

Online Supplemental Material for Outlier Accommodation with Semiparametric Density Processes

1 Proof of Concept

In order to assess the ability of our method to accurately identify outliers using our semiparametric density process, we perform a simulation study that serves as a proof of concept. The simulation study reflects the key data attributes of our motivating example, while simplifying certain aspects that are specific to the snow core we analyze.

The simulation study focuses on the ability of our method to correctly identify observations not generated by the underlying physical process of snow densification. Data for the simulation study is generated under three separate scenarios that mirror aspects of our data application. For each scenario, 1,000 depth-density observations are generated from the mean function and then a portion of the simulated density measurements (20%) are shifted up or down, according to the scenario. The mean function is a piece-wise linear function with an intercept of 0.4, slopes of 0.01 between 0-m and 20-m, 0.005 between 20-m and 80-m, and 0.0005 between 80-m and 150-m, and normally-distributed random error with mean 0 and standard deviation 0.025. In the first outlier setting, the density of the observations is reduced by 20 – 100%; in the second setting, 80% of the observations near the surface (before 50-m) are increased by 0.15 – 0.3, 20% of the observations near the surface are decreased by 0.15 – 0.3, and the observations deeper than 50-m are reduced by 20 – 100%; in the third setting, 70% of the observations near the surface (before 50-m) are increased by

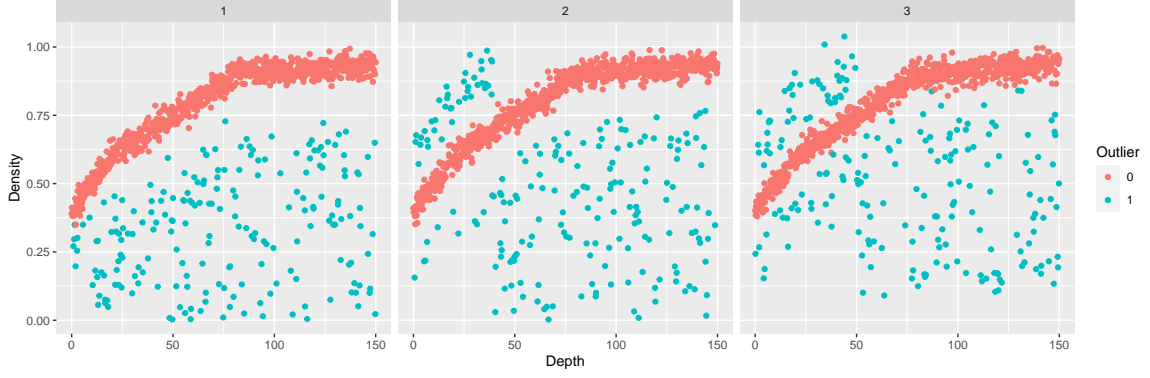


Figure 1: Simulated data sets for each of the three scenarios

0.1 – 0.3, 30% of the observations near the surface are decreased by 0.1 – 0.3, and the observations deeper than 50-m are reduced by 10 – 90%. A sample data set under each of the 3 settings can be seen in Figure 1.

To test the ability of our method to correctly identify outliers, we simulate data under each of the scenarios put forth above and use our two component mixture model to estimate the probability that each observation is an outlier. We classify an observation as an outlier if it has less than a 80% probability of coming from the the mean function, since 80% is the known proportion of observations coming from the mean function unaltered. We perform 1,000 simulations for each of the three settings and calculate the average number of the 200 outliers that are correctly identified as outliers and the average number of the 800 non-outliers that are misclassified as outliers. The results of the simulation can be seen in Table 1.

Based on the results of the simulation study, we can conclude that our method is successful in identifying outliers. The mean number of outliers identified in each of the three settings is roughly 200, which is how many observations were actually outliers. Also, the number of non-outliers that were misclassified as outliers is relatively small

Table 1: Mean number of outliers identified and observations misclassified as outliers for 1,000 simulations

	Outliers Identified	Non-outliers Misclassified
Scenario 1	199.992	6.328
Scenario 2	200.000	12.198
Scenario 3	199.704	16.115

(6.3, 12.2, 16.1) compared to the sample size (800) for each setting. Thus, our method has a low misclassification rate that varies somewhat depending on the complexity of the outlier distribution.

2 Extended Model Fitting Results and Posterior Analysis

The mean component of the mixture model is defined by an intercept α and several β coefficients. Estimates of these coefficients, along with 95% Credible Intervals, are given in Table 2.

An intercept of -0.0953 corresponds to a surface density of 0.4367 g/cm^3 , which agrees well with the data. Also, at a depth of 160-m, the estimated snow density is 0.9095 g/cm^3 , which is within one standard error of ρ_I (SE: 0.0082 g/cm^3). The starting and ending points of the posterior mean curve serve as indications that the function is modeling the data appropriately.

To check that the posterior samples of the final model converged to the true param-

Table 2: Coefficients for the mean component of the final model with 95% credible intervals

Parameter	Mean	2.5%	97.5%
α	-0.0953	-0.1006	-0.0900
β_1	0.8439	0.8354	0.8533
β_2	0.3261	0.3153	0.3354
β_3	1.5288	1.5166	1.5422
β_4	1.4804	1.4565	1.5042
β_5	0.4814	0.4388	0.5220
β_6	0.2323	0.1718	0.3019

eters, we examine a traceplot of each of the variables. The traceplot of the deviance is given in Figure 2, which shows that the model is exploring the posterior space well. Noticeable correlation does exist between the traceplots of the β coefficients, but that is merely a consequence of using an I-spline as the basis function expansion. As with any spline, a small change in the function between two knots will result in a change in the function between two other knots due to the constraints imposed on the splines (continuity, differentiability, and concavity). Similar issues with proper mixing in generalized linear models is a well-known phenomenon. Thus, we look to more formal tests to determine whether the posterior samples are stationary and mixing well.

To check the stationarity of the posterior samples, Geweke’s convergence diagnostic is computed for each variable (Geweke et al., 1991). Geweke’s diagnostic compares the first 10% of the posterior draws to the last 50% of the posterior draws and tests the null hypothesis that they come from the same distribution. The test statistic generated

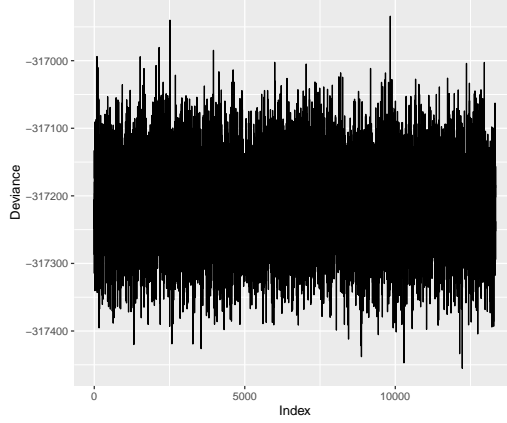


Figure 2: Traceplot of the deviance of the model

for each variable is a standard Z-score, thus if $|z| > 1.96$ for any variable we reject the hypothesis of posterior stationarity at an α level of 0.05. The magnitude of the Geweke Z-scores for all the variables is less than 1.96, indicating that our posterior samples are coming from stationary distributions.

The Heidelberg and Welch (HW) diagnostic is also calculated to test for posterior stationarity and validation of posterior means ([Heidelberger and Welch, 1981, 1983](#)). The HW diagnostic consists of two parts: the first tests the null hypothesis that the posterior draws of each variable come from a stationary distribution. Given the draws do come from a stationary distribution, the second part implements a halfwidth test to ensure that the draws estimate the posterior mean with adequate precision. We perform the first test at an α level of 0.05 and all parameters pass, meaning each of the chains appears to have converged. For the second part of the HW diagnostic, the halfwidth test is performed by calculating half the width of a $(1 - \alpha)\%$ confidence interval about the mean and comparing it to the mean. If this calculated ratio is below some ϵ then the chain is said to be estimating the posterior mean with adequate precision. All of the variables in the model pass this test with ϵ set at 0.1. Thus,

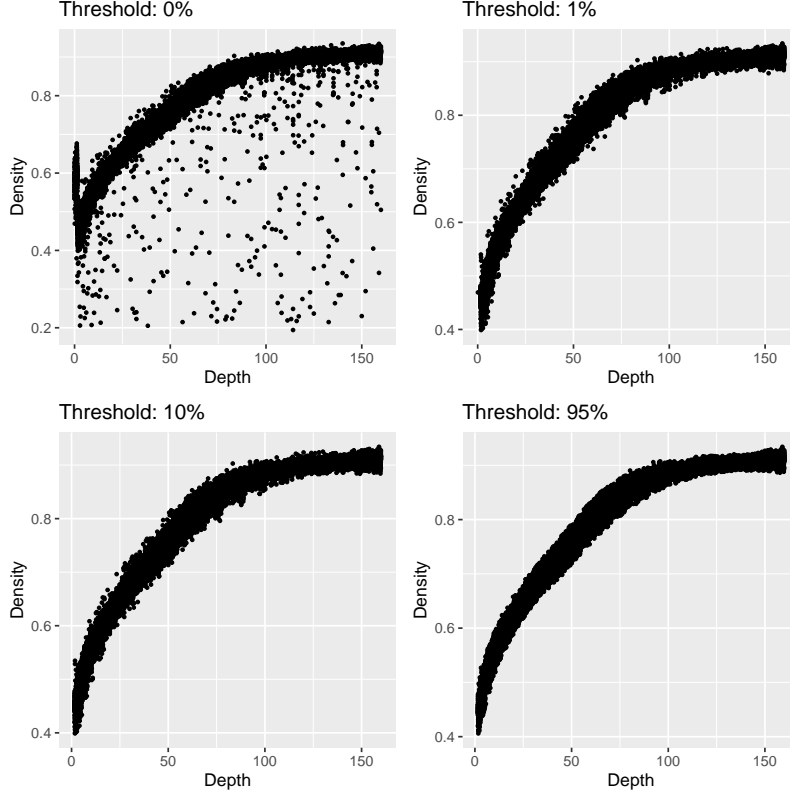


Figure 3: Plots of the cleaned data with differing snow density probability thresholds

despite apparent poor mixing among the β coefficients based on the traceplots, the posterior samples obtained appear to have converged and are estimating the posterior means with sufficient precision.

Because one of the benefits of the manner in which we apply our novel methodology in this study is the ability to perform probability-based data cleaning, we clean the raw dataset using 4 different snow density probability thresholds, as seen in Figure 3. The snow density probability thresholds used are 0%, 1%, 10%, and 95%, and the number of observations removed from each dataset are 0, 832, 878, and 1153, respectively.

3 Model Comparison and Selection

In our model comparison, we look at five modelling questions: (1) What is the ideal number of knots in the I-spline? (2) Do the data justify a heteroscedastic variance component in the mean function? (3) Does a two-component mixture model improve the fit of the model? (4) Should the mean function assume Normally distributed errors or t -distributed errors? (5) What is the ideal resolution for the grid of points used in the semiparametric prior density process?

Due to the relatively large number of observations in the dataset, the I-spline of the mean function could have dozens of interior knots without overfitting the data. That said, the curve we are trying to fit is quite simple and should not need many knots to fit the data well. We compare models with varying numbers of evenly-spaced knots and models with knots chosen based on previous knowledge. This simple model comparison is performed by comparing the Bayesian information criterion (BIC) and visual fits to the data. After this basic model comparison, we decide to use a model with 3 knots spaced evenly over the domain of depth measurements, at the 25th, 50th, and 75th depth percentiles. Somewhat surprisingly, 3 evenly-spaced knots fit the data better than 3 knots placed at the critical densities mentioned in Section 2.1. Therefore, knots are placed at approximately 40-m, 79-m, and 120-m.

Modelling questions (2) - (4) are addressed by comparing the model fit of several models whose deviance information criterion (DIC), mean deviance (\bar{D}), and complexity measure (P_d) values are reported in Table 3. As can be seen, including heteroscedastic errors does improve the fit of the model, but not nearly as much as including a two-component mixture model of a Normal distribution with a physically-motivated

Table 3: DIC, mean deviance, and size complexity of each model fit

Model	Variances	Error Dist	DIC	\bar{D}	P_d
Non -	Homoscedastic	Normal	-202723	-202733	10
		t	-293149	-293158	10
Mixture	Heteroscedastic	Normal	-208749	-208762	13
		t	-305374	-305385	11
SPDP	Homoscedastic	Normal	-300974	-306160	5186
		t	-302747	-305889	3142
Mixture	Heteroscedastic	Normal	-315201	-317214	2013
		t	-311148	-314532	3383

mean function and our semiparametric prior density process (SPDP). Also, the errors of the mean function are better explained with a Normal distribution than with a t -distribution with $df = 4$.

Interestingly, assuming that the errors follow a t -distribution, rather than a Normal distribution, drastically improves the fits of the non-mixture models, but the trend is not evident in the case of the mixture models and is even reversed in the case of the heteroscedastic mixture models. The results of the model fitting suggest that this phenomenon may be explained by the fact that having an explicit outlier distribution acts as a robust measure against outliers, thus nullifying the need to assume a robust error distribution. That said, a clear drawback of the mixture models is the increased model complexity relative to the non-mixture models, evident in Table 3. Despite the drastic increase in model complexity, the mixture models are nonetheless superior due to the great increase in model fit, as evidenced by the much lower mean deviances. Also, by using the mixture models as a robust measure against outliers rather than

assuming t -distributed errors to do so, we avoid altering inference on the physical model.

We use DIC for model comparison despite the known drawbacks of this method. Unfortunately, due to the size and complexity of the model, Monte Carlo standard error estimates for DIC using repeated model fitting described by [Zhu and Carlin \(2000\)](#) are infeasible for this model comparison. Each mixture model compared needs to run for over 24 hours to obtain enough posterior samples after convergence for model comparison, and thus obtaining hundreds of DIC samples for each model is simply impractical. Thus, based on the results shown in [Table 3](#), we can only say that the two-component mixture model of a heteroscedastic Normal Distribution and our semiparametric prior density process is marginally best. Whether the difference in DIC between the model with Normally distributed errors and the model with t distributed errors is statistically significant is unknown, but we proceed with the model assuming Normally distributed errors.

In addition to the models compared in [Table 3](#), we compare mixture models with differing grid resolutions for the semiparametric prior density process. The grid resolutions we look at include 3, 5, and 7 evenly-spaced knot locations in the snow density dimension and 3, 5, 7, and 9 evenly-spaced knot locations in the depth dimension, resulting in test grids with as few as 9 knots and as many as 63. The differences in DIC between these models, with respect to the model with the smallest DIC, are given in [Table 4](#). The model fit improves as the number of knots increases, but the increasing model complexity eventually leads to higher DIC values. Based on the results of the model fitting, we proceed with a grid that has 3 knot locations across the range of depth at 0, 80, and 160-m and 9 knot locations across the range of snow

Table 4: Differences in DIC between models with differing grid resolutions

Number		Density			
of Knots		3	5	7	9
Depth	3	823.5	1083.6	261.8	0.0
	5	815.6	450.2	1073.9	107.4
	7	849.6	411.8	229.6	90.0

density at 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, and 1.0 g/cm³. We choose this grid because it marginally has the best DIC value and because the coarser grid, relative to the 5×9 and 7×9 grids, reduces the computational time needed to fit the final model. Some grid resolutions finer than those presented in Table 4 are also explored, but it is found that they do not significantly improve the model, we therefore proceed with the 3×9 grid.

4 Application to Other Cores

To further illustrate the merit of our method, we apply it to an additional three cores to show that the method is general enough to be easily applied to other cores, despite the numerous decisions that were seemingly core-specific.

We perform further analyses on three additional core density records: one core (NGT27-B21) collected as part of the Alfred Wegner Institute’s North Greenland Traverse (NGT) (Wilhelms, 1996), one core (FB9801) collected from the pre-site surveying for the European Project for Ice Coring in Antarctica (EPICA) Dronning Maud Land deep ice core campaign (Oerter et al., 1999, 2000), and one core (SEAT2010-

3) collected as part of the Satellite Era Accumulation Traverse (SEAT) in the West Antarctic interior (Burgener et al., 2013). The SEAT core consists of density measurements at 2-cm resolution derived from volume and weight measurements of core sections while the other two cores use a gamma-ray attenuation technique (similar to that used on the WDC06A core) with resolutions of 1-3 mm. The choice of cores span a variety of spatial locations, total depths, and logging resolutions.

The first additional core we analyze is a high-resolution, deep core, like the WAIS core we use to motivate our methodology, but we don't have access to the raw data. Thus, the data available has already undergone some pre-processing that has removed most of the outliers we would expect to see far below the mean function. The second additional core is a high-resolution, shallow core, and we have access to the raw measurements. The final core isn't high-resolution and is also a shallow core. Since the third core isn't high-resolution we don't expect there to be as many outliers since the density measurements are averaged over a greater length. The maximum depths of the three cores are 96-m, 27.8-m, and 15.8-m, respectively, and the average depth between measurements is 1.3-mm, 5.8-mm, and 24.6-mm, respectively. These cores were selected for analysis because they are each distinct in their resolution, depth, and apparent contamination level. Thus, the flexibility of our modelling approach can be seen in these analyses.

One modelling difference of note is that the additional cores don't have distinct hardened surface layers like the WAIS core in our motivating example, which makes the cutoff parameter and second mixing parameter unnecessary. Thus, we remove both of these parameters and leave the other aspects of the model remain unchanged.

To summarize the analyses of the additional cores, scatterplots of the posterior prob-

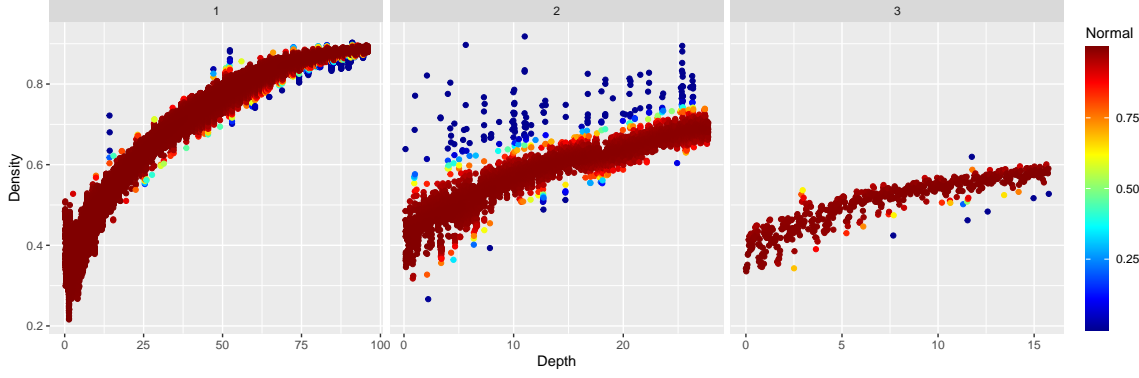


Figure 4: Scatterplots of the additional cores with the probability that an observation comes from the mean function

abilities of coming from the snow densification process for each core can be seen in Figure 4. Despite the pre-processing that each of these cores has undergone, there are clearly still some observations that likely aren't representative of the physical process of interest and should be considered outliers.

For each of the additional cores analyzed, we calculate the proportion of their observations that would be discarded for any given snow density probability threshold and plot the results in Figure 5. As a point of reference, at a threshold of 95%, the number of observations identified as outliers for each core is 841, 455, and 36, respectively. As can be seen, the trend in proportion and number of observations identified as outliers as a function of snow density probability for each core is similar to the others. Differences in the proportion of outliers can likely be attributed to the fact that each core came from a different location, was not obtained via the same data collection/measurement methodology, and underwent a different level of pre-processing prior to our analysis. The performance of our methodology when applied to these three cores suggests that it can be applied to the majority of snow cores, that don't exhibit deviations from monotonicity due to melt-refreeze layers, with very few

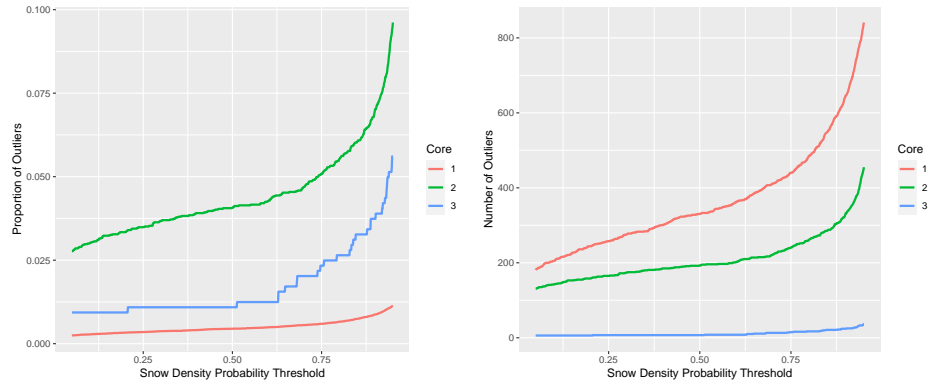


Figure 5: The proportion (Left) and number (Right) of observations that would be removed for a given snow density probability threshold from 5% to 95%, by core modifications.