Published in final edited form as:

Stat Med. 2017 May 10; 36(10): 1638–1654. doi:10.1002/sim.7216.

Detection of Smoothly Distributed Spatial Outliers, with Applications to Identifying the Pulmonary Distribution of Parenchymal Hyperresponsiveness

Andrew L. Thurman^a, Jiwoong Choi^b, Sanghun Choi^b, Ching-Long Lin^b, Eric A. Hoffman^c, Chang Hyun Lee^d, and Kung-Sik Chan^{a,*}

^aDepartment of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242

^bDepartment of Mechanical and Industrial Engineering, University of Iowa, Iowa City, IA 52242

^cDepartment of Radiology and Department of Biomedical Engineering, University of Iowa, Iowa City, IA 52242

^dDepartment of Radiology and Epidemiology, Seoul National University Hospital, 28 Yeongeondong, Jongno-gu, Seoul 110-744, South Korea

Abstract

Methacholine challenge tests are used to measure changes in pulmonary function that indicate symptoms of asthma. In addition to pulmonary function tests, which measure global changes in pulmonary function, computed tomography (CT) images taken before and after administration of methacholine provide local air volume changes at individual acinar units. Some of the acini may have extreme air volume changes relative to the global average, indicating hyperresponsiveness, and those extreme values may occur in clusters. We propose a Gaussian mixture model with a spatial smoothness penalty to improve prediction of hyperresponsive locations that occur in spatial clusters. A simulation study provides evidence that the spatial smoothness penalty improves prediction under different data-generating mechanisms. We apply this method to CT data from Seoul National University Hospital on five healthy and ten asthmatic subjects.

Keywords

asthma; GAM; methacholine challenge; mixture; panel of random fields

1. Introduction

Parenchymal hyperresponsiveness (PHR) is often quantitated based on parenchymal provocation tests such as a methacholine challenge test. As the methacholine challenge test is relatively safe and its effects easily reversible by bronchodilators, it is the parenchymal provocation test of choice. Results from a methacholine challenge test are generally quantified by the minimal concentration of methacholine, denoted by PC_{20} , that effects a

*Correspondence to: Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242. kung-sik-chan@uiowa.edu.

20% reduction of the baseline forced expiratory volume in 1 second (FEV1). While PC₂₀ provides an invaluable global assessment of PHR, it is desirable to assess whether PHR may be heterogeneous or homogeneous over the entire lung. And if the former is true, it is pertinent to quantify the spatial variation in PHR. With the recent advance of pulmonary image registration enabling a one-one mapping between voxels, from two lung images, that pertain to the same entity in the lung, we aim to develop a new method for identifying regions of PHR in a lung, based on (logarithmic) air volume changes at the voxel level, computed from two images, one taken at pre- and another at post-methacholine challenge. The proposed method is motivated by the following considerations: (i) a particular voxel subject to PHR would have atypical air volume change from pre- to post-methacholine challenge, and (ii) the collection of such voxels may comprise several spatially contiguous regions. Thus, for each subject, the air volume change from pre-to-post methacholine challenge is a random process, indexed by the voxel, that fluctuates about a fixed mean level, to be corrupted by spatially correlated outliers over regions of PHR (if they exist). As a first step for exploring the spatial variation of PHR in a lung, we shall ignore detailed modeling of the spatial correlation in the data. Rather, the main goal is to identify the PHR region by developing an algorithm motivated by a working model that assumes that for any voxel in a region not subject to PHR, the (log) air volume change has a marginal normal distribution with a subject-specific constant mean and fixed but unknown variance σ_0^2 . But for any voxel in a region subject to PHR, the air volume change has a marginal normal distribution with the same subject-specific constant mean and another fixed but unknown variance σ_1^2 , with $\sigma_1 \gg \sigma_0$. Note that while the mean is constant over the lung and subject-specific, the two variances are postulated to be identical across subjects. The dummy variable of the unknown PHR region is then modeled as a spatial Bernoulli process with its log odds equal to some subject-specific smooth function. The working model assumes that conditional on knowing the true PHR region, then air volume changes are independent random variables. We develop a quasi-likelihood estimation procedure with data from a panel of subjects. Using the proposed method, we can reconstruct the region of PHR. Moreover, we can quantitate the evidence of PHR in terms of the probability of PHR as a function of voxel and its integral over the entire lung as a global measure of PHR. Simulations reported herein show that the proposed method provides robust identification of PHR regions, for several data generation mechanisms that may violate the assumptions of the working model. We apply the proposed method to estimate the spatial distributions of PHR for a set of five healthy and ten asthmatic subjects, based on their lung images taken at pre- and post-methacholine challenge.

2. Data

Computed tomography (CT) scans for 5 healthy and 10 asthmatic subjects were acquired at total lung capacity (TLC) of pre- and post-methacholine (PrM and PoM, respectively) challenges (Table 1) in the supine position. Furthermore, the PrM and PoM CT scans were followed by respective pulmonary function tests (PFTs) [forced expiratory volume in 1 second (FEV1) and forced vital capacity (FVC)] in the upright position. For the CT scans, the Siemens Medical Solutions (Forchheim, Germany) scanner was employed. The scanner parameters were set to 1.5-mm thickness, a pitch of 0.45 at 15 effective mAs with 140 kV and a reconstruction matrix of a 512×512 matrix using B30f kernel. A semi-automatic

software Apollo (VIDA Diagnostics, Coralville, Iowa) was utilized to process the lung volumetric CT image for segmentations of airways, lobes, lungs, and vessels.

This study was approved by the institutional review board of Seoul National University Hospital. Asthmatic subjects were recruited through the Asthma and Allergic Diseases Center of Seoul National University Hospital. They were diagnosed by physicians using the guidelines of the U.S. National Asthma Education and Prevention Program [1], including a positive methacholine provocation test result, i.e. a decrease ($> 20\%$) in FEV1 from the baseline value after inhalation of 4 mg/mL or less of methacholine. To obtain % predicted values of FEV1 and FVC, the normal values were estimated by the equations of Hankinson et al. [2], along with the multiplication of racial factor (0.88) [3].

3. Models

Suppose that for a set of N subjects, we collect data $X_{ij} \in \mathbb{R}$ located at spatial location $v_{ij} \in \mathbb{R}^d$, $i = 1, \dots, n_j$, $j = 1, \dots, N$. We assume $E(X_{ij}) = \mu_j \in \mathbb{R}$ is the same at every location for subject j , but that there are some locations i with large deviations $X_{ij} - \mu_j$. Assuming data from different spatial locations are independent and the probability of a large deviation for subject j , $\gamma_j \in [0, 1]$, is the same at all locations, one model for such data is a two-component Gaussian mixture model with the probability density function

$$f_{X_{ij}}(x_{ij}) = \gamma_j \phi(x_{ij}; \mu_j, \sigma_1) + (1 - \gamma_j) \phi(x_{ij}; \mu_j, \sigma_0), \quad (1)$$

where $\phi(\cdot; \mu, \sigma)$ is the probability density function of the normal distribution with mean μ and standard deviation σ . In this case, $\sigma_1 > \sigma_0$, so that regions with large deviations are characterized by a large variance.

Model (1) allows the probability of a large deviation to vary from subject-to-subject, but if we suspect that there are regions of large deviations, it could fail to capture important spatial aspects of the data. To account for regional differences in these probabilities, we allow a different parameter $\gamma_j(v_{ij})$ at every location v_{ij} , giving the more general probability density function

$$f_{X_{ij}}^*(x_{ij}) = \gamma_j(v_{ij}) \phi(x_{ij}; \mu_j, \sigma_1) + (1 - \gamma_j(v_{ij})) \phi(x_{ij}; \mu_j, \sigma_0). \quad (2)$$

Because we expect nearby spatial locations to have similar probability of large deviation, we impose a smoothness constraint on $\gamma_j(v)$. We parametrize the function $\gamma_j(v)$ as

$$\log[\gamma_j(v)/\{1 - \gamma_j(v)\}] = g_j(v) \quad (3)$$

and assume $g_j(v)$ is a thin-plate spline [4, 5]; that is, the function g_j is subject to a penalty $P_j(g_j)$ on its second-order derivatives, see the Appendix below. The penalty $P_j(g_j)$ is defined as the integral of the sum of all squared mixed second derivatives of g_j over the space of locations. This smoothness constraint will produce similar values at $g_j(u)$ and $g_j(v)$ when $\|u - v\|_2$ is small. There is no loss of generality in assuming the g_j to be thin-plate splines as their penalized estimates to be introduced in the next section can be shown to be thin-plate splines, even if the functions are relaxed to be smooth functions. For a broader discussion of smoothing splines, see [4].

The constant subject-specific mean assumption in model (2) is motivated by the following consideration. If a lung is non-hyperresponsive everywhere, then the air-volume change is expected to be of zero mean per each voxel. In practice, due to variation in a subject's effort in achieving total lung capacity in the pre- and post-methacholine challenge test scans, there is likely air lung volume changes between the two scans. Under such a more realistic situation, for a homogeneously non-hyperresponsive lung, the per-voxel air volume change may have a constant subject-specific mean. Hence, an ideal model for the spatially distributed air-volume change in a healthy, i.e., homogeneously non-hyperresponsive lung is constant mean plus noise of constant variance. Thus, the working model specifies a constant mean plus noise model with the noise of zero mean and spatially varying variance to account for possible departure from the ideal case of an homogeneously non-hyperresponsive lung.

4. Estimation

The log-likelihood function corresponding to (2) is

$$\ell(\theta) = \sum_{j=1}^N \sum_{i=1}^{n_j} \log \{ f_{x_{ij}}^* (x_{ij}) \}, \quad (4)$$

where θ is the vector of all model parameters. We aim to maximize the objective function

$$\ell(\theta) - \sum_{j=1}^N \lambda_j P_j(g_j), \quad \text{where } \lambda_j > 0 \text{ is a tuning parameter controlling the smoothness of } g_j, j = 1, \dots, N. \text{ We employ a generalized expectation-maximization (GEM) algorithm [6, 7] to maximize this function and estimate } \theta.$$

Assume that λ_j is fixed, $j = 1, \dots, N$. Let $Z_{ij} \in \{0, 1\}$ be an unobserved class label at spatial location v_{ij} indicating whether that location is a large deviation ($Z_{ij} = 1$) or normal deviation ($Z_{ij} = 0$). The complete data log-likelihood function corresponding to (4) is

$$\ell^c(\theta) = \sum_{j=1}^N \sum_{i=1}^{n_j} [Z_{ij} \log \{ \gamma_{ij} \phi(x_{ij}; \mu_j, \sigma_1) \} + (1 - Z_{ij}) \log \{ (1 - \gamma_{ij}) \phi(x_{ij}; \mu_j, \sigma_0) \}]. \quad (5)$$

Thus, for the E-step, we compute $Z_{ij}^{(t)} = E_{\theta(t)} [Z_{ij} | X_{ij} = x_{ij}]$, which yields

$$Z_{ij}^{(t)} = \frac{\gamma_{ij}^{(t)} \phi(x_{ij}; \mu_j^{(t)}, \sigma_1^{(t)})}{\gamma_{ij}^{(t)} \phi(x_{ij}; \mu_j^{(t)}, \sigma_1^{(t)}) + (1 - \gamma_{ij}^{(t)}) \phi(x_{ij}; \mu_j^{(t)}, \sigma_0^{(t)})}. \quad (6)$$

For the M-step, we need to obtain the maximizer $\theta^{(t+1)}$ of

$$Q(\theta; \theta^{(t)}) = \sum_{j=1}^N \sum_{i=1}^{n_j} \left[Z_{ij}^{(t)} \log\{\gamma_{ij} \phi(x_{ij}; \mu_j, \sigma_1)\} + (1 - Z_{ij}^{(t)}) \log\{(1 - \gamma_{ij}) \phi(x_{ij}; \mu_j, \sigma_0)\} \right] - \sum_{j=1}^N \lambda_j P_j(g_j).$$

(7)

This optimization problem can be separated into two steps. The first step updates $\mu_j^{(t+1)}$, $j = 1, \dots, N$, $\sigma_1^{(t+1)}$, and $\sigma_0^{(t+1)}$, and the second step updates $\gamma_{ij}^{(t+1)}$, $i = 1, \dots, n_j$, $j = 1, \dots, N$.

In the first step, we find $\mu_j^{(t+1)}$, $j = 1, \dots, N$ that maximize (7) given the current values $\sigma_1^{(t)}$ and $\sigma_0^{(t)}$. The closed-form solution for $\mu_j^{(t+1)}$ is

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n_j} w_{ij}^{(t)} x_{ij}}{\sum_{i=1}^{n_j} w_{ij}^{(t)}},$$

where $w_{ij}^{(t)} = (\sigma_1^{(t)})^{-2} Z_{ij}^{(t)} + (\sigma_0^{(t)})^{-2} (1 - Z_{ij}^{(t)})$. Now, given new values $\mu_j^{(t+1)}$, $j = 1, \dots, N$, the maximizers of (7) over σ_1^2 and σ_0^2 are

$$(\sigma_0^2)^{(t+1)} = \frac{\sum_{i=1}^{n_j} (1 - Z_{ij}^{(t)}) (x_{ij} - \mu_j^{(t+1)})^2}{\sum_{i=1}^{n_j} (1 - Z_{ij}^{(t)})}$$

and

$$(\sigma_1^2)^{(t+1)} = \frac{\sum_{i=1}^{n_j} Z_{ij}^{(t)} (x_{ij} - \mu_j^{(t+1)})^2}{\sum_{i=1}^{n_j} Z_{ij}^{(t)}}.$$

In the second step, we maximize (7) over $\{\gamma_{ij}, i = 1, \dots, n_j, j = 1, \dots, N\}$. If we isolate the terms of (7) that depend on $\{\gamma_{ij}, i = 1, \dots, n_j, j = 1, \dots, N\}$, we obtain the objective function

$$\sum_{i=1}^{n_j} \left[Z_{ij}^{(t)} \log(\gamma_{ij}) + (1 - Z_{ij}^{(t)}) \log\{(1 - \gamma_{ij})\} \right] - \sum_{j=1}^N \lambda_j P_j(g_j).$$

For each j , this function has the form of the penalized log-likelihood function corresponding to a generalized additive model (GAM) in the binomial family with a logit link, so it can be maximized using software for fitting GAMs. We repeat computation of (6) and maximization of (7) iteratively until a convergence criterion is achieved.

Recall that θ is the vector of all model parameters. To estimate $\lambda_j, j = 1, \dots, N$ ideally, we would estimate θ for different fixed values of $\lambda_j, j = 1, \dots, N$ and estimate them using cross-validation, generalized cross-validation, or minimization of an information criterion [4].

Instead, at each iteration, we update $\lambda_j^{(t+1)}, j = 1, \dots, N$ based on the current value $\hat{\theta}^{(t+1)}$ using generalized cross-validation, so that the estimate of λ_j is $\lambda_j^{(t+1)}$ at convergence, which leads to a computationally efficient estimation scheme.

We estimate the parameters of the simpler model (1) by substituting the density function $f_{X_{ij}}$ for $f_{X_{ij}}^*$ in (4). The penalties $P(g_j)$ no longer apply, and we maximize the log-likelihood function directly. The E-step and updates of $\mu_j^{(t+1)}, j = 1, \dots, N, (\sigma_0^2)^{(t+1)}$, and $(\sigma_1^2)^{(t+1)}$ remain the same, and the update of $\gamma_j^{(t+1)}$ simplifies to

$$\gamma_j^{(t+1)} = n_j^{-1} \sum_{i=1}^{n_j} Z_{ij}^{(t)}.$$

5. Simulation

5.1. Simulation Setup

We compared the finite sample properties of these methods through a simulation study. Here $N = 1$, and we suppress the subject index j . Two sample sizes n were compared. The data set with a smaller sample size was the 32×32 regular grid $\{1, 2, \dots, 31, 32\} \times \{1, 2, \dots, 31, 32\}$ with $n = 1,024$, and the data set with a larger sample size was the 64×64 regular grid $\{1, 1.5, \dots, 31.5, 32\} \times \{1, 1.5, \dots, 31.5, 32\}$ with $n = 4,096$. Data were generated from three different models.

First, we generated data from a model where the latent variables $Z_i, i = 1, \dots, n$ had the spatially-varying probability functions $\gamma_{small}(\cdot)$ when $n = 1,024$ and $\gamma_{large}(\cdot)$ when $n = 4,096$ shown in Figure 1. Given Z_i the response variable X_i was normally-distributed with mean $\mu = 0$ and variance $\sigma_0^2 = 1$ when $Z_i = 0$ or variance $\sigma_1^2 = 10$ when $Z_i = 1$. We refer to this

simulation setting as *Model 1*. The true values of σ_i^2 , $i = 0, 1$ were chosen to mimic the relative magnitude of their counterparts in the real application reported below.

Next, we introduced spatial correlation to the data X_i . Define the data with normal deviations as $X_h = \{X_i : Z_i = 0\}$ and the data with large deviations as $X_H = \{X_i : Z_i = 1\}$. The latent variables Z_i , $i = 1, \dots, n$ were generated as in Model 1. Using the same parameters μ , σ_0^2 , and σ_1^2 as in Model 1, we simulated X_h and X_H as Gaussian random fields [8] with constant mean μ at all locations and covariance functions $C_h(u, v) = \sigma_0^2 \exp(-\phi^{-1} \|u - v\|_2)$ and $C_H(u, v) = \sigma_1^2 \exp(-\phi^{-1} \|u - v\|_2)$, respectively, where $\|u - v\|_2$ is the Euclidean distance between locations u and v . We refer to this simulation setting as *Model 2*.

In our third simulation setting, instead of characterizing large deviations in terms of a large variance, we characterized large deviations as locations with exceptionally large or small means relative to the overall mean. The data X_i , $i = 1, \dots, n$ were simulated from independent normal distributions with the constant variance $\sigma_0^2 = 1$ but spatially-varying mean functions μ_{small} and μ_{large} shown in Figure 2. We refer to this simulation setting as *Model 3*.

We simulated 1,000 data sets from these simulation settings and estimated the parameters using two methods. The first method estimates μ , σ_0^2 , σ_1^2 , and γ under (1), so that $Z_1, \dots, Z_n$ are independent and identically-distributed (IID) with a common γ . We refer to this as the *IID method*. The second method estimates μ , σ_0^2 , σ_1^2 , and γ_i , $i = 1, \dots, n$ under (2), where γ_i is subject to the smoothness constraint in (3). We refer to this as the *spline method*.

In addition to parameter estimation, we obtain soft thresholding and hard thresholding predictions of the latent variables $Z_1, \dots, Z_n$ to predict which areas are prone to large deviations. The *soft thresholding predictions* are the values given by (6) at convergence, and they are denoted $\hat{Z}_i$, $i = 1, \dots, n$. The *hard thresholding predictions* are $\hat{Z}_i^H$ if $\hat{Z}_i > .5$ and $\hat{Z}_i^H = 0$ otherwise.

5.2. Simulation Results: Model 1

Figure 3 plots the simulated data and latent variables for a random data set with $n = 4,096$ generated from Model 1 with soft thresholding predictions and estimated probability functions for the IID and spline methods. The spline method predictions appear to characterize the region of large deviations more precisely than the IID method predictions, and the spline method estimates of γ_i capture the region of large deviations while the IID method estimates are constant throughout the entire domain.

Table 2 gives values of bias, standard deviation, and root mean squared error (RMSE) for estimates under the IID and spline methods under Model 1. The IID method estimates of μ tended to be less biased and have smaller standard deviations and RMSEs compared to the spline method, yet the IID estimates of σ_0^2 and σ_1^2 tended to be more biased and have larger standard deviations and RMSEs compared to the spline method. As we expected, when the

sample size was increased, the RMSE values decreased for all parameter estimates for both methods.

For each simulated data set, we computed the proportion of correct hard thresholding predictions of the latent variables,

$$n^{-1} \sum_{i=1}^n I\{\hat{Z}_i^H = Z_i\},$$

where $I\{A\}$ is the indicator function of event A . To assess the soft thresholding prediction performance, we computed the RMSE values of the soft thresholding predictions,

$$\left\{ n^{-1} \sum_{i=1}^n (\hat{Z}_i - Z_i)^2 \right\}^{1/2}.$$

Table 4 contains percentiles of the distributions of these two quantities. The spline method was able to correctly predict more of the latent variables than the IID method in both the small sample (median values .9570 vs. .9209, respectively) and large sample (median values .9648 vs. .9194, respectively) settings. Further, the spline method tended to have much smaller RMSE values for soft thresholding predictions than the IID method in both settings (median values .1676 vs. .2606 in the small sample and .1542 vs. .2628 in the large sample, respectively).

5.3. Simulation Results: Model 2

Figure 4 plots the simulated data and latent variables for a random data set with $n = 4,096$ generated from Model 2 with soft thresholding predictions and estimated probability functions for the IID and spline methods. As in Model 1, the spline method soft thresholding predictions and estimates of γ_i appear to characterize the region of large deviations more precisely than the IID method, but for both methods, some of regions are falsely identified to contain large deviations.

Table 3 gives values of bias, standard deviation, and RMSE for parameter estimates under Model 2. The RMSE values for $\hat{\mu}$ and $\hat{\sigma}_0^2$ were smaller for the IID method, and the RMSE values for $\hat{\sigma}_1^2$ were smaller for the spline method under both sample sizes. Under the large sample setting, the RMSE values decreased for all parameter estimates for both methods, except in the case of $\hat{\sigma}_0^2$ for the spline method. This may be due to Monte Carlo error.

Table 5 presents the prediction performance of the two methods under Model 2, analogous to Table 4. As in Table 4, the spline method outperforms the IID method in both hard thresholding and soft thresholding prediction performance. While the addition of spatial correlation in Model 2 did not modify the prediction performance of the IID method, prediction performance of the spline method was greatly affected, as the proportion of

correct hard thresholding predictions were smaller and RMSE values of soft thresholding predictions were larger in Table 5 than Table 4.

5.4. Simulation Results: Model 3

The data from Model 3 were generated under a misspecified model, where large deviations are characterized by different mean values rather than large variances. Still, we can attempt to identify locations whose values significantly deviate from the overall mean. Figure 5 plots the simulated data and latent variables for a random data set with $n = 4,096$ generated from Model 3 with soft thresholding predictions and estimated probability functions for the IID and spline methods. As in Models 1 and 2, the spline method soft thresholding predictions and estimates of γ_i appear to characterize the region of large deviations more precisely than the IID method.

Because Model 3 is a mixture model with different means and a constant variance and the IID and spline methods estimate parameters of a mixture model with constant mean and different variances, the parameters of Model 3 and estimates for the IID and spline methods are not directly comparable, so we focused instead on prediction.

Table 6 presents the prediction performance of the two methods under Model 3, analogous to Tables 4 and 5. Again, the spline method outperforms the IID method in both hard thresholding and soft thresholding prediction performance. While 97% of the spline method hard thresholding predictions are correct, only about 71% of the IID method hard thresholding predictions are. Also, the RMSE values for soft thresholding predictions in Table 6 for the spline method are 1/3 the size of the values for the IID method (.1593 vs. .4852 for the small sample and .1469 vs. .4877 for the large sample, respectively).

6. Application: Determination of Parenchymal Hyperresponsiveness in the Human Lung

6.1. Computed Tomography Data

The objective of our study was to identify locations with highly variable changes in air volume from PrM to PoM. To make these measurements unitless and to symmetrize the distribution, we chose the log of the ratio of PoM to PrM air volume as the response variable, X_{ij} , for location i in subject j . These measurements are plotted for three subjects in Figure 6. The top row contains data for a healthy lung, and the bottom two rows contain data for asthmatic lungs. Note that while hyperresponsiveness is a symptom of asthma, it may also result from allergic or non-allergic respiratory disorders, see [9]. Consequently, both healthy (i.e. non-asthmatic) and asthmatic lungs can both experience hyperresponsiveness, but this behavior should be more pronounced in asthmatic lungs than healthy lungs, c.f. Figure 6, which shows the extensive hyperresponsiveness in the asthmatic lung images although a moderately hyperresponsive zone (a red zone) can also be seen on the result of a healthy lung image (top-left).

6.2. Results

We analyzed these data under the IID and spline methods defined in Section 5. As in Section 3, we allow separate mean parameters μ_j and probability functions $\gamma_j(\cdot)$ for each subject, $j = 1, \dots, n$, but we estimate common variance parameters σ_0^2 and σ_1^2 for all subjects. Table 7 gives estimates of $\hat{\mu}_j$ for both methods, $\hat{\gamma}_j$ for the IID method, and a comparable parameter

$\bar{\gamma}_j = n_j^{-1} \sum_{i=1}^{n_j} \hat{\gamma}_{ij}$ for the spline method, which is the average probability of PHR across the lung. The two methods produced similar values of $\hat{\mu}_j$, but $\hat{\gamma}_j$ and $\bar{\gamma}_j$ were quite different. In eleven of fifteen subjects, the spline method estimated a higher probability of PHR than the IID method. In particular, the probabilities of PHR were larger for all healthy subjects under the spline method. The variance parameters, σ_0^2 and σ_1^2 , were estimated to be 0.0053 and 0.0448 for the IID method and 0.0042 and 0.0444 for the spline method, respectively. These values are similar, although both are smaller for the spline method.

Figures 7 and 8 provide plots of the soft thresholding predictions, $\hat{Z}_i$, for the IID and spline methods, respectively. Compared to the IID method, the spline method predicted vaster PHR regions for the healthy subject and more clearly separated PHR and non-PHR regions for the asthmatic subjects. Under the IID method, the upper regions of the asthmatic subjects are colored light blue or teal, where predictions of $\hat{Z}_i$ are around .5, whereas under the spline method, the dark blue regions indicate more definitive predictions of non-PHR locations.

Hard thresholding predictions (figures not shown) also highlight differences between these two methods. For the asthmatic subjects, the spline method appears to have segmented the lungs into non-PHR upper regions and PHR lower regions, whereas the IID method produced a speckled appearance because the soft thresholding predictions were more ambiguous in the upper regions.

Figures 9 and 10 plot the values of $\hat{\gamma}_j$ for the IID method and $\hat{\gamma}_{ij}$ for the spline method, respectively. Under the IID method, the asthmatic subjects had relatively higher probability of PHR than the healthy subjects, yet we could not distinguish which regions were prone to PHR. Alternatively, the spline method identified the PHR regions in dark red and the non-PHR regions in dark blue. Although both methods allow subject-to-subject differences in the abundance of PHR locations, only the spline method can identify regional variation within each subject.

7. Discussion

One advantage gained through the proposed statistical modeling is the ability to compare parameter estimates across subjects. For instance, the mean $\bar{\gamma}_j^{\text{spline}}$ over the whole lung quantifies the overall extent of pulmonary hyperresponsiveness for the j th subject. Although we did not perform comparisons across subjects because there was a small sample size (5 healthy subjects, 10 asthmatic subjects), the proposed methodology allows a clinician performing these comparisons in given a large enough study, for instance, to establish the normal range of the mean pulmonary hyperresponsiveness, and assess factors influencing

the variation of pulmonary hyperresponsiveness. Besides computing the mean hyperresponsiveness over the whole lung, the (sub)-lobar mean hyperresponsiveness may be computed for identifying regions which are most affected. Moreover, such quantitative estimates can be potentially useful in assessing the effectiveness of a treatment on pulmonary hyperresponsiveness. Also, there is visual appeal to the probability surfaces. The images can be examined to understand regions of individual lungs that are prone to hyperresponsiveness. Moreover, by establishing a statistical model that allows the degree of smoothness to be selected by the data and estimates the probability of hyperresponsiveness voxel by voxel, we can avoid having to identify and outline PHR regions on an ad hoc basis and allow the model to elicit regions of interest based on the estimated probability surface.

There is an extensive literature on outlier detection. See [10] for a recent review. In particular, much attention has been given to developing fast algorithms for detecting spatial outliers by calibrating a local statistic of some non-spatial attribute against a reference population of the statistic. For instance, hyperresponsive voxels may be detected as spatial outliers in the squared air volume change using methods outlined in [10, Section 5.3]. However, the existing methods for spatial outlier detection do not consider spatial correlation in the spatial outliers, which is the focus of our proposed method and with which we can borrow strength across voxels in estimating regions of hyperresponsiveness.

We have proposed a mixture model that utilizes spatial information to detect outliers, in particular, spatial regularization via a thin-plate spline that imposes smoothness on the spatially-varying probability function of the latent variables. In a simulation study, we compared the performance of the proposed method to the method with a constant probability function. Even under model misspecification, the spatially-varying probability function allowed us to obtain better predictions of locations of outliers and maps of the latent variables with less spatial heterogeneity. In a study of CT images of the human lung, we found that we were able to segment the lungs into homogeneous regions of outliers and non-outliers.

In addition to estimating the parameters of our model, we might be interested to obtain standard errors of our estimates. A popular and straightforward method to obtain standard errors for parameter estimates is the parametric bootstrap [11]. Unfortunately, there is a large computational cost because the GEM algorithm must be repeated a large number of times. Many alternative methods for standard error estimation exist [7], which are less costly in computation but do not generally incorporate the penalty we apply to the likelihood function.

Here we have proposed a model that contains a *dense* probability function in the sense that every location in every subject has an estimated probability between zero and one. We might expect that certain subjects do not have any outliers because their lungs are healthy, so that all locations for those subjects have probability exactly zero. One idea is to impose a group lasso [12] penalty, which could produce zero probabilities at all locations for some subjects.

Our model assumes a constant mean for each subject. We could allow the mean to be a function of covariates and further allow the regression coefficients to vary by region. For

example, in this study, the methacholine challenge test was intended to provoke bronchoconstriction, and a subsequent bronchodilator treatment was intended to reverse bronchoconstriction. We may expect there to be regions more responsive or less responsive to bronchodilation, and we could model this behavior using different regression coefficients for different mixture components. Air volume changes from pre-bronchodilator to post-bronchodilator in a highly responsive region should have strong negative correlations with air volume changes from pre-methacholine to post-methacholine, indicating the ability of the bronchodilator to reverse the effect of methacholine. We leave the work discussed in this section as future work.

Acknowledgments

Contract/grant sponsor: The authors gratefully acknowledge support by the U.S. National Institutes of Health (NIH U01 HL114494 and NIH RO1 HL089897).

Appendix: Thin-plate Splines

We now provide some details about thin-plate spline, first in the simple case of conditionally normal response and $d = 2$. Write $\mathbf{v} = (x, z) \in \mathbb{R}^2$. Consider then the regression model

$$y_i = g(x_i, z_i) + e_i,$$

where e_i , $i = 1, 2, \dots, n$, are iid normal errors. The regression function g then models both the main and interaction effects of x and z . It is assumed that g is a smooth function that admits finite second derivatives. We estimate g by minimizing the following penalized least squares:

$$\sum_{i=1}^n \{y_i - g(x_i, z_i)\}^2 + \lambda J(g),$$

where

$$J(g) = \iint_{\mathbb{R}^2} \left\{ \frac{\partial^2 g}{\partial x^2} \right\}^2 + 2 \left\{ \frac{\partial^2 g}{\partial x \partial z} \right\}^2 + \left\{ \frac{\partial^2 g}{\partial z^2} \right\}^2 dx dz,$$

and $\lambda > 0$ is the smoothing (tuning) parameter to be determined from the data. For a fixed λ , the solution to the preceding penalized least squares problem is a thin plate spline that is a higher-dimensional analogue of a natural cubic spline. We now describe a set of basis functions for the thin plate splines. Recall $\mathbf{v}$ denote a generic vector $(x, z)^T$ in the $x - z$ plane and the observed covariate vectors by $\mathbf{v}_1, \dots, \mathbf{v}_n$. The thin-plate-spline bases are generated

by the function $\eta(r) = \frac{1}{16\pi} r^2 \log(r)$, $r > 0$ and $\eta(0) = 0$. Then the functions

$\eta(|\mathbf{v} - \mathbf{v}_i|) = \eta(\sqrt{(x - x_i)^2 + (z - z_i)^2})$ form part of a set of basis functions for the thin plate

splines. Besides these basis functions, monomials in v of degree less than 2, i.e. 1, x , z constitute the remaining basis functions. Thus, a thin plate spline can be written as a linear combination of these basis functions:

$$g(v) = g(x, z) = a_1 + a_2x + a_3z + \sum_{i=1}^n \delta_i \eta(|v - v_i|).$$

However, as in the case of natural cubic spline, the coefficients of the basis functions satisfy the constraint $T\delta = 0$ where

$$T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ v_1 & v_2 & \cdots & v_n \end{pmatrix}, \quad \delta = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_n \end{pmatrix}.$$

For the case of a general $d \geq 1$ and $g = g(t_1, \dots, t_d)$, the following more general roughness measure may be used in lieu of $\mathcal{K}(g)$, which relates to the integrated squared partial derivatives of order m :

$$J_{m,d}(g) = \int \cdots \int_{\mathbb{R}^d} \sum_{\nu_1 + \cdots + \nu_d = m} \frac{m!}{\nu_1! \cdots \nu_d!} \left(\frac{\partial^m g}{\partial t_1^{\nu_1} \cdots \partial t_d^{\nu_d}} \right)^2 dt_1 \cdots dt_d$$

where $2m > d$; this constraint is necessary to rule out some undesirable properties of the roughness measure [13, Section 7.9]. The basis functions of the thin plate spline are then generated by the $\eta(r)$ function that is now proportional to $r^{2m-d} \log(r)$ if there are even number of covariates in the interaction term, and otherwise proportional to r^{2m-d} . Notice for $m = 2$ and $d = 1$, we get back the natural cubic spline basis functions, namely $|x - x_j|^3$. The full generality of thin plate splines is efficiently implemented in the R package *mgcv*.

The preceding discussion can be readily lifted to the setting of generalized linear model, with the conditional mean response linked to a smooth function, in which case the penalized least squares is replaced by the penalized likelihood. See [5, 4] for details.

References

1. National Heart Lung and Blood Institute, National Asthma Education and Prevention Program. Expert Panel Report 3: guidelines for the diagnosis and management of asthma. Bethesda, MD: NIH Publication No. 07-4051; 2007.
2. Hankinson JL, Odencrantz JR, Fedan KB. Spirometric reference values from a sample of the general U.S. population. American Journal of Respiratory and Critical Care Medicine. 1999; 159:179–187. [PubMed: 9872837]
3. Hankinson JL, Kawut SM, Shahar E, Smith LJ, Stukovsky KH, Barr RG. Performance of American Thoracic Society - recommended spirometry reference values in a multiethnic sample of adults: the Multi-Ethnic Study of Atherosclerosis (MESA) lung study. Chest. 2010; 137:138–145. [PubMed: 19741060]

4. Wood, S. Generalized Additive Models: An Introduction with R. Chapman & Hall; Boca Raton: 2006.
5. Wood SN. Thin plate regression splines. *Journal of Royal Statistical Society*. 2003; B 65:95–114.
6. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*. 1977; 39:1–38.
7. McLachlan, GJ., Krishnan, T. The EM Algorithm and Extensions. John Wiley & Sons; Hoboken: 2008.
8. Gaetan, C., Guyon, X. Spatial Statistics and Modeling. Springer; New York: 2010.
9. Brusasco V, Crimi E. Methacholine provocation test for diagnosis of allergic respiratory diseases. *Allergy*. 2001; 56:1114–1120. [PubMed: 11736739]
10. Schubert E, Zimek A, Kriegel HP. Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection. *Data Mining and Knowledge Discovery*. 2014; textbf28:190–237.
11. Efron, B., Tibshirani, RJ. An Introduction to the Bootstrap. Chapman & Hall; Boca Raton: 1993.
12. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*. 2006; 68:49–67.
13. Green, PJ., Silverman, BW. Nonparametric regression and generalized linear models. London: Chapman and Hall; 1994.

Figure 1.
Simulated data: plots of the probability functions $\gamma_{small}(\cdot)$ and $\gamma_{large}(\cdot)$ for the latent variables, Z_i , $i = 1, \dots, n$, in Model 1 and Model 2.

Figure 2. Simulated data: plots of the mean functions $\mu_{small}(\cdot)$ and $\mu_{large}(\cdot)$ for the data, X_i , $i = 1, \dots, n$, in Model 3.

Figure 3. Simulation results: for a randomly chosen simulated data set ($n=4,096$) from Model 1, the observed data (X), the latent variables (Z), soft thresholding predictions of latent variables under the IID method ($\hat{Z}_{IID}$) and spline method ($\hat{Z}_{spline}$), and estimates of the probability function of large deviation $\gamma(\cdot)$ under the IID method ($\hat{\gamma}_{IID}$) and the spline method ($\hat{\gamma}_{spline}$)

Figure 4. Simulation results: for a randomly chosen simulated data set ($n=4,096$) from Model 2, the observed data (X), the latent variables (Z), soft thresholding predictions of latent variables under the IID method ($\hat{Z}_{IID}$) and spline method ($\hat{Z}_{spline}$), and estimates of the probability function of large deviation $\gamma(\cdot)$ under the IID method ($\hat{\gamma}_{IID}$) and the spline method ($\hat{\gamma}_{spline}$)

Figure 5. Simulation results: for a randomly chosen simulated data set ($n=4,096$) from Model 3, the observed data (X), the mean function (μ), soft thresholding predictions of latent variables under the IID method ($\hat{Z}_{IID}$) and spline method ($\hat{Z}_{spline}$), and estimates of the probability function of large deviation $\gamma(\cdot)$ under the IID method ($\hat{\gamma}_{IID}$) and the spline method ($\hat{\gamma}_{spline}$)

Figure 6.

Computed tomography (CT) data: log ratio of post-methacholine to pre-methacholine air volume for the Seoul National University Hospital CT imaging data. The color scale is analogous to Figures 1 – 5, with dark blue indicating the smallest values and dark red indicating largest values. All subjects are colored according to the same scale. The top row is a healthy subject and the bottom two rows are asthmatic subjects, respectively. Each column gives a different angle for each subject.

Figure 7. Computed tomography (CT) data: IID method soft thresholding predictions of high variance locations for the Seoul National University Hospital CT imaging data. The color scale is analogous to Figures 1 – 5, with dark blue indicating the smallest values and dark red indicating largest values. All subjects are colored according to the same scale. The top row is a healthy subject and the bottom two rows are asthmatic subjects, respectively. Each column gives a different angle for each subject.

Figure 8.

Computed tomography (CT) data: spline method soft thresholding predictions of high variance locations for the Seoul National University Hospital CT imaging data. The color scale is analogous to Figures 1 – 5, with dark blue indicating the smallest values and dark red indicating largest values. All subjects are colored according to the same scale. The top row is a healthy subject and the bottom two rows are asthmatic subjects, respectively. Each column gives a different angle for each subject.

Figure 9.

Computed tomography (CT) data: IID method estimates of probabilities of parenchymal hyperresponsiveness, $\hat{\gamma}_j$, in the Seoul National University Hospital CT imaging data. The color scale is analogous to Figures 1 – 5, with dark blue indicating the smallest values and dark red indicating largest values. The top row is a healthy subject and the bottom two rows are asthmatic subjects, respectively. Each column gives a different angle for each subject.

Figure 10.

Computed tomography (CT) data: spline method estimates of the spatially-varying probability function of parenchymal hyperresponsiveness, $\hat{\chi}_f(\cdot)$, in the Seoul National University Hospital CT imaging data. The color scale is analogous to Figures 1 – 5, with dark blue indicating the smallest values and dark red indicating largest values. The top row is a healthy subject and the bottom two rows are asthmatic subjects, respectively. Each column gives a different angle for each subject.

Table 1

Demographic and pulmonary function tests (PFT) information of five healthy and ten asthmatic subjects. For numeric variables, mean (SD) and p-values from a two-sample Wilcoxon test, and for sex, count (%) and p -value for χ^2 association test are given. BMI, FEV1 and FVC denote body mass index, forced expiratory volume in 1 second, and forced vital capacity, respectively. Races of all subjects were Asian (South Korean).

FEV1 (FVC) % predicted values were calculated by subtracting post-methacholine FEV1 (FVC) % predicted values from pre-methacholine FEV1 (FVC) % predicted values.

	Healthy	Asthma	p-value
<i>Demography</i>			
Age	29.2 (3.3)	48.6 (12.0)	0.008
BMI (kg/m ²)	23.9 (2.9)	24.5 (1.8)	0.951
Gender (Female)	0.0 (0.0)	6.0 (0.6)	0.094
<i>Pre-methacholine</i>			
FEV1 % predicted	85.3 (11.8)	90.4 (15.8)	0.244
FVC % predicted	75.4 (9.0)	94.3 (14.2)	0.019
FEV1/FVC $\times$ 100	92.3 (4.9)	75.7 (7.5)	0.005
<i>Post-methacholine</i>			
FEV1 % predicted	76.1 (17.0)	67.3 (12.4)	0.371
FVC % predicted	67.4 (10.5)	77.2 (12.0)	0.165
FEV1/FVC $\times$ 100	91.9 (11.5)	69.0 (8.1)	0.005
<i>Hyperresponsiveness of FEV1 and FVC</i>			
FEV1 % predicted	9.1 (6.1)	23.1 (8.1)	0.005
FEV1 % predicted	8.0 (4.6)	17.2 (9.4)	0.040

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2
Simulation results: bias, standard deviation (SD), and root mean squared error (RMSE) for parameter estimates under the IID method and the spline method applied to data generated from Model 1

Estimate	Method	<i>n</i> = 1,024				<i>n</i> = 4,096			
		Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD
$\hat{\mu}$	IID	-0.0003	0.0364	0.0364	-0.0001	0.0181	0.0181	0.0181	0.0181
	spline	0.0012	0.0591	0.0591	0.0007	0.0314	0.0314	0.0314	0.0314
$\hat{\sigma}_0^2$	IID	-0.0688	0.0728	0.1002	-0.0581	0.0365	0.0686	0.0686	0.0686
	spline	-0.0380	0.0482	0.0614	-0.0275	0.0243	0.0367	0.0367	0.0367
$\hat{\sigma}_1^2$	IID	-1.5240	1.9239	2.4537	-1.5447	0.9067	1.7909	1.7909	1.7909
	spline	0.4617	1.6058	1.6700	0.5129	0.7573	0.9143	0.9143	0.9143

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

Simulation results: bias, standard deviation (SD), and root mean squared error (RMSE) for parameter estimates under the IID method and the spline method applied to data generated from Model 2

Estimate	Method	$n = 1,024$				$n = 4,096$			
		Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD
$\hat{\mu}$	IID	-0.0025	0.1511	0.1511	0.0063	0.1449	0.1450		
	spline	-0.0007	0.1844	0.1844	0.0110	0.1825	0.1828		
$\hat{\sigma}_0^2$	IID	-0.1057	0.1255	0.1640	-0.0937	0.1092	0.1439		
	spline	-0.1194	0.1136	0.1647	-0.1319	0.1028	0.1672		
$\hat{\sigma}_1^2$	IID	-1.8181	3.2253	3.7010	-1.7687	2.9552	3.4428		
	spline	-0.8905	3.1609	3.2824	-1.1944	2.8734	3.1105		

Simulation results: percentiles of the distributions of proportion of correct hard thresholding predictions (Prop. correct) and root mean squared error (RMSE) for soft thresholding predictions under the IID method and the spline method applied to data generated from Model 1

Table 4

Rate	Method	<i>n</i> = 1,024				<i>n</i> = 4,096			
		25th	50th	75th	25th	50th	75th	25th	75th
Prop. correct	IID	0.9160	0.9209	0.9248	0.9167	0.9194	0.9214		
	spline	0.9512	0.9570	0.9629	0.9614	0.9648	0.9680		
RMSE	IID	0.2546	0.2606	0.2683	0.2599	0.2628	0.2662		
	spline	0.1585	0.1676	0.1769	0.1501	0.1542	0.1587		

Simulation results: percentiles of the distributions of proportion of correct hard thresholding predictions (Prop. correct) and root mean squared error (RMSE) for soft thresholding predictions under the IID method and the spline method applied to data generated from Model 2

Table 5

Rate	Method	<i>n</i> = 1,024				<i>n</i> = 4,096			
		25th	50th	75th		25th	50th	75th	
Prop. correct	IID	0.9111	0.9199	0.9287	0.9102	0.9191	0.9258		
	spline	0.9316	0.9473	0.9570	0.9338	0.9495	0.9603		
RMSE	IID	0.2497	0.2635	0.2771	0.2530	0.2642	0.2783		
	spline	0.1722	0.1917	0.2210	0.1677	0.1908	0.2234		

Simulation results: percentiles of the distributions of proportion of correct hard thresholding predictions (Prop. correct) and root mean squared error (RMSE) for soft thresholding predictions under the IID method and the spline method applied to data generated from Model 3

Table 6

Rate	Method	<i>n</i> = 1,024				<i>n</i> = 4,096			
		25th	50th	75th	25th	50th	75th	25th	75th
Prop. correct	IID	0.6140	0.7148	0.7812	0.6688	0.7100	0.7463		
	spline	0.9626	0.9678	0.9727	0.9702	0.9727	0.9753		
RMSE	IID	0.4394	0.4852	0.5404	0.4637	0.4877	0.5106		
	spline	0.1492	0.1593	0.1696	0.1411	0.1469	0.1526		

Computed tomography (CT) data: parameter estimates for the Seoul National University Hospital CT imaging data under the IID method and the spline method. $\hat{\gamma}_j^{\text{spline}}$ is the average of $\hat{\gamma}_{i,j}^{\text{spline}}$, $i = 1, \dots, n_j$ for subject j .

Severity	Subject	$\hat{\mu}_j^{\text{IID}}$	$\hat{\mu}_j^{\text{spline}}$	$\hat{\gamma}_j^{\text{IID}}$	$\hat{\gamma}_j^{\text{spline}}$
Healthy	1	-0.0709	-0.0525	0.2101	0.2831
	2	0.0417	0.0367	0.0984	0.1427
	3	0.0690	0.0694	0.0501	0.1334
	4	-0.0110	-0.0029	0.1978	0.2704
	5	0.0012	0.0051	0.1868	0.2229
Mild	1	0.2709	0.2776	0.7301	0.5942
	2	0.1465	0.1395	0.5791	0.4964
	3	0.0668	0.0573	0.3200	0.3304
	4	0.0455	0.0357	0.1976	0.2336
	1	0.1748	0.1820	0.4780	0.4494
Moderate	2	0.1212	0.1149	0.3986	0.3982
	3	0.0665	0.0675	0.1513	0.2273
	4	-0.0234	-0.0210	0.2937	0.3460
	5	-0.0349	-0.0360	0.1643	0.2186
Severe	1	0.1276	0.1194	0.3516	0.3624