Knorr-Held and Raßer Implementation Summary - Sam Leonard

Here's a quick summary of my duplication of the MCMC described by Knorr-Held and Raßer in their paper *Bayesian Detection of Clusters and Discontinuities in Disease Maps*. I assume the reader is familiar with their work, so I start with a summary of my results implementing their MCMC. The second half describes in brief the details of their algorithm.

Part 1: MCMC Details and Results

After $1,000,000$ burn-in I ran the MCMC for $100,000,000$ iterations, saving every $10,000^{\text{th}}$ iteration. The starting number of cluster centers k was randomly selected from $\{1, ..., 544\}$, and the k cluster centers were selected at random from the 544 available regions.

Observed heights for this dataset varied from .14 to 2.4. μ was initially selected at random from a uniform distribution on the log of this range. The initial value for σ^2 was drawn from an inverse gamma distribution with $\alpha = 1$ and $\beta = .01$.

With k and G_k selected, the vector H_k was then drawn from a gamma approximation the of $\prod L(y_i|h_j)p(h_j)$. The exact form of this approximation is decribed in the birth step in part 2.

I calculated the autocorrelation for each of the relative risk samples in each region. For a lag of one, median autocorrelation was .03, with a maximum of .5. For all but 7 of the 544 regions, the autocorrelation for a lag of five was less than .1.

The table below summarizes acceptance rates observed by the authors and in my own implementation.

Step	Author Acceptance	Leonard Acceptance
Birth	.24	.23
Death	.24	.23
Shift	.21	.20
Switch	.41	.40
Height	.98	.44 or .68

The two acceptance rates for the height step reflect different methods of calculating this rate. Each $h_j^* \in H^*$ is accepted or rejected separately. The .44 acceptance rate was calculated treating each individual h_j^* as an attempt to a new height. Thus each time the height step was selected there were k attempts at new heights, and .44 is the proportion of these that were accepted.

However, we could also treat any change in the vector H overall as an accepted height step. In other words, if at least one $h_i^* \in H^*$ was accepted, the vector H was altered and thus we could consider the height step as being accepted. This thinking is the .68 acceptance rate.

In either case, the acceptance rate for the height step is the only rate markedly different from the authors' implementation. As of right now I am not certain of the cause. Given that my results are nearly identical to the authors' in the crucial components of this algorithm I do not consider this difference in acceptance rate to be significant.

The range of posterior median estimates for the relative risk I obtained were nearly identical to the authors' results. I obtained estimates between .64 and 1.41, while the authors had estimates that varied between .65 and 1.42.

The posterior median for k was 40, identical to what the authors obtained given $c = .02$. The posterior distribution of the number of clusters k is presented in the graph below, along with its trace plot. This posterior plot is unsmoothed and drawn directly from the incidence of k values in the $10,000$ samples obtained. Nevertheless, the posterior distribution is essentially identical: unimodal but not symmetric, values for $k > 40$ being slightly favored.

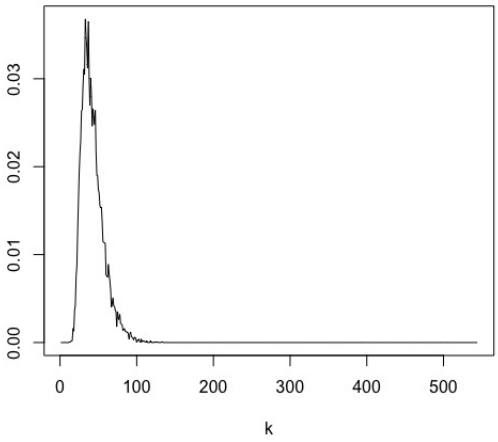

Figure 1: Posterior Distribution of k

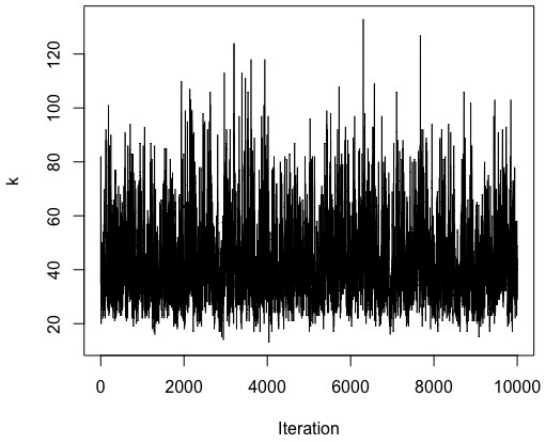

Figure 2: Trace Plot for k

Figure 3 displays the median height for each region, ignoring clustering. Plots of median heights for each region are nearly identical to the authors' plot. Regions of relative risk above 1.2 in Mecklenburg-West Pomerania (northeast), Saarland (southwest), and Franconia (east-central region directly south of lowest risk) correspond to the authors' results.

West Berlin (southernmost region of elevated risk in the northeast) has a conspicuously high relative risk compared to its neighbors, precisely the result the authors obtained. I obtained median risk of 1.22 for this, identical to the authors' result of 1.22.

Kiel (tiny region in the very north) is another region with median risk higher than its immediate neighbors. This run produced a median risk of 1.15 for this region, similar to the authors' result of 1.13.

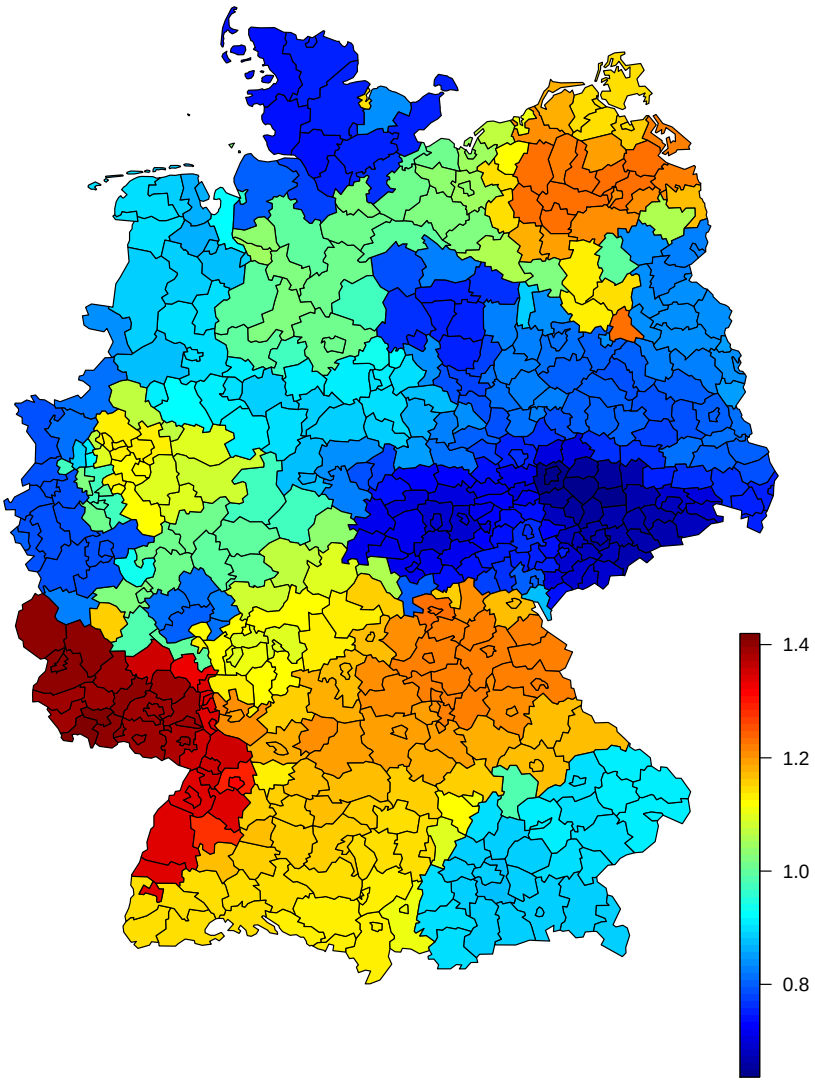

Figure 3: Median Height by Region

I used Lau and Green's (2007) extension of Binder's loss function to obtain an optimal clustering configuration. This optimal clustering is found by finding the cluster configuration $\hat{C}$ which minimizes over all observed clusters the function:

$$E(L(C,\hat{C})|y) = \sum_{i,j \in n} a\mathbb{I}[\hat{c}_i \neq \hat{c}_j]p(c_i = c_j|y) + b\mathbb{I}[\hat{c}_i = \hat{c}_j]p(c_i \neq c_j|y)$$

In this notation for regions i and j, $c_i = c_j$ if i and j are in the same cluster for a given partition C. Using $a = b = .5$, the optimal clustering configuration is plotted below in Figure 4. This configuration consists of 78 Shade in this plot does not correspond to magnitude of relative risk, the map is shaded simply to distinguish between clusters.

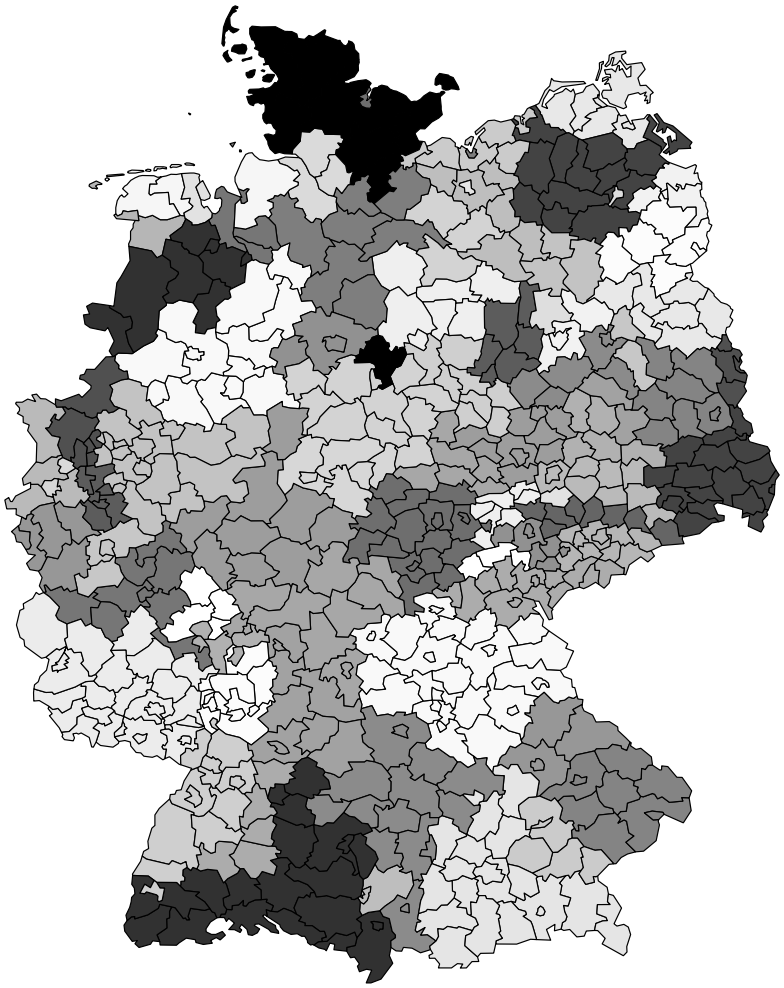

Figure 4: Lau/Green Optimal Clustering

This plot is clearer if we maintain this cluster configuration, but color each region by the median height taken from the list of median heights from each region in the cluster. This is figure 5.

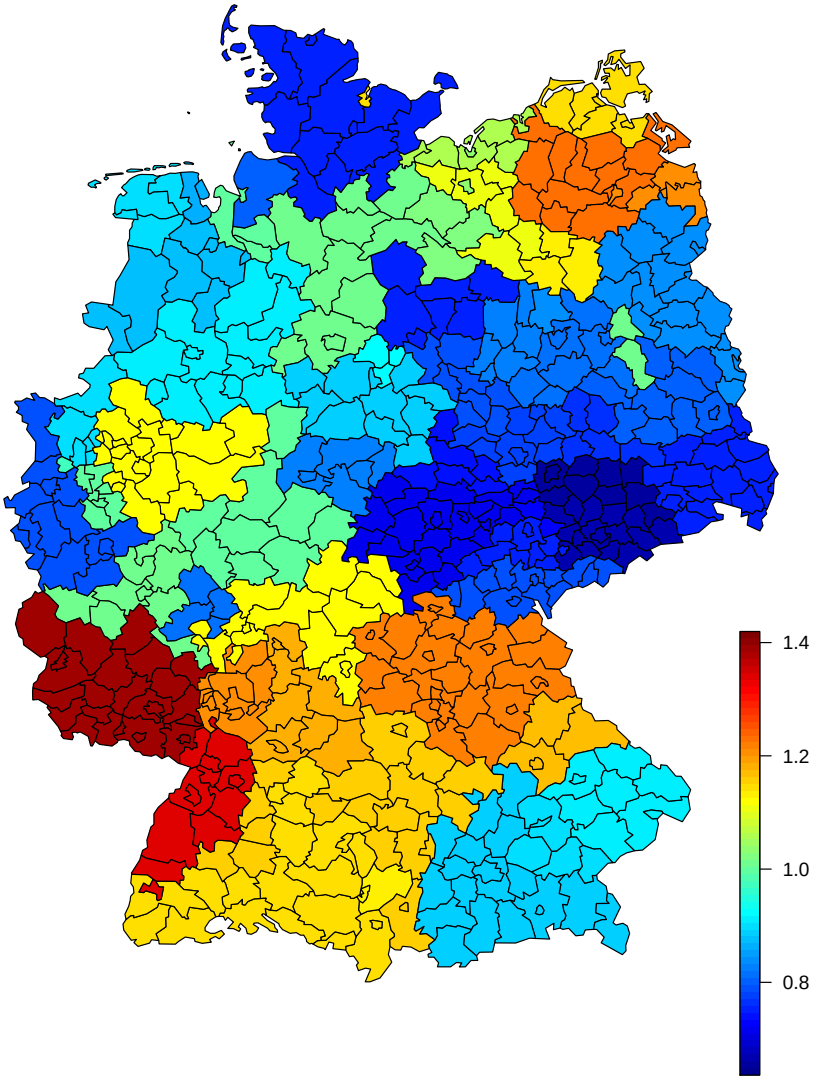

Figure 5: Optimal Clustering Colored by Median Height

Figures 6 and 7 display the pairwise probability two regions belong in the same cluster. Axis labels are omitted due to the number of regions involved. Figure 6 orders regions from $1 \rightarrow 544$. In Figure 7 the regions are reordered to correspond to the optimal clustering obtained by the Lau/Green function.

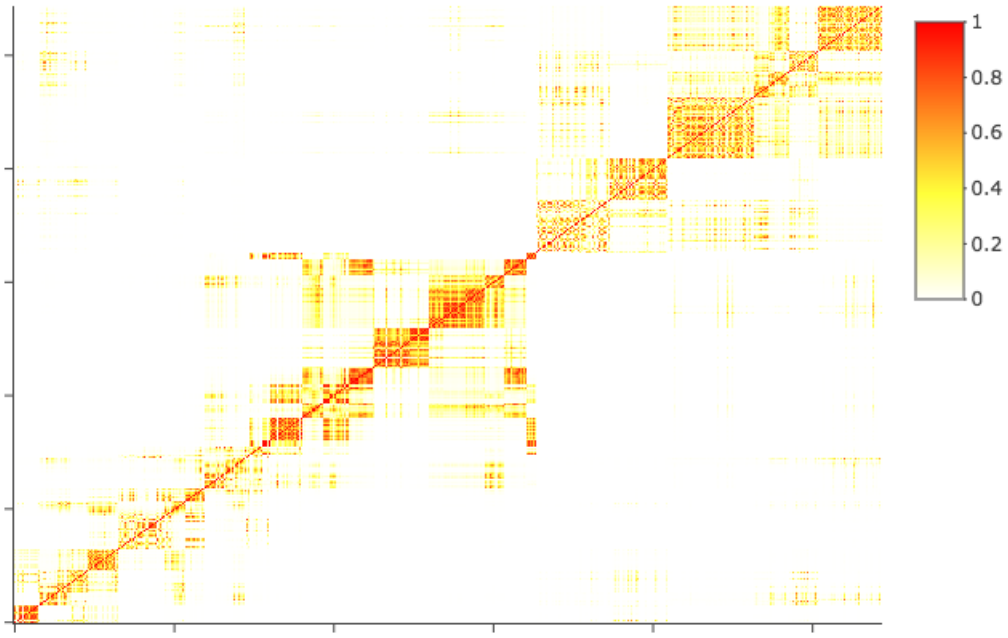

Figure 6: Posterior Probability Regions i, j in Same Cluster; Order of Regions Preserved

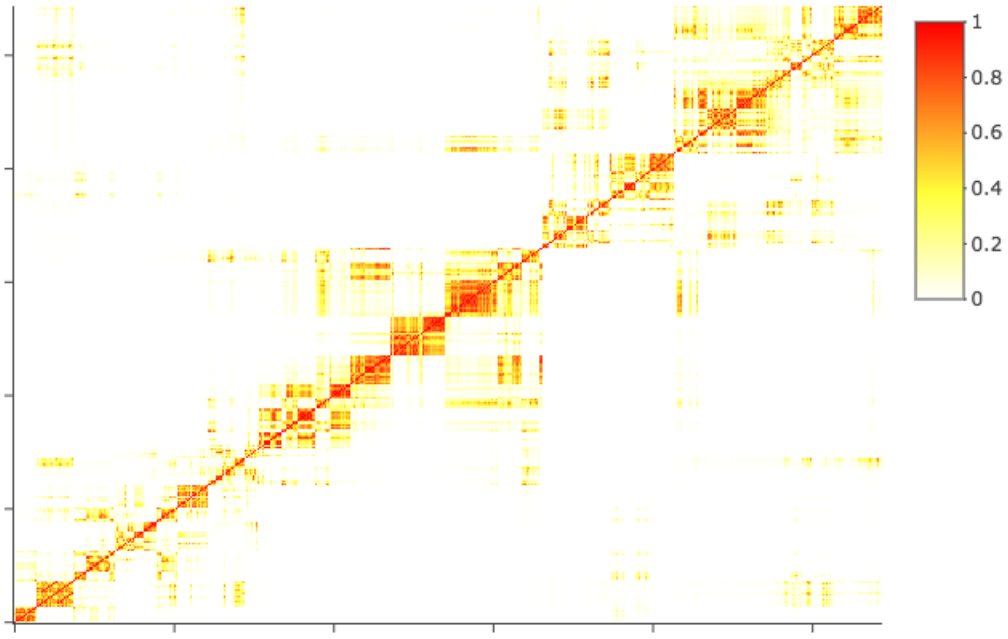

Figure 7: Posterior Probability Regions i, j in Same Cluster; Reordered According to Lau/Green

Part 2: Model Description

Likelihood

Consider a geographic area divided into n regions. A country with n districts, for example. For each of these regions we have two data points: y_i, $i \in 1, ..., n$, is the observed number of individuals suffering from a given disease in region i, and the expected number of cases in each region e_i.

The authors' model partitions a geographic region into k clusters, and assumes that the relative risk h_j, $j \in 1, ..., k$, is constant within each cluster. Let H_k denote the vector of relative risks. Clusters are denoted C_j, and the number of clusters k is unknown.

The authors assume each y_i has a Poisson distribution with $\lambda = e_i h_j$. Therefore, given k clusters, a corresponding cluster configuration, and a relative risk h_j for each cluster, the likelihood of observed data y_i and e_i is:

$$L(y|H_k, k, G_k) = \prod_{j=1}^{k} \prod_{i \in C_j} \frac{(e_i h_j)^{y_i}}{y_i!} e^{-e_i h_j}$$

Clustering

Consider again a geographic area divided into n regions. The authors define the distance between regions i and j as the minimum number of borders one must cross in order to travel from the interior of region i to the interior of region j.

To form a clustering configuration, select without replacement k regions from the set of n regions. Preserve the order in which this selection occurs. These selected regions are cluster centers, and they completely determine a cluster configuration. Let G_k represent this vector of cluster centers. A cluster configuration is formed by assigning each region that is not a cluster center to its closest cluster center.

The distance between a non-center and two separate cluster centers could be equal. Such regions are assigned to the cluster center that appears first in the vector of cluster centers G_k.

Priors

The authors assume a truncated geometric prior for k: $p(k) \propto (1-c)^k$ for some fixed c. In their analysis of the German dataset $c = .02$.

The vector G_k is assumed to be uniform across all possible lists of cluster centers. Given k there are $\binom{n}{k}$ possible selections for cluster centers, and $k!$ orderings of these k cluster centers. Thus $Pr(G_k|k) = \frac{1}{\binom{n}{k} * k!} = \frac{(n-k)!}{n!}$.

Heights are assumed to be symmetrically distributed on a lognormal distribution with with $p(\mu)$ proportional to the whole real line and $p(\sigma^2)$ an inverse gamma distribution. In their analysis of the German data set, $p(\sigma^2) = \text{Inv-Gam}\ (1, .01)$.

MCMC

Each iteration of their MCMC executed one of six steps: birth, death, shift, switch, height, and hyperparameter. The probability each step is selected is $(.4, .4, .05, .05, .05, .05)$ respectively.

Birth: Given k cluster centers and a corresponding vector G_k there are $n - k$ regions that are not cluster centers. Choose one at random, and insert it at random into G_k. This is the proposed vector of cluster centers. Draw a proposed height h_j^* for this new region from a gamma distribution:

$$\alpha = \Big(\sum_{i \in C_j^*} y_i \Big) + \frac{(\mu + \frac{1}{2}\sigma^2)^2}{(e^{\sigma^2} * (e^{\sigma^2} - 1) * e^{2\mu})^2}$$

$$\beta = \Big(\sum_{i \in C_j^*} e_i \Big) + \frac{(\mu + \frac{1}{2}\sigma^2)}{(e^{\sigma^2} * (e^{\sigma^2} - 1) * e^{2\mu})^2}$$

Here C_j^* represents the new, birthed cluster. This distribution is an approximation of the full conditional $p(h_j|\cdot)$.

Only regions whose heights have changed will influence the likelihood. So, after some simplification, the birth step is accepted with probability:

$$r = \min\Big\{1, \frac{L(y_i^*|H_k^*, \mu, \sigma^2, k, G_k^*)}{L(y_i^*|H_k, \mu, \sigma^2, k, G_k)} * \frac{p(h^*|\mu, \sigma^2)}{q(h^*|\mu, \sigma^2)}\Big\}$$

Here $p(h^*)$ is the density of a lognormal distribution evalauted at h^*, while $q(h^*)$ is the density of the approximating gamma distribution at h^{**}. And y_i^* represents those clusters whose heights changed due to the birth of a new cluster.

Death: Select a cluster at random from your list of k cluster centers G_k. Eliminate this cluster center and its corresponding height. You are left with a proposal configuration of $k-1$ clusters and $k-1$ heights.

This step is the reverse of a birth step, its acceptance probability is the inverse of the birth step acceptance probability.

Shift: In the list of cluster centers, let $n(G_k)$ be the number of cluster centers that are adjacent to regions which are not in G_k. These cluster centers are eligible for the shift step.

Select one of these eligible cluster centers at random and call it g_j. Let $m(g_j)$ be the number of clusters adjacent to g_j that are not cluster centers. Select one of these regions at random and call it g^*. Replace g_j with g^* to form a new cluster configuration G_k^*.

The shift step is accepted with probability

$$r = \min\Big\{1, \frac{L(y|H_k^*, \mu, \sigma^2, G_k^*)}{L(y|H_k^*, \mu, \sigma^2, G_k^*)} * \frac{n(G_k)m(g_j)}{n(G_k^*)m(g^*)}\Big\}$$

Switch: From the list of k cluster centers, select two at random. Switch their order in the list of cluster centers G_k to create a new vector G_k^*. Because order matters in this list, this will create a new cluster configuration C_j^*.

The authors do not make clear if the heights corresponding to these switched cluster centers are also swapped. However, in my runs of their MCMC leaving the vector of cluster centers in its original order yields a low acceptance rate, about .05, while switching the heights to follow the cluster centers yields an acceptance rate in their range, about .41. Therefore in my implementation I switched the heights corresponding to the switched cluster centers.

Note that the switch step is symmetric with respect to its jumping probability. The probability of jumping from G_k to G_k^* is identical to jumping from G_k^* to G_k. Therefore the probability of accepting the switch step depends only on its likelihood ratio.

Hyperparameters: The hyperparameters μ and σ^2 are updated by a Gibbs step:

$$\mu|\cdot \sim N\Big(\frac{1}{k}\sum_{j=1}^{k}\log(h_j), \frac{1}{k}\sigma^2\Big)$$

$$\sigma^2|\cdot \sim \text{IG}\Big(a + \frac{k}{2}, b + \frac{1}{2}\sum_{j=1}^{k}\big(\log(h_j) - \mu\big)^2\Big)$$