Robust Bivariate Error Detection in Skewed Data with Application to Historical Radiosonde Winds

Amanda S. Hering, PhD
Assistant Professor
Department of Applied Mathematics and Statistics

COLORADOSCHOOLOFMINES
EARTH • ENERGY • ENVIRONMENT

Big Data and the Environment Workshop
Buenos Aires, Argentina
November 12, 2015

Golden, Colorado and CSM Campus

Collaboration With:

Ying Sun (KAUST)

Doug Nychka (NCAR–IMAGe) and
Joey Comeaux (NCAR–CISL)

Ashley Bell Anderson, former CSM
Master's student
Josh Browning, a current CSM
Ph.D. student

What is a Radiosonde?

- A small, expendable instrument package that is suspended below a 2 meter wide balloon filled with hydrogen or helium.
- Sensors on the radiosonde measure geopotential height, pressure, temperature, and dew point.
- By tracking the position of the radiosonde in flight, information on wind speed and direction is also obtained.

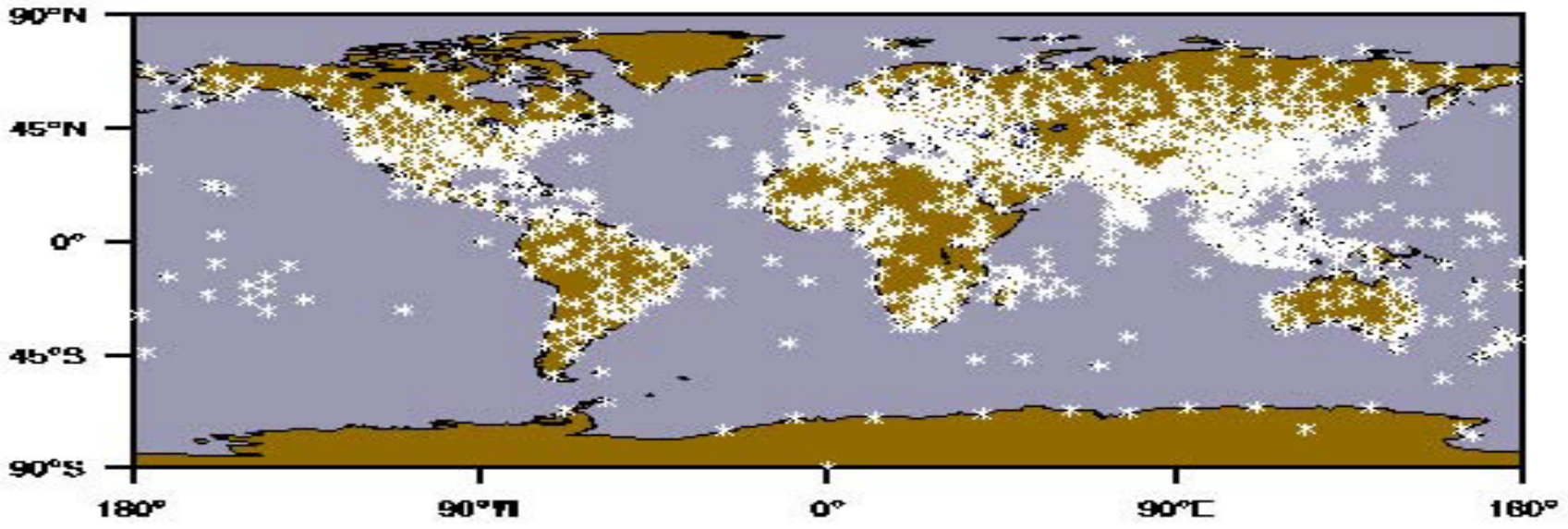
NWS Photo

Radiosonde Launches

Currently, balloons are launched twice a day at 00 UTC and 12 UTC at 700 sites worldwide.

Considering, (1) number of launch sites, (2) launches per day, (3) pressure levels recorded, and (4) the number of years, the NCAR radiosonde archive contains between

50 and 90 million soundings.



Radiosonde Data

The radiosonde archives contain the only measured data of the Earth's upper atmosphere.

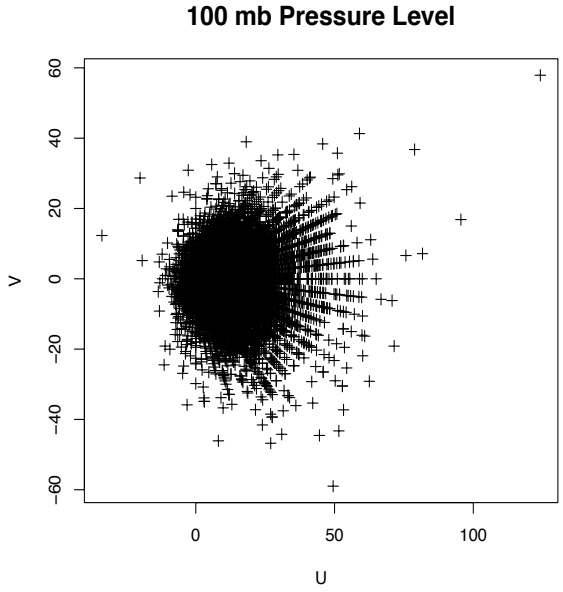
The data is currently used for:

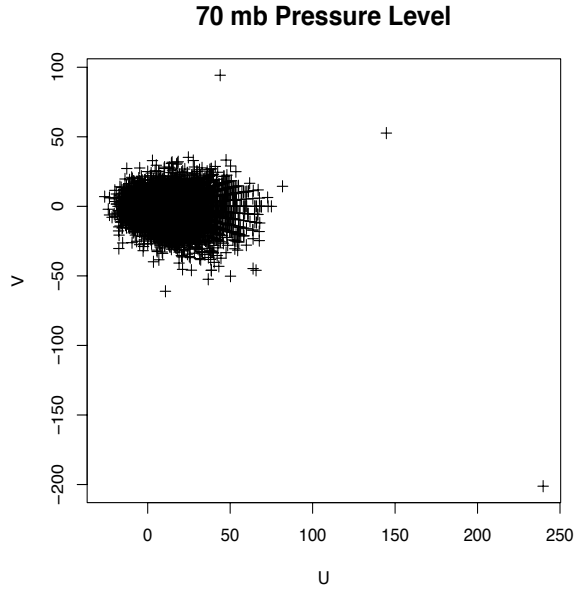
- Ground truth for satellite data
- Input for computer-based weather prediction models (it is **required** to produce weather forecasts!!)
- Weather and climate research
- Input to data assimilation products, such as NCEP reanalyses
- Boundary conditions in global and regional climate models

Problems with the Historical Record

Collected since the 1920's, these older records:

- Many errors (e.g., transmission, key punch, balloon motion)
- QC methods for current launches cannot be applied
  - For example, DART is used as a background model to validate the current radiosonde observations
- **Goal:** Automatically identify random errors, not extreme values or systematic errors.

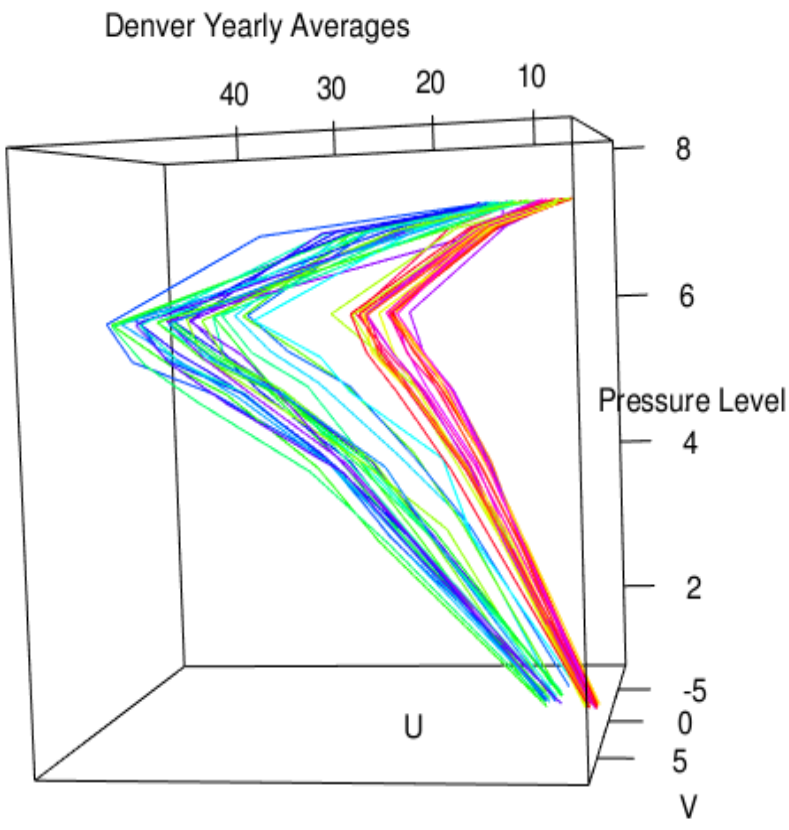




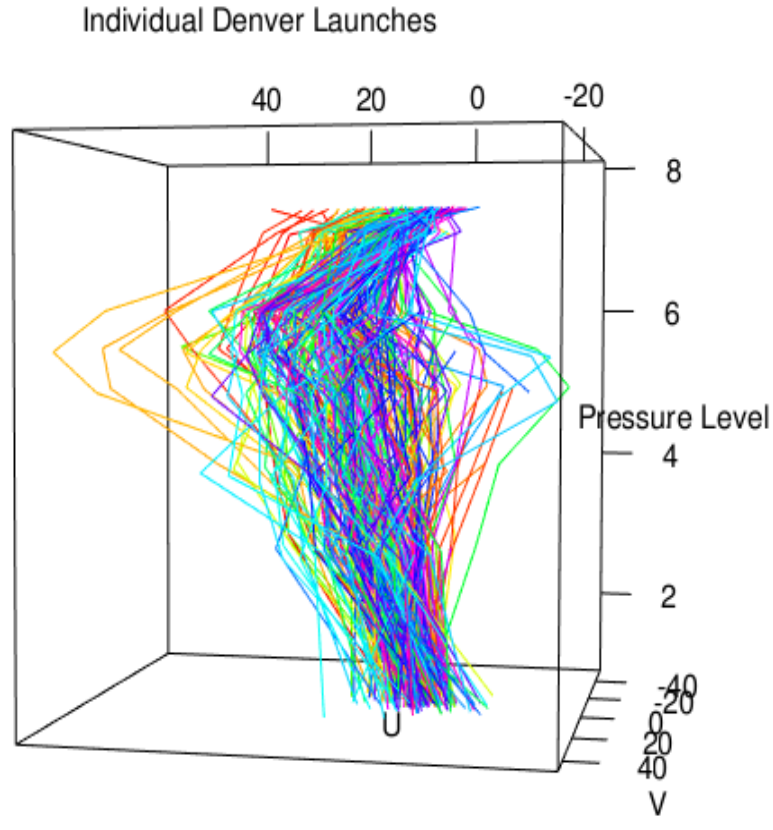
Denver Station Launches

50 Years of Launches from 1962 to 2011, Almost 36,000 Launches:

(a) Yearly Averages



(b) 230 Launches



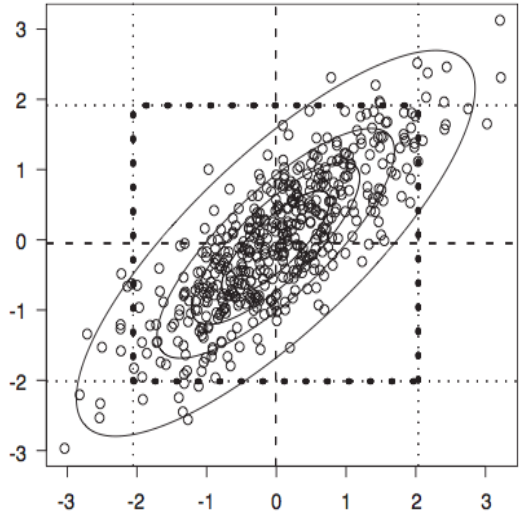
Since outliers do not occur in all directions with equal chance, MVN is not the best model to base outlier detection on.

Method I: Bivariate Normal (BVN)

- Filzmoser et al. (2005) assumes the underlying data generating process follows a multivariate normal distribution.
- Uses Mahalanobis distance to measure the distance of observations from the center of the distribution.
- Thus, for a p-dimensional multivariate sample, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, the Mahalanobis distance is

$$MD_i = ((\mathbf{x}_i - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}))^{1/2}.$$

- Then, $MD_i^2 \sim \chi_p^2$.



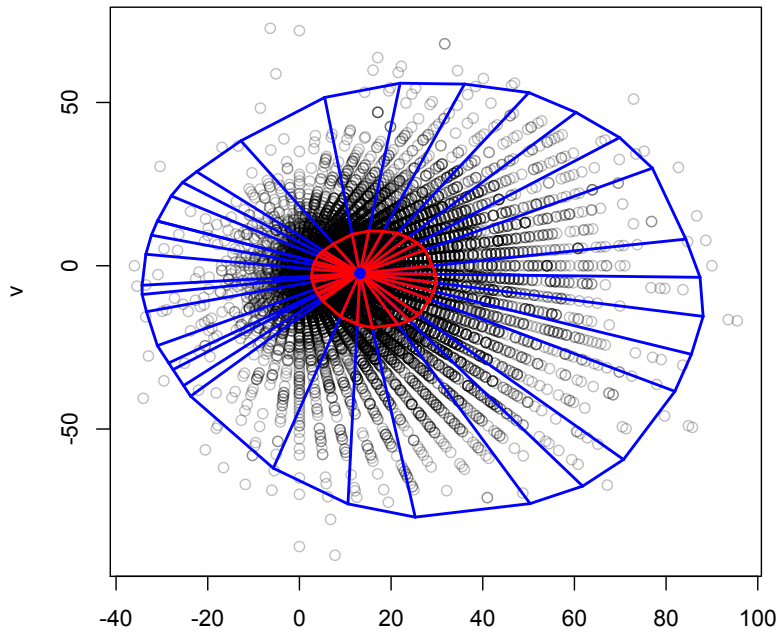
Method I: Bivariate Normal (BVN)

- The presence of outliers can strongly affect estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in MD_i, so this should be done robustly.
- Use the Minimum Covariance Determinant (MCD) (Rousseeuw 1985) determined by that subset of observations of size h which minimizes the determinant of the sample covariance matrix, computed from only these h points.
- They compare the upper tails of the empirical distribution function of the robust squared Mahalanobis distances with the theoretical χ^2_p distribution.
- They adjust the threshold for classifying an outlier based on the sample size and number of variables.

Filzmoser, P., Reimann C., and Garrett R. G. (2005) "Multivariate outlier detection in exploration geochemistry," *Computers and Geosciences*, 31: 579–587.

Method II: Bivariate Depth or Bagplot

1. For each pressure level, compute Tukey's depth values for each observation, $\mathbf{x}_j$, denoted $HD_i(\mathbf{x}_j)$.
2. Find the maximum $HD_i(\mathbf{x}_j)$, named the median.
3. Draw a convex hull that contains the 50% of points whose depths are the largest.
4. Find the distance from the median to each vertex on the convex hull.
5. Multiply the distance by 3 to obtain an outer convex hull.
6. Outliers fall outside of this fence. (Rousseuw et al. 1999)



Method III: Multivariate Skew-t

A p-dimensional random vector, $\mathbf{Y}$, following a multivariate skew-t (MST) distribution can be described by 4 sets of parameters (Azzalini and Capitanio 2003):

$$\mathbf{Y} \sim MST_p(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu), \quad \text{where}$$

- $\boldsymbol{\xi}$ is $p \times 1$, $\approx$ the center
- $\boldsymbol{\Omega}$ is $p \times p$, $\approx$ the variance-covariance matrix
- $\boldsymbol{\alpha}$ is $p \times 1$, $\approx$ the skewness in each dimension
- ν is scalar, $\approx$ the heaviness of the tails

The MVN is a special case when $\boldsymbol{\alpha} = (0, \ldots, 0)$ and $\nu = \infty$.
The lower ν is, the heavier the tails are.

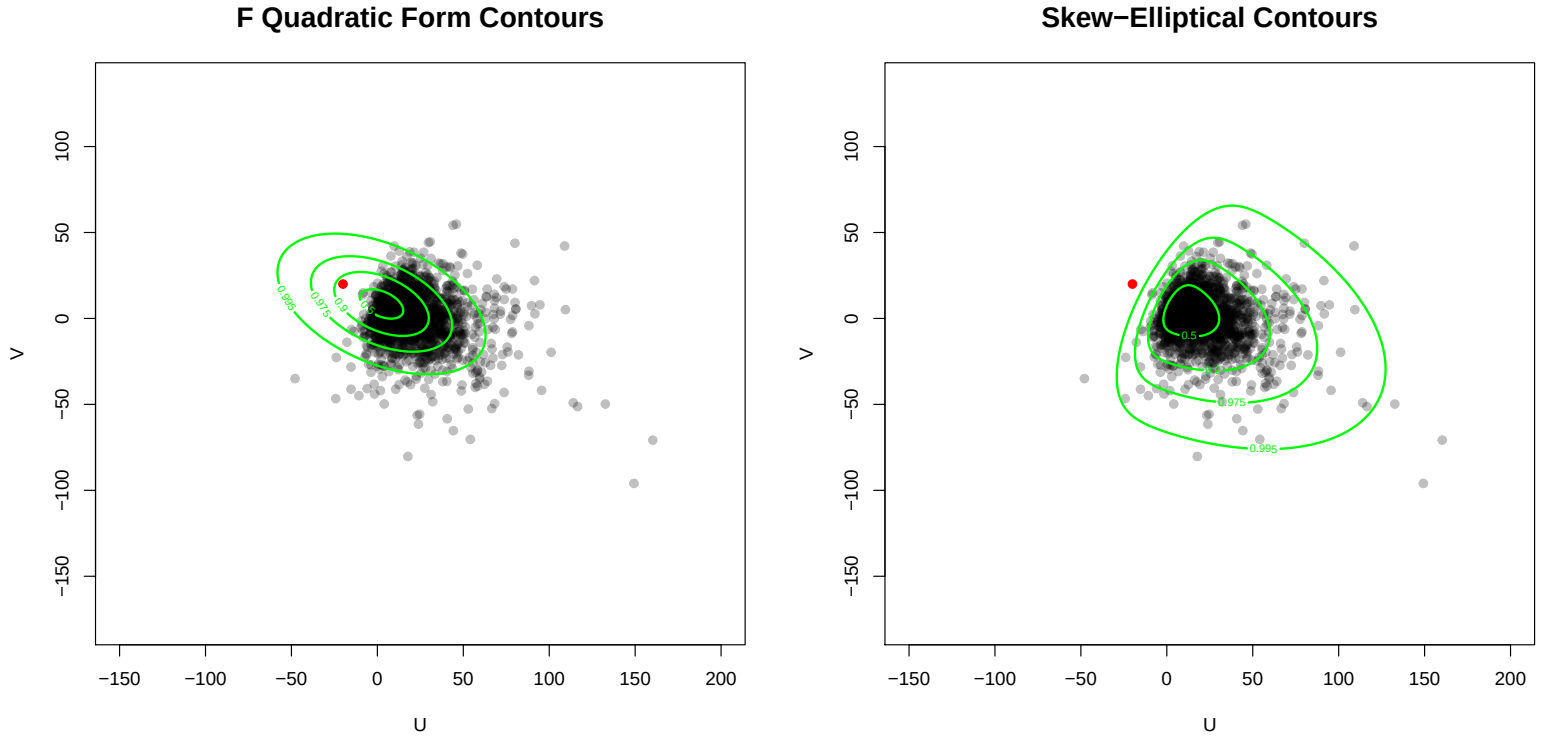
Method III: Multivariate Skew-t—Distance-Based

- The quadratic form, $(Y - \boldsymbol{\xi})^T \boldsymbol{\Omega}^{-1} (Y - \boldsymbol{\xi}) \sim p \cdot F(p, \nu)$.
- We apply a method similar to the MVN approach of Filzmoser et al. (2005), but we replace the robust Mahalanobis distances based on the MVN to the MST distances.
- However, we found that $(Y - \boldsymbol{\xi})^T \boldsymbol{\Omega}^{-1} (Y - \boldsymbol{\xi})$ is very sensitive to misspecification of ν, and the presence of outliers in the data strongly affects the estimation of ν, making it lower than it should be.

Two Problems:

- The contours should be elliptically skewed.
- We need robust estimates of the parameters in the ST distribution.

Method III: Multivariate Skew-t—Skew-Elliptical Contours



Find the region $R_{BST} \subset \mathbb{R}^2$ of smallest geometrical size such that $P(\mathbf{Y} \in R_{BST}) = (1 - \alpha)$.

The solution must be of the type

$$R_{BST} = \{\mathbf{y} : f_{BST}(\mathbf{y}; \boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu) \geq f_0\},$$

where f_{BST} is the pdf of the BST, and f_0 is a value ensuring that $P(\mathbf{Y} \in R_{BST}) = (1 - \alpha)$.

Robust Estimation of Multivariate Skew-t

- Let $\boldsymbol{\theta} \in \mathbb{R}^d$ represent the vector of d parameters.
- We develop an M-estimator, which takes the following general form for a single parameter:

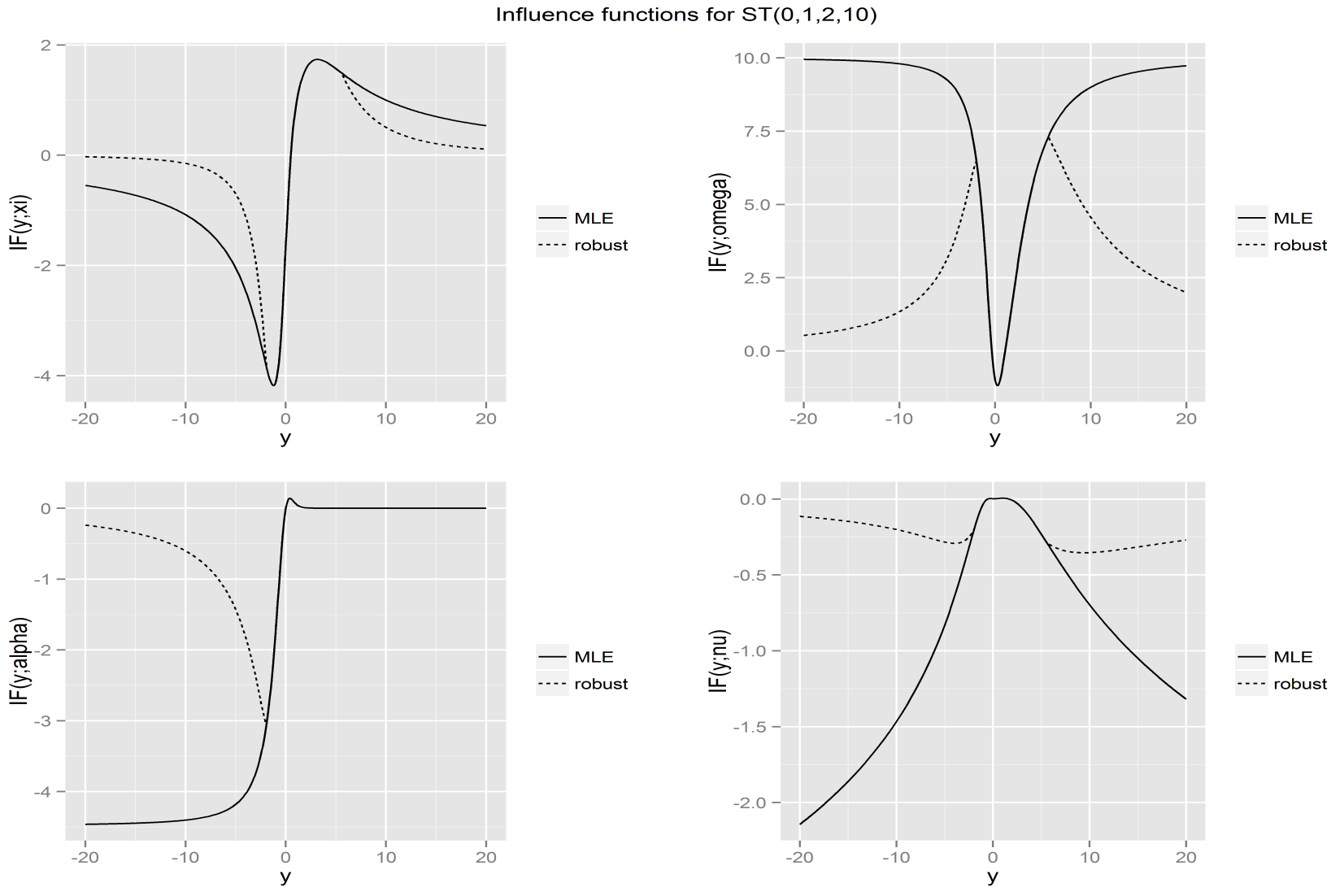
$$\hat{\theta}_j = \arg\min_{\theta_j} \left(\sum_{i=1}^{n} \rho(\mathbf{y}_i, \boldsymbol{\theta}) \right),$$

  where $\rho(\mathbf{y}, \boldsymbol{\theta})$ can be set to $-\log(f(\mathbf{y}, \boldsymbol{\theta}))$ to obtain the MLE.
- The derivatives of the negative log-likelihood for all parameters of the skew-t distribution can be obtained except ν, which must be numerically differentiated.
- An estimator is generally considered robust if its influence function is bounded.

We propose a redescending M-estimator, which has an influence function that is non-decreasing near the origin but decreases to 0 as $|\mathbf{y}| \to \infty$.

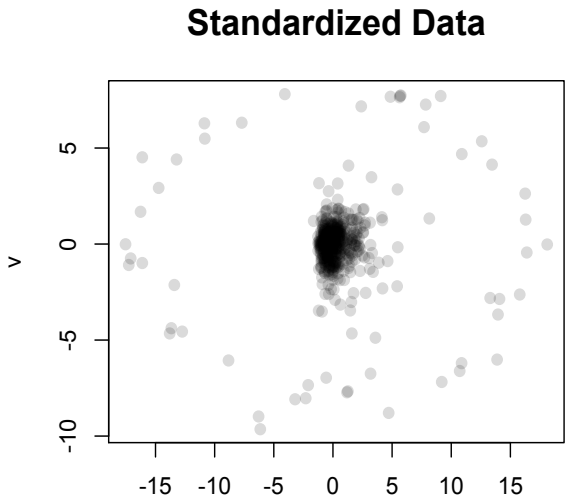
$$\rho_{\text{robust}}(\mathbf{y}, \boldsymbol{\theta}) = \begin{cases} \rho(\mathbf{y}, \boldsymbol{\theta}) & \text{if } \rho(\mathbf{y}, \boldsymbol{\theta}) \leq k, \\ 2k - ke^{-\rho(\mathbf{y}, \boldsymbol{\theta})/(k+1)} & \text{if } \rho(\mathbf{y}, \boldsymbol{\theta}) > k \end{cases}$$

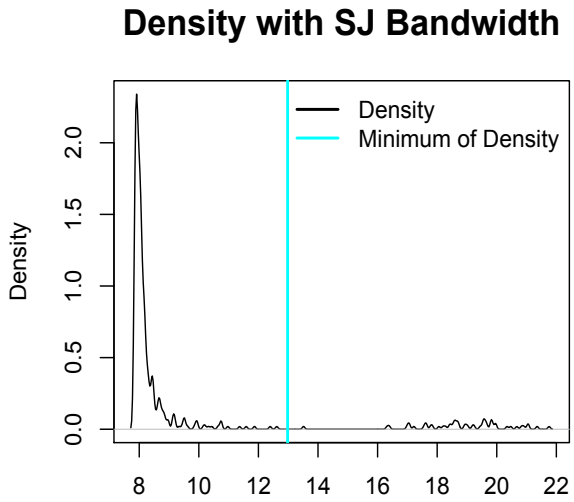
Robust Estimation of Multivariate Skew-t

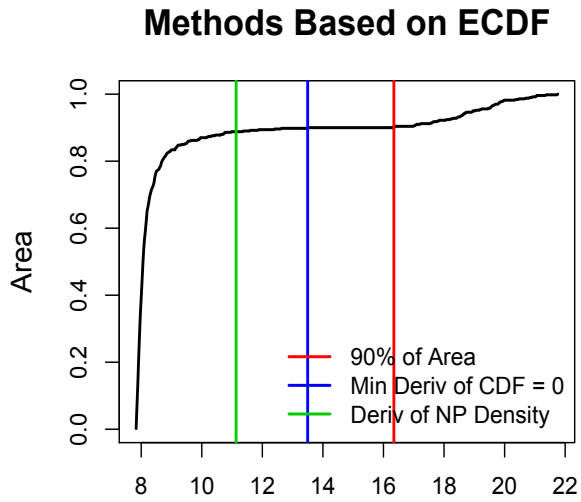


Methods for Selecting k

- (a) The minimum of the nonparametric (NP) density of the non-robust likelihood values of the observations.
- (b) The $90th$ quantile of the empirical cumulative distribution function (ECDF) of the non-robust likelihood values of the observations.
- (c) The minimum value of the ECDF such that the derivative is zero for the longest stretch.
- (d) From first and second derivatives of the cumulative NP density such that their changes are within prespecified tolerances.
- (e) A fixed choice of $k = 10$, which was found to work well in many situations in prior simulations studies.







Simulation to Study Parameter Estimates for a Chosen k

- 250 datasets simulated for each combination of settings.
- The center parameter is $\boldsymbol{\xi} = (16.51, -1.22)'$.
  - Skewness:
    - SYM = Symmetric with $\boldsymbol{\alpha} = (0, 0)^T$ and $\nu = \infty$
    - OBS = Observed with $\boldsymbol{\alpha} = (1, -1)^T$ and $\nu = 10$
    - EXT = Extreme with $\boldsymbol{\alpha} = (6, 0)^T$ and $\nu = 5$
  - Variability:
    - Five values for $\boldsymbol{\Omega}_i$ increasing in variability from $i = 1$ to 5.
  - Percent Contamination:
    - The data has been contaminated with 0, 5, or 10% outliers.
  - Type of Outliers:
    - (a) All around the main cloud of points or (b) in the direction of the tail of the distribution.
  - Sample Sizes:
    - Sample sizes $n \in \{100, 250, 500, 750, 1000\}$ were generated.

Simulation to Study Parameter Estimates for a Chosen k

1. Let $\mathbf{X} = (\mathbf{u}, \mathbf{v})' \in \mathbb{R}^{n \times 2}$ be the contaminated simulated data, so $\mathbf{X} \sim ST_2(\boldsymbol{\xi}, \boldsymbol{\Omega}_i, \boldsymbol{\alpha}, \nu)$.
2. Center and scale the data as follows:

$$\mathbf{Y} = \mathbf{A}(\mathbf{x} - \mathbf{a}) = \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{a},$$

   where $\mathbf{a} = (M_u, M_v)'$ be the vector of medians of $\mathbf{u}$ and $\mathbf{v}$, and

$$\mathbf{A}^{-1} = \begin{pmatrix} MAD(\mathbf{u}) & 0 \\ 0 & MAD(\mathbf{v}) \end{pmatrix}.$$

3. Then, $\mathbf{Y} \sim ST_2(\mathbf{A}(\boldsymbol{\xi} - \mathbf{a}), \mathbf{A}\boldsymbol{\Omega}_i\mathbf{A}, \boldsymbol{\alpha}, \nu)$.
4. Then, one of 5 methods is applied to select k.
5. The parameters are estimated and saved for each of the 5 choices of k above.

Simulation to Study Parameter Estimates for a Chosen k

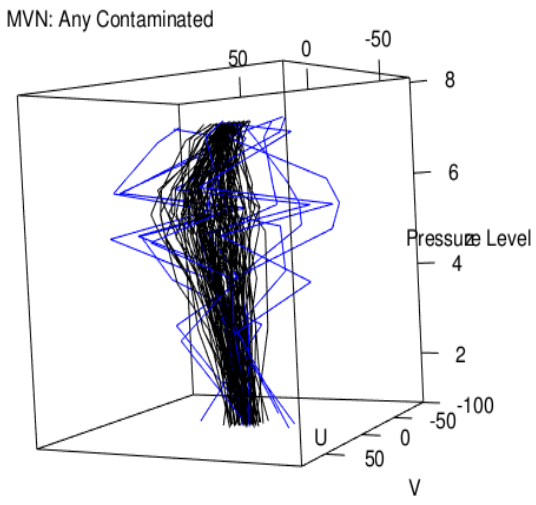
- Some variation across n with larger sample sizes leading to larger values of k.
- However, methods (d) and (e) are comparable.

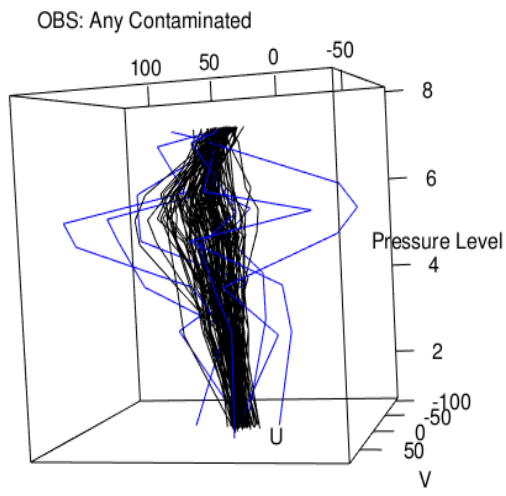
		Average MSE of Parameters Excluding ν				
		10% Angle				
	$\mathbf{\Omega}_i$	(a)	(b)	(c)	(d)	(e)
EXT	1	0.97	1.07	1.00	0.66	0.53
	2	0.84	0.80	0.99	0.64	0.49
	3	0.69	0.88	0.74	0.54	0.48
	4	1.18	1.15	1.22	0.79	0.60
	5	1.15	1.26	1.17	0.77	0.61
		Median of Estimated ν				
		10% Angle				
	$\mathbf{\Omega}_i$	(a)	(b)	(c)	(d)	(e)
EXT	1	1.84	1.67	1.85	2.10	2.51
	2	1.83	1.69	1.82	2.08	2.49
$\nu = 5$	3	1.87	1.65	1.89	2.12	2.50
	4	1.81	1.69	1.82	2.06	2.50
	5	1.77	1.70	1.80	2.07	2.50
		Average k Chosen				
		10% Angle				
	$\mathbf{\Omega}_i$	(a)	(b)	(c)	(d)	(e)
EXT	1	14.43	15.61	15.47	11.65	10
	2	14.70	15.45	16.21	11.78	10
	3	13.58	15.74	14.59	11.47	10
	4	15.42	15.36	16.14	11.84	10
	5	15.84	15.39	16.36	11.84	10

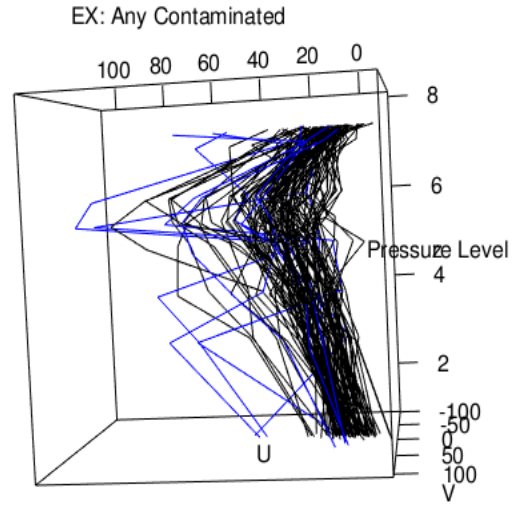
Outlier Simulation Design

1. Simulate 500 launches with one of 3 types of skewness:
   - MVN: no skewness
   - OBS: similar skewness to what is observed in the data
   - EX: more extreme skewness than observed in the data
2. Contaminate the launches with no outliers or with 5% or 10% of one of 3 types of outliers:
   - An entire launch
   - Random higher levels of a launch
   - Any random level within a launch
3. Apply each of 10 methods to the simulated data:
4. Record the percentage of
   - TP = True Positives (correctly flagged outliers), high
   - FP = False Positives (incorrectly flagged non-outliers), low
5. Repeat steps 1-4 1,000 times.

Any Level Contaminated

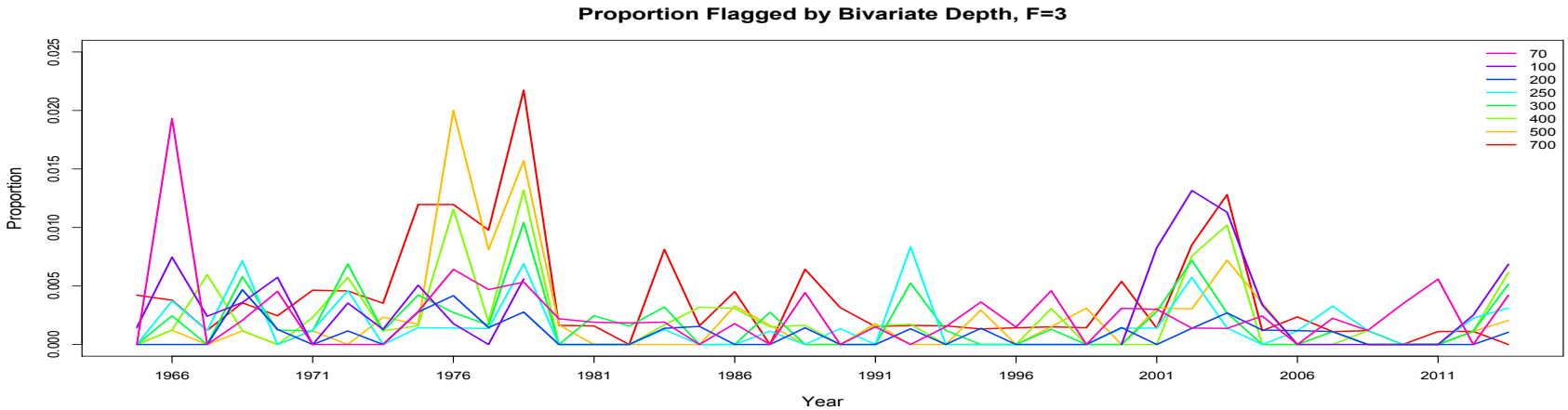


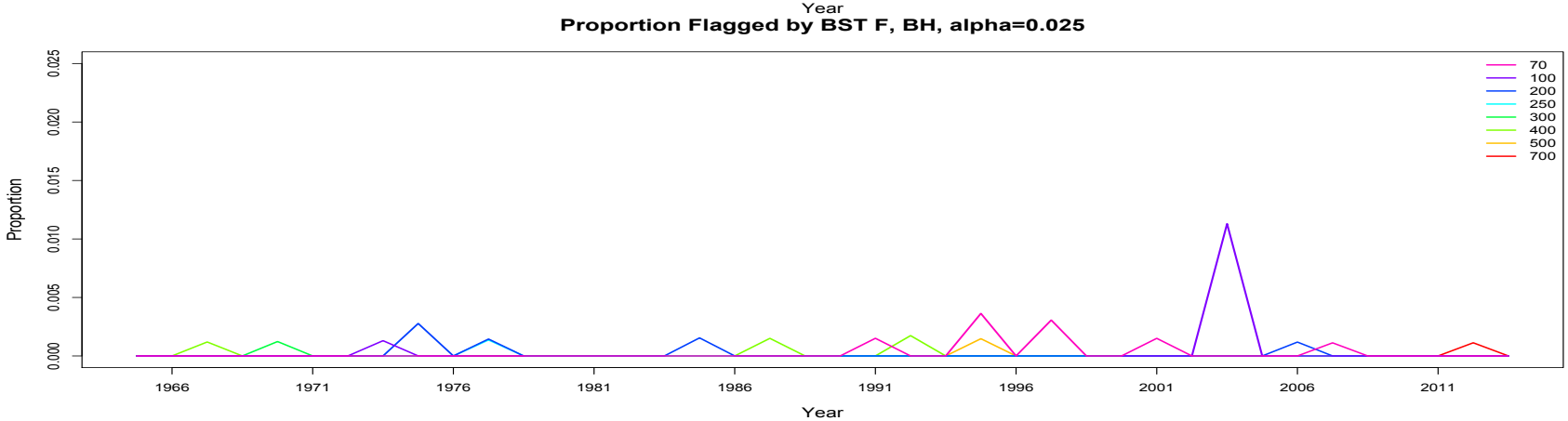


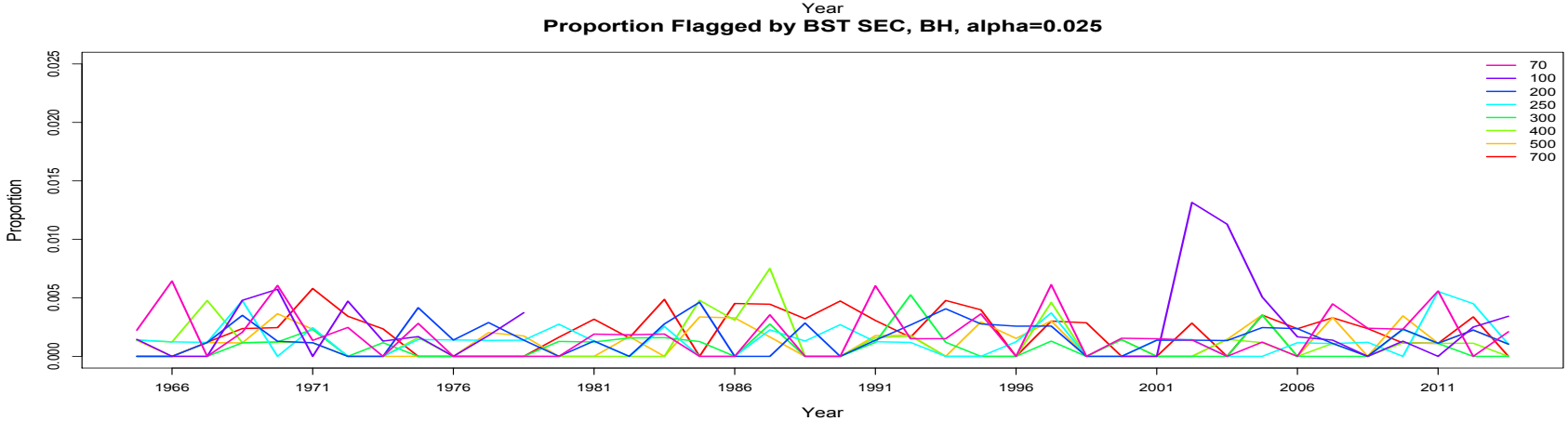


Distribution		BVN	Depth	BST (SEC)		
				true	mle	robust
MVN	TP	92.7	93.1	100	93.4	99.9
	FP	0.22	0.30	2.50	0.01	1.62
OBS	TP	92.9	95.9	99.8	88.5	99.8
	FP	0.22	0.30	2.49	0.15	2.68
EX	TP	88.5	93.2	98.6	68.6	99.1
	FP	0.91	1.11	2.45	0.46	3.61

Case Study: Denver Station

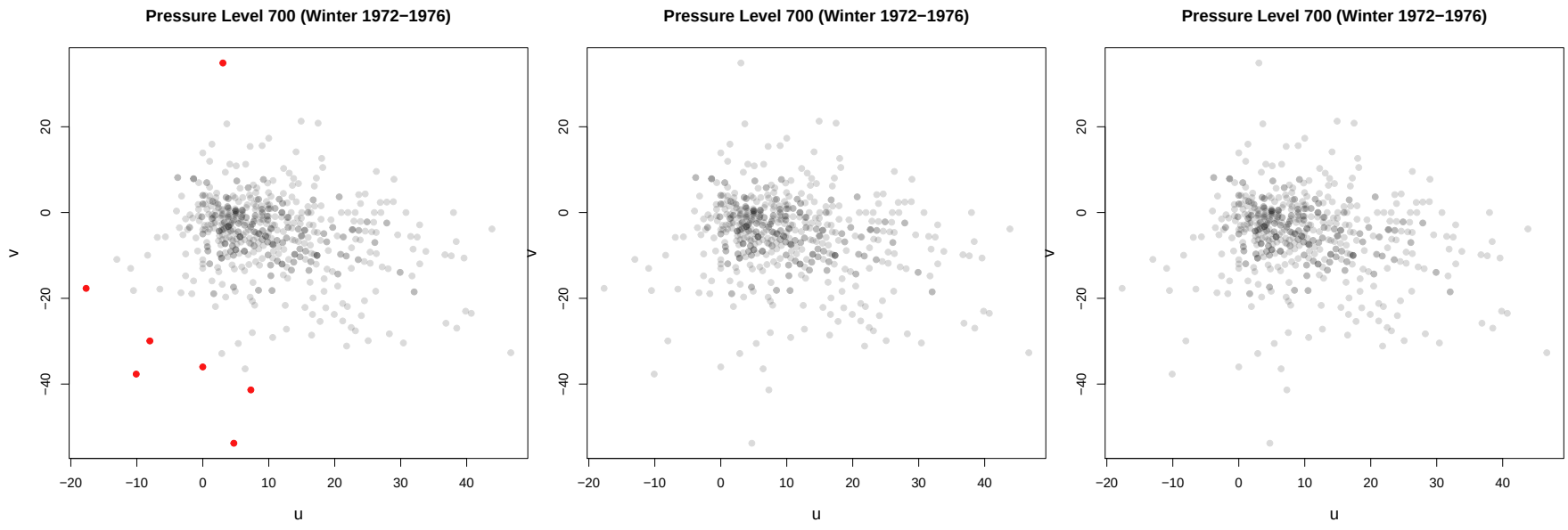






Case Study: Denver Station

Depth **BST F** **BST SEC**



- The depth methods flags too many outliers.
- The SEC method picks up some outliers that the F method does not.

Future Work

- Develop a multivariate outlier detection method for the entire vertical column, but parameterize $\mathbf{\Omega}$ to reflect the structure in the pressure levels; otherwise, we would have 169 parameters for 8 pressure levels.
- Time: the method must be fast and automatic to process the millions of records in the archive quickly
- Systematic/Climatic Changes: we want to minimize the number of observations that we run through the method that have different data generating processes
  - Windows of observations
  - Diurnal variability

Thanks for your attention!!