectation of complete log-likelihood, i.e., the Q-function of the linear system, is given by,

$$\begin{aligned} Q(\boldsymbol{\gamma},\lambda) &= E_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi},\boldsymbol{\gamma}^{(\text{old})},\lambda^{(\text{old})}}[\log(p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s},\lambda)p(\boldsymbol{s}|\boldsymbol{\gamma}))] \\ &= E_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi},\boldsymbol{\gamma}^{(\text{old})},\lambda^{(\text{old})}}[\log p(\boldsymbol{y}|\boldsymbol{\Phi},\boldsymbol{s},\lambda) + \log p(\boldsymbol{s}|\boldsymbol{\gamma})] \end{aligned}$$

where $E_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi},\boldsymbol{\gamma}^{(\text{old})},\lambda^{(\text{old})}}[\cdot]$ denotes the conditional expectation with respect to the posterior, and $\cdot^{(\text{old})}$ denotes the parameter estimated in the previous iteration. Notice that the log-terms of $\boldsymbol{\gamma}$ and λ are separated in $Q(\boldsymbol{\gamma},\lambda)$. Thus we can divide the Q-function into two simplifying sub-Q-functions.

The sub-Q-funciton for $\boldsymbol{\gamma}$ is:

$$\begin{aligned} Q(\boldsymbol{\gamma}) =& E_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi},\boldsymbol{\gamma}^{(\text{old})},\lambda^{(\text{old})}}[\log(p(\boldsymbol{s}|\boldsymbol{\gamma}))] \\ \propto& -\log|\boldsymbol{\Sigma}_0| - \text{Tr}[\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{s}} + \boldsymbol{\mu}_{\boldsymbol{s}}\boldsymbol{\mu}_{\boldsymbol{s}}^{\text{T}})]. \end{aligned} \tag{4}$$

As every γ_i is independent, the derivative of $Q(\boldsymbol{\gamma})$ with respect to γ_i is given by,

$$\frac{\partial Q(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}_i} = -\frac{N}{2\boldsymbol{\gamma}_i} + \frac{1}{2\boldsymbol{\gamma}_i^2}\text{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{s}}^i + \boldsymbol{\mu}_{\boldsymbol{s}}^i(\boldsymbol{\mu}_{\boldsymbol{s}}^i)^{\text{T}}],$$

where $\boldsymbol{\mu}_{\boldsymbol{s}}^i \in \mathbb{R}^{N\times 1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}}^i \in \mathbb{R}^{N\times N}$ are the ith group in $\boldsymbol{\mu}_{\boldsymbol{s}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}}$ respectively. Specifically, by using Python notations we define $\boldsymbol{\mu}_{\boldsymbol{s}}^i = \boldsymbol{\mu}_{\boldsymbol{s}}[(i-1)N+1:iN]$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}}^i = \boldsymbol{\Sigma}_{\boldsymbol{s}}[(i-1)N+1:iN,(i-1)N+1:iN]$.

By setting the derivative be zero, the update rule for γ_i, i.e., the $\boldsymbol{\gamma}$-value of component i, can be obtained

$$\gamma_i \leftarrow ((\boldsymbol{\mu}_{\boldsymbol{s}}^i)^{\text{T}}\boldsymbol{\mu}_{\boldsymbol{s}}^i + \text{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{s}}^i])N^{-1},\ i=1,\ldots,N.$$

To estimate λ, the sub-Q-funciton is given by

$$\begin{aligned} Q(\lambda) \propto& -TN\log\lambda - \frac{1}{\lambda}E_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi},\boldsymbol{\gamma}^{(\text{old})},\lambda^{(\text{old})}}[\|\ \boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{s}\ \|_2^2] \\ =& -TN\log\lambda - \frac{1}{\lambda}(\|\ \boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\mu}_{\boldsymbol{s}}\ \|_2^2 \\ &+ E_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi},\boldsymbol{\gamma}^{(\text{old})},\lambda^{(\text{old})}}[\|\ \boldsymbol{\Phi}(\boldsymbol{s}-\boldsymbol{\mu}_{\boldsymbol{s}})\ \|_2^2]) \\ =& -TN\log\lambda - \frac{1}{\lambda}(\|\ \boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\mu}_{\boldsymbol{s}}\ \|_2^2 + \text{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{\Phi}^{\text{T}}\boldsymbol{\Phi}]). \end{aligned}$$

The derivative of the $Q(\lambda)$ with respect to λ is given by

$$\frac{\partial Q(\lambda)}{\partial\lambda} = -\frac{TN}{\lambda} + \frac{1}{\lambda^2}(\|\ \boldsymbol{y}-\boldsymbol{\Phi}\boldsymbol{\mu}_{\boldsymbol{s}}\ \|_2^2 + \text{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{\Phi}^{\text{T}}\boldsymbol{\Phi}]).$$

By setting the derivative to zero, the update rule for λ then can be obtain

$$\lambda \leftarrow (TN)^{-1}(\|\ \boldsymbol{y}-\boldsymbol{\Phi}\mu_{\boldsymbol{s}}\ \|_2^2 + \text{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{s}}\boldsymbol{\Phi}^{\text{T}}\boldsymbol{\Phi}]).$$

B.2 Logistical discrete system

As the integral in the Q function can be analytically solved in the logistical system, we apply variational EM to estimate the hyper-parameters by optimizing an approximate low bound of the marginal likelihood. Based on the lower bound for the joint distribution $\boldsymbol{s}$ and $\boldsymbol{y}$ Eq. 3, the Q-function of the logistical system can be written as:

$$Q(\boldsymbol{\xi},\boldsymbol{\gamma}) = E_{\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi},\boldsymbol{\gamma}^{(\text{old})},\boldsymbol{\xi}^{(\text{old})}}[\log(h(\boldsymbol{s},\boldsymbol{\xi})p(\boldsymbol{s}|\boldsymbol{\gamma}))].$$

This Q-function also can be separated w.r.t $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$, and the sub-Q-funciton for $\boldsymbol{\gamma}$ is:

$$Q(\boldsymbol{\gamma}) \propto -\log|\boldsymbol{\Sigma}_0| - \text{Tr}[\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\Sigma}_{\boldsymbol{s}} + \boldsymbol{\mu}_{\boldsymbol{s}}\boldsymbol{\mu}_{\boldsymbol{s}}^{\text{T}})].$$

This sub-Q-funciton is the same with Eq. 4, thus the update rule for $\boldsymbol{\gamma}$ In the logistical discrete system is the same with the linear continue system.

Finally, the sub-Q-funciton for the variational parameter $\boldsymbol{\xi}$ is:

$$Q(\boldsymbol{\xi}) \propto \sum_{n=1}^{TN}\frac{\log\sigma(\xi_n)-\xi_n}{2} - \pi(\xi_n)(\boldsymbol{\Phi}_n(\Sigma_{\boldsymbol{s}}+\boldsymbol{\mu}_{\boldsymbol{s}}\boldsymbol{\mu}_{\boldsymbol{s}}^{\text{T}})\boldsymbol{\Phi}_n^{\text{T}} - \xi_n^2).$$

By letting the derivative of $Q(\boldsymbol{\xi})$ over ξ_n to be zero, we can get the learning rule for ξ_n:

$$\xi_n \leftarrow \sqrt{\boldsymbol{\Phi}_n(\Sigma_{\boldsymbol{s}}+\boldsymbol{\mu}_{\boldsymbol{s}}\boldsymbol{\mu}_{\boldsymbol{s}}^{\text{T}})\boldsymbol{\Phi}_n^{\text{T}}},\ n=1,\ldots,TN.$$

APPENDIX C
DERIVATION OF EFFICIENT MATRIX OPERATIONS

In this section, we give the derivation of two efficient operations on diagonal-block matrix, Section C.1 for matrix multiplication, Section C.2 for matrix inverse, and Section C.3 for .

C.1 Theorem 3

Proof. Without loss of generality, let $\mathbf{A} \in \mathbb{R}^{nk\times mk}$ and $\mathbf{B} \in \mathbb{R}^{mk\times lk}$ denote two diagonal-block matrix. The sub-matrix of $\mathbf{A}$ is $\mathbf{A}_{ij} \in \mathbb{R}^{k\times k}$, who is a diagonal matrix with all diagonal entries having the same value entry a_{ij}. Analogously, matrix $\mathbf{B}$, sub-matrix $\mathbf{B}_{ij}$, and entry b_{ij} have the same relationship.

Based on the property of block matrix, the product $\mathbf{C} \in \mathbb{R}^{nk\times lk}$ is a block matrix and its sub-matrix $\mathbf{C}_{ij} = \sum_{r=1}^{m}\mathbf{A}_{ir}\mathbf{B}_{rj}$, for i=1 : n, j=1 : l. And $\mathbf{C}_{ij}$ is a diagonal matrix with all diagonal entries having the same value $c_{ij} = \sum_{r=1}^{m}a_{ir}b_{rj}$. Thus, $\mathbf{C}$ is a diagonal-block matrix according to the Definition 1, and $\bar{\mathbf{C}}$ is the block projective matrix of $\mathbf{C}$ according to the Definition 2.

For each entry of $\bar{\mathbf{C}} \in \mathbb{R}^{n\times l}$, $\bar{\mathbf{C}}_{ij} = c_{ij} = \sum_{r=1}^{m}a_{ir}b_{rj} = \sum_{r=1}^{m}\bar{\mathbf{A}}_{ir}\bar{\mathbf{B}}_{rj}$. Therefore, $\bar{\mathbf{C}} = \bar{\mathbf{A}}\ \bar{\mathbf{B}}$. □

C.2 Theorem 4

Proof. Without loss of generality, let square matrix $\mathbf{A} \in \mathbb{R}^{nk\times nk}$ denote a diagonal-block matrix. The adjugate matrix of $\mathbf{A}$, $\text{adj}(\mathbf{A})$, is also a diagonal-block matrix because that if an entry is zero in the matrix $\mathbf{A}$ the corresponding entry who has the same location in adjugate matrix $\text{adj}(\mathbf{A})$ must be also zero.

The inverse matrix of $\mathbf{A}$ can be calculated by $\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A})$, where the determinant $\det(\mathbf{A})$ is a scalar. Thus the inverse matrix $\mathbf{A}^{-1}$ has the same structure with $\text{adj}(\mathbf{A})$, i.e., the inverse matrix $\mathbf{A}^{-1}$ is also a diagonal-block matrix.

Let $\bar{\mathbf{A}} \in \mathbb{R}^{n\times n}$ be the block projective matrix of square matrix $\mathbf{A}$. Based on eigendecomposition rules, we have $\mathbf{A} = \mathbf{Q}_1\boldsymbol{\Lambda}_1\mathbf{Q}_1^{-1}$ and $\bar{\mathbf{A}} = \mathbf{Q}_2\boldsymbol{\Lambda}_2\mathbf{Q}_2^{-1}$, where $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ are two diagonal matrices whose diagonal elements are the corresponding eigenvalues, i.e., $\boldsymbol{\Lambda}_1$=diag($\boldsymbol{\lambda}_1$) and $\boldsymbol{\Lambda}_2$=diag($\boldsymbol{\lambda}_2$), where vector $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are the eigenvalues of $\mathbf{A}$ and $\bar{\mathbf{A}}$, respectively.

Let $\boldsymbol{\lambda}_2(i)$ denote the i-th eigenvalue in $\boldsymbol{\lambda}_2$. We have $\bar{\mathbf{A}}\boldsymbol{v}_2(i) = \boldsymbol{\lambda}_2(i)\boldsymbol{v}_2(i)$, where $\boldsymbol{v}_2(i)$ is the corresponding eigenvector. According to the Theorem 3, we have $\mathbf{A}\boldsymbol{v}_1(i) = \boldsymbol{\lambda}_2(i)\boldsymbol{v}_1(i)$, where $\boldsymbol{v}_1(i) \in \mathbb{R}^{nk\times k}$ is a diagonal-block matrix and its block projective matrix is $\boldsymbol{v}_2(i)$. That is to say, $\boldsymbol{\lambda}_2(i)$ is a k-tuply repeated eigenvalue to $\mathbf{A}$, i.e., the entries of $\boldsymbol{\lambda}_2$ repeat k times in $\boldsymbol{\lambda}_1$. Therefore, $\boldsymbol{\Lambda}_2$ is the block projective matrix of $\bar{\boldsymbol{\Lambda}}_1$, i.e., $\bar{\boldsymbol{\Lambda}}_1 = \boldsymbol{\Lambda}_2$. As $\boldsymbol{\Lambda}_1$ and $\boldsymbol{\Lambda}_2$ are both diagonal matrices, their inverse matrices satisfy $(\boldsymbol{\Lambda}_1^{-1}) = \boldsymbol{\Lambda}_2^{-1}$. Analogously, we can also have $\bar{\boldsymbol{Q}}_1 = \boldsymbol{Q}_2$.

The inverse matrices of $\mathbf{A}$ and $\bar{\mathbf{A}}$ are $\mathbf{A}^{-1} = \mathbf{Q}_1\mathbf{\Lambda}_1^{-1}\mathbf{Q}_1^{-1}$ and $(\bar{\mathbf{A}})^{-1} = \mathbf{Q}_2\mathbf{\Lambda}_2^{-1}\mathbf{Q}_2^{-1}$, respectively. Based on the conclusions of $\bar{\boldsymbol{Q}}_1 = \boldsymbol{Q}_2$ and $\mathbf{\Lambda}_2^{-1} = \overline{\mathbf{\Lambda}_1^{-1}}$, we have $\overline{\mathbf{A}^{-1}} = (\bar{\mathbf{A}})^{-1}$ according to the Theorem 3. □

C.3 Apply efficient matrix operations to logistical discrete system

There is an inapplicable problem for the two efficient matrix operations in logistical discrete system. The two efficient matrix operations is designed for matrix with diagonal-block structure because such matrix can be effectively compressed as block projective matrix without any information loss (see Fig. 5 in the main paper). However, the logistical discrete system introduce variational parameter vector $\boldsymbol{\xi}$, and the corresponding matrix $\boldsymbol{\pi}(\boldsymbol{\xi})$ is not a diagonal-block matrix because the diagonal entries in its square sub-matrices are determined by $\boldsymbol{\xi}$ and not the same. Since $\boldsymbol{\pi}(\boldsymbol{\xi})$ is a part of both $\mathbf{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\boldsymbol{\xi})\mathbf{\Phi}$ and the covariance matrix $\mathbf{\Sigma}_s$, $\mathbf{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\boldsymbol{\xi})\mathbf{\Phi}$ and $\mathbf{\Sigma}_s$ become not diagonal-block matrices. Thus, the efficient matrix operations no longer apply to calculating the multiplication $\mathbf{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\xi)\mathbf{\Phi}$ and the inverse $\mathbf{\Sigma}_s^{-1}$ directly.

Based on above analysis, we can see the inapplicable problem is caused by the introduced variational parameter $\boldsymbol{\xi}$, and the problem will be addressed if we can modify $\boldsymbol{\xi}$ so as to shape $\boldsymbol{\pi}(\boldsymbol{\xi})$ being a diagonal-block matrix. By doing so, both $\mathbf{\Phi}^{\mathrm{T}}\boldsymbol{\pi}(\xi)\mathbf{\Phi}$ and $\mathbf{\Sigma}_s$ will keep diagonal-block structure and the efficient matrix operations can be applied. Toward this end, we model the variational parameter $\boldsymbol{\xi} \in \mathcal{R}^{TN}$ being with the following particular group structure in accelerated algorithm,

$$\boldsymbol{\xi} = \big[\underbrace{\hat{\xi}_1, \ldots, \hat{\xi}_1}_{N}, \underbrace{\hat{\xi}_2, \ldots, \hat{\xi}_2}_{N}, \ldots, \underbrace{\hat{\xi}_T, \ldots, \hat{\xi}_T}_{N}\big]^{\mathrm{T}} \tag{5}$$

Specially, we let the N variational parameters for the N dynamic data at time $t, t \in [1, \ldots, T]$ be the same and equal to $\hat{\xi}_t$, i.e., $\boldsymbol{\xi}_{(t-1)N+i} = \hat{\xi}_t, \forall i \in [1, \ldots, N]$. After such modelling, the corresponding matrix $\boldsymbol{\pi}(\boldsymbol{\xi})$ becomes a diagonal-block matrix and the two efficient matrix operations can apply to the logistical discrete system.

From Taylor expansion perspective, the variational parameter vector $\boldsymbol{\xi}$ consists of the contact points where the lower bound approximation functions are tangent to the original functions [1]. Thus, the modelling for $\boldsymbol{\xi}$ in accelerated algorithm is actually to expand the variational lower bound functions, Eq. 2, on T points, instead of TN in original algorithm. Here, decreasing contact points may influence adversely the variational approximation because the lower bound functions are not expended on the most suitable points. However, experimental results of accelerated algorithm show the influence is limited.

In the case of new modelling for $\boldsymbol{\xi}$ in accelerated algorithm, the learning rule for variational parameter becomes

$$\hat{\xi}_t \leftarrow \frac{1}{N}\sum_{n=(t-1)N+1}^{tN} \sqrt{\mathbf{\Phi}_n(\Sigma_{\boldsymbol{s}} + \boldsymbol{\mu}_{\boldsymbol{s}}\boldsymbol{\mu}_{\boldsymbol{s}}^{\mathrm{T}})\mathbf{\Phi}_n^{\mathrm{T}}},\ t = 1, \ldots, T. \tag{6}$$

There is no change in learning rules for other hyper-parameters.

APPENDIX D
PLOT THE SPIKE-AND-SLAB PRIOR

In this section, we clarify how to plot the spike-and-slab prior distribution $p(\boldsymbol{s})$, as shown in Fig. 2 in the main text, based on an auxiliary function of the prior, Eq. 7 in the main text.

The Eq. 7 in the main text is not the probability density function (PDF) of the spike-and-slab prior $p(\boldsymbol{s})$, but is an auxiliary function being proportional to the prior. The prior actually cannot be calculated analytically because of the intractable limit operation in the Eq. 7 when $a \to 0$ and $b \to 0$. Instead of using PDF to represent the prior, we uncover and illustrate the shape of the prior by introducing an auxiliary function which is proportional to the PDF of prior and easy to calculate, as shown in the right side of Eq. 7 in the main text. The auxiliary function has a same shape with the prior because it is proportional to the prior PDF. Therefore, the prior distribution can be illustrated by plotting the auxiliary function, as shown in Fig. 2 in the main text.

To plot the auxiliary function in a 3-D space, there are two steps, sampling and interpolating. In sampling step, we random sample a large amount of points (i.e., (x, y)) in a 2-D feasible region of the auxiliary function, i.e., $\{(x, y)||x| < l, |y| < l, x \neq 0, y \neq 0\}$, which is a square region without the points on every axis. Then we calculate the function value (i.e., z) on every sampled point though feeding the point to the auxiliary function, i.e., the right side of Eq. 7 in the main text. After standardizing the summation of all function values z to one, we obtain the final data samples (i.e., (x, y, z)) outlining the auxiliary function.

In step two, we employ a smoother interpolating method in MATLAB toolbox to form a continuous 2D surface from the discrete data samples. This surface is then plotted in Fig. 2 in the main text, from which we find two nice properties of the prior distribution to induce group sparsity of parameters.

REFERENCES

[1] C. M. Bishop, *Pattern Recognition and Machine Learning*. Berlin: Springer, 2006.

Major Differences between our Preliminary AAAI Paper and the New TPAMI Submission

A preliminary version for part of the current work was presented as a regular paper at the *32nd AAAI Conference on Artificial Intelligence (AAAI'18)*. This new TPAMI submission is a "more than 70% (more than 13 out of 18 pages were newly added or updated)" substantial revision of our preliminary conference publication.

The key differences between the two are summarized as follows:

(1) In the current version, we provided a theoretical supporting for the proposed active surveillance method. We first revealed that the proposed prior can enforce group sparse patterns by presenting its two important properties of being concave and monotonic, and heavy-tailed (please see Section 3.2.1). Next, from a geometric perspective, we showed that the proposed measure is capable of determining the monitoring priority of a component for dynamics prediction (please see Section 3.2.3).

(2) We enhanced empirical validations for the proposed method. We compared the proposed method with three new state-of-the-art methods, namely: FrameSense, MNEP, and MPME. We further added two new real-world datasets, i.e., temperature measurements in Intel Berkeley Lab and thermal diffusion on UltraSPARC T1 microprocessor. Besides, we proposed a novel evaluate criterion, i.e., cumulative mean square error, and redesigned our validation scheme. We rewrote entirely the experimentation section (please see Section 4).

(3) We newly added an Appendix to describe the technical details of the proposed method. The Appendix includes the derivation of posterior distribution (Section A), the derivation of hyper-parameters estimation (Section B), the derivation of efficient matrix operations (Section C), and plots of the spike-and-slab prior (Section D). This appendix was not officially included in the previous conference paper (please see the Appendix).

(4) We entirely rewrote the section on related work, by surveying and adding many related studies and giving a new taxonomy to reorganize the existing studies as related to active surveillance. Besides, we further clarified the novelty of the proposed method (please see Section 2).

(5) We completely rewrote the sections of abstract, introduction, and conclusion in the new version, respectively, to make them consistent with the aforementioned improvements, which are now entirely different from those in the preliminary conference paper (please see Abstract, and Introduction and Conclusion in Section 1 and 5, respectively).

(6) We newly added a discussion section, i.e., the connection to block sparse Bayesian learning, to distinguish the group sparse Bayesian modelling in this submission from the existing block sparse Bayesian learning methods (please see Section 5.1).

We enclose the preliminary AAAI conference paper in the following pages for your further reference.

The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Group Sparse Bayesian Learning for Active Surveillance on Epidemic Dynamics

Hongbin Pei,[1,2] **Bo Yang,**[1,2*] **Jiming Liu,**[3] **Lei Dong**[4]
[1]College of Computer Science and Technology, Jilin University, Changchun, China
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
[3]Department of Computer Science, Hong Kong Baptist University, Hong Kong
[4]Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing, China

Abstract

Predicting epidemic dynamics is of great value in understanding and controlling diffusion processes, such as infectious disease spread and information propagation. This task is intractable, especially when surveillance resources are very limited. To address the challenge, we study the problem of active surveillance, i.e., how to identify a small portion of system components as sentinels to effect monitoring, such that the epidemic dynamics of an entire system can be readily predicted from the partial data collected by such sentinels. We propose a novel measure, the γ value, to identify the sentinels by modeling a sentinel network with row sparsity structure. We design a flexible group sparse Bayesian learning algorithm to mine the sentinel network suitable for handling both linear and non-linear dynamical systems by using the expectation maximization method and variational approximation. The efficacy of the proposed algorithm is theoretically analyzed and empirically validated using both synthetic and real-world data.

1 Introduction

Predicting epidemic dynamics is of great value in understanding and controlling diffusion processes. Diffusion phenomena, such as infectious disease spread and information propagation, exist widely in the real world. By the notion of a dynamical system, which is a powerful tool for characterizing dynamics (Brunton, Proctor, and Kutz 2016), the task of epidemic dynamics prediction is to estimate ensuing system states using given current states. Therefore, the foundation of this task is surveillance, which is to monitor and report the current system states in a timely manner.

However, in practice, it is often very challenging to monitor the overall components in a system because diffusion phenomena often cross over very large spatiotemporal ranges, e.g., spreading of infectious diseases in a country (Polgreen et al. 2009), air contaminant diffusion in a large city (Zheng, Liu, and Hsieh 2013), and hot topics/meme forwarding on social media (Chen et al. 2013). For large-scale diffusion phenomena, timely and all-around monitoring is hard and infeasible, especially when available surveillance resources are very limited.

*Corresponding author. ybo@jlu.edu.cn

Owing to a lack of systematic deployment of already limited surveillance resources, disease surveillance has long suffered from low reporting rates, biased sampling, and lengthy reporting time-lags (Gerardo-Giorda et al. 2013). For instance, Tengchong city, a malaria endemic region in China, has 18 towns (consisting of 221 villages), 167,964 households, and 658,207 residents that are distributed in a mountainous area of 5,845 square kilometers. From 2005 to 2011 in Tengchong, 7,835 confirmed malaria cases were reported, but the Tengchong Centers for Disease Control (CDC), the local disease surveillance team, could only afford to have a few staff members conducting the time-consuming case surveys.

Active surveillance is a promising strategy to address the challenge of limited surveillance resources, in which epidemic dynamics is predicted by proactively monitoring a relatively small number of sentinel components, whose state data are collected to achieve a good trade-off between prediction accuracy and surveillance cost. The key to implementing active surveillance is to determine, in a dynamical system, which components are important for dynamics prediction and how to identify them. This is a non-trivial task because the interaction structure among components, which characterizes the dynamics mechanism, is usually hidden and heterogenous, such as the social contact network in charge of disease spread (Yang et al. 2017).

Here, we address this challenge by proposing a novel importance measure, the γ value, to determine how important a component is to predicting epidemic dynamics. Based on the measure, we develop a backward-selection algorithm designated the sentinel network mining algorithm (SNMA for short) to mine a sentinel network. The sentinel network encodes the influence relationship from sentinel components to the overall system components, which is a row sparse network that contains only the influential links emitted from the sentinels. With the discovered sentinel network, one can predict the future overall system dynamics by only monitoring and feeding the sentinels' states into a system function.

Different from existing works, we model the task of sentinel network mining as a group sparse learning problem and propose an effective and flexible Bayesian learning algorithm for various dynamical systems, including the most widely used linear continuous system and logistic discrete system. The expectation-maximization (EM) and variational

approximation methods are employed to infer the posterior distribution of the sentinel network. Particularly, we solved the scalability problem by means of proposing more efficient multiplication and inverse operations on diagonal-block metrics. Validations and comparisons were performed on both synthetic and real-world data, which show that the proposed method outperforms existing methods.

We summarize the main contributions of this paper as follows: (1) We propose a novel measure, the γ value, to identify sentinel components by modeling the sentinel network with row sparsity structure; (2) We design an effective and flexible group sparse Bayesian learning algorithm to discover the sentinel network; (3) We solve the scalability and generality problem of this algorithm to a certain degree; (4)We develop a comprehensive framework for active surveillance and validate it by performing extensive experiments on both synthetic and real-world data.

2 Related works

In the literature, most of the optimal sensor placement studies identify the sentinels' location via the framework of the budgeted maximum coverage problem (BMCP) (Khuller, Moss, and Naor 1999), which aims to maximize a specific objective by finding a set of components with the minimum budget (the budget is often defined as the set size). Examples include the timely detection of contaminated water (Krause et al. 2008) and early detection of the outbreak of Weblogs (Leskovec et al. 2007). As BMCP is NP-hard, heuristic algorithms like sub-modular maximization are often employed. Note that designing the objective functions of BMCP relies heavily on the known interaction structure among components. When the underlying interaction structure cannot be observed directly, such as the social media network in charge of information diffusion and the social contact network for disease spread, the current methods proposed within the BMCP framework cannot be used (Gomez Rodriguez, Leskovec, and Krause 2010).

In spatial statistics, Gaussian processes (GPs) can effective represent the spatial correlation and uncertainty of the sensed field. Based on GPs, one can adopt general information criteria, typically such as mutual information (GPs-MI), to greedily select sentinels (Krause, Singh, and Guestrin 2008; Hoang et al. 2014). In this framework, both the GPs and the information criteria are model-free. They neglect the mechanism of data generation. As a result, the prior knowledge of the phenomena [e.g., the susceptible-infectious-recovered (SIR) model for infectious disease (Dimitrov and Meyers 2010)] is difficult to integrate into the method. If such available prior knowledge can be adequately incorporated, the performance of learning and prediction will be significantly improved, as shown in our experiments.

In addition to the epidemic dynamics data, some works turn to other types of available data to help identify the sentinels. For instance, socio-economic data is leveraged to estimate the malaria infection risk caused by imported cases, and further to implement an effective active surveillance plan (Yang et al. 2014). Traffic data and environmental data are used to infer real-time air quality, and further determine the best deployment locations of new air contaminant monitoring stations (Hsieh, Lin, and Zheng 2015). It is difficult to reuse these customized methods on other types of active surveillance applications, if the domain data or domain knowledge they require are not available.

3 Active surveillance framework

For epidemic dynamics, we now propose the framework of active surveillance. It consists of three main steps.

Step 1: collect epidemic dynamics data in N components of interest.

Step 2: mine the sentinel network from the data. In the network, the number of sentinel components (i.e., sentinel nodes) k is according to a budget.

Step 3: with the sentinel network, predict future epidemic dynamics of the N components based on the data collected from the k sentinel components.

The last two steps constitute the foundation of the framework, and we will elaborate them in following sections.

3.1 Problem formulation

Consider a diffusion among N components in a dynamical system. Let matrix $\mathbf{D} \in \mathbb{R}^{T\times N}$=$[\mathbf{D}_1, \cdots, \mathbf{D}_T]^{\mathrm{T}}$ be the epidemic dynamics during a time window $[1, T]$. Specifically, $\mathbf{D}_t$=$[\mathbf{D}_{t,1}, \cdots, \mathbf{D}_{t,N}]$, where each entry $\mathbf{D}_{t,i}$ denotes the state of component i at time t, and it may be a real number (e.g., the number of newly infected cases in the city i) or a Boolean value (e.g., whether a news is posted in the blog i). Let $\mathbf{D}^{\boldsymbol{s}} \in \mathbb{R}^{T\times N}$ denote the surveillance data collected by k sentinel components. Specifically, $\mathbf{D}^{\boldsymbol{s}}_{t,i}$ is equal to $\mathbf{D}_{t,i}$ when component i is a sentinel, and empty otherwise.

Let $f(\mathbf{D}^{\boldsymbol{s}}_t; \mathbf{S})$ be the dynamical system function achieving the dynamics prediction. Let matrix $\mathbf{S} \in \mathbb{R}^{N\times N}$ denote a sentinel network, a set of key parameters in the system function. It depicts the influential relationship from the sentinels to all components, where each link $\mathbf{S}_{i,j}$ encodes the effect of sentinel i on component j by its weight. Thus, $\mathbf{S}$ is a row sparse matrix only containing links emitted from the k sentinels. Now, the active surveillance can be formulated as to predict the future components' states based on the surveillance data $\mathbf{D}^{\boldsymbol{s}}$ and the sentinel network $\mathbf{S}$:

$$\mathbf{D}_{t+1} \approx \hat{\mathbf{D}}_{t+1} = f(\mathbf{D}^{\boldsymbol{s}}_t; \mathbf{S}). \tag{1}$$

From Eq. 1, two computational issues need to be addressed for the goal of active surveillance:

I) Sentinel identification: How to identify the sentinels from all components and mine the sentinel network $\mathbf{S}$ according to a given budget from the dynamics $\mathbf{D}$?

II) Sentinel prediction: How to predict the future dynamics $\mathbf{D}_{t+1}$ from the current surveillance data $\mathbf{D}^{\boldsymbol{s}}_t$ based on the discovered $\mathbf{S}$?

3.2 Sentinel identification

Our basic idea is intuitive: *In a dynamical system, the components having little influence on others are unimportant for predicting others' states, while those exerting a heavy influence on others dominate the system dynamics and should be selected as sentinels.* In terms of the sentinel network $\mathbf{S}$, one

can determine whether a component is important or not by inferring row sparsity. That is, unimportant components are associated with sparse rows in $\mathbf{S}$, in which zeros are much more than non-zeros; on the other hand, important ones are associated with non-sparse rows. Figure1 shows an illustration by taking a linear dynamical system as an example.

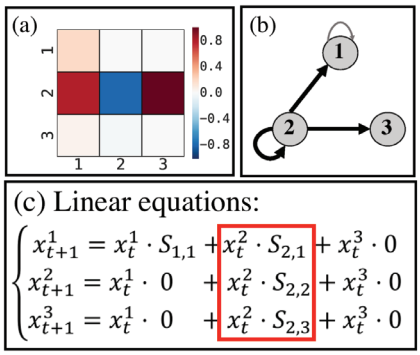


Figure 1: Unimportant = row sparse. (a) Sentinel network $\mathbf{S}$; (b) the graph of $\mathbf{S}$; (c) the equations of a linear dynamical system, where the component 2 dominates the system. Unimportant component 3 is associated with a sparse row.

Based on this idea, we propose a novel index, the γ value, to measure components' importance in predicting the epidemic dynamics: a component is important if it is important in both prior and posterior structures of a sentinel network. Specifically, the γ value is defined as the data-dependent hyper-parameter of the prior of the sentinel network, and also that reflecting the profiles of the posterior of the sentinel network:

$$\gamma_i = ((\boldsymbol{\mu}_{\boldsymbol{s}}^i)^{\mathrm{T}} \boldsymbol{\mu}_{\boldsymbol{s}}^i + \mathrm{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{s}}^i])N^{-1}, \tag{2}$$

where γ_i is the γ value of component i. In the following, we elaborate this importance measure from the perspectives of prior and posterior.

3.2.1 Prior perspective From the basic idea, a sentinel network is desired to have a row sparse structure. Thus, we adopt a zero-mean multivariate Gaussian prior for each row:

$$p(\mathbf{S}_{i,\cdot}|\gamma_i) \sim \mathcal{N}(\mathbf{0}, \gamma_i \mathbf{I}_N),\ i = 1, \ldots, N \tag{3}$$

where vector $\mathbf{S}_{i,\cdot} \in \mathbb{R}^N$ denotes the ith row of the sentinel network, and $\mathbf{I}_N \in \mathbb{R}^{N \times N}$ is an identity matrix. By doing so, γ_i controls the diversity of row i from a zero vector.

For conciseness, we vectorize matrix $\mathbf{S}$, i.e., let $\boldsymbol{s} = \mathrm{vec}(\mathbf{S}^{\mathbf{T}})$, where operator $\mathrm{vec}(\cdot)$ denotes the vectorization of the input matrix by stacking its columns into a column vector. By doing so, vector $\boldsymbol{s} \in \mathbb{R}^{N \times N}$ consists of N groups of length N, where each group is associated with a row in $\mathbf{S}$. Now, the row sparse structure of $\mathbf{S}$ is equal to the group sparse structure of $\boldsymbol{s}$. In terms of the prior on $\mathbf{S}$ (Eq. 3), the prior over $\boldsymbol{s}$ is

$$p(\boldsymbol{s}|\boldsymbol{\gamma}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0), \tag{4}$$

where vector $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_N)^{\mathrm{T}}$ and the covariance matrix $\boldsymbol{\Sigma}_0 \in \mathbb{R}^{N^2 \times N^2}$ is a diagonal matrix:

$$\boldsymbol{\Sigma}_0 = \begin{bmatrix} \gamma_1 \mathbf{I}_N & & \\ & \ddots & \\ & & \gamma_N \mathbf{I}_N \end{bmatrix}. \tag{5}$$

As mentioned before, the links sent from a component reflect its effect on the system dynamics. Now, the links sent from i in $\mathbf{S}$ (i.e., the entries of group i in $\boldsymbol{s}$) are tied together and controlled by a common data-dependent hyper-parameter γ_i. This hyper-parameter is a type of automatic relevance determination (ARD) mechanism (MacKay 1995): when γ_i is small, the group i in $\boldsymbol{s}$ is sparse and vice versa; when the group i is sparse, the links sent from i are very weak in $\mathbf{S}$. That is, component i is unimportant and can be pruned out without losing much in prediction accuracy.

3.2.2 Posterior perspective The γ value also reflects the profile of the posterior of the sentinel network. We model the sentinel network for two kinds of dynamical systems widely used to characterize diffusion phenomena in the real world: a linear continuous system and a logistical discrete system.

Likelihood of linear system. Starting from the linear continuous system, we first give the likelihood function and illustrate the pre-processing of dynamics data. The system function of a linear continuous system is

$$\mathbf{Y} = \mathbf{X}\mathbf{S} + \mathbf{V}, \tag{6}$$

where $\mathbf{Y} = \mathbf{D}_{2:T+1} \in \mathbb{R}^{T \times N}$ and $\mathbf{X} = \mathbf{D}_{1:T} \in \mathbb{R}^{T \times N}$ are both extracted from the dynamics data $\mathbf{D}$. $\mathbf{Y}$ is the epidemic dynamics later than $\mathbf{X}$ one time-unit. Specifically, $\mathbf{Y}_{t-1} = \mathbf{X}_t = \mathbf{D}_t$ and $\mathbf{V}$ is a Gaussian noise matrix.

For convenience, we further transform Eq. 6 into a vector form, $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{s} + \mathbf{v}$, where vector $\boldsymbol{y} = \mathrm{vec}(\mathbf{Y}^{\mathrm{T}}) \in \mathbb{R}^{TN \times 1}$, $\boldsymbol{s} = \mathrm{vec}(\mathbf{S}^{\mathbf{T}}) \in \mathbb{R}^{N^2 \times 1}$, and $\mathbf{v} = \mathrm{vec}(\mathbf{V}^{\mathrm{T}}) \in \mathbb{R}^{TN \times 1}$. The matrix $\boldsymbol{\Phi} = \mathrm{Kron}(\mathbf{X}, \mathbf{I}_N) \in \mathbb{R}^{TN \times N^2}$, where the operator $\mathrm{Kron}(\cdot,\cdot)$ represents the Kronecker product of two input matrices. Now, based on the Gaussian noise assumption, the likelihood of the linear continuous system can be given as

$$p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{s}, \lambda) \sim \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{s}, \lambda \mathbf{I}), \tag{7}$$

where λ denotes the noise level and $\mathbf{I}$ is an identity matrix.

Likelihood of logistical system. For the logistical discrete system, the entry in the dynamics data $\mathbf{D}_{t,i}$ is represented by a Boolean value, 0 or 1, indicating whether component i is "infected" at time t. After the same data pre-processing, we adopt a Bernoulli distribution over each entry of $\boldsymbol{y}$, i.e., $\boldsymbol{y}_n$:

$$p(\boldsymbol{y}_n = 1|\boldsymbol{\Phi}_n, \boldsymbol{s}) = \sigma[\boldsymbol{\Phi}_n \boldsymbol{s}],\ n = 1, \ldots, TN, \tag{8}$$

where $\sigma[\boldsymbol{\Phi}_n \boldsymbol{s}] = 1/(1 + e^{-\boldsymbol{\Phi}_n \boldsymbol{s}})$ denotes the sigmoid function, and $\boldsymbol{\Phi}_n$ is the nth row of $\boldsymbol{\Phi}$. Then, the likelihood of the dynamics can be written as

$$p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{s}) = \prod_{n=1}^{TN} \sigma[\boldsymbol{\Phi}_n \boldsymbol{s}]^{\boldsymbol{y}_n} (1 - \sigma[\boldsymbol{\Phi}_n \boldsymbol{s}])^{1 - \boldsymbol{y}_n}. \tag{9}$$

Now, based on the aforementioned prior and likelihood we have the following conclusion about the posterior:

Theorem 1. *For both the linear system and logistical system, the posteriors of the sentinel network are a Gaussian or an approximate Gaussian:* $p(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\Phi}, \Theta) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{s}}, \boldsymbol{\Sigma}_{\boldsymbol{s}})$:

(1) Linear system: hyper-parameters set $\Theta = \{\boldsymbol{\gamma}, \lambda\}$,

$$\boldsymbol{\mu}_{\boldsymbol{s}} = \lambda^{-1} \boldsymbol{\Sigma}_{\boldsymbol{s}} \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{y},\ \boldsymbol{\Sigma}_{\boldsymbol{s}}^{-1} = \boldsymbol{\Sigma}_0^{-1} + \lambda^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi};$$

(2) Logistical system: hyper-parameters set $\Theta = \{\boldsymbol{\gamma}, \boldsymbol{\xi}\}$,

$$\boldsymbol{\mu}_{\boldsymbol{s}} = 2^{-1} \boldsymbol{\Sigma}_{\boldsymbol{s}} \boldsymbol{\Phi}^{\mathrm{T}} (2\boldsymbol{y} - \mathbf{1}),\ \boldsymbol{\Sigma}_{\boldsymbol{s}}^{-1} = \boldsymbol{\Sigma}_0^{-1} + \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\pi}(\boldsymbol{\xi}) \boldsymbol{\Phi},$$

where $\boldsymbol{\xi}=[\xi_1,\ldots,\xi_{TN}]^{\mathrm{T}}$ *denotes variational parameters, and* $\boldsymbol{\pi}(\boldsymbol{\xi})\in\mathbb{R}^{TN\times TN}$ *is a diagonal matrix. Specifically,* $\boldsymbol{\pi}(\boldsymbol{\xi})_{n,n}=-\frac{1}{2\xi_n}(\sigma[\xi_n]-\frac{1}{2})$.

Proof. See the Supporting Information[1] □

Estimation of hyper-parameters. Theorem 1 gives the posteriors of the two kinds of systems (i.e., the Gaussian mean $\boldsymbol{\mu_s}$ and covariance matrix $\boldsymbol{\Sigma_s}$). In the following, we study how to iteratively update the hyper-parameters in the posteriors, because they cannot be obtained in a closed form. By treating the $\boldsymbol{s}$ as hidden variables, we can employ the EM method to estimate the hyper-parameters set Θ. When the integral in the Q function can be analytically solved (in the linear system), the EM obtains a well-formed solution to estimating the hyper-parameters; otherwise, we apply variational EM to estimate the hyper-parameters by optimizing an approximate low bound of the posterior (in the logistical system). Furthermore, the EM method can guarantee the convergence of the estimation.

Here, we only give the learning rules for the hyper-parameters since the limited space (derivation details can be found in the Supporting Information). In the linear continuous system, the learning rule for the noise parameter λ is

$$\lambda\leftarrow(TN)^{-1}(\|\boldsymbol{y}-\boldsymbol{\Phi}\mu_{\boldsymbol{s}}\|_2^2+\mathrm{Tr}[\boldsymbol{\Sigma_s}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}]). \tag{10}$$

In the logistical discrete system, we update the variational parameter vector $\boldsymbol{\xi}$ by

$$\xi_n\leftarrow\sqrt{\boldsymbol{\Phi}_n(\Sigma_{\boldsymbol{s}}+\boldsymbol{\mu_s}\boldsymbol{\mu_s}^{\mathrm{T}})\boldsymbol{\Phi}_n^{\mathrm{T}}},\ n=1,\ldots,TN. \tag{11}$$

It is extremely interesting that in both the linear and logistical systems the learning rule for the $\boldsymbol{\gamma}$ value is the same,

$$\boldsymbol{\gamma}_i\leftarrow((\boldsymbol{\mu_s^i})^{\mathrm{T}}\boldsymbol{\mu_s^i}+\mathrm{Tr}[\boldsymbol{\Sigma_s^i}])N^{-1},\ i=1,\ldots,N, \tag{12}$$

where the vector $\boldsymbol{\mu_s^i}\in\mathbb{R}^N$ and matrix $\boldsymbol{\Sigma_s^i}\in\mathbb{R}^{N\times N}$ denote the ith group of $\boldsymbol{\mu_s}$ and $\boldsymbol{\Sigma_s}$, respectively.

Intuitively, there are two terms that contribute to the $\boldsymbol{\gamma}$ value according to its learning rule, Eq .12. The first term is the inner product term $(\boldsymbol{\mu_s^i})^{\mathrm{T}}\boldsymbol{\mu_s^i}$, which denotes the sum of the squares of the mean weights of the overall links sent from node i in the sentinel network. In other words, it characterizes the *influence strength* of component i. The second term is the trace term $\mathrm{Tr}[\boldsymbol{\Sigma_s^i}]$, which denotes the variance of the posterior estimation on links sent from node i; that is to say, it features the *influence uncertainty* of component i. In summary, a larger $\boldsymbol{\gamma}$ value corresponds to a node that has many links with large and diverse influences on other nodes.

On the whole, the $\boldsymbol{\gamma}$ value is an index by which the importance of a component for predicting the epidemic dynamics of an entire system can be measured. This index integrates the profiles of both the prior and posterior of a sentinel network. For a trivial component in the system, its $\boldsymbol{\gamma}$ value will tend to be zero due to the ARD mechanism during Bayesian learning. For an important component, whose $\boldsymbol{\gamma}$ value larger than zero, its $\boldsymbol{\gamma}$ value could indicate its monitoring priority. A component with a larger $\boldsymbol{\gamma}$ value exerts a great influence on other components, and its state is important to monitor for making a prediction.

Based on the $\boldsymbol{\gamma}$ value, we propose a backward-selection algorithm called the SNMA, as shown in Algorithm 14. It starts with all N components of interest and removes one component at a time until only k components are left (k is according the budgets). The component that is removed should be chosen as the one with the minimum $\boldsymbol{\gamma}$ value. The form of backward-selection algorithm is theoretically guaranteed to pick a optimal subset of components if the system perturbation is small enough (Couvreur and Bresler 2000). A trade-off between accuracy and budget is practically necessary: the more sentinels are selected, the more predictive accuracy is expected, while more cost is needed.

Algorithm 1: SNMA

Input: epidemic dynamics $\mathbf{D}$, quantity of components of interest N, quantity of sentinels k;
Output: posterior structure of sentinel network, i.e., mean vector $\boldsymbol{\mu_s}$, covariance matrix $\boldsymbol{\Sigma_s}$;

```
Pre-processing: extract y and Φ from D;
Randomly initialize γ, λ (the linear) or ξ (the logistical);
L ← N;
while L > k do
    while γ is not converged do
        // The optimization step
        update μ_s and Σ_s via Theorem 1;
        update γ, λ (the linear) or γ, ξ (the logistical) via
          Eq. 12,10 and 11;
    end
    find the minimum entry i in the vector γ
      // The selection step
    update Φ, μ_s, Σ_s, λ γ, ξ through pruning out the
      entries of ith group in them;
    L ← L − 1;
end
```

3.3 Sentinel prediction

Once we have obtained the posterior structure of the sentinel network, the epidemic dynamics of the overall system, $\mathbf{D}$, can be predicted based on the surveillance data $\mathbf{D}^{\boldsymbol{s}}$. Let $\mathbf{D}_*^{\boldsymbol{s}}$ be a new set of surveillance data, where only the values on k sentinels' locations are kept and the rest are empty. As mentioned above, we obtain $\boldsymbol{\Phi}_*^{\boldsymbol{s}}$ through the data preprocessing. Then, a predictive distribution over the following system states $\boldsymbol{y}_*$ is given by

$$p(\boldsymbol{y}_*|\boldsymbol{\Phi}_*^{\boldsymbol{s}},\boldsymbol{y},\boldsymbol{\Phi})=\int p(\boldsymbol{y}_*|\boldsymbol{\Phi}_*^{\boldsymbol{s}},\boldsymbol{s})p(\boldsymbol{s}|\boldsymbol{y},\boldsymbol{\Phi})d\boldsymbol{s}. \tag{13}$$

Linear continuous system. In this case, the integral in Eq. 13 is a Gaussian convolution (refer to the proof of Theorem 1), whose analytical solution is a Gaussian. Then, we have

$$p(\boldsymbol{y}_*|\boldsymbol{\Phi}_*^{\boldsymbol{s}},\boldsymbol{y},\boldsymbol{\Phi})\sim\mathcal{N}(\mu_{y_*},\sigma_{y_*}^2)$$

with parameters $\mu_{y_*}=\boldsymbol{\Phi}_*^{\boldsymbol{s}}\boldsymbol{\mu_s}$, $\sigma_{\boldsymbol{y_*}}^2=\lambda+\boldsymbol{\Phi}_*^{\boldsymbol{s}}\boldsymbol{\Sigma_s}(\boldsymbol{\Phi}_*^{\boldsymbol{s}})^{\mathrm{T}}$.

Logistical discrete system. The predictive distribution of discrete data is a Bernoulli distribution. By substituting the

[1]The supporting information of this paper, *Supporting Information: Group Sparse Bayesian Learning for Active Surveillance on Epidemic Dynamics*, have posted on arXiv.org.

term of the posterior of $\boldsymbol{s}$ in Eq. 13 for the variational approximation posterior given in Theorem 1, we have the following predictive distribution:

$$p(\boldsymbol{y}_{*n} = 1|\boldsymbol{\Phi}_*^{\boldsymbol{s}}, \boldsymbol{y}, \boldsymbol{\Phi}) \approx \int p(\boldsymbol{y}_{*n} = 1|\boldsymbol{\Phi}_*^{\boldsymbol{s}}, \boldsymbol{s}) q(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\Phi}) d\boldsymbol{s},$$

where $n = 1 \cdots N$.

However, the integral in this approximation is a convolution of a logistical function with a Gaussian distribution, which cannot be analytically integrated. By introducing an error function, it can be approximated as a re-parameterized logistical function (Maragakis et al. 2008):

$$\int p(\boldsymbol{y}_*{=}1|\boldsymbol{\Phi}_*^{\boldsymbol{s}}, \boldsymbol{s}) q(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{\Phi}) d\boldsymbol{s} \approx (1 + e^{-\tau \boldsymbol{\Phi}_*^{\boldsymbol{s}} \boldsymbol{\mu}_{\boldsymbol{s}}})^{-1}, \quad (14)$$

where $\tau = (1 + \frac{\pi}{8} \boldsymbol{\Phi}_*^{\boldsymbol{s}} \boldsymbol{\Sigma}_{\boldsymbol{s}} (\boldsymbol{\Phi}_*^{\boldsymbol{s}})^{\mathrm{T}})^{-\frac{1}{2}}$. A very small approximation error is guaranteed in theory (Maragakis et al. 2008).

3.4 Scaling up

In the SNMA, the most expensive operations are the matrix multiplication ($\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$) and the matrix inverse ($\boldsymbol{\Sigma}_{\boldsymbol{s}}^{-1}$) for calculating the posterior parameters in Theorem 1. Considering the large size of $\boldsymbol{\Phi} \in \mathbb{R}^{TN \times N^2}$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}} \in \mathbb{R}^{N^2 \times N^2}$, the time complexity of one iteration will be $O(N^5 \times \max(T, N))$, making it infeasible to handle real-world problems. We now propose two fast matrix operations to solve the scalability problem. The time complexity after scaling up is reduced to $O(N^2 \times \max(T, N))$, which is faster than the competitors in the experiments.

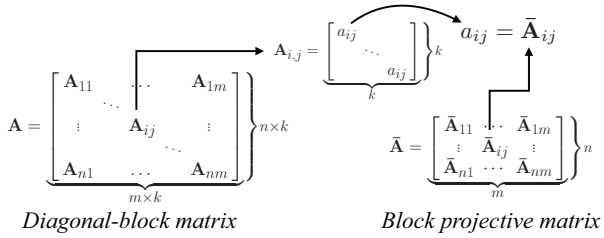


Figure 2: Illustration of the relationship between a diagonal-block matrix and block projective matrix.

Definition 1. *(Diagonal-block matrix). Matrix* $\mathbf{A} \in \mathbb{R}^{nk \times mk}$ *is called a diagonal-block matrix if it consists of* n *by* m *square sub-matrices, and each sub-matrix* $\mathbf{A}_{i,j} \in \mathbb{R}^{k \times k}$ *is a diagonal matrix with all diagonal entries having the same value* a_{ij}, i=$1 : n, j$=$1 : m$.

Definition 2. *(Block projective matrix). Let* $\mathbf{A} \in \mathbb{R}^{nk \times mk}$ *be a diagonal-block matrix. Matrix* $\bar{\mathbf{A}} \in \mathbb{R}^{n \times m}$ *is called the block projective matrix of* $\mathbf{A}$ *if each entry* $\bar{\mathbf{A}}_{i,j} = a_{ij}$, i=$1 : n, j$=$1 : m$.

The structure of the two matrices and their relationship is shown in Fig. 2. The block projective matrix $\bar{\mathbf{A}}$ is much smaller than its original diagonal-block matrix $\mathbf{A}$. In the meantime, it contains the overall distinct elements in $\mathbf{A}$. Based on the definitions, we have the following theorems, which can significantly reduce the computational cost of the two kinds of operations.

Theorem 2. *Let* $\mathbf{A}, \mathbf{B}$ *be two diagonal-block matrices with the same size of sub-matrix. The product* $\mathbf{AB} = \mathbf{C}$ *is a diagonal-block matrix, and those block projective matrices satisfy* $\bar{\mathbf{A}}\bar{\mathbf{B}} = \bar{\mathbf{C}}$.

Proof. See the Supporting Information. □

Theorem 3. *Let* $\mathbf{A}$ *be a square diagonal-block matrix. Its inverse matrix* $\mathbf{A}^{-1}$ *is also a diagonal-block matrix and satisfies* $(\mathbf{A}^{-1}) = (\bar{\mathbf{A}})^{-1}$.

Proof. See the Supporting Information. □

In the SNMA, $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}_{\boldsymbol{s}}$ are two diagonal-block matrices. When we alternatively calculate the multiplication ($\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}$) and the matrix inverse ($\boldsymbol{\Sigma}_{\boldsymbol{s}}^{-1}$) on the block projective matrices $\bar{\boldsymbol{\Phi}} \in \mathbb{R}^{T \times N}$ and $\bar{\boldsymbol{\Sigma}}_{\boldsymbol{s}} \in \mathbb{R}^{N \times N}$ via the two theorems, the time complexity of one iteration can be reduced by 3 orders of magnitude (i.e., N^3). Note that this scaling-up technique not only solves the difficulty faced by our algorithm, but can also address other tasks involving diagonal-block matrices, such as the computational obstacle of a multiple measurement vector (MMV) model in compressed sensing (Zhang and Rao 2011). To process further large-scale data in practice, the SNMA can be readily parallelized by adopting the group testing strategy, which has been used for parallel feature selection (Zhou et al. 2014).

3.5 Embedding non-linear dynamical models

The proposed framework is flexible and can be readily extended to various dynamical systems by mean of embedding basic functions, as long as the dynamical system functions can be represented as the combinations of basic functions. Each basic function, $\psi_i(x), i \in [1, \cdots, m]$, can be an arbitrary function, e.g., polynomial, $\psi_i(x)$= $x^2 + x$; trigonometric, $\psi_i(x)$= $\sin(x)$; or others.

We illustrate the embedding technology based on the linear system Eq. 6 as an example, where $\mathbf{X}_{t,i}$ denotes the state of component i at time t. Let vector $\Psi(\mathbf{X}_{t,i}) = [\psi_1(\mathbf{X}_{t,i}), \cdots, \psi_m(\mathbf{X}_{t,i})] \in \mathbb{R}^m$ be the mapping of $\mathbf{X}_{t,i}$ through m basic functions. Let vector $\Psi(\mathbf{X}_t) = [\Psi(\mathbf{X}_{t,1}), \cdots, \Psi(\mathbf{X}_{t,N})] \in \mathbb{R}^{mN}$ denote the mapping of all components at time t, and matrix $\Psi(\mathbf{X}) = [\Psi(\mathbf{X}_1); \cdots; \Psi(\mathbf{X}_T)]^{\mathrm{T}} \in \mathbb{R}^{T \times mN}$ denote the mapping of all components during the time window. Then, the system function Eq.6 can be extended as $\mathbf{Y} = \Psi(\mathbf{X})\dot{\mathbf{S}} + \mathbf{V}$, where the augmented sentinel network $\dot{\mathbf{S}} \in \mathbb{R}^{mN \times N}$ characterizes the interactions from the mN mapping to the N components. This extended equation can also represent non-linear dynamical systems if non-linear basic functions are adopted, such as the SIR model, a well-studied and widely adopted disease spread model. Moreover, the basic functions of the SIR model can be constructed based on its reproduction matrix form (Wallinga, van Boven, and Lipsitch 2010).

To integrate the above extended system function into the current SNMA, just let $\boldsymbol{\Phi} = \mathrm{Kron}(\Psi(\mathbf{X}), \mathbf{I}_N) \in \mathbb{R}^{TN \times mN^2}$ and $\boldsymbol{s} = \mathrm{vec}(\dot{\mathbf{S}})^{\mathrm{T}}$. Meanwhile, the group size needs be changed from N to mN. The rest of the steps are the same as the SNMA shows in Algorithm 14. For the logistical system, the extended process is analogous.

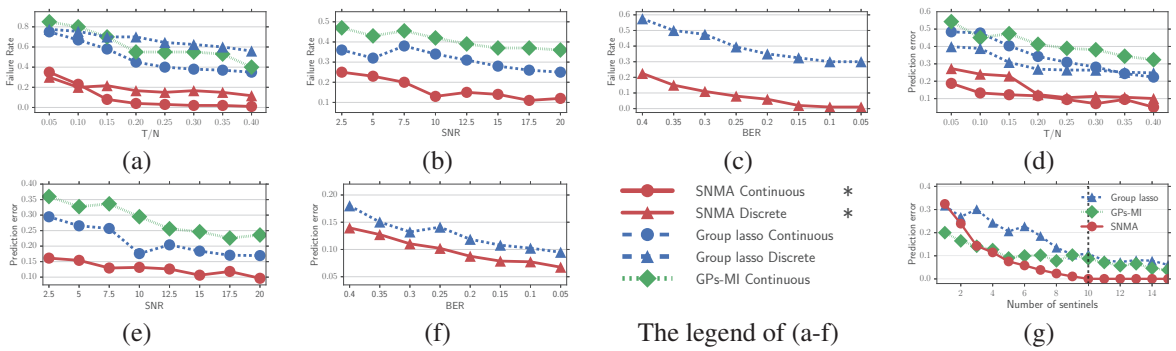


Figure 3: Results of experiments on synthetic data. ($*$ denotes our methods).

4 Validations

Comparative Study. We validate the framework on both synthetic and three real-world data. The two most related methods are selected as competitors: group Lasso and GPs-MI. Group Lasso (Meier, Van De Geer, and Bühlmann 2008) is a typical method for group sparse learning, which is similar to that addressed by our proposed group sparse Bayesian learning algorithm. To achieve the aim of active surveillance, we use the proposed framework and replace the group sparse Bayesian learning with group lasso.

GPs-MI is a popular sensor placement method (Krause, Singh, and Guestrin 2008; Hoang et al. 2014), which is similar to the task of active surveillance addressed by our work. It outperforms the placement methods based on experiment design, such as A-, D-, and E-optimal designs. As Gaussian processes cannot directly work on discrete data, GPs-MI is only applied on experiments with continuous data.

Evaluation. Two criteria are adopted to evaluate the performance of the methods. Failure rate is used to measure whether the a method can discover the sentinels, i.e., the problem of sentinel identification. Failure rate is the percentage of wrong sentinels' locations given by a method, $1 - |\mathbf{\Gamma} \cap \hat{\mathbf{\Gamma}}|/|\mathbf{\Gamma}|$, where $\mathbf{\Gamma}$ is the set of the ground truth sentinel locations in, and $\hat{\mathbf{\Gamma}}$ is the set discovered by, a method.

We use *root-mean-square error* (RMSE) to quantify the prediction error, i.e., the problem of sentinel prediction. Specifically, RMSE=$\frac{||\mathbf{Y}-\hat{\mathbf{Y}}||_2}{TN}$, where $\hat{\mathbf{Y}}$ is the predicted epidemic dynamics based on the surveillance data.

4.1 Validations on synthetic data

Synthetic Data Generation. We generate synthetic data by imagining diffusion processes taking place in a linear continuous system or logistical discrete system. There are three steps in total: (1) Random generation of a ground truth $\boldsymbol{\gamma}$-value vector with 500 entries, where 100 entries are sampled from $\mathcal{N}(0, 10)$ and the other 400 from $\mathcal{N}(0, 0.1)$, i.e., 100 sentinels and 400 trivial components; (2) Random sampling of a ground truth sentinel network $\mathbf{S}$ via the prior Eq.3 based on the ground truth $\boldsymbol{\gamma}$ value; (3) Based on the $\mathbf{S}$, simulation of the epidemic dynamics via the linear system Eq. 6 or the logistical system Eq. 8 embedding a quadratic basic function $\phi(x) = x^2 + x$.

Environment setting. The comparisons are conducted under various data volumes and noise levels. We use the value T/N to denote the ratio of data volume to the number of parameters to estimate. *Signal-to-noise ratio* (SRN) and *bit error rate* (BER) are adopted to denote noise levels in the linear system and logistical system, respectively.

We adopt a 5-fold cross-validation strategy in experiments on synthetic data. We firstly identify 100 sentinels from the training data via the three methods. For each method, the average failure rate of sentinel identification is shown in Figs. 3 (a-c). Then, we evaluate the performance of sentinel prediction by feeding the surveillance data (collected on the 100 discovered sentinels) to the corresponding prediction model of each method, such as Eq. 13 for the proposed method. The average prediction error is shown in Figs.3(d-f). The results show that the proposed methods are superior to GPs-MI and group lasso on both failure rate and prediction error.

We evaluate the trade-off between the prediction accuracy and surveillance cost in the linear system by setting different number of sentinels as shown in Fig.3(g). In this experiment, there are only 10 ground truth sentinels in $\mathbf{S}$. We give the number of sentinels k (x-axis), and then evaluate the prediction error of each method (y axis). Intuitively, the more sentinels that are selected, the better the accuracy that can be obtained. However, as indicated, the improvement in prediction becomes negligible when k is over 10. Note that GPs-MI shows a better performance only when k is very small.

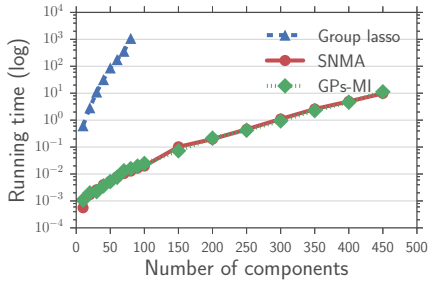


Figure 4: Comparison on running time. y-axis is the time (seconds) on a log scale. x-axis denotes the size of a system.

Fig. 4 presents the average running time of one iteration of the three methods on a PC with a 3.4GHz CPU and 8GB memory. Since GPs-MI employ forward greedy strategy (fast when only a few sentinels) but SNMA use backward selection method (fast when many sentinels), it's fair to evaluate them by comparing the running time of one iteration of each algorithm. The results show group lasso is the slowest, and GPs-MI and SNMA are almost at the same level.

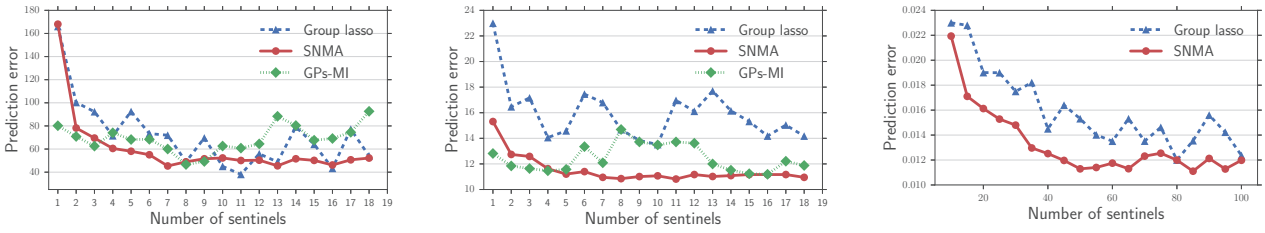


Figure 5: Comparisons on the real-world epidemic dynamics. x-axis is the number of selected sentinels and y-axis denotes the prediction error (RMSE). (a) 2009 Hong Kong H1N1 flu pandemic. (b) 2005-2009 Tengchong malaria outbreak. (c) The dynamics of hot words cascading in Baidu Tieba.

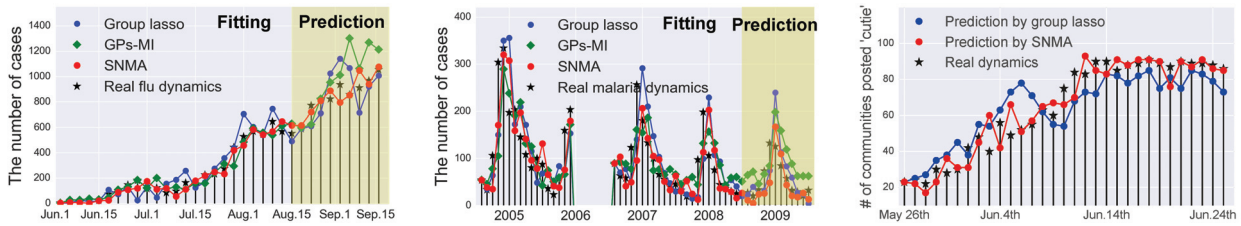


Figure 6: Model fittings and sentinel predictions for epidemic dynamics. (a) 2009 Hong Kong H1N1 flu pandemic. (b) 2005-2009 Tengchong malaria. (c) The dynamics of hot word "cutie" in the Baidu Tieba.

4.2 Validations on real diffusion data

In real cases, the failure rate cannot be evaluated because the ground truth sentinel network is unknown.

4.2.1 2009 Hong Kong H1N1 flu pandemic The cases report data of 2009 Hong Kong H1N1 influenza epidemic, was provided by Centre for Health Protection (CHP), Department of Health, Government of the Hong Kong Special Administrative Region. During this pandemic, the first imported case of human swine influenza (HSI) was confirmed on May 1, 2009. As of Sep. 2010, there were over 36,000 confirmed cases of HSI, among which about 290 were severe cases and over 80 of them died (CHP 2010). We consider the epidemic dynamics for 105 days since the disease onset in Jun. 1st 2009 based on the confirmed cases of H1N1 infection reported by CHP (www.chp.gov.hk), which gives the spatial position and infection times of each case.

Hong Kong consists of 18 administrative districts, i.e., 18 components. By merging the confirmed cases during 105 days, 3 days as a basic infectious period, we obtain the dynamics of N=18 and T=35. We set the dynamics from Jun.1 to Aug.15 as training data and the one from Aug. 15 to Sep. 15 as test data. According to a pre-given k, sentinel districts of Hong Kong can be identified via the methods (SNMA use the linear system setting). Then, based on the sentinels, we predict the newly cases on test data as shown in Fig 5 (a). Obviously, SNMA outperforms the competitors. Fig. 6 (a) shows a case of fittings and predictions for the dynamics (8 districts are selected as sentinels). The sentinel network of Hong Kong and the sentinels' spatial distribution can be found in Supporting Information.

4.2.2 2005-2009 Tengchong malaria outbreak Tengchong City, Yunnan Province, China, has 18 towns, and 658,207 residents that are distributed in a wide area of 5,845 km^2 in 2011. Because of the suitable climate for mosquito habitats, Tengchong has a quite serious malaria outbreak. Five years' (2005-2009) monthly malaria cases data at the town level, were collected by Tengchong CDC and can be obtained from the annual reports of National Institute of Parasitic Disease, China CDC. By eliminating the missing data from 2005 Jun. to Dec., we get malaria dynamics, which contains $N = 18$ components and $T = 53$ months.

We set the malaria dynamics from 2005 to 2008 as training data and the one during 2009 as test data. Similar to the Hong Kong case, we identify the sentinel towns according to a pre-given k and predict the dynamics on the test data, as shown in Fig 5 (b). Once again, SNMA achieves the best sentinel prediction in most cases. Fig. 6 (b) shows a case of fittings and predictions for the malaria dynamics with 7 sentinels. The sentinel network of Thengchong and the sentinels' distribution is in Supporting Information.

4.2.3 Hot words diffusion in Baidu Tieba Baidu Tieba (tieba.baidu.com), one of the largest online community platforms in China, is a collection of thousands of active topic-specific communities. Tieba users can post any hot words (vocabularies that are widely used in Tieba during a short period) in any communities. Hot words often present a diffusion phenomena in Baidu Tieba: a hot word first appears in only a few forums, and then is posted gradually in many other forums by the users who are active in multiple forums. Thus, Tieba can be regarded as a logistical discrete dynamical system, where communities are components, hot words are contagions, and the infected state of a component is 0 or 1. GPs-MI cannot be applied to this case because it's base on Gaussian model and cannot directly work on discrete data.

We tracked the dynamics of 11 independent hot words cascading among the top-100 active communities in Baidu Tieba from Apr. 2014 to Oct. 2015 (18 months). Only the dynamics during the words' bursting period is preserved, and the total bursting period of the words is 738 d, i.e., N = 100, T=738 by using 1 d as a time-unit. Here, we split the training and test data in term of the hot words, i.e., we alternatively set the dynamics of one hot word as test data and the rest be the training data. Then, we identify the sen-

tinel communities in Baidu Tieba by the methods (GPs-MI cannot be applied in discrete data; SNMA use the logistic system setting). Fig 5 (c) shows the results of sentinel prediction. Fig. 6 (c) shows that the dynamics of the hot word "cutie" is successfully predicted based on the data from 66 sentinel communities.

Summarily, SNMA outperform GPs-MI and group lasso in the most experiments. Our method has two obvious advantages. (1) Our method is model-based and can readily integrate prior knowledge, which makes it more effective and easier to train. (2) Our method is more robust against noise and insufficient data owing to the Bayesian framework that can effectively handle the uncertainty from both data and model.

5 Conclusions

In this work, we addressed the challenge of epidemic dynamics prediction in cases in which surveillance resources are very limited. We proposed a novel importance measure, the γ value, by modeling a sentinel network with row sparse structure and presented an effective and flexible group sparse Bayesian learning algorithm for mining the sentinel network in two kinds of widely used dynamical systems. With the discovered sentinel network, the overall epidemic dynamics can be predicted based on partial data only collected by the few sentinels. Moreover, we significantly reduced the computational complexity of the algorithm and extended it to various nonlinear systems using basic function embedding technology. We validated the proposed framework by a set of experiments on both synthetic and real-world datasets.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grants 61373053, 61572226, Jilin Province Natural Science Foundation under grant 20150101052JC.

References

Brunton, S. L.; Proctor, J. L.; and Kutz, J. N. 2016. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences* 113(15):3932–3937.

Chen, Y.; Amiri, H.; Li, Z.; and Chua, T.-S. 2013. Emerging topic detection for organizations from microblogs. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 43–52. Dublin, Ireland: ACM.

2010. Summary report on the surveillance of adverse events following HSI immunisation and expert group's comment on the safety of hsi vaccine in Hong Kong. Technical report, Centre for Health Protection, Hong Kong.

Couvreur, C., and Bresler, Y. 2000. On the optimality of the backward greedy algorithm for the subset selection problem. *SIAM Journal on Matrix Analysis and Applications* 21(3):797–808.

Dimitrov, N. B., and Meyers, L. A. 2010. Mathematical approaches to infectious disease prediction and control. Technical report.

Gerardo-Giorda, L.; Puggioni, G.; Rudd, R. J.; Waller, L. A.; and Real, L. A. 2013. Structuring targeted surveillance for monitoring disease emergence by mapping observational data onto ecological process. *Journal of The Royal Society Interface* 10(86):20130418.

Gomez Rodriguez, M.; Leskovec, J.; and Krause, A. 2010. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1019–1028. Washington, USA: ACM.

Hoang, T. N.; Low, K. H.; Jaillet, P.; and Kankanhalli, M. 2014. Nonmyopic ϵ-bayes-optimal active learning of gaussian processes. In *Proceedings of the 24th International Conference on Machine Learning*, 739–747. Beijing, China: IMLS.

Hsieh, H.-P.; Lin, S.-D.; and Zheng, Y. 2015. Inferring air quality for station location recommendation based on urban big data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 437–446.

Khuller, S.; Moss, A.; and Naor, J. S. 1999. The budgeted maximum coverage problem. *Information Processing Letters* 70(1):39–45.

Krause, A.; Leskovec, J.; Guestrin, C.; VanBriesen, J.; and Faloutsos, C. 2008. Efficient sensor placement optimization for securing large water distribution networks. *Journal of Water Resources Planning and Management* 134(6):516–526.

Krause, A.; Singh, A.; and Guestrin, C. 2008. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research* 9(Feb):235–284.

Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 420–429.

MacKay, D. J. 1995. Probable networks and plausible predictions - a review of practical bayesian methods for supervised neural networks. *Network: Computation in Neural Systems* 6(3):469–505.

Maragakis, P.; Ritort, F.; Bustamante, C.; Karplus, M.; and Crooks, G. E. 2008. Bayesian estimates of free energies from nonequilibrium work data in the presence of instrument noise. *The Journal of Chemical Physics* 129(2):024102.

Meier, L.; Van De Geer, S.; and Bühlmann, P. 2008. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(1):53–71.

Polgreen, P. M.; Chen, Z.; Segre, A. M.; Harris, M. L.; Pentella, M. A.; and Rushton, G. 2009. Optimizing influenza sentinel surveillance at the state level. *American journal of epidemiology* 170(10):1300–1306.

Wallinga, J.; van Boven, M.; and Lipsitch, M. 2010. Optimizing infectious disease interventions during an emerging epidemic. *Proceedings of the National Academy of Sciences* 107(2):923–928.

Yang, B.; Guo, H.; Yang, Y.; Shi, B.; Zhou, X.; and Liu, J. 2014. Modeling and mining spatiotemporal patterns of infection risk from heterogeneous data for active surveillance planning. In *Proceedings of the the 28th AAAI Conference on Artificial Intelligence*, 493–499. Quebec, Canada: AAAI.

Yang, B.; Pei, H.; Chen, H.; Liu, J.; and Shang, X. 2017. Characterizing and discovering spatiotemporal social contact patterns for healthcare. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(8):1532–1546.

Zhang, Z., and Rao, B. D. 2011. Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning. *IEEE Journal of Selected Topics in Signal Processing* 5(5):912–926.

Zheng, Y.; Liu, F.; and Hsieh, H.-P. 2013. U-air: when urban air quality inference meets big data. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1436–1444. Chicago, USA: ACM.

Zhou, Y.; Porwal, U.; Zhang, C.; Ngo, H. Q.; Nguyen, X.; Ré, C.; and Govindaraju, V. 2014. Parallel feature selection inspired by group testing. In *Advances in Neural Information Processing Systems*, 3554–3562.