Department of Mathematics
TUM School of Computation, Information and Technology
Technical University of Munich

Robust Score Matching for Graphical Models

Richard Schwank

Thesis for the attainment of the academic degree

Master of Science

at the TUM School of Computation, Information and Technology of the Technical University of Munich

Supervisor:
Prof. Mathias Drton, Ph. D.

Advisors:
Andrew McCormack, Ph. D.

Submitted:
Munich, 02 May 2024

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

Munich, 02 May 2024 Richard Schwank

Zusammenfassung

Eine Schwierigkeit bei der multivariaten Datenanalyse ist es, die Normalisierungskonstante von hochdimensionalen Modellen numerisch zu berechnen. Score Matching ist eine Schätzmethode, die die Berechnung der Normalisierungskonstante durch partielle Integration der logarithmischen Dichte vermeidet. Der klassische Score Matching Schätzer ist jedoch nicht sehr robust gegenüber Datenverzerrungen. Diese Masterarbeit präsentiert eine robustere Variante basierend auf dem geometrischen Median-von-Mittelwerten. Das Anwendungsziel der vorgestellten Methode ist multivariate Abhängigkeitsanalyse durch Graphische Modelle, besonders im Falle von hochdimensionalen Daten. Die Robustheit der neuen Methode gegenüber Verzerrung von ganzen Messungen wird sowohl theoretisch als auch praktisch gezeigt. Die Anwendung des geometrischen Median-von-Mittelwerten in diesem von hochdimensionalen asymmetrischen Verteilungen geprägten Kontext zeigt außerdem eine willkommene Eigenschaft auf: der geometrische Median-von-Mittelwerten scheint sich dem gewöhnlichen Mittelwert anzunähern, wenn die Vektorkomponenten nicht allzu statistisch abhängig sind.

Abstract

A challenge in fitting statistical models to multivariate data is the curse of dimensionality when computing the normalizing constant. Score matching is an estimation paradigm that avoids computing the normalizing constant through strategic integration by parts on the gradient of the log density. However, when applied to data sets with outliers, the basic version of score matching struggles. This thesis presents a more robust score matching procedure built on the geometric median-of-means. The primary application is multivariate dependency recovery in graphical models, with special attention to high-dimensional applications. Robustness of the new procedure with respect to casewise corruption is confirmed through simulations and theoretical guarantees. Further, employing the geometric median-of-means in this high-dimensional setting with asymmetric underlying distributions reveals it has a favorable property: the geometric median-of-means seems to approach the mean with increasing dimension when components aren't too dependent.

Contents

1 Introduction

In modern machine learning scenarios, inputs frequently have a large number of features. There is often inter-dependence between features, which a model should ideally capture. *Probabilistic graphical models* can identify and represent such dependency in an interpretable manner.

Often, any single feature only depends on a (relatively) small number of other features. This so-called *sparsity* must be kept in mind when fitting a model. In fact, sparsity only permits model fitting in areas like genetics, where the number of features is often greater than the number of samples due to cost constraints.

Finally, data collection in practice faces unpredictable challenges like instrument errors, unexpected events and human error. Manually cleaning the resulting measurements becomes very hard or impracticable for large numbers of features. Consequently, *robust* model fitting procedures are sought.

To summarize, this thesis aims to find robust estimators for sparse graphical models.

We consider the large graphical model class of *pairwise interaction models* introduced in section 2.1. Their estimation is challenging due to the curse of dimensionality when computing the normalizing constant, especially in iterative methods. The *score matching paradigm* avoids this problem. As detailed in section 2.2, the parameter of interest θ can be obtained by minimizing a quadratic form $\frac{1}{2}\theta^\top \Gamma_0 \theta + g_0^\top \theta$, where Γ_0 and g_0 are means under the true parameter θ_0.

Previous work [LDS16] and [YDS19] has extended this formulation with L1 regularisation to account for sparsity.

To additionally achieve robustness, this thesis proposes to estimate the means Γ_0 and g_0 by a *median-of-means* procedure from the sample population. The idea of the classical median-of-means is to divide a sample of real numbers into blocks, compute the sample means on the blocks, and then aggregate these block means with the univariate median. This is to strike a balance between robustness from the median and unbiasedness from the block means. It has gained attention recently in part due to its excellent concentration properties. Since Γ_0 and g_0 are objects in high dimensions, we consider multivariate extensions of the univariate median-of-means in section 2.4. We settle on the *geometric median-of-means*.

The first contribution of this thesis is to extend performance guarantees from [LDS16] and [YDS19] to the estimator based on the geometric median-of-means. Particularly, robustness is shown by allowing for corruption of a small number of observations. The result can be found in section 4.2 with work leading up to it in section 3.3.

Second, this thesis sheds some light on the geometric median-of-means itself, which is applied to parameter tuning in section 4.4. Concretely, evidence is presented in section 3.2 that the geometric median gravitates towards the mean in high dimensions when components aren't too dependent. Further, in section 3.1 the breakdown point of the geometric median-of-means is revisited under less extreme corruption assumptions.

Finally, simulations are carried out in section 4.5 to confirm the newly established theoretical guarantees. We find that the estimator based on the geometric median-of-means performs on par with the version based on the sample mean from [YDS19] in an uncorrupted setting and outperforms the sample mean version under casewise corruption.

Notation Lower case letters (e.g. β, p) denote scalars or (column) vectors (e.g. x, μ), while the upper case can denote matrices (e.g. V, Γ, Θ) or random variables (e.g. X). One exception is K, which denotes the number of blocks in the median-of-means. Random observations lie in $\mathbb{R}^m$ and the quantities to estimate are in $\mathbb{R}^p$ (usually $p \in \{m^4, m^2\}$). For a matrix $A \in \mathbb{R}^{a \times b}$, we adopt $A_{i,:} := (A_{i1}, \ldots, A_{ib})$. When identifying A as a vector, we write $\mathrm{vec}(A) := (A_{1,:}, \ldots, A_{a,:})$. Some special matrices and vectors we use are the $m \times m$

identity matrix I_m and the jth Cartesian coordinate vector $e_j \in \mathbb{R}^m$, that is $(e_j)_k = 1$ for $k = j$ and zero otherwise.

$\|\cdot\|_1, \|\cdot\|_2$ and $\|\cdot\|_\infty$ denote the Manhattan, Euclidean and maximum vector norm respectively. We sometimes identify matrices as vectors in this thesis, i.e. $\|\cdot\|_2$ particularly expresses the Frobenius norm. If not specified otherwise, $\|\cdot\|$ refers to $\|\cdot\|_2$.

Matrix norms have three bars. For a matrix $A \in \mathbb{R}^{a\times b}$, we use $|||A|||_{\infty,\infty} = \max_{i=1,\dots,a} \sum_{j=1}^{b} |A_{ij}|$. Further, if $a = b$, we write $\operatorname{diag}(A)$ for the $a \times a$ matrix satisfying $(\operatorname{diag}(A))_{ij} = A_{ij}$ if $i = j$ and zero otherwise. Further we write $\operatorname{tr}(A)$ for the sum of all entries in $\operatorname{diag}(A)$.

For a function $f\colon \mathbb{R}^p \to \mathbb{R}$, the partial derivative with respect to x_j is denoted by $\partial_j f$. The gradient is denoted by ∇f and the Laplace operator by Δf. Absolute derivatives are written as $\frac{\mathrm{d}}{\mathrm{d}x_j} f$, which we especially need for taking derivatives of the interaction terms from pairwise interaction models in a unified notation. Repeated derivatives with respect to the same variable are denoted as $\partial_j^{(r)} f$ and $\frac{\mathrm{d}^{(r)}}{\mathrm{d}x_k} f$.

For a random vector X, the expectation is denoted by $\mathrm{E}[X]$ and the variance-covariance matrix by $\mathrm{Var}[X]$. For a univariate random variable X, we further write $\mathrm{sd}[X] := \sqrt{\mathrm{Var}[X]}$ and $\mathrm{skew}[X] := \mathrm{E}\left[((X - \mathrm{E}[X])/\mathrm{sd}[X])^3\right]$.

We write $\propto$ to denote equivalence up to (problem-specific) constants. Further, for $a, b \in \mathbb{R}^p$, the vector $a \circ b$ is defined as $(a \circ b)_i := a_i \cdot b_i$. A frequently used index set is $[a] := \{1, \dots, a\}$ for a natural number a. For the cardinality of a set A, we write $\#A$. For a real number b, the numbers $\lfloor b \rfloor$ and $\lceil b \rceil$ denote the closest smaller (greater) integer to b. The signum function $\operatorname{sign}(c)$ is 1 if $c > 0$, zero if $c = 0$ and -1 otherwise. Finally, we make use of the Landau symbols $O(p)$ and $o(p)$.

2 Background

2.1 Graphical models

Graphical modeling means the endeavor to capture dependencies between observations in a *graph* [DM17; Lau04]. Presented with a data matrix where rows are independent observations, the goal is to find a dependence network describing dependence between the columns. That said, the graphs fitting the models in this thesis are encapsulated in the zero-structure of the model's matrix parameter. Hence, one can also understand the matter of this thesis as an estimation problem for a matrix parameter.

Example 2.1.1 (Dependence between industry sectors)**.** *Consider the task of relating revenue across industry sectors. For instance, increased turnover in the chemical industry can boost both large-scale freight transport and machinery demand. However, since the freight industry demands minimal machinery and machines are relatively lightweight, there isn't a direct relationship between freight movement and machinery production. Nonetheless, both industries' revenues might correlate due to stimulation from the chemical industry. Graphical modeling addresses this absence of direct influence despite potential overall correlation.*

Consider an m-dimensional random vector X. Let $G = (V, E)$ be an undirected graph with vertices $V = \{1, \dots, m\}$ and an undirected edge set E. For disjoint subsets A, B, C of V we say that B *separates* A *and* C if every path from a node in A to a node in C passes through a node in B. In this case, we write $\langle A, C \mid B \rangle_G$. We say that X possesses the *(global) Markov property* with respect to G, if $\langle A, C \mid B \rangle_G$ implies that X_A is conditionally independent of X_C given X_B. There are different notions of the Markov property, which are equivalent under mild conditions by the Hammersley-Clifford theorem, see [Lau04].

Since any distribution is global Markov with respect to the full graph (as there are no separation statements), we seek a *faithful* graph. The idea is to require that the reverse of the Markov implication should hold (see [DM17]). With the Markov property above, if X_A is conditionally independent of X_C given X_B, then $\langle A, C \mid B \rangle_G$ must hold for any disjoint subsets A, B, C of V.

Example 2.1.2 (Dependency between industry sectors ctd.)**.** *We assumed that turnover in the freight industry and in machinery production are independent up to influence by the chemical industry. Further, we assumed direct dependencies between the chemical and freight industry as well as between the chemical industry and machinery production. The corresponding faithful graph is displayed in fig. 2.1.*

This thesis treats *pairwise interaction models* ([DM17, Sect. 3.5], [LDS16]). We consider the following model, although extensions will be discussed where appropriate. The model consists of densities p_Θ on $\mathbb{R}^m$ which are assumed to be proportional to

$$p_\Theta(x) \propto \exp\left(\sum_{1 \le j \le k \le m} \Theta_{jk} t_{jk}(x_j, x_k) - \psi(\Theta) + b(x) \right), \qquad x \in \mathbb{R}^m. \tag{2.1}$$

The natural parameter space consists of all collections of real values $\{\Theta_{jk} \mid 1 \le j \le k \le m\}$ such that p_Θ is normalizable.

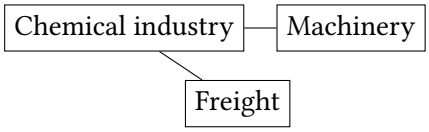


Figure 2.1 Graph for industry revenue example.

Variables $i < j$ appear jointly in the density p_Θ if and only if $\Theta_{ij} \neq 0$. We consequently define the graph G_Θ on $\{1, \dots, m\}$ to have an edge from i to j if and only if $\Theta_{ij} \neq 0$.

The Hammersley-Clifford theorem (see [DM17; Lau04]) supports the intuition that p_Θ is Markov with respect to G_Θ. If the interaction functions t are non-degenerate, G_Θ is faithful as well.

Example 2.1.3. *We consider one classical pairwise interaction model and one slight extension to eq.* (2.1)*:*

(a) The Gaussian graphical model (GGM) *has density p_Θ proportional to* $\exp(-\frac{1}{2}x^\mathsf{T}\Theta x)$ *on* $\mathbb{R}^m$ *(i.e. it is multivariate Gaussian with mean zero). The interaction matrix Θ is also referred to as the* precision matrix. *Since Θ is symmetric, $t_{jk} = x_j x_k$ for $j < k$ and $t_{jj} = \frac{1}{2}x_j^2$.*

A well known theorem states that the margins X_i and X_j are conditionally independent given all other variables if and only if $\Theta_{ij} = 0$ ([Lau04]). This theorem is evidence to the fact that the graph G_Θ from earlier is Markov and faithful to p_Θ.

(b) The square root graphical model *from [IRD16] aims to provide a flexible multivariate generalization of the univariate exponential distribution. It's supported on the positive orthant and assumed to have a density proportional to*

$$p_\Theta \propto \exp\left(-\sqrt{x}^\mathsf{T}\Theta\sqrt{x} + 2\eta^\mathsf{T}\sqrt{x}\right) \qquad (x \in \mathbb{R}^m_+),$$

where the parameter η is added to account for mean shifts. If η is known to be zero, the model is also referred to as centered *square root graphical model.*

As turnovers of a company are very seldom negative, the square root graphical model could fit the introductory example 2.1.1 better than for instance a Gaussian graphical model. For its practical usefulness, we include this model in the simulation study in section 4.5.

However, restricting to non-negative data requires some extra care at the boundaries. As is shown in [YDS19], this doesn't effect estimation too much and some details are given in the next section. Yet, to simplify exposition, restricted domains aren't treated formally in this thesis.

2.2 Score matching

Score matching is a general-purpose estimation technique introduced by A. Hyvärinen in [Hyv05]. It's particularly useful if the normalizing constant of a density is intractable. This section starts with a general introduction to score matching in section 2.2.1 and continues with a treatment of score matching for pairwise interaction models in section 2.2.2.

2.2.1 General score matching

Assume a statistical model $(p_\theta)_{\theta\in\Theta}$ of densities relative to Lebesgue's measure on $\mathbb{R}^m$ and that data is drawn from the distribution with parameter θ_0 (unknown) such that

(A1) $p_\theta > 0$ for all $\theta \in \Theta$ (needed to take logarithms).

(A2) $\partial_i^{(2)} p_\theta(x)$ is continuous for all $1 \le i \le m$ and $\theta \in \Theta$.

(A3) $\mathrm{E}_\theta\big[\, \|\nabla \log(p_\theta(X))\|^2 \,\big] < \infty$ for all $\theta \in \Theta$ (where ∇ is taken w.r.t. x).

(A4) $p_{\theta_0}(x) \cdot \partial_i \log(p_\theta(x)) \to 0$ as $\|x\| \to \infty$ for all $1 \le i \le m$ and $\theta \in \Theta$.

We initially consider the *Fisher divergence* given by

$$d_F(\theta_0, \theta) := \frac{1}{2}\int_{\mathbb{R}^m} p_{\theta_0}(x)\|\nabla \log(p_\theta(x)) - \nabla \log(p_{\theta_0}(x))\|^2 \,\mathrm{d}x \overset{(A3)}{<} \infty.$$

It is easy to see that $d_f(\theta_0, \theta) = 0$ if and only if $\theta = \theta_0$. Indeed, assumption (A1) with continuity from (A2) implies $\nabla \log(p_{\theta_0}(x)) = \nabla \log(p_\theta(x))$ for all $x \in \mathbb{R}^m$. This entails equality of $\log(p_\theta)$ and $\log(p_{\theta_0})$ up to an additive constant, which in turn is fixed to zero by the requirement that densities integrate to one.

Given observations from the true distribution p_{θ_0}, it seems natural to minimize an empirical version of the Fisher divergence. While it is promising that the divergence is an expectation under the true distribution, the second gradient explicitly depends on the unknown parameter θ_0. The beauty of the score matching approach is we can eliminate this dependence through integration by parts. First, note

$$d_F(\theta_0, \theta) = \frac{1}{2} \mathrm{E}_{\theta_0}\left[\|\nabla \log(p_\theta(X))\|^2 \right] - \sum_{1 \le i \le m} \int_{\mathbb{R}^m} p_{\theta_0}(x)\, \partial_i \log(p_{\theta_0}(x))\, \partial_i \log(p_\theta(x))\, \mathrm{d}x + \text{const},$$

where the constant doesn't depend on θ. After using the chain rule on $\partial_i \log(p_{\theta_0}(x))$ (which cancels $p_{\theta_0}(x)$), we can apply integration by parts to the second term. The boundary term vanishes by assumption (A4), also see the proof of theorem 1 in [Hyv05].

$$\int_{\mathbb{R}^m} p_{\theta_0}(x)\, \partial_i \log(p_{\theta_0}(x))\, \partial_i \log(p_\theta(x))\, \mathrm{d}x = \int_{\mathbb{R}^m} \partial_i p_{\theta_0}(x)\, \partial_i \log(p_\theta(x))\, \mathrm{d}x = -\mathrm{E}_{\theta_0}\left[\partial_i^2 \log(p_\theta(X)) \right].$$

All in all, we have identified $d_F(\theta_0, \theta)$ up to constants in θ as

$$d_F(\theta_0, \theta) = \frac{1}{2} \mathrm{E}_{\theta_0}\left[\|\nabla \log(p_\theta(X))\|^2 \right] + \mathrm{E}_{\theta_0}[\, \Delta \log(p_\theta(X)) \,] + \text{const}. \tag{2.2}$$

Provided with observations from p_{θ_0}, we can estimate these expectations for any θ by sample averages and search for a θ that minimizes the empirical divergence. This procedure is illustrated for a specific statistical model in the next section.

We can now see why score matching works even if the density's integration constant is intractable: In eq. (2.2), the integration constant contained in p_θ first gets additively separated by the logarithms and then eliminated by taking derivatives with respect to the data parameters through ∇ and Δ.

2.2.2 Score matching for pairwise interaction models

We examine how score matching plays out for the *pairwise interaction model* from eq. (2.1). Additionally, assume A1 to A4 from the last section. The true parameter is denoted by Θ_0.

To write the score matching loss eq. (2.2) more compactly, we first set $\Theta_{jk} := \Theta_{kj}$ when $1 \le k < j \le m$. Further, similar to [LDS16], we introduce the $m \times m$ matrices $V_j^{(r)}$ via

$$\left(V_j^{(r)}\right)_{kl}(x) := \begin{cases} \frac{\mathrm{d}^{(r)}}{\mathrm{d}x_k} t_{kl}(x_k, x_l) & j = k \le l \\ \frac{\mathrm{d}^{(r)}}{\mathrm{d}x_k} t_{lk}(x_l, x_k) & j = k > l \\ 0 & \text{otherwise} \end{cases} \qquad (r \in \{1, 2\};\ 1 \le j, k, l \le m). \tag{2.3}$$

It remains to compute derivatives of $\log(p_\Theta(x))$. As explained in the last paragraph of section 2.2.1, the integration constant of p_Θ gets annulled since it does not depend on x. The symmetrization of Θ is useful:

$$\partial_j^{(r)} \log(p_\Theta(x)) = \sum_{l \ge j} \Theta_{jl} \frac{\mathrm{d}^{(r)}}{\mathrm{d}x_j} t_{jl}(x_j, x_l) + \sum_{l < j} \Theta_{lj} \frac{\mathrm{d}^{(r)}}{\mathrm{d}x_j} t_{lj}(x_l, x_j) + \partial_j^{(r)} b(x) = \sum_{l \ge j} \Theta_{jl} \frac{\mathrm{d}^{(r)}}{\mathrm{d}x_j} t_{jl}(x_j, x_l) +$$

$$\sum_{l < j} \Theta_{jl} \frac{\mathrm{d}^{(r)}}{\mathrm{d}x_j} t_{lj}(x_l, x_j) + \partial_j^{(r)} b(x) = \left(\mathrm{vec}\left(V_j^{(r)}(x) \right) \right)^\mathsf{T} \cdot \mathrm{vec}\,(\,\Theta\,) + \partial_j^{(r)} b(x) \qquad (r \in \{1, 2\};\ 1 \le j \le m).$$

Since considering the symmetrized Θ as a vector $\mathrm{vec}\,(\,\Theta\,) \in \mathbb{R}^{m^2}$ will simplify score matching notation generally, we set $\theta := \mathrm{vec}\,(\,\Theta\,) \in \mathbb{R}^{m^2}$.

Take $r = 1$ in the above, define a $m \times m^2$ matrix $V(x)$ as $V_{j,:}(x) := \mathrm{vec}\left(V_j^{(1)}(x) \right)^\mathsf{T}$ and we obtain

$$\|\nabla \log(p_\Theta(x))\|^2 = \theta^\mathsf{T} V^\mathsf{T}(x) V(x) \theta + 2 \nabla b^\mathsf{T}(x) V(x) \theta + \|\nabla b(x)\|^2.$$

To find the Laplacian of $\log(p_\Theta(x))$, we set $r = 2$ and identify

$$\Delta \log(p_\Theta(x)) = \left(\sum_{1 \le j \le m} \operatorname{vec}\left(V_j^{(2)}(x) \right) \right)^\mathsf{T} \theta + \Delta b(x).$$

We established the following quadratic form for the ingredients of the score matching loss:

$$\frac{1}{2}\|\nabla \log(p_K(x))\|^2 + \Delta \log(p_K(x)) = \frac{1}{2}\theta^\mathsf{T} V^\mathsf{T}(x) V(x)\theta + \left(V^\mathsf{T}(x)\nabla b(x) + \sum_{1 \le j \le m} \operatorname{vec}\left(V_j^{(2)}(x) \right) \right)^\mathsf{T} \theta$$
$$=: \frac{1}{2}\theta^\mathsf{T}\Gamma(x)\theta + g(x)^\mathsf{T}\theta. \quad (2.4)$$

$\Gamma(x)$ is a symmetric $m^2 \times m^2$ matrix and even block diagonal by the structure of $V(x)$. Taking expectation with respect to p_{Θ_0}, by eq. (2.2) we find the following up to constants in Θ:

$$d_F(\Theta_0, \Theta) \propto \frac{1}{2}\theta^\mathsf{T}\Gamma_0\theta + g_0^\mathsf{T}\theta, \quad (2.5)$$

where $\Gamma_0 = \mathrm{E}_{\Theta_0}[\,\Gamma(X)\,]$ and $g_0 := \mathrm{E}_{\Theta_0}[\,g(X)\,]$. Recall from the last section, that the right hand side of eq. (2.5) is minimized if and only if $\Theta_0 = \Theta$.

Building a first score matching estimator Assume we are given observations $x_1, \dots, x_n \in \mathbb{R}^m$ from p_{Θ_0}. The preceding theory suggests to plug estimates $\hat{\Gamma}, \hat{g}$ of Γ_0, g_0 into the right hand side of eq. (2.5) and minimize with respect to Θ.

Concretely, consider $\hat{\Gamma} := \frac{1}{n}\sum_{i=1}^n \Gamma(x_i)$ and $\hat{g} := \frac{1}{n}\sum_{i=1}^n g(x_i)$ and (numerically) solve (if the minimizer exists)

$$\hat{\Theta} := \operatorname{argmin}_{\Theta \in \mathbb{R}^{m\times m} \text{ symmetric}} \quad \frac{1}{2}\operatorname{vec}(\Theta)^\mathsf{T}\hat{\Gamma}\operatorname{vec}(\Theta) + \hat{g}^\mathsf{T}\operatorname{vec}(\Theta). \quad (2.6)$$

This first draft of a score matching estimator is extended to account for sparse models and corrupt data in section 4.1. The coming example explicitly gives $\Gamma(x)$ and $g(x)$ in two instances and discusses solutions to the optimization problem in eq. (2.6).

Example 2.2.1. *We continue our study in example 2.1.3:*

(a) Recall that the Gaussian graphical model *has density proportional to* $\exp(-\frac{1}{2}x^\mathsf{T}\Theta x)$ *on* $\mathbb{R}^m$*, where* Θ *is the precision matrix. Assumptions (A1) and (A2) are clear, (A3) reduces to the existence of moments. Finally, (A4) holds since the negative exponential decreases faster than any polynomial.*

To recall, $t_{jj}(x_j, x_j) = -\frac{1}{2}x_j^2$ *and* $t_{jk} = -x_j x_k$ *for* $j < k$*. Consequently,* $(V_j^{(1)})_{j,:} = -x^\mathsf{T}$ *and zero in other rows. Hence, the diagonal blocks of* $\Gamma(x)$ *are all equal to* xx^T*. Further,* $(V_j^{(2)})_{j,:} = -e_j^\mathsf{T}$ *and the the base measure* b *has zero derivative. Compactly:*

$$\Gamma(x) = I_m \otimes xx^\mathsf{T} \qquad\qquad g(x) = -\operatorname{vec}(I_m).$$

We conclude that Γ_0 *consists of* m *blocks all equal to* Σ_0*, the covariance matrix.* g_0 *is simply equal to* $-\operatorname{vec}(I_m)$ *in absence of randomness.*

We come to the estimator in eq. (2.6)*. From* $\Gamma(x)$ *and* $g(x)$ *above, we find* $\hat{\Gamma} = I_m \otimes (\frac{1}{n}\sum_{i=1}^n x_i x_i^\mathsf{T})$ *and* $\hat{g} = -\operatorname{vec}(I_m)$*. First, we consider eq.* (2.6) *without the symmetry restriction on* Θ*. In this case, we simply take derivatives and set to zero. If* $n \ge m$*, the sample covariance matrix is a.s. positive definite, and we can a.s. invert* $\hat{\Gamma}$ *to find*

$$\hat{\Theta} = \hat{\Gamma}^{-1} \cdot (-\hat{g}) = \left(I_m \otimes (\frac{1}{n}\sum_{i=1}^n x_i x_i^\mathsf{T}) \right)^{-1} \cdot (-\hat{g}) = \left(\frac{1}{n}\sum_{i=1}^n x_i x_i^\mathsf{T} \right)^{-1}.$$

There are two things to note. First, the score matching estimator coincides with the maximum likelihood estimator (MLE). Second, by symmetry of the sample covariance matrix, our result is symmetric and thus also solves eq. (2.6) *with the symmetry constraint on* Θ*.*

(b) We continue examining the square root graphical model *from example 2.1.3 (b). In contrast to the pairwise interaction models supported on the whole of $\mathbb{R}^m$, the integration by parts in section 2.2.1 would result in non-trivial boundary terms. A possible solution is to multiply the model densities with a known function h of the data that forces densities to zero close to the boundary. This approach is discussed in detail in [YDS19]. One gets a similar quadratic form to eq.* (2.5) *with Γ_0 and g_0 now depending on the function h.*

The additional parameter η can be incorporated into the score matching framework without problems, only calculations typically require case distinctions.

Due to the added parameter η, now $\Gamma(x)$ has m blocks of size $(m+1)\times(m+1)$ and $g(x)$ is a sequence of m vectors $g_j(x)$ in $\mathbb{R}^{m+1}$. The blocks are, for $1 \le j \le m$:

$$\Gamma_j(x) = \frac{h(x_j)}{x_j}\begin{pmatrix}-\sqrt{x}\\ 1\end{pmatrix}\begin{pmatrix}-\sqrt{x}^\top & 1\end{pmatrix}, \qquad g_j(x) = \frac{h(x_j) - 2h'(x_j)x_j}{2x_j^{3/2}}\begin{pmatrix}\sqrt{x}\\ -1\end{pmatrix} - \frac{h(x_j)}{2x_j}e_j,$$

where e_j is the jth Cartesian coordinate vector in $\mathbb{R}^{m+1}$. The formulas are taken from example 5.4 in [YDS19] and adapted to different sign conventions for g (compare 2.5 with equation 10 in chapter 3 of [YDS19]). Again, $\hat{\Gamma} := \frac{1}{n}\sum_{i=1}^n \Gamma(x_i)$ and $\hat{g} := \frac{1}{n}\sum_{i=1}^n g(x_i)$.

For the estimator in eq. (2.6)*, it's important to keep the symmetry constraint in mind. One multiplies out the quadratic form first, then identifies Θ_{ji} with Θ_{ij} for $i \neq j$ and finally groups all occurrences of Θ_{ij} together. Concretely, with $I := (i-1)(m+1)$, $J := (j-1)(m+1)$, $S_{-I} := [m(m+1)] \setminus \{I+j\}$ and $S_{-J} := [m(m+1)] \setminus \{J+i\}$, one obtains:*

$$\Theta_{ij}^2\left(\frac{\hat{\Gamma}_{I+j,I+j} + \hat{\Gamma}_{J+i,J+i}}{2}\right) + \Theta_{ij}\left(\frac{1}{2}\left(\hat{\Gamma}_{I+j,S_{-I}}\cdot \operatorname{vec}(\Theta)_{S_{-I}} + \hat{\Gamma}_{J+i,S_{-J}}\cdot \operatorname{vec}(\Theta)_{S_{-J}}\right) + \hat{g}_{I+j} + \hat{g}_{J+i}\right),$$

where some terms equal to zero by the block structure of $\hat{\Gamma}$ were omitted. Note that $\Gamma_{I+j,S_{-I}}\cdot \operatorname{vec}(\Theta)_{S_{-I}}$ simplifies further due to the block structure of $\hat{\Gamma}$. One can now minimize the new quadratic form in $(\Theta_{ij})_{i\le j}$. In this thesis, the minimization for $(\Theta_{ij})_{i\le j}$ is performed numerically by coordinate descent to be introduced later. The naive approach $\hat{\Gamma}^{-1}\cdot(-\hat{g})$ ignoring the symmetry constraint doesn't yield a symmetric result in general, since the blocks of $\hat{\Gamma}$ explicitly depend on j and the optimization decouples. This is not surprising given that $\Gamma(x)$ was derived under explicit assumption of symmetry in Θ.

2.3 High-dimensional problems and sparsity

In many modern statistical problems there is a large numbers of parameters to estimate. Sometimes, for example in gene expression problems, there is even a larger number of parameters than the number of observations due to cost constraints. This imbalance between parameter dimension and number of observations is typically what makes a problem to be considered *high-dimensional*. A well-known textbook in this domain is [Wai19].

Traditional statistical methods like linear regression break down in high-dimensional scenarios. To some degree, this is not surprising given that there isn't enough "information" for all the parameters to be determined.

Fortunately, there is a low-dimensional structure in many real-world problems. For example, in a regression problem with ten thousand predictors, it is often reasonable to expect that the vast majority of predictors contributes little to the outcome. In the graphical modeling framework, it is often the case that any node only interacts with very few neighboring nodes or that there are very few interactions in general. The assumption that the vast majority of parameters is zero (i.e. not contributing), is referred to as *sparsity*.

A well known example how sparsity can be exploited in a linear regression context is the celebrated *LASSO* by R. Tibshirani [Tib96]. The linear least squares loss is additionally penalized with the L1 norm of the regression weights β, so that the loss becomes

$$\frac{1}{2N}\|X\beta - y\|^2 + \lambda\|\beta\|_1,$$

where X is the predictor matrix, y the observed outcome vector and N the number of observations. The real number $\lambda > 0$ serves as a tuning parameter to determine how "sparse" the resulting minimizer $\hat{\beta}$ should be.

The non-smoothness of the L1 norm indeed leads to sparse results. In turn, it also complicates minimization of the loss. The important observation can be made that the LASSO optimization problem with a single predictor β_1 has an exact solution. More concretely, define the *soft-thresholding operator* $S_\lambda(x) := \text{sign}(x)(|x| - \lambda)_+$. Then, the one dimensional problem has the minimizer $S_\lambda(\frac{1}{N}x^\top y)$. Background on the lasso and its theory can be found in the textbook [HTW15].

This observation suggests *coordinate descent*, a numerical optimization scheme for LASSO-type optimization problems: all components of a multivariate β are cycled through and in each step all but one component is held constant. The remaining component is then updated via soft-thresholding. The numerical algorithms in section 4.5 are built on coordinate descent.

High-dimensionality is of particular concern for graphical modeling, since the possible number of edges in a graph scales quadratically with the number of vertices. To account for this, L1 regularization is applied to the score matching loss in section 4.1.

2.4 Robust estimation

Many estimation methods are adversely affected by outliers in the data set. This thesis develops a robust score matching estimator by utilizing the *median-of-means* principle. As the name suggests, observations are first grouped into blocks, on which the mean is computed. These block means are then aggregated by a median. The concept will be discussed in more detail in section 2.4.2.

What makes an estimator "robust" is arguably subjective and problem-dependent. We focus on robustness with respect to data perturbations. A general intuition is that a few corrupt observations should not be able to overpower the estimation procedure entirely. For background on robust statistics, [Mar+19] is recommended.

This intuition leads to the concept of the *breakdown point*. Informally, this is the fraction of samples that, if tampered with arbitrarily, can result in arbitrarily derailed estimates.

Definition 2.4.1 (Breakdown point)**.** *Let $X \in \mathbb{R}^{n\times p}$ be an n-sample of a p-dimensional random variable and $t_n\colon \mathbb{R}^{n\times p} \to \mathbb{R}^p$ be an estimation procedure. We define the* breakdown point at sample X *by*

$$\varepsilon^*(t_n, X) := \min_{1\le m\le n}\left\{\frac{m}{n}\colon \sup_{Y_m}\|t_n(X) - t_n(Y_m)\| = \infty\right\},$$

where Y_m can differ arbitrarily from X in m data rows.

If the breakdown point doesn't depend on X, we simply write $\varepsilon^*(t_n)$.

The breakdown point is a simple-to-understand (and often simple-to-compute) quantity, but it can be too pessimistic as it allows "infinite" corruption. As this is especially noticeable in the median-of-means framework, we consider the *breakdown probability* in section 3.1. There are more nuanced measures of robustness like *influence functions* (e.g. chapter 3 in [Mar+19]), however they are out of scope for this thesis.

Before discussing some robust estimators, we take a closer look at outliers itself. [Alq+09] surveys different outlier models and collects some important references.

The breakdown point assumes *rowwise* corruption (also referred to as *casewise* corruption or the Tukey-Huber corruption model), where any observation (a row in the data matrix) is either corrupted or not. This form of corruption is for example appropriate, when an unexpected event influences many sensors at once. A more nuanced framework is the *cellwise* corruption, where only one sensor could be deemed malfunctioning in a whole data row.

A further distinction is made between sources of corruption. Often, corruption is assumed to be at random like in the Tukey-Huber model, however in some contexts the corruption originates from an intelligent attacker (*adversarial* corruption).

A related framework is that of heavy tails. In this case, outliers are model-inherent, in that the data generating process is known to produce extreme values from time to time. When extremes occur, standard estimation procedures usually also produce extreme values. One motivation behind the median-of-means principle is to find estimators that are more concentrated in a heavy-tailed setting. See [LM19] for a survey of estimation under heavy tails.

Score matching for pairwise interaction models requires estimation of the matrix Γ_0 and the vector g_0 (see section 2.2.2), both of which are high-dimensional. The main focus of this section is therefore to review multivariate instances of the median-of-means principle. We start with multivariate medians and then generalize to median-of-means.

2.4.1 Multivariate medians

The univariate median has a desirable robustness property: moving an already maximal data point to infinity does not alter the median at all. It's therefore a natural idea to generalize the univariate median to the multivariate setting in order to obtain multivariate robust estimators. Plenty of generalizations for the univariate median have been proposed, each with their own merits and flaws.

There are many desirable properties for a median in higher dimensions (see introduction of [Sma90]). We can ask the median to be the center of symmetry for symmetric distributions. Additionally, just like the univariate median is equivariant under monotone transformations, we can require equivariance under symmetry preserving transformations. Finally, the multivariate definition should reduce to the univariate median in the one-dimensional case.

We additionally require some properties specific to the score matching setting:

(R1) To estimate the score matching design matrix Γ, we require the median of choice to **preserve positive (semi)definiteness** such that the resulting optimization problem to be well-posed.

(R2) The median of choice needs to be **computationally feasible** in high dimensions.

(R3) Finally, the median of choice should have some **robustness** properties, for example a decent breakdown point.

In what follows, three median concepts are discussed. Namely, the *geometric* median, the *Tukey* median and the componentwise median. These and more medians are surveyed in [Sma90].

The geometric median

This median generalizes the fact that the univariate median can be obtained as the minimizer of the mean absolute deviation:

Definition 2.4.2 (Geometric median)**.** *Let $x_1, \dots, x_n \in \mathbb{R}^p$. Any $y \in \mathbb{R}^p$ minimizing $D_n(y) := \sum_{i=1}^n \|x_i - y\|_2$ is defined to be a* geometric median. *If the minimizer is unique, we write* $\operatorname{Med}(x_1, \dots, x_n)$.

Lemma 2.4.3 (Existence and uniqueness of the geometric median)**.** *Let $x_1, \dots, x_n \in \mathbb{R}^p$. Then, there exists at least one geometric median of $x_1, \dots, x_n$. If observations don't fall on a line, it is unique.*

Proof. Clearly, $D_n(y) \geq 0$ and D_n is continuous. We observe $D_n(y) \geq \sum_{i=1}^n |\|y\|_2 - \|x_i\|_2|$ by the inverse triangle inequality, which implies $D_n(y) \to \infty$ as $\|y\|_2 \to \infty$. Existence of the geometric median follows as continuous functions attain a minimum on compact sets. If observations fall on a line, the problem reduces to the univariate median which need not be unique. That this is the only case where uniqueness is violated is shown in [MD87]. □

This median is also referred to as *spatial* median or *L1* median. The term *geometric* median likely stems from fact that the euclidean norm is used. It also possesses an intuitive geometric interpretation: If a geometric median y doesn't fall one one of the data points x_i, we have

$$\nabla D_n(y) = 0 \quad \Leftrightarrow \quad \sum_{i=1}^n \frac{y - x_i}{\|y - x_i\|} = 0. \tag{2.7}$$

In other words: if all data points are projected onto the unit sphere around y, then y must be the barycenter of the projected points (recall that the *barycenter* of some points in euclidean space is their arithmetic mean). Because of the projection, it doesn't matter how far away a data point is from y - only the direction matters. This is an intuitive explanation why the geometric median is robust against outliers.

We consider our additional requirements R1 to R3 and find that the geometric median satisfies all of them.

(R1) Preserving positive (semi)definiteness Solving eq. (2.7) for the geometric median y, we find

$$y = \frac{1}{\sum_{i=1}^n 1/\|y - x_i\|} \sum_{i=1}^n \frac{x_i}{\|y - x_i\|}. \tag{2.8}$$

Hence, if the geometric median y is not equal to on one of the data points x_i, it lies in the convex hull of the data points. Should it fall on one of the data points, it trivially lies in the convex hull as well.

If x_i were positive (semi)definite $m^2 \times m^2$ matrices as in $\Gamma(x)$, we would treat them as vectors in $\mathbb{R}^{m^4}$ for taking the geometric median. When we reinterpret the geometric median in $\mathbb{R}^{m^4}$ as a $m^2 \times m^2$ matrix, it is positive (semi)definite as a convex combination of positive (semi)definite matrices.

(R2) Computational feasibility The relation in eq. (2.8) immediately suggests a simple fixed point algorithm to find the geometric median: an initial estimate is plugged into the right hand side to find a new estimate and this process is repeated with the new estimate until convergence.

This algorithm is referred to as *Weiszfeld's algorithm*. It works quite well in practice, however it can get stuck on data points x_i. Slight modifications of Weiszfeld's algorithm to remedy this issue with convergence guarantees have been discussed in the literature and readily implemented in numerous R packages.

(R3) Robstness As the authors show in [LR91] (Theorem 2.2), the high breakdown point of the univariate median generalizes to the geometric median:

Lemma 2.4.4 (Breakdown point of the geometric median)**.** *Let $X \in \mathbb{R}^{n \times p}$. The geometric median has a sample-independent breakdown point of*

$$\varepsilon^*(\mathrm{Med}_n) := \varepsilon^*(\mathrm{Med}, X) = \frac{\lfloor (n+1)/2 \rfloor}{n}.$$

Remark 2.4.5. *When corruption is below the breakdown point, the geometric median remains bounded by definition. Two examples for how the geometric median is affected by outliers in this scenario are presented in the appendix.*

To summarize, the geometric median fits the requirements R1 to R3 very well, which is why it is chosen as the outer median for the median-of-means in this thesis.

The Tukey median

This median generalizes the intuition that half of the data points should lie to the "left" of the median and the other half to the "right".

Definition 2.4.6 (Tukey median)**.** *Let $x_1, \dots, x_n \in \mathbb{R}^p$. Define $\mathcal{H}$ to be the set of all closed half-spaces in $\mathbb{R}^p$ and let*

$$D(y) := \inf_{H \in \mathcal{H}:\, y \in H} \#\{i = 1, \dots, n \mid x_i \in H\}/n$$

be the Tukey depth *of $y \in \mathbb{R}^p$* in the data set. *Any y with maximal depth is called a* Tukey median.

A point y with depth $1/2$ would collect at least half the data points x_i in any half-space that contains y on the boundary. Although a point of depth $1/2$ is not guaranteed, this intuition justifies the term "median". The Tukey median is also referred to as *halfspace median* ([Sma90]).

Uniqueness of the Tukey median typically cannot be expected. A common approach is to define the mean of all points with maximal Tukey depth to be "the" Tukey median ([DG92], p. 1809).

We come to the additional requirements:

(R1) Preserving positive (semi)definiteness As discussed for the geometric median, it suffices that the median lies within the convex hull of all data points.

Lemma 2.4.7. *Any Tukey median lies within the convex hull of $x_1, \dots, x_n$.*

Proof. In short, any point outside of the convex hull has depth zero, while the depth of any data point is at least $1/n$. Hence, no point outside of the convex hull can have maximal depth.

More concretely, denote the convex hull of $x_1, \dots, x_n$ by C. Let $y \in \mathbb{R}^p \setminus C$. We show that y has depth 0.

Let y_0 be the projection of y onto C. By assumption, $y_0 \neq y$. Let $v := y_0 - y \neq 0$ and let P be the hyperplane through y spanned by the orthogonal complement of v.

Assume, P intersects C at point c and denote $w := c - y$. For $\lambda \in (0, 1)$ let $y_\lambda := y_0 + \lambda(c - y_0)$. As $w \perp v$, we have $c \neq y_0$ and thus $y_\lambda \neq y_0$ for all λ. We find

$$\|y_\lambda - y\|^2 = \|(1-\lambda)v + \lambda w\|^2 \overset{v \perp w}{=} (1-\lambda)^2\|v\|^2 + \lambda^2\|w\|^2 \quad \Rightarrow \frac{\mathrm{d}\|y_\lambda - y\|^2}{\mathrm{d}\lambda}\Big|_{\lambda=0} = -2\|v\|^2 < 0.$$

This means, there is a small lambda such that $\|y_\lambda - y\| < \|y_0 - y\|$, which contradicts the projection property of y_0.

We conclude that P *doesn't* intersect C. Consider the halfspace H with boundary P that does not contain C. It contains y but none of x_i (as $C \cap H = \emptyset$), which means $D(y) = 0$. □

(R2) Computational feasibility That the Tukey median typically is not unique already hints at a harder computational problem. There are modern algorithms, which are for example implemented in the R-package `TukeyRegion` [BM23]. Since one has to expect a computational complexity of roughly $O(n^p)$ (see [LMM19]), these algorithms struggle in high dimensions.

(R3) Robustness The Tukey median possesses robustness properties, however they seem to vary from case to case. While the breakdown point can be as high as $1/3$ (in the limit for centrally symmetric distributions), a lower bound only guarantees a breakdown point of at least $1/(p+1)$ (see [DG92], [Sma90]).

To summarize, the Tukey median is an interesting contender for the outer median in the median-of-means, yet the geometric median fits the requirements R1 to R3 better.

The coordinatewise median

It is a natural idea to define a higher dimensional median by applying the univariate median coordinatewise. Explicitly, the i-th coordinate of the median of $x_1, \dots, x_n$ would be $\text{median}(x_{1i}, \dots, x_{ni})$ for $1 \leq i \leq p$.

This construction is usually not considered to be a multivariate "median", since it is not invariant under orthogonal transformations. Further, it doesn't preserve positive definiteness **(R1)**.

However, one can argue that it is very efficient to compute **(R2)** and that is excellently robust **(R3)**, in particular to *cellwise corruption*.

We don't consider the coordinatewise median further theoretically, mainly because the score matching optimization problem often fails to be well-posed as the design matrix is typically not positive definite. This is illustrated by one numerical experiment in section 4.5.1. Yet, paired with positive definiteness guarantees, the coordinatewise median could be an attractive tool.

2.4.2 Robust mean estimation via median-of-means

We need to estimate the means $\mathrm{E}_{\theta_0}[\,\Gamma(X)\,]$ and $\mathrm{E}_{\theta_0}[\,g(X)\,]$ for score matching. The medians discussed so far offer attractive robustness properties, however as to be expected from the univariate case, medians are typically biased estimators for the mean. To rectify this, one divides the n data points into K blocks $B_1, \dots, B_K$ and first computes the block means $\hat{\mu}_i = \frac{1}{|B_i|}\sum_{j \in B_i} x_j$ for $1 \leq i \leq K$. In the last step, one applies a median to the block means $\hat{\mu}, \dots, \hat{\mu}_k$, hence the name *median-of-means*. See [LSC21] for the original sources and contemporary results on the univariate median-of-means, notably also under corruption. We can expect the distribution of the block means to be rather symmetric by the central limit theorem. As multivariate medians usually reduce to the mean in symmetric settings, the bias should decrease significantly.

Next to robustness, its concentration properties are another strong motivation to consider the median-of-means. To illustrate, we treat the univariate case. Recall that a univariate random variable X with mean μ is called *sub-Gaussian*, if there exists $\sigma > 0$ such that $\mathrm{E}[\,\exp(\lambda(X-\mu))\,] \leq \exp(\sigma^2\lambda^2/2)$. The Chernoff bound leads to

$$\mathrm{P}\left(\,|X-\mu| \geq t\,\right) \leq 2e^{\frac{-t^2}{2\sigma^2}} \qquad \Leftrightarrow \qquad \mathrm{P}\left(\,|X-\mu| \geq \sigma\sqrt{2}\sqrt{\log(2) + \log\left(\frac{1}{\delta}\right)}\right) \leq \delta,$$

which is often called *sub-Gaussian tail* behavior. It ensures that X strongly concentrates around μ and is the basis for many statistical error bounds ([Wai19]).

The empirical mean, arguably the most popular mean estimator, only achieves sub-Gaussian tail behavior on an independent identically distributed sample if the underlying distribution *itself* is sub-Gaussian. The median-of-means on the other hand can achieve sub-Gaussian tail around the mean *even* for distribution that only possess a second moment (see [LSC21]). In heavy-tailed situations, the median-of-means can therefore produce a mean estimate with sub-gaussian tails.

In fact, the (univariate) median-of-means can even guarantee sub-Gaussian concentration under corruption, be it at the cost of a degraded constant (see [LSC21]).

We now formally define a multivariate median-of-means. As the geometric median suited the requirements R1 to R3 best, we choose it as the outer median. Fittingly, the procedure has been termed *geometric median-of-means* in [Min15], where it was first introduced.

Definition 2.4.8 (Geometric median-of-means)**.** *Let $x_1, \dots, x_n \in \mathbb{R}^p$ and $1 \leq K \leq n/2$ be the number of blocks. We assume that K divides n. Define the partition $\mathcal{B}(K,n) = \{B_1, \dots, B_K\}$ and corresponding block means via*

$$B_j := \{(j-1)\cdot n/K + 1, \dots, j \cdot n/K\} \qquad \textit{and} \qquad \hat{\mu}_j := \frac{1}{n/K}\sum_{i \in B_j} x_i \qquad (1 \leq j \leq K).$$

The block means are then aggregated by the geometric median into

$$\mathrm{GMoM}_K[\, x_1, \ldots, x_n \,] := \mathrm{Med}\,(\hat{\mu}_1, \ldots, \hat{\mu}_K)\,.$$

Remark 2.4.9. *The restriction $K \leq n/2$ only excludes the special case of the geometric median ($K = n$). In this case, the term median-of-means is not very fitting. Additionally, concentration analysis often requires excluding the median (e.g. [LSC21] and theorem 3.3.2).*

We start by investigating the robustness of the geometric median-of-means in terms of its breakdown point. As the author notes in [Mat21] on p.46, the breakdown point of the univariate median-of-means equals $\lceil K/2 \rceil / n$, or equivalently $\lfloor (K+1)/2 \rfloor / n$ in the style of [LR91]. We show that this generalizes to the geometric median-of-means:

Lemma 2.4.10 (Breakdown point of the geometric median-of-means)**.** *Let $X \in \mathbb{R}^{n \times p}$ and $1 \leq K \leq n/2$ that divides n. The geometric median-of-means with K blocks has a sample-independent breakdown point of*

$$\varepsilon^*(\mathrm{GMoM}_K[\, n \,]) := \varepsilon^*(\mathrm{GMoM}_K, X) = \frac{\lfloor (K+1)/2 \rfloor}{n}.$$

Proof. If strictly less than $\lfloor (K+1)/2 \rfloor / n$ of the samples are altered, that is less than $\lfloor (K+1)/2 \rfloor$ samples in absolute terms, strictly less than $\lfloor (K+1)/2 \rfloor / K$ of the block means can be corrupted. Lemma 2.4.4 implies that the outer geometric median cannot be corrupted arbitrarily in this case and therefore $\varepsilon^*(\mathrm{GMoM}_K, X) \geq \lfloor (K+1)/2 \rfloor / n$.

We show that this bound is tight. Let $\mu_1, \ldots, \mu_K \in \mathbb{R}^p$ be the block means as in definition 2.4.8. By lemma 2.4.4, there exists a sequence $\left(\mu_1^{(m)}, \ldots, \mu_K^{(m)}\right)_{m \in \mathbb{N}}$, such that

$$\#\{i \in [K] : \mu_i^{(m)} \neq \mu_i\} \leq \lfloor (K+1)/2 \rfloor \qquad \forall m \in \mathbb{N} \tag{2.9}$$

and

$$\sup_{m \in \mathbb{N}} \| \mathrm{Med}\,(\mu_1, \ldots, \mu_K) - \mathrm{Med}\left(\mu_1^{(m)}, \ldots, \mu_K^{(m)}\right) \| = \infty. \tag{2.10}$$

We construct a matrix sequence $X^{(m)} \in \mathbb{R}^{n \times p}$ by setting the rows to

$$X^{(m)}_{j,:} := \begin{cases} \frac{n}{K} \cdot \mu^{(m)}_{Kj/n} - \sum_{i=j-n/K+1}^{j-1} X_{i,:} & \text{if } j \bmod (n/K) = 0 \\ X_{j,:} & \text{otherwise} \end{cases} \qquad j = 1, \ldots, n.$$

Even though it looks as if $X^{(m)}$ could differ in up to K rows from X (i.e. from the first case of its definition), note that $\mu_i^{(m)} = \mu_i$ for some $i \in [K]$ and $m \in \mathbb{N}$ implies $X^{(m)}_{in/K,:} = X_{in/K,:}$. Hence, by (2.9), we conclude $X^{(m)}$ differs by at most $\lfloor (K+1)/2 \rfloor$ data rows from X.

Further, by construction of $X^{(m)}$, we have that

$$\mathrm{GMoM}_K\left[\, X^{(m)} \,\right] = \mathrm{Med}\left(\mu_1^{(m)}, \ldots, \mu_K^{(m)}\right),$$

which by (2.10) implies

$$\sup_{m \in \mathbb{N}} \| \mathrm{GMoM}_K[\, X \,] - \mathrm{GMoM}_K\left[\, X^{(m)} \,\right] \| = \infty.$$

□

The breakdown point of the geometric median-of-means is noticeably decreased compared to the geometric median through division by the sample size n. The reason is simply that a single outlier is enough to corrupt an entire block mean, as demonstrated in the proof above. However, the block structure also has an advantage: a block doesn't get "more" corrupted if it contains two or more outliers and can therefore neutralize more than one outlier. This added robustness is not quantified in the breakdown point, which is why we consider the *breakdown probability* in section 3.1.

When it comes to concentration, the geometric median-of-means can mostly live up to the univariate expectations. As the authors note in [Dev+16], the geometric median-of-means fails to achieve optimal multivariate sub-Gaussian concentration. Very recently, improved (albeit yet still slightly suboptimal) bounds for large class of distributions were shown in [MS23]. Optimal sub-Gaussian tails are achieved by so-called *median-of-means tournaments* introduced in [LM16]. These tournaments are computationally not tractable, although approximations and variants have been proposed (see the survey article [LM19]). To summarize, we should not expect optimal sub-Gaussian tails from the geometric median-of-means, yet can still hope for very good concentration properties. This thesis features a multivariate concentration result under corruption in section 3.3.

Remark 2.4.11. *The assumption that K divides n and that blocks are connected subsets of $\{1, \dots, n\}$ can be relaxed. Similar to U-statistics, one could consider all subsets of a fixed cardinality $2 \leq J \leq n$:*

$$\operatorname{Med}\left(\left\{\frac{1}{J}\sum_{j\in S_j} x_j : S_j \subset \{1,\dots,n\} \text{ satisfies } \#S_j = J\right\}\right).$$

Since it's computationally infeasible to consider all subsets with cardinality J, one would randomly sample a large number of them. Note that the theoretical analysis gets harder since the block means aren't guaranteed to be independent anymore. Some results for the univariate case are presented in section 2.6 of [Min19]. To keep analysis simple, we stick to definition 2.4.8.

3 Results on the geometric median-of-means

This chapter contains three results that enable proving a robustness result for a score matching estimator (section 4.2) and that provide us with intuition on the geometric median (of means), which we apply in section 4.4.

3.1 Breakdown probability of the geometric median-of-means

If an intelligent attacker can compromise ε percent of data points in a median-of-means procedure, they would aim to spread their attack on as many different blocks as possible, since one corrupt sample in a block is enough to alter the block mean arbitrarily. In this scenario, one should ideally work to limit ε by the breakdown point of $\lfloor (K+1)/2 \rfloor / n$ found earlier.

If instead of an adversarial source, we assume a random source of corruption (e.g. malfunctioning sensors), it's likely that two corruptions fall into the same bin. In this case, the median-of-means tolerates more corrupt samples than $\lfloor (K+1)/2 \rfloor$.and the breakdown point could be too pessimistic.

The goal of this section is to answer the following question: *If a fixed proportion of samples is corrupted at random, what is the probability that at least $\lfloor (K+1)/2 \rfloor$ blocks are affected, in which case the geometric median-of-means can diverge uncontrollably?*

Definition 3.1.1 (Breakdown probability of the geometric median-of-means)**.** *Let n, K and $\mathcal{B}(K,n) = \{B_1, \dots, B_K\}$ as in definition 2.4.8. For $S \subset [n]$ define $U_{\mathcal{B}(K,n)}(S)$ to be the number of unique bins that S intersects, i.e.*

$$U_{\mathcal{B}(K,n)}(S) := \#\{k \in [K] : B_k \cap S \neq \emptyset\}.$$

Let $\varepsilon \in [0, 1/2)$ such that εn is a whole number. Define $\mathcal{S}_{\varepsilon n}(n) := \{S \subset [n] : \#S = \varepsilon n\}$. As sets don't distinguish the order of their elements, we have $\#\mathcal{S}_{\varepsilon n}(n) = \binom{n}{\varepsilon n}$.

Let S be distributed according to the uniform distribution on $\mathcal{S}_{\varepsilon n}(n)$. We define the breakdown probability of the geometric median-of-means under uninformed corruption *as*

$$\mathrm{BdPr}_\varepsilon(\mathrm{GMoM}_K[\,n\,]) := \mathrm{P}\left(U_{\mathcal{B}(K,n)}(S) \geq \lfloor (K+1)/2 \rfloor \right).$$

Note that finding the breakdown probability is a purely combinatorial problem. Consequently, it applies to other median-of-means like the univariate version as well.

We now establish an explicit formula for the breakdown probability. The proof uses *generating functions* of combinatorial sequences (see chapter 5 of [Bee15] for an introduction). The correctness of the formula is numerically verified in fig. 3.1.

Theorem 3.1.2. *Let n, K, ε be like in definition 3.1.1. Then,*

$$\mathrm{BdPr}_\varepsilon(\mathrm{GMoM}_K[\,n\,]) = \frac{1}{\binom{n}{\varepsilon n}} \sum_{k=\lfloor (K+1)/2 \rfloor}^{K} \binom{K}{k} \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} \binom{l\frac{n}{K}}{\varepsilon n}.$$

Proof. Using $U_{\mathcal{B}(K,n)}$ and $\mathcal{S}_{\varepsilon n}(n)$ from definition 3.1.1, we define the counts $I_{\varepsilon n}(= b\,; \mathcal{B}(c,d))$ and $I_{\varepsilon n}(\geq b\,; \mathcal{B}(c,d))$ as

$$\#\left\{S \in \mathcal{S}_{\varepsilon n}(n) : U_{\mathcal{B}(c,d)}(S) = b\right\} \qquad \text{and} \qquad \#\left\{S \in \mathcal{S}_{\varepsilon n}(n) : U_{\mathcal{B}(c,d)}(S) \geq b\right\},$$

where $1 \leq b, c \leq d$ are natural numbers such that c divides d. Further, let $K^* := \lfloor (K+1)/2 \rfloor$.

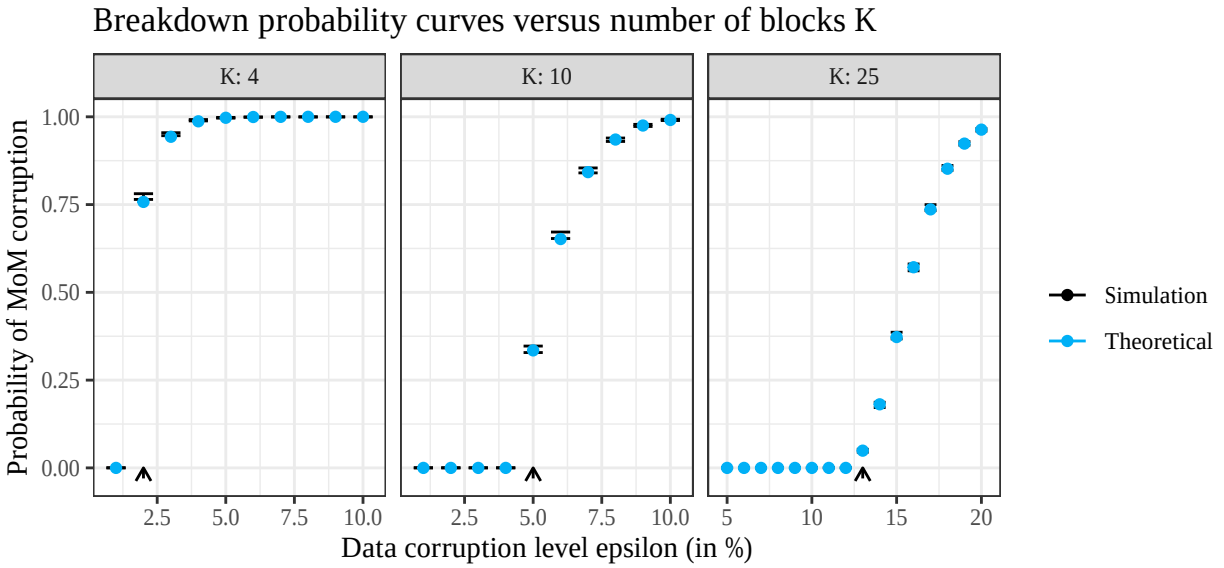


Figure 3.1 For $n = 100$ and $K \in \{4, 10, 25\}$, the breakdown probability $\mathrm{BdPr}_\varepsilon(\mathrm{GMoM}_K[\,n\,])$ is simulated for a range of corruption levels ε and reported in black with a 95% Agresti-Coull binomial confidence interval from 10^4 observations. The corresponding theoretical values from theorem 3.1.2 are plotted as blue dots. The breakdown point is marked with an arrow at the bottom. The simulation confirms the correctness of theorem 3.1.2. We further see that as K increases, the probability at the breakdown point (=arrow) decreases and grows more slowly thereafter. In summary, the breakdown point is most relevant when K is small.

Since all index sets of corrupted rows are assumed to be equally likely and as noted $\#\mathcal{S}_{\varepsilon n}(n) = \binom{n}{\varepsilon n}$, it suffices to show that the double sum in the theorem statement equals $I_{\varepsilon n}(\geq K^*\,; \mathcal{B}(K,n))$. It holds that

$$I_{\varepsilon n}(\geq K^*\,; \mathcal{B}(K,n)) = \sum_{k=K^*}^{K} I_{\varepsilon n}(= k\,; \mathcal{B}(K,n)) = \sum_{k=K^*}^{K} \sum_{\{B_{i_1},\ldots,B_{i_k}\}\subset\mathcal{B}(K,n)} I_{\varepsilon n}(= k\,; \bigcup_{j=1}^{k} B_{i_j}) =$$

$$\sum_{k=K^*}^{K} I_{\varepsilon n}(= k\,; \mathcal{B}(k, \frac{kn}{K})) \cdot \sum_{\{B_{i_1},\ldots,B_{i_k}\}\subset\mathcal{B}(K,n)} 1 \quad = \sum_{k=K^*}^{K} I_{\varepsilon n}(= k\,; \mathcal{B}(k, \frac{kn}{K})) \cdot \binom{K}{k}.$$

We calculate the (ordinary) generating function of $I_{\varepsilon n}(= k\,; \mathcal{B}(k, \frac{kn}{K}))$ in the subscript parameter. The symbolic argument will be denoted by z. Note that $I_m(= k\,; \mathcal{B}(k, \frac{kn}{K}))$ can equivalently be phrased as the number of choices to distribute m indistinguishable balls into k boxes with n/K slots each such that each box contains at least one ball. As there are $\binom{n/K}{l}$ choices to put l balls into n/K slots and there must be between one and n/K balls in each box, the generating function for each box is given by $\sum_{l=1}^{n/K} \binom{n/K}{l} z^l$. Hence the generating function of all k boxes combined (i.e. of $I_m(= k\,; \mathcal{B}(k, \frac{kn}{K}))$) is given by

$$\left(\sum_{l=1}^{n/K} \binom{n/K}{l} z^l\right)^k = \left(-1 + (1+z)^{n/K}\right)^k = \sum_{l=0}^{k} \binom{k}{l} (-1)^{k-l} (1+z)^{ln/K}.$$

As the subscript parameter m equals εn, we must find the coefficient of $z^{\varepsilon n}$ in this generating function. By linearity, this reduces to the coefficient of $z^{\varepsilon n}$ in $(1+z)^{ln/K}$, which equals $\binom{ln/K}{\varepsilon n}$ by the binomial formula. Putting everything together yields the claim. □

From the visualization in fig. 3.1 we can deduce that for fixed n,

- the breakdown probability at the breakdown point decreases as K increases. (Note the important exception $K = n$ where the breakdown probability equals one at the breakdown point).

- the breakdown probability grows more slowly beyond the breakdown point as K increases.

To summarize, the breakdown point is most relevant for the robustness of the median-of-means when the number of blocks K is small. For larger K, the median-of-means can even withstand corruption slightly above the breakdown point with high probability, when the corruption happens randomly.

3.2 Population bias of the geometric median as a mean estimator

From the univariate setting, we expect the geometric median to be a biased estimator of the mean. Surprisingly, it seems like this bias can decrease with the ambient dimension when the components aren't too dependent.

This section consists of an informal (yet mostly rigorous) derivation of this phenomenon when all components are fully independent. The derivation leads to eq. (3.4), an expansion of the geometric median in the ambient dimension. Intriguingly, the expansion features the skewness of the component distribution.

We consider a univariate distribution F with mean μ that satisfies the following conditons:

- F has a density that allows exchanging expectation and derivative (e.g. an **absolutely continuous density**).
- F has finite **fourth moments**.
- It holds $\inf_{m\in\mathbb{R}} E_2(m) > 0$ where $E_2(m) := \mathrm{E}_F\left[\,(X-m)^2\,\right]$. This ensures that F **doesn't concentrate essentially at** a single point.

For $p \in \mathbb{N}_{\geq 2}$, let X_p denote a random vector in $\mathbb{R}^p$ with components that are iid according to F. Define the *geometric median*

$$\mathrm{Med}_p(F) := \mathrm{argmin}_{m\in\mathbb{R}^p} \mathrm{E}\left[\,\|X_p - m\|_2\,\right].$$

This is the population version of definition 2.4.2. Existence is guaranteed similar to 2.4.3 since second moments exist (also see chapter 3 of [MNO10] for an alternative definition with relaxed moment assumptions). Uniqueness follows since the independent components prohibit concentration on a line almost surely (see [MD87]).

We start by finding an expression for $\mathrm{Med}_p(F)$. Since the geometric median is unique and since the loss is symmetric under the iid assumption, we can conclude that all p components of $\mathrm{Med}_p(F)$ are equal and are denoted by $m_p \in \mathbb{R}$. The definition of $\mathrm{Med}_p(F)$ above hence reduces to a univariate problem. Under our assumptions on F, we can exchange the derivative with respect to $m \in \mathbb{R}$ with the expectation and find the following expression for m_p after setting the derivative to zero:

$$0 = \mathrm{E}\left[\frac{\sum_{i=1}^p (m_p - X_i)}{\|X - m_p\|_2}\right] = m_p\,\mathrm{E}\left[\frac{1}{\|X-m_p\|_2}\right] - \mathrm{E}\left[\frac{X_1}{\|X-m_p\|_2}\right] \Leftrightarrow$$

$$m_p = \mu + \mathrm{E}\left[\frac{1}{\|X-m_p\|_2}\right]^{-1} \mathrm{E}\left[\frac{X_1-\mu}{\|X-m_p\|_2}\right]. \quad (3.1)$$

To show that m_p is close to μ in high dimensions, it remains to examine the second term. We use the notation $E_2(m)$ from the assumptions on F. By applying Jensen's inequality to $1/\sqrt{x}$ in the inverse expectation, we bound the absolute value of the second term in eq. (3.1) by

$$\sqrt{p\,E_2(m_p)}\left|\mathrm{E}\left[\frac{X_1-\mu}{\|X-m_p\|_2}\right]\right| = \left|\mathrm{E}\left[\frac{X_1-\mu}{\sqrt{\frac{1}{p}\sum_{i=1}^p \frac{(X_i-m_p)^2}{E_2(m_p)}}}\right]\right| =: |\varepsilon_p(m_p)|.$$

As $\mathrm{E}\left[\, (X_i - m_p)^2 / E_2(m_p) \,\right] = 1$ by the iid assumption, we expect the denominator in $\varepsilon_p(m_p)$ to be very close to one by the law of large numbers. A natural next step is to conduct a first order Taylor expansion of $1/\sqrt{x}$ around one. As the components of X are independent and $\mathrm{E}[\, X_1 - \mu \,] = 0$, the approximation simplifies drastically:

$$\varepsilon_p(m_p) \approx \mathrm{E}\left[(X_1 - \mu) \left(1 - \frac{1}{2} \left(\frac{1}{p} \sum_{i=1}^{p} \frac{(X_i - m_p)^2}{E_1(m_p)} - 1 \right) \right) \right] = \frac{-1}{2p} \mathrm{E}\left[(X_1 - \mu) \frac{(X_1 - m_p)^2}{\mathrm{E}\left[\, (X_1 - m_p)^2 \,\right]} \right] =: \frac{1}{p} \cdot t_1(m_p). \quad (3.2)$$

This approximation indicates that the deviation of m_p from μ is of order $1/p$ as $p \to \infty$. There are two concerns with this conclusion:

(a) $t_1(m_p)$ also depends on p through m_p, which could in theory affect the rate $1/p$.

(b) The higher order Taylor terms also contain terms with $1/p$, which could diverge in infinite sum or at least change the coefficient $t_1(m_p)$ for the order $1/p$.

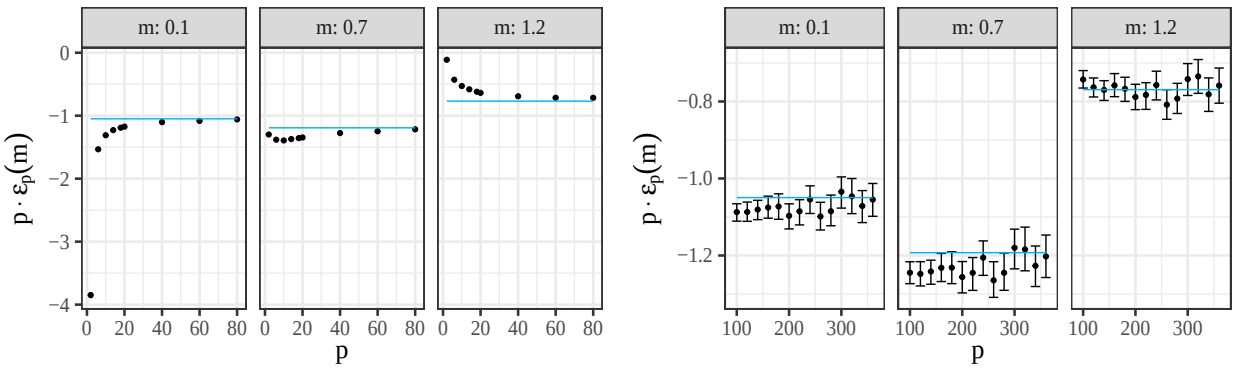
We begin by addressing (a). Applying Cauchy-Schwarz, we obtain

$$|t_1(m_p)| \leq \frac{\mathrm{sd}[\, F \,]}{2} \cdot \frac{\sqrt{\mathrm{E}\left[\, (X_1 - m_p)^4 \,\right]}}{E_2(m_p)} \leq \frac{\mathrm{sd}[\, F \,]}{2} c_F,$$

where $c_F := \sup_{m \in \mathbb{R}} f(m) := \sup_m \sqrt{\mathrm{E}\left[\, (X_1 - m)^4 \,\right]} / E_2(m)$ only depends on the distribution F. Under our assumptions on F, we have $c_F < \infty$. To see this, first note that $\mathrm{E}\left[\, (X_1 - m)^k \,\right]$ is a continuous function of m for $k = 2, 4$ as fourth moments are assumed to exist. The additional assumption $\inf_{m \in \mathbb{R}} E_2(m) > 0$ ensures that f is continuous in m. As both the numerator and the denominator of f are $O(m^2)$, we find $\lim_{m \to \pm\infty} f(m) = 1$. Therefore, f attains a finite maximum by continuity.

To summarize, $t_1(m_p)$ stays bounded when $p \to \infty$. This addresses our concerns regarding (a).

Concern (b) is harder to address as it involves the interplay of infinitely many higher-order Taylor terms. We consider some numerical evidence that eq. (3.2) may capture the order $1/p$ correctly in standard cases. As we don't know m_p, we confirm by simulation that $\lim_{p \to \infty} p\, \varepsilon_p(m) = t_1(m)$ for a range of values $m \in \mathbb{R}$ and for F being the standard exponential distribution. The results are shown in fig. 3.2. It's hypothesized that the alternating sign in the derivative of $1/\sqrt{x}$ plays a crucial role.



(a) Dimension $p < 100$

(b) Dimension $p \geq 100$

Figure 3.2 Numerical validation that $\lim_{p \to \infty} p \cdot \varepsilon_p(m) = t_1(m)$ for $m \in \mathbb{R}$ and $F = \mathrm{Exp}(1)$. We simulate $p \cdot \varepsilon_p(m)$ for a range of dimensions p and $m = \{0.1, 0.7, 1, 1.2\}$. Each black dot represents the mean of 10^3 Monte Carlo simulations with a corresponding bootstrap confidence interval. The horizontal lines state the values of $t_1(m)$. We observe that $p \cdot \varepsilon_p(m)$ indeed approaches $t_1(m)$ when p is large.

We are now more confident that

$$|m_p - \mu| \leq \frac{1}{p} t_1(m_p) + o(1/p) \qquad (p \to \infty). \tag{3.3}$$

With some more intuition, we can guess an explicit expansion of m_p. First, note that eq. (3.3) with the fact that t_1 is bounded implies $m_p \to \mu$ as $p \to \infty$. For p large, we can therefore replace $t_1(m_p)$ by $t_1(\mu)$. Next, we can expect $\mathrm{E}\big[\, 1/\|X - m_p\|_2 \,\big]^{-1} \approx \sqrt{p\, E_2(m_p)}$ for large p with the law of large numbers in mind. This means, we can replace "$\leq$" by "$=$" in eq. (3.3) and and hypothesize that

$$m_p = \mu + \frac{t_1(\mu)}{p} + o(1/p) = \mu - \frac{\mathrm{sd}[\, F \,]}{2} \,\mathrm{skew}[\, F \,]\, \frac{1}{p} + o(1/p) \qquad (p \to \infty). \tag{3.4}$$

This expansion is supported by numerical experiments for three different distributions F (see fig. 3.2).

The appearance of the skewness is particularly intriguing, since according to statistical folklore the sign of the skewness determines whether the median lies to the right or to the left of the mean in the univariate case. Here $\mathrm{skew}[\, F \,] > 0$ implies $m_p < \mu$ up to first order, which is in line with the univariate intuition.

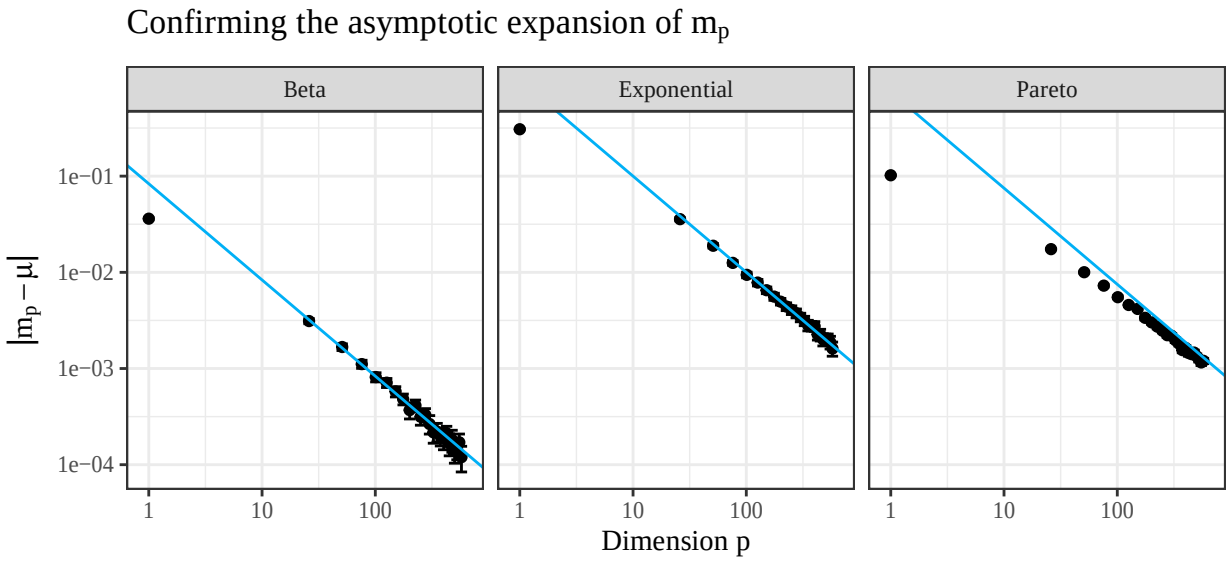


Figure 3.3 Validation of expansion 3.4 for Beta(5,1) (negatively skewed), Exp(1) and Pareto(1,5) distributions. For a range of $p \leq 600$, the quantity of interest m_p is simulated from 10^3 observations. Plotted together on double logarithmic scale are $|m_p - \mu|$ with bootstrap confidence intervals from 100 observations as well as the exact value of $|t_1(\mu)|/p$. Even though the Pareto observations converge rather slowly, the simulation clearly supports the expansion in eq. (3.4). It would have been interesting to consider Pareto distributions with shape less than 4 to test the assumption of fourth moments, however the even slower convergence and increased variance would have consumed a lot more computational resources.

To make the derivation in this section rigorous, it remains to make approximation in eq. (3.2) precise. This is left as an open problem. What makes it hard is the fact that $1/\|X - m\|_2$ doesn't have higher moments (also see lemma 1 in [MNO10]) and that the approximation remainder depends on the unknown value of m_p.

3.3 Concentration of the geometric median-of-means under corruption

The median-of-means principle was advertised for its good concentration and robustness properties in section 2.4.2. Here, we quantify this claim and show that the geometric median-of-means can concentrate

tightly around the mean even if a portion of the sample is corrupted arbitrarily. Unsurprisingly, the concentration gets worse when the amount of corrupt samples increases. Figure 3.4 plots the constant in the concentration inequality versus the corruption parameter τ.

As discussed in section 2.4.2, we should not expect an optimal multivariate sub-Gaussian concentration result. Some details are provided in the remark after the theorem statement.

The concentration result is a variant of corollary 4.1 in [Min15]. As in [Min15], let

$$\psi(\alpha, p) := (1-\alpha)\log\left(\frac{1-\alpha}{1-p}\right) + \alpha\log\left(\frac{\alpha}{p}\right) \tag{3.5}$$

and for $0 < \alpha < 1/2$ let

$$C_\alpha := (1-\alpha)\sqrt{\frac{1}{1-2\alpha}}. \tag{3.6}$$

We base the proof on the following robustness result on the geometric median of independent estimators (from [Min15] Remark 3.1.a)

Lemma 3.3.1 (Minsker, 2015). *Let $\mu \in \mathbb{R}^p$ and $\hat{\mu}_1, \ldots, \hat{\mu}_k \in \mathbb{R}^p$ be a collection of independent estimators of μ. Let the hyperparameters $0 < \alpha < 1/2$, $0 < p < \alpha$ and $\varepsilon > 0$ be such that*

$$\mathrm{P}\left(\|\hat{\mu}_j - \mu\|_2 > \varepsilon\right) \leq p \qquad \forall j \in J,$$

where $J \subset \{1, \ldots, K\}$ has cardinality at least $(1-\tau)K$, and $\tau < \frac{\alpha - p}{1-p}$. Then

$$\mathrm{P}\left(\|\operatorname{Med}(\hat{\mu}_1, \ldots \hat{\mu}_k) - \mu\|_2 > C_\alpha \varepsilon\right) \leq e^{-K(1-\tau)\psi\left(\frac{\alpha-\tau}{1-\tau}, p\right)}.$$

In the following theorem, τ is a parameter that quantifies the amount of corruption.

Theorem 3.3.2 (Concentration of GMoM, possibly under corruption). *Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be independent samples from a p-dimensional random variable X. Assume the covariance Σ exists. Fix a confidence level of $0 < \delta \leq 1$. We allow for up to $\tau(\lfloor \log(1/\delta)/\psi(0.25, 0.125) \rfloor + 1)$ samples to be arbitrarily corrupted, where $0 \leq \tau < 1/2$. Split the samples into K blocks of equal size $\lfloor \frac{n}{K} \rfloor$, where*

$$K = K(\delta, \tau) := \left\lfloor \frac{\log(1/\delta)}{(1-\tau)\psi\left(\frac{(1/2-\tau)^2}{1-\tau}, \frac{1}{2}\left(\frac{1}{2}-\tau\right)^2\right)} \right\rfloor + 1.$$

Further, define

$$c(\tau) := \frac{2 \cdot (3/4 - \tau^2)}{(1/2-\tau)\sqrt{1/2 - 2\tau^2}\sqrt{(1-\tau)\psi\left(\frac{(1/2-\tau)^2}{1-\tau}, \frac{1}{2}\left(\frac{1}{2}-\tau\right)^2\right)}}.$$

If for the confidence level δ it holds that $K \leq n/2$, then

$$\mathrm{P}\left(\|\operatorname{GMoM}_K\left[x_1, \ldots, x_{K \cdot \lfloor n/K \rfloor}\right] - \mathrm{E}[X]\|_2 > c(\tau)\sqrt{\log\left(\frac{4}{(1-\tau)^2}\frac{1}{\delta}\right)\frac{\operatorname{tr}(\Sigma)}{n}}\right) \leq \delta.$$

Proof. The main step of this proof is applying Lemma 3.3.1 to the block means $\hat{\mu}_1, \ldots, \hat{\mu}_K$. We start by fixing α, p and ε in the Lemma. Consider the following choices depending on the corruption parameter τ:

$$\begin{aligned} p(\tau) &:= \frac{1}{2}\left(\frac{1}{2} - \tau\right)^2 \\ \alpha(\tau) &:= 2p(\tau) + \tau = \tau^2 + \frac{1}{4} \\ \varepsilon(\tau) &:= \sqrt{\frac{2K \operatorname{tr}(\Sigma)}{n\, p(\tau)}}. \end{aligned}$$

It remains to verify that these choices can satisfy the conditions in Lemma 3.3.1. To choose the set J, first note that $K(\delta, \cdot)$ is an increasing function. By assumption, at most $\tau K(\delta, 0)$ samples are corrupted. So, the proportion of corrupted blocks is at most

$$(\tau K(\delta, 0))/K(\delta, \tau) = \tau(K(\delta, 0)/K(\delta, \tau)) \leq \tau \cdot 1 = \tau.$$

Therefore, we can set J to be the set of uncorrupted blocks.

To show the probabilistic bound for all blocks $j \in J$, we set $\mu := \mathrm{E}[\, X \,]$ and assume w.l.o.g. that $j = 1$. Using the fact that $\lfloor n/K \rfloor^{-1} \leq 2K/n$ due to $K \leq n/2$, we find

$$\mathrm{E}\left[\, \|\hat{\mu}_1 - \mu\|_2^2 \,\right] = \frac{1}{\lfloor n/K \rfloor^2} \sum_{i,j=1}^{\lfloor n/K \rfloor} \mathrm{E}\left[\, (X_i - \mu)^T (X_j - \mu) \,\right] =$$

$$\frac{1}{\lfloor n/K \rfloor^2} \sum_{i=1}^{\lfloor n/K \rfloor} \mathrm{E}\left[\, (X_i - \mu)^T (X_i - \mu) \,\right] = \frac{\mathrm{E}\left[\, \|X - \mu\|^2 \,\right]}{\lfloor n/K \rfloor} \leq \frac{2K}{n} \operatorname{tr}(\Sigma).$$

The probabilistic bound now follows from Chebycheff's inequality, where everything but $p(\tau)$ cancels.

For the second condition, check

$$\frac{\alpha(\tau) - p(\tau)}{1 - p(\tau)} = \frac{2\,p(\tau) + \tau - p(\tau)}{1 - p(\tau)} = \frac{p(\tau)}{1 - p(\tau)} + \frac{\tau}{1 - p(\tau)} > 0 + \frac{\tau}{1} = \tau.$$

To simplify notation, let

$$\hat{\mu} := \mathrm{GMoM}_K\left[\, x_1, \ldots, x_{K \cdot \lfloor n/K \rfloor} \,\right] = \mathrm{Med}(\hat{\mu}_1, \ldots, \hat{\mu}_K).$$

By lemma 3.3.1, we have established

$$\mathrm{P}\left(\|\hat{\mu} - \mu\|_2 > C_{\alpha(\tau)} \varepsilon(\tau) \right) \overset{3.3.1}{\leq} e^{-K(1-\tau)\psi\left(\frac{\alpha(\tau)-\tau}{1-\tau}, p(\tau)\right)}. \tag{3.7}$$

We start by simplifying the exponent in the right hand side of (3.7) for our choice of $K, \alpha(\tau)$ and $p(\tau)$. We drop the dependency on τ for simplicity. First, note that

$$K = \left\lfloor \frac{\log(1/\delta)}{(1-\tau)\psi\left(\frac{2p}{1-\tau}, p\right)} \right\rfloor + 1,$$

which allows the following simplifications:

$$K(1-\tau)\psi\left(\frac{\alpha - \tau}{1 - \tau}, p\right) = \left(\left\lfloor \frac{\log(1/\delta)}{(1-\tau)\psi\left(\frac{2p}{1-\tau}, p\right)} \right\rfloor + 1 \right) (1-\tau)\psi\left(\frac{2p}{1-\tau}, p\right)$$

$$\overset{\text{for some } c \in [0,1)}{=} \left(\frac{\log(1/\delta)}{(1-\tau)\psi\left(\frac{2p}{1-\tau}, p\right)} - c + 1 \right) (1-\tau)\psi\left(\frac{2p}{1-\tau}, p\right) =$$

$$\log(1/\delta) + (1 - c)(1-\tau)\psi\left(\frac{2p}{1-\tau}, p\right) \overset{c \in [0,1)}{\geq} \log(1/\delta) + 0 = \log(1/\delta).$$

Since the negative of the initial term is the exponent, we can bound the right hand side of (3.7) by

$$e^{-K(1-\tau)\psi\left(\frac{\alpha-\tau}{1-\tau}, p\right)} \leq e^{-\log(1/\delta)} = e^{\log(\delta)} = \delta.$$

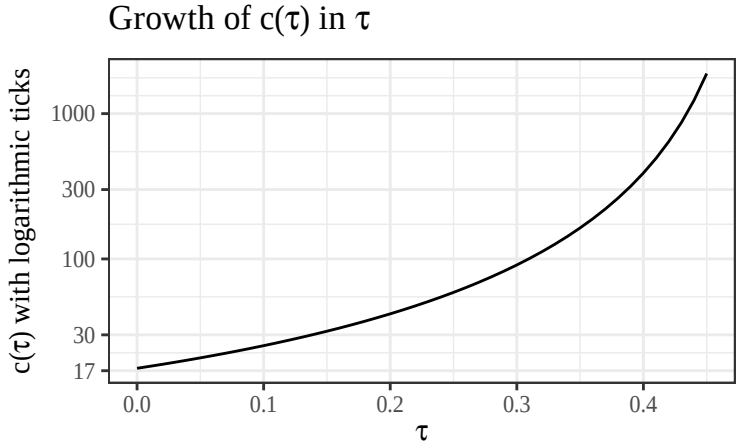


Figure 3.4 Logarithmic plot of the constant in theorem 3.3.2.

All that remains is to simplify $C_\alpha \varepsilon$:

$$C_\alpha \varepsilon = C_\alpha \sqrt{\frac{2K \operatorname{tr}(\Sigma)}{np}} = \frac{C_\alpha \sqrt{2}}{\sqrt{p}\sqrt{(1-\tau)\psi\left(\frac{\alpha-\tau}{1-\tau}, p\right)}} \cdot \sqrt{K \cdot (1-\tau)\psi\left(\frac{\alpha-\tau}{1-\tau}, p\right)} \cdot \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} \leq$$

$$c(\tau)\sqrt{\left(\frac{\log(1/\delta)}{(1-\tau)\psi\left(\frac{2p}{1-\tau}, p\right)} + 1\right) \cdot (1-\tau)\psi\left(\frac{2p}{1-\tau}, p\right)} \cdot \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} =$$

$$c(\tau)\sqrt{\log(1/\delta) + (1-\tau)\psi\left(\frac{2p}{1-\tau}, p\right)} \cdot \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} \overset{\text{First term of } \psi \text{ negative}}{\leq}$$

$$c(\tau)\sqrt{\log(1/\delta) + (1-\tau)\frac{2p}{1-\tau}\log\left(\frac{2p}{(1-\tau)p}\right)} \cdot \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} \overset{p \leq 1}{\leq}$$

$$c(\tau)\sqrt{\log(1/\delta) + 2\log\left(\frac{2}{1-\tau}\right)} \cdot \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} = c(\tau)\sqrt{\log\left(\frac{4}{(1-\tau)^2 \cdot \delta}\right)} \cdot \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}}.$$

□

Remark 3.3.3. *The concentration inequality in theorem 3.3.2 looks similar to the univariate sub-Gaussian concentration from section 2.4.2. This results in decent concentration, yet the mean of multivariate Gaussians concentrates even better (equation (3.1) in [LM19]):*

$$\mathrm{P}\left(\|\bar{X} - \mu\|_2 > \sqrt{\frac{\operatorname{tr}(\Sigma)}{n}} + \sqrt{2\lambda_{max}}\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta,$$

where $\bar{X}$ is the empirical mean of n iid Gaussians with mean μ and covariance matrix Σ having maximal eigenvalue λ_{max}. Note that other than in theorem 3.3.2, the "dimension" $\operatorname{tr}(\Sigma)$ appears separated from the confidence level δ. As discussed at the end of section 2.4.2, improved concentration results closer to the concentration of the multivariate Gaussian mean have been proven very recently in [MS23] for a large class of distributions.

Interestingly, the coordinatewise median concentrates similarly to theorem 3.3.2, see remark 4.1 (c) in [Min15].

4 A robust score matching estimator for sparse graphical models

4.1 Definition of the estimator

This section defines a score matching estimator for the pairwise interaction model from eq. (2.1) that is robust and well-suited for high-dimensional problems with sparse interaction matrices. A first score matching estimator for the pairwise interaction model was already defined in eq. (2.6), however it's built on the non-robust empirical mean and has no special properties in high-dimensional problems. We therefore extend the estimator in eq. (2.6) in two main ways.

First, we regularize eq. (2.6) with the L1 norm, as this has been fruitful for many high-dimensional problems. Additionally, to increase robustness, we replace the empirical means in eq. (2.6) by geometric median-of-means, as introduced in section 2.4.2.

Let $x_1, \dots, x_n \in \mathbb{R}^m$ be an iid sample from the pairwise interaction model from eq. (2.1) with interaction matrix Θ_0 satisfying the additional assumptions from section 2.2.2. Again, we symmetrize such that $\Theta_0 \in \mathbb{R}^{m \times m}$ and denote $\theta_0 := \operatorname{vec}(\Theta_0) \in \mathbb{R}^{m^2}$. Further, we re-use $\Gamma\colon \mathbb{R}^m \to \mathbb{R}^{m^2 \times m^2}$ and $g\colon \mathbb{R}^m \to \mathbb{R}^{m^2}$ from eq. (2.4). Finally, let $K \leq n/2$ be a natural number. Similarly to 3.3.2, we truncate the data matrix until the number of rows is divided by K.

A first approach to incorporate the two modifications to eq. (2.6) discussed above would be to define

$$\hat{\Gamma}_K := \mathrm{GMoM}_K\left[\, \Gamma(x_1), \dots, \Gamma(x_{K \cdot \lfloor n/K \rfloor}) \,\right], \qquad \hat{g}_K := \mathrm{GMoM}_K\left[\, g(x_1), \dots, g(x_{K \cdot \lfloor n/K \rfloor}) \,\right] \tag{4.1}$$

and then for some $\lambda > 0$ minimize

$$\frac{1}{2}\theta^\top \hat{\Gamma}_K \theta + \hat{g}_K^\top \theta + \lambda \|\theta\|_1.$$

However, $\hat{\Gamma}_K$ is only guaranteed to be positive semidefinite, not positive definite. If there is a vector $\tilde{\theta}$ in its kernel and if $\hat{g}_K^\top \tilde{\theta} + \lambda \|\tilde{\theta}\| < 0$, the optimization problem is unbounded from below (consider $a\tilde{\theta}$ for $a > 0$).

If n is large enough relative to m, the estimator $\hat{\Gamma}_K$ is often positive definite. Note that $\hat{\Gamma}_K$ is a convex combination of $(\Gamma(x_i))_i$ for $i = 1, \dots, K \cdot \lfloor n/K \rfloor$, which was one of the reasons why we chose the geometric median (see R1 in section 2.4.1). In the multivariate Gaussian case (see example 2.2.1 (a)), each block of $\hat{\Gamma}_K$ is thus a convex combination of $(x_i x_i^\top)_i$. Only when less than m weights are nonzero can the convex combination be indefinite with positive probability (see [Gup71]).

Still, especially since we are interested in the high-dimensional case where n is not necessarily large relative to m, it's important to ensure that $\hat{\Gamma}_K$ is guaranteed to be positive definite.

Following [YDS19], we introduce a *diagonal multiplier* $\beta > 0$ and define the final estimator as

Definition 4.1.1. *Using $\hat{\Gamma}_K$ and $\hat{g}_K$ from eq. (4.1), we define for $\beta, \lambda > 0$*

$$\hat{\theta}(K, \beta, \lambda) := \operatorname{argmin}_{\theta \in \mathbb{R}^{m^2}} \quad \frac{1}{2}\theta^\top \left(\hat{\Gamma}_K + \beta \cdot \operatorname{diag}(\hat{\Gamma}_K)\right) \theta + \hat{g}_K^\top \theta + \lambda \|\theta\|_1.$$

Remark 4.1.2. *To simplify the theoretical analysis, we follow [LDS16] and do not require $\hat{\theta}$ to be symmetric in definition 4.1.1 (in contrast to eq. (2.6)). The algorithms in section 4.5 solve the symmetrized problem as shown in 2.2.1 (b).*

For theoretical simplicity, the L1 regularization is placed on all components of θ equally. Since the diagonal of Θ_0 and the parameter η in the square root graphical model typically aren't sparse, they aren't penalized

in the implementation. This also clears a small inconsistency: taking 4.1.1 literally, off-diagonal entries of Θ_0 would appear twice in $\|\theta\|_1$ and thus receive twice the weight compared to the diagonal (which as explained, isn't penalized at all in practice).

To see why the minimization problem in definition 4.1.1 has a positive definite design matrix and in turn why $\hat{\theta}(K, \beta, \lambda)$ hast at least one solution (for uniqueness see theorem 4.2.2), consider in the notation of section 2.2.2 that for $i \in \{0, \dots, m-1\}$ and $j \in [m]$

$$(\Gamma(x))_{im+j,im+j} = \|V(x)_{:,\,im+j}\|^2 = ((V_i^{(1)})_{ij})^2 = \begin{cases} \left(\frac{\mathrm{d}t_{ij}(x_i,x_j)}{\mathrm{d}x_i}\right)^2 & \text{if } i \le j \\ \left(\frac{\mathrm{d}t_{ji}(x_j,x_i)}{\mathrm{d}x_i}\right)^2 & \text{otherwise} \end{cases}.$$

In most cases of interest, $\frac{\mathrm{d}t_{ij}(x_i,x_j)}{\mathrm{d}x_i} \neq 0$ almost surely (e.g. in the Gaussian case $\frac{\mathrm{d}t_{ij}(x_i,x_j)}{\mathrm{d}x_i} = x_j$). In this case, $\Gamma(x)$ has a positive diagonal almost surely. As the geometric median-of-means is a convex combination of its arguments, the diagonal of $\hat{\Gamma}_K$ is also positive almost surely. As $\hat{\Gamma}_K$ is positive semidefinite (since the Γ-matrices are positive semidefinite), we conclude that $\hat{\Gamma}_K + \beta \cdot \mathrm{diag}(\hat{\Gamma}_K)$ is a positive definite matrix almost surely in standard cases.

4.2 Performance guarantee under corruption

The aim of this section is to prove that $\hat{\theta}(K, \beta, \lambda)$ from definition 4.1.1 approximates the true parameter well with high probability *even if* a part of the sample is corrupted arbitrarily.

Again, consider the pairwise interaction model p_Θ in eq. (2.1) with the additional assumptions from section 2.2.2 and symmetrized interaction matrix $\Theta \in \mathbb{R}^{m \times m}$ and $\theta := \mathrm{vec}(\Theta) \in \mathbb{R}^{m^2}$. True parameters are denoted by Θ_0 and θ_0. We re-use notation from section 2.2.1 and section 2.2.2.

Definition 4.2.1. *Define d_{θ_0} to be the maximum degree of any nodes in G_{θ_0}, i.e. the maximum number of non-zero off-diagonal entries in any column of Θ_0. Let $c_{\theta_0} := |||\Theta_0|||_{\infty,\infty}$.*

Write $S(\theta)$ for the support of a parameter vector θ, i.e. for $\{i \in m^2 : \theta_i \neq 0\}$. Abbreviate $S_0 := S(\theta_0)$.

Further, if Γ_{0,S_0S_0} is invertible, set

$$c_{\Gamma_0} := \left|\left|\left|(\Gamma_{0,S_0S_0})^{-1}\right|\right|\right|_{\infty,\infty}.$$

Finally, we say Γ_0 satisfies the irrepresentability condition *with incoherence parameter $\alpha \in (0, 1]$ and edge set S_0, if*

$$\left|\left|\left|\Gamma_{0,S_0^cS_0}(\Gamma_{0,S_0S_0}^{-1})\right|\right|\right|_{\infty,\infty} \le (1-\alpha).$$

Recall from section 2.2.2 that θ_0 minimizes $\theta^\mathsf{T}\Gamma_0\theta + g_0^\mathsf{T}\theta$. In definition 4.1.1, the minimization is L1 regularized and (Γ_0, g_0) are replaced by random quantities. The following theorem ensures that when the random deviation from (Γ_0, g_0) is small, the minimization problem from definition 4.1.1 admits a unique minimizer which is reasonably close to θ_0. The theorem was first proven in [LDS16]; the idea to include a diagonal multiplier was first reported in [YDS19].

Theorem 4.2.2 (Lin et al.)**.** *Suppose, Γ_{0,S_0S_0} is invertible and satisfies the irrepresentability condition with incoherence parameter α. Assume*

$$\|\left(\hat{\Gamma}_K + \beta \cdot \mathrm{diag}(\hat{\Gamma}_K)\right) - \Gamma_0\|_\infty < \varepsilon_1, \qquad \|\hat{g}_K - g_0\|_\infty < \varepsilon_2,$$

and $d_{\theta_0}\varepsilon_1 \le \alpha/(6c_{\Gamma_0})$. If

$$\lambda > \frac{3(2-\alpha)}{\alpha}\max(c_{\theta_0}\varepsilon_1, \varepsilon_2),$$

then it holds that he minimizer $\hat{\theta}(K, \beta, \lambda)$ in definition 4.1.1 is unique with $S(\hat{\theta}(K, \beta, \lambda)) \subset S_0$ and satisfies

$$\|\hat{\theta}(K, \beta, \lambda) - \theta_0\|_\infty \le \frac{c_{\Gamma_0}}{2-\alpha}\lambda.$$

Remark 4.2.3. *Theorem 4.2.2 can readily be adapted to include additional parameters like η in the square root graphical model. Only d_{θ_0} needs to be slightly altered to account for the extra parameters. See chapter 6.1 in [YDS19] for details.*

To guarantee maximal deviations by $\varepsilon_1, \varepsilon_2$ in theorem 4.2.2 with high probability under corruption, we extend the previous concentration result of the geometric median-of-means from theorem 3.3.2 to allow for a diagonal multiplier. Set $b := \beta \cdot \mathrm{vec}(I_m)$ in the following lemma to obtain the diagonal multiplier as in definition 4.1.1. The technical treatment of the diagonal multiplier in the following proof is similar to theorems 15 to 17 in [YDS19], however since we don't assume normality, the restriction on the diagonal multiplier depends on the underlying distribution.

Lemma 4.2.4. *Let $x_1, \dots, x_n \in \mathbb{R}^p$ be independent samples from a p-dimensional random variable X. Assume the covariance Σ exists. Fix a confidence level of $0 < \delta \leq 1$. We allow for up to $\tau(\lfloor \log(1/\delta)/\psi(0.25, 0.125) \rfloor + 1)$ samples to be arbitrarily corrupted, where $0 \leq \tau < 1/2$. Split the samples into K blocks of equal size $\lfloor \frac{n}{K} \rfloor$, where $K = K(\delta, \tau)$ as in theorem 3.3.2. Further, let $c(\tau)$ as in theorem 3.3.2.*

Assume that $\mathrm{tr}(\Sigma) > 0$ and let $b \in \mathbb{R}^p$ such that

$$\|b\|_\infty \leq \frac{1}{1 + (\|\mathrm{E}[X]\|_2 / \sqrt{2\,\mathrm{tr}(\Sigma)})\sqrt{n/K}}.$$

If for the confidence level δ it holds that $K \leq n/2$, then

$$\mathrm{P}\left(\|(1+b) \circ \mathrm{GMoM}_K[x_1, \dots, x_{K \cdot \lfloor n/K \rfloor}] - \mathrm{E}[X] \|_\infty > 2 \cdot c(\tau) \sqrt{\log\left(\frac{4}{(1-\tau)^2}\frac{1}{\delta}\right)\frac{\mathrm{tr}(\Sigma)}{n}} \right) \leq \delta,$$

where $\circ$ denotes elementwise multiplication.

Proof. To simplify notation, let $\mu := \mathrm{E}[X]$ and

$$\hat{\mu} := \mathrm{GMoM}_K[x_1, \dots, x_{K \cdot \lfloor n/K \rfloor}], \qquad t := c(\tau)\sqrt{\log\left(\frac{4}{(1-\tau)^2}\frac{1}{\delta}\right)\frac{\mathrm{tr}(\Sigma)}{n}}.$$

We show the implication

$$\|\hat{\mu} - \mu\|_2 \leq t \Rightarrow \|b \circ \hat{\mu}\|_2 \leq t. \tag{4.2}$$

If the left hand side of eq. (4.2) holds, we find (recall $t > 0$ since $\mathrm{tr}(\Sigma) > 0$)

$$\|\hat{\mu}\|_2 \leq \|\mu\|_2 + t = t\left(\frac{\|\mu\|_2}{t} + 1\right) \Leftrightarrow \frac{1}{1 + \|\mu\|_2/t}\|\hat{\mu}\| \leq t \tag{4.3}$$

We can use eq. (4.3) for the right hand side of eq. (4.2). Recalling the definitions of $\alpha(\tau), p(\tau)$ and $\varepsilon(\tau)$ from the proof of theorem 3.3.2 as well as the fact that the end of said proof can be rephrased as $C_{\alpha(\tau)}\varepsilon(\tau) \leq t$, we find

$$\begin{aligned}
\|b \circ \hat{\mu}\|_2 \leq \frac{1}{1 + (\|\mathrm{E}[X]\|_2 / \sqrt{2\,\mathrm{tr}(\Sigma)})\sqrt{n/K}}\|\hat{\mu}\|_2 &\overset{\sqrt{p(\tau)} \leq 1}{\leq} \\
\frac{1}{1 + \sqrt{p(\tau)}(\|\mu\|_2 / \sqrt{2\,\mathrm{tr}(\Sigma)})\sqrt{n/K}}\|\hat{\mu}\|_2 &\overset{C_{\alpha(\tau)} \geq 1}{\leq} \\
\frac{1}{1 + \|\mu\|_2/(C_{\alpha(\tau)}\varepsilon(\tau))}\|\hat{\mu}\| \leq \frac{1}{1 + \|\mu\|_2/t}\|\hat{\mu}\| &\overset{\text{eq. (4.3)}}{\leq} t.
\end{aligned}$$

This bound proves eq. (4.2) which allows us to deduce the inclusion of events

$$\{\|\hat{\mu} - \mu + b \circ \hat{\mu}\|_2 > 2t\} \subset \{\|\hat{\mu} - \mu\|_2 + \|b \circ \hat{\mu}\|_2 > 2t\} \overset{\text{eq. (4.2)}}{\subset} \{\|\hat{\mu} - \mu\|_2 > t\}.$$

Hence, by inclusion of events $\{\|\cdot\|_\infty \geq 2t\} \subset \{\|\cdot\|_2 \geq 2t\}$ and theorem 3.3.2

$$\mathrm{P}\left(\, \|(1+b)\circ\hat{\mu}-\mu\|_\infty > 2t \,\right) \leq \mathrm{P}\left(\, \|\hat{\mu}-\mu\|_2 > t \,\right) \overset{\text{theorem 3.3.2}}{\leq} \delta.$$

□

Combining lemma 4.2.4 and theorem 4.2.2, we can prove the desired performance guarantee:

Theorem 4.2.5. *Let $x_1, \ldots, x_n \in \mathbb{R}^m$ be independent samples from a pairwise interaction model p_{Θ_0} satisfying the additional assumptions from section 2.2.2. Assume that Γ_{0,S_0S_0} is invertible and satisfies the irrepresentability condition with incoherence parameter α. Further, suppose $\Sigma_{\Gamma_0} := \mathrm{Var}_{\Theta_0}(\Gamma(X)) < \infty$ and $\Sigma_{g_0} := \mathrm{Var}_{\Theta_0}(g(X)) < \infty$.*

Fix a confidence level $0 < \delta \leq 1$. We allow for up to $\tau(\lfloor \log(1/\delta)/\psi(0.25, 0.125)\rfloor + 1)$ samples to be arbitrarily corrupted, where $0 \leq \tau < 1/2$. Split the samples into K blocks of equal size $\lfloor \frac{n}{K} \rfloor$, where $K = K(\delta, \tau)$ as in theorem 3.3.2. Further, let $c(\tau)$ as in theorem 3.3.2. Assume $\mathrm{tr}(\Sigma_{\Gamma_0}) > 0$ and let

$$0 \leq \beta \leq \frac{1}{1 + (\|\Gamma_0\|_2/\sqrt{2\,\mathrm{tr}(\Sigma_{\Gamma_0})})\,\sqrt{n/K}}.$$

Finally, with constants and notation from definition 4.2.1, if

$$n > \left(\frac{24 d_{\theta_0} c_{\Gamma_0} c(\tau)}{\alpha}\right)^2 \log\left(\frac{4}{(1-\tau)^2}\frac{1}{\delta}\right)\mathrm{tr}(\Sigma_{\Gamma_0})$$

$$\lambda > \frac{6c(\tau)(2-\alpha)}{\alpha}\sqrt{\log\left(\frac{4}{(1-\tau)^2}\frac{1}{\delta}\right)\frac{1}{n}} \cdot \max\left(2c_{\theta_0}\sqrt{\mathrm{tr}(\Sigma_{\Gamma_0})}, \sqrt{\mathrm{tr}(\Sigma_{g_0})}\right),$$

then with probability at least $1-2\delta$, the estimator $\hat{\theta}(K,\beta,\lambda)$ in definition 4.1.1 is unique with $S(\hat{\theta}(K,\beta,\lambda)) \subset S_0$ and satisfies

$$\|\hat{\theta}(K,\beta,\lambda) - \theta_0\|_\infty \leq \frac{c_{\Gamma_0}}{2-\alpha}\lambda.$$

Proof. Define

$$\varepsilon_1 := 4c(\tau)\sqrt{\log\left(\frac{4}{(1-\tau)^2}\frac{1}{\delta}\right)\frac{\mathrm{tr}(\Sigma_{\Gamma_0})}{n}} \qquad \varepsilon_2 := 2c(\tau)\sqrt{\log\left(\frac{4}{(1-\tau)^2}\frac{1}{\delta}\right)\frac{\mathrm{tr}(\Sigma_{g_0})}{n}}.$$

Treating $\hat{\Gamma}_K + \beta\,\mathrm{diag}(\hat{\Gamma}_K)$ by lemma 4.2.4 and $\hat{g}_K$ by theorem 3.3.2 (together with the inclusion of events $\{\|\cdot\|_\infty > \text{const}\} \subset \{\|\cdot\|_2 > \text{const}\}$), we find by applying the union bound that with probability at least $1-2\delta$

$$\|\hat{\Gamma}_K + \beta\,\mathrm{diag}(\hat{\Gamma}_K)\|_\infty \leq \varepsilon_1/2 < \varepsilon_1, \qquad \|\hat{g}_K - g_0\|_\infty \leq \varepsilon_2/2 < \varepsilon_2.$$

Further, the growth condition on n ensures $d_{\theta_0}\varepsilon_1 \leq \alpha/(6c_{\Gamma_0})$ and by construction $\lambda > 3(2-\alpha)\max(c_{\theta_0}\varepsilon_1, \varepsilon_2)/\alpha$. The claim thus follows from theorem 4.2.2. □

To to compare theorem 4.2.5 with the results in [YDS19], set $\delta := m^{3-\tau}$ as done in the proofs in [YDS19], where τ is the tuning parameter in [YDS19] not to be confused with the corruption parameter τ in this thesis. Then, the requirements on n and λ read similarly, apart from the $\mathrm{tr}(\Sigma_\cdot)$ terms. Some discussion of this discrepancy is provided in chapter 5.

4.3 Asymptotic bias induced by the diagonal multiplier

Following [YDS19], we introduced the *diagonal multiplier* β in section 4.1 to ensure positive definiteness of the score matching design matrix. Since the diagonal multiplier increases the diagonal of $\hat{\Gamma}_K$, we must expect some form of bias in the estimate $\hat{\theta}(K, \beta, \lambda)$. As discussed in section 4.1, altering $\hat{\Gamma}_K$ can especially make sense in high-dimensional regimes.

In this section we aim to understand the consequences of "overbiasing", i.e. what happens to the estimates when $\hat{\Gamma}_K$ was already positive definite and no diagonal multiplier was necessary. Experiments in [YDS19] show that overbiasing can *improve* support recovery performance. Here, a drawback of overbiasing is presented.

To simplify we assume $\lambda = 0$, i.e. we don't consider regularization. However, by continuity of the solution paths, the results carry over qualitatively for small $\lambda > 0$.

Lemma 4.3.1. *Let $\hat{\Gamma}_K \in \mathbb{R}^{m^2 \times m^2}$ be a positive definite matrix, $\hat{g}_K \in \mathbb{R}^{m^2}$ and $\beta > 0$. For β large enough, we have on the support of $\hat{g}_K$ that*

$$\operatorname{sign}\left(\hat{\theta}(K, \beta, \lambda = 0)_i\right) = -\operatorname{sign}\left((\hat{g}_K)_i\right) \qquad (i \in S(\hat{g}_K)).$$

Proof. To simplify notation, let $A := \hat{\Gamma}_K$, $A_\beta := \hat{\Gamma} + \beta \cdot \operatorname{diag}(\hat{\Gamma})$ and $a := -\hat{g}_K$. With this notation, definition 4.1.1 reads

$$\hat{\theta}(K, \beta, \lambda = 0) = \operatorname{argmin}_{\theta \in \mathbb{R}^{m^2}} \quad \frac{1}{2}\theta^\top A_\beta \theta - a^\top \theta.$$

Since A is positive definite (and $\beta > 0$), the matrix $\beta \operatorname{diag}(A)$ is positive definite and thus A_β is also positive definite. We can therefore conclude that $\hat{\theta}(K, \beta, \lambda = 0) = A_\beta^{-1} a$ by smooth optimization theory.

We show that $\hat{\theta}(K, \beta, \lambda = 0) \to (\beta \operatorname{diag}(A))^{-1} a$ as $\beta \to \infty$ using bound (7.5) in VII.7.2 of [AE06]:

$$\begin{aligned}\|A_\beta^{-1} a - (\beta \operatorname{diag}(A))^{-1} a\| &\le \|A_\beta^{-1} - (\beta \operatorname{diag}(A))^{-1}\| \, \|a\| \le \\ 2\| (\beta \operatorname{diag}(A))^{-1} \|^2 \, \|A_\beta - \beta \operatorname{diag}(A)\| \, \|a\| &= 2\|\frac{1}{\beta} \operatorname{diag}(A)^{-1}\|^2 \, \|A\| \, \|a\| = \\ &\frac{2}{\beta^2}\| \operatorname{diag}(A)^{-1}\|^2 \, \|A\| \, \|a\| \xrightarrow{\beta \to \infty} 0.\end{aligned}$$

Since A is positive definite and $\beta > 0$,

$$\operatorname{sign}\left(\left((\beta \operatorname{diag}(A))^{-1} a\right)_i\right) = \operatorname{sign}(a_i) = -\operatorname{sign}(\hat{g}_K) \qquad \forall 1 \le i \le p.$$

The convergence implies that the signs of $\hat{\theta}(\beta)$ and $-\hat{g}_K$ are the same for β large, unless $\operatorname{sign}\left((\hat{g}_K)_i\right) = 0$. □

It's best to illustrate lemma 4.3.1 with a concrete example:

Example 4.3.2. *We consider the* square root graphical model *from example 2.2.1 (b). It will turn out that the estimated interaction matrix is biased towards positive entries. This means, the estimated density is biased to contain $-\sqrt{x_i x_j}$ over $\sqrt{x_i x_j}$ in the exponent, which downweights amplifying effects between x_i and x_i.*

Consider $h(x) := x^2$, the original choice due to Hyvärinen as discussed in section 2.2 of [YDS19]. Plugging this choice into the formulas for $\Gamma(x)$ and $g(x)$ in example 2.2.1 (b) and recalling that $x \in \mathbb{R}_+^m$, we conclude $\hat{g}_K < 0$ in the parameter Θ (i.e. everything but the last coordinate corresponding to η_j) for all choices of K. Lemma 4.3.1 implies that all *interactions are estimated to be positive when β is large enough - even if the true interaction is negative or zero.*

A simulation study confirming the bias is presented in fig. 4.1.

We have seen in example 4.3.2 that the diagonal multiplier can bias the parameters of a pairwise interaction model significantly when it's chosen too large. Further, note that the convergence rate in the proof

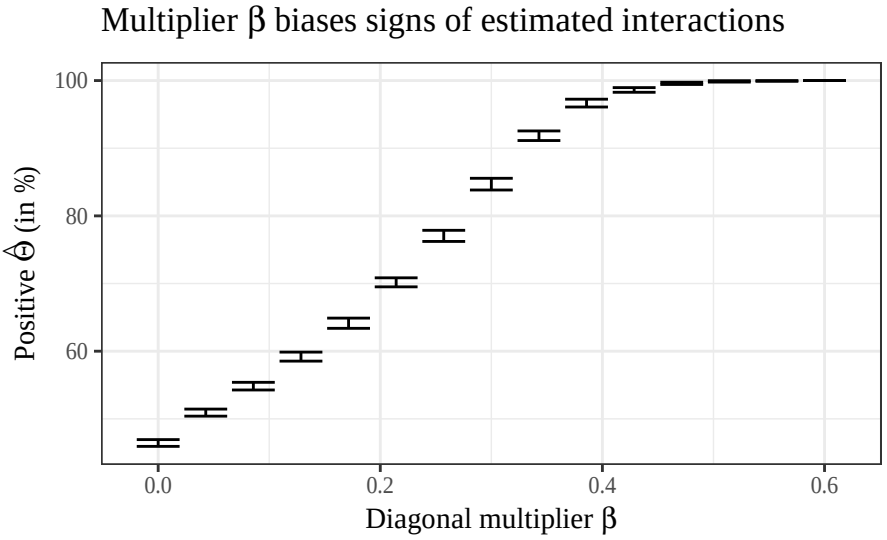


Figure 4.1 A centered square root graphical model with interaction matrix $\Theta \in \mathbb{R}^{10\times 10}$ was constructed such that a random selection of interaction strengths appeared both as positive and negative off-diagonal entry in Θ. The model was then re-estimated from $n = 400$ observations using $\hat{\theta}(K = 1, \beta, \lambda = 0)$ in 100 Monte Carlo simulations for a range of diagonal multipliers. Reported in black are bootstrap confidence intervals for the percentage of positive off-diagonal entries in Θ. We see that rather quickly the vast majority of interactions are estimated to be positive, even though in reality only half of them are.

of lemma 4.3.1 is *quadratic* in the β, which substantiates the caution towards large values of the diagonal multiplier β.

As shown in section 5.2 of [YDS19], one can at least prevent a bias on parameters similar to η in the square root graphical model, since it suffices to apply the diagonal multiplier to the entries in Γ that correspond to the interaction matrix Θ. In the experiment reported in fig. 4.1, the square root graphical model was *centered* (i.e. $\eta = 0$ was known), hence η didn't play a role since it was "estimated" perfectly.

4.4 Practical choice of hyperparameters

The estimator $\hat{\theta}(K, \beta, \lambda)$ contains three hyperparameters. We discuss some practical considerations on how to choose these.

4.4.1 Number of blocks K

The (geometric) median-of-means is parametrized by the number of blocks K. We already defined a choice for K in theorem 3.3.2 under corruption, however this choice was engineered to guarantee good concentration with probability $1 - \delta$. In practice, one would be more interested in choosing K such that the procedure achieves a low error rate, for example low *mean squared error* (MSE).

The *mean squared error* of a p-dimensional estimator T with respect to the true parameter θ_0 decomposes into *variance* and *squared bias* like

$$\mathrm{MSE}(T; \theta_0) := \mathrm{E}_{\theta_0}\left[\, \|T - \theta_0\|_2^2 \,\right] = \mathrm{tr}\left(\mathrm{Var}[T]\right) + \|\, \mathrm{E}_{\theta_0}[\, T \,] - \theta_0\|^2.$$

This section investigates how the number of blocks K influences the variance and squared bias of the geometric median-of-means as an estimator for the population mean. In a next step, we fuse variance and bias together to gain insight into the MSE depending on K. In a simulation, the best choice of K for estimating Γ in the Gaussian graphical model from 2.2.1 (a) is found for different scenarios.

The more formal parts of the derivation apply asymptotic theory for the (geometric) median-of-means, which hasn't been discussed in the thesis so far. Further, we find instances of the reduced bias phenomenon from section 3.2.

Note that there is *no* corruption assumed in this section. To treat corruption in the MSE framework would require some assumptions on the corruption distribution and a general statement seems hard to obtain. Further note that a decent MSE on the geometric median-of-means guarantees reasonable estimation of Γ and g - whether this carries over to a reasonable $\hat{\theta}(K, \beta, \lambda)$ also depends on the diagonal multiplier β and the L1 regularisation with λ.

Variance We begin with the variance of the *univariate* median-of-means. Consider the case that both K *and* n/K are large, i.e. that we can apply the central limit theorem both to the block means *and* to the outer median. The block means then roughly follow $\hat{\mu}_j \sim N(\mu, \frac{K}{n}\sigma^2)$. The asymptotic distribution of the median is given in the following

Lemma 4.4.1 (Example 24 in [Pol84])**.** *Let f be a density on $\mathbb{R}$ with respect to Lebesgue measure such that $f(m) > 0$ for the median m. With independent observations $x_1, \dots, x_K$ from f for K odd, we have*

$$\sqrt{K} \cdot (\text{median}(x_1, \dots, x_K) - m) \xrightarrow{d} N(0, (4f(m)^2)^{-1}).$$

From lemma 4.4.1, we deduce that the median-of-means is roughly normal with mean μ and variance $\frac{1}{K}\frac{\pi}{2}(\frac{K}{n}\sigma^2) = \frac{\pi}{2}\frac{\sigma^2}{n}$, which is *independent* of K.

A rigorous asymptotic analysis of the univariate median-of-means in section 2.5 of [Min19] finds the same asymptotic variance $\pi\sigma^2/2$ - under additional assumptions including a growth restriction on K to ensure that n/K is large enough.

To summarize: if both K and n/K are large, we expect that the variance of the univariate median-of-means *doesn't* depend on K and should be higher than the variance of the sample mean by a factor of $\pi/2$.

This intuition is confirmed in a small simulation study for three different univariate distributions in fig. 4.2. The black horizontal line marks the variance of the sample mean scaled by $\pi/2$ and we see that variances are close to the line in the middle of the plot. Additionally, we see that if K is large (the case excluded in the central limit theorm from [Min19]), the underlying distribution determines the variance curve. Under the exponential distribution, the median-of-means has favorable variance (comparable to the mean), the variance in the bimodal case is larger than the variance of the mean.

When we increase the ambient dimension and look at the geometric median-of-means, the qualitative picture stays the same. However, there are two points to notice.

First, the asymptotic efficiency of the geometric median for Gaussian distributions improves with the ambient dimension. Concretely, what was $\pi/2$ in the univariate case converges to one as the dimension grows to infinity. This effect was first reported and quantified in [Bro83].

Second, we'll observe that if the underlying distribution is exponential, the variance of the geometric median is now *better* than the variance of the mean. This improvement is not unheard of (e.g. for the univariate double-exponential distribution, the median has a smaller variance than the sample mean), still it's a bit unexpected. There is asymptotic theory for the geometric median (see [MNO10]) so the variance is "known", however it requires computing difficult expectations. Understanding under which conditions the geometric median has strictly better asymptotic variance than the mean is an open problem.

We consider the same experiment from fig. 4.2 in dimension $p = 10$. The exponential and the bimodal distribution were extended to a random vector by defining the components to be iid, and for the Gaussian a random covariance matrix was drawn. The asymptotic variance of the geometric median-of-means in dimension $p = 10$ is only by a factor of $65536/(19845\pi) \approx 1.05$ higher than the variance of the mean (see [Bro83]). Again, the experiment supports this intuition for the variance of the geometric median-of-means as shown in fig. 4.3.

There is another slight difference as compared to the univariate case. While in fig. 4.2 the variance *rose* from the mean ($K = 1$) to the black line as K increased, in dimension $p = 10$ the variance *jumps* above the black line and then descends on the line as K increases (see fig. 4.3). This seems to be a more general pattern, however it's unclear why the jump happens.

Concerning the choice of K, it can thus make sense to avoid this jump in variance and restrict to $K \geq 10$.

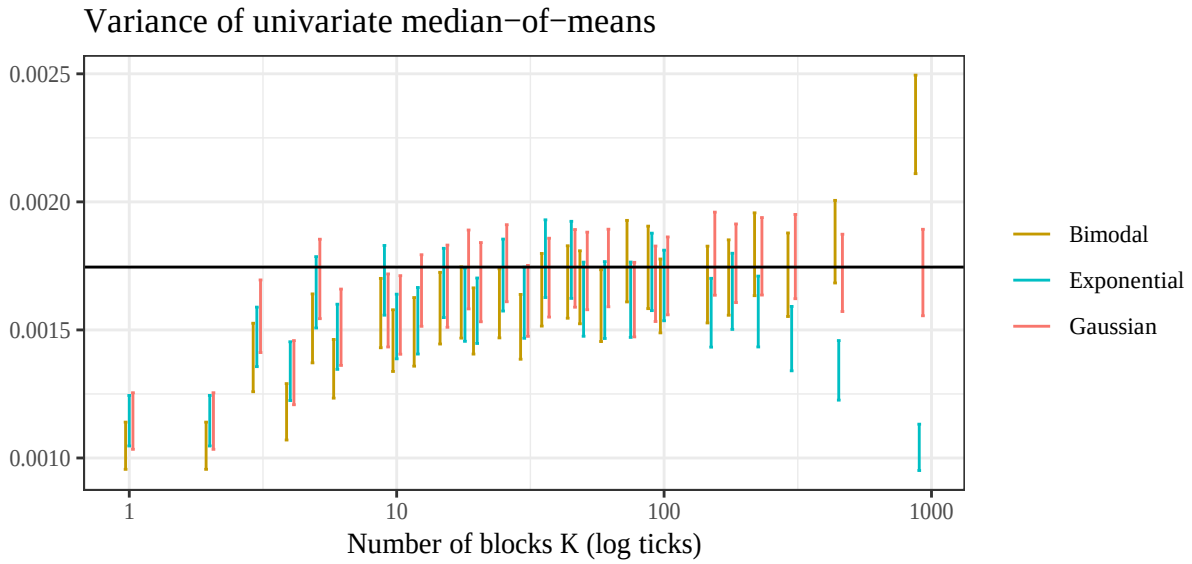


Figure 4.2 $n = 900$ (roughly 1000 but with more divisors) samples from three distributions were drawn. The distributions were the `Bimodal` (Equal mixture between standard Gaussians with mean one and minus one respectively standardized to have variance one), the standard `exponential` and the standard `Gaussian`. Then, the median-of-means was computed with the number of blocks K ranging through all divisors of n. This was repeated in 10^3 Monte Carlo runs. Bootstrap confidence intervals for the variance of all estimators are reported. The black horizontal line marks the variance of the empirical mean (i.e. $1/900$) scaled by $\pi/2$. We see that when both K and n/K are large, the variances approach the theoretical line. When K is large, the variance curve depends on the underlying distribution. While the variance decreases with growing K for the exponential distribution, it increases for the bimodal distribution.

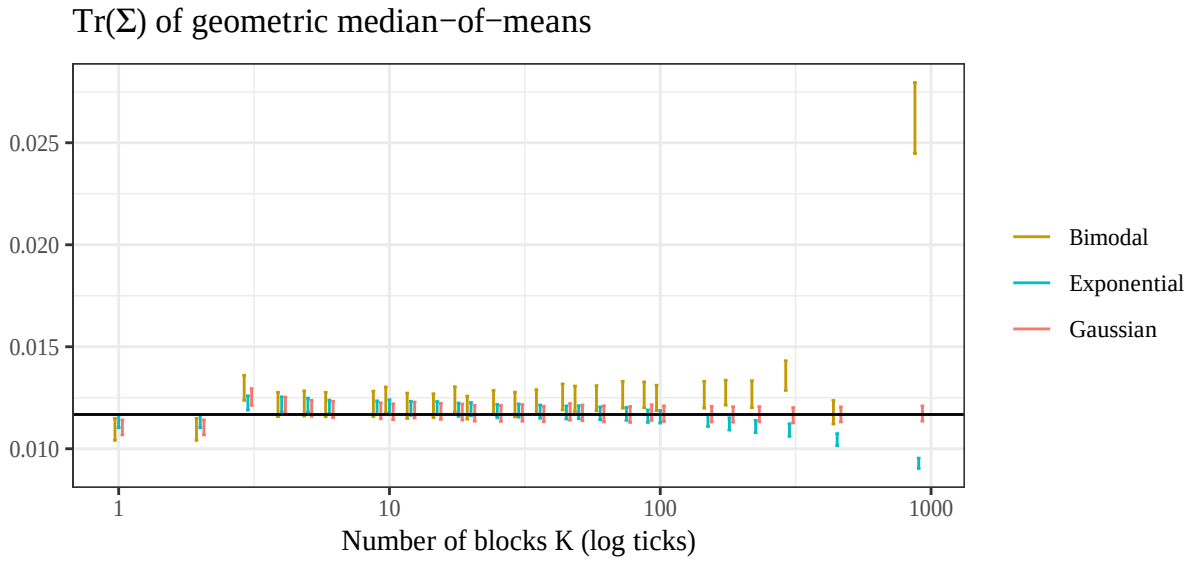


Figure 4.3 A repeat of fig. 4.2 in ambient dimension $p = 10$, thus using the *geometric* median-of-means. The exponential and bimodal random vectors are 10 iid copies of the corresponding univariate distribution, and the Gaussian is based on a random covariance matrix with trace p (to fit the other two). We see that the variances group around the asymptotic theoretical horizontal line when both K and n/K are large with the fit being noticably better in the Gaussian and exponential case. Further, note that for the exponential distribution, the geometric median ($K = n$) has a *smaller* variance than the sample mean.

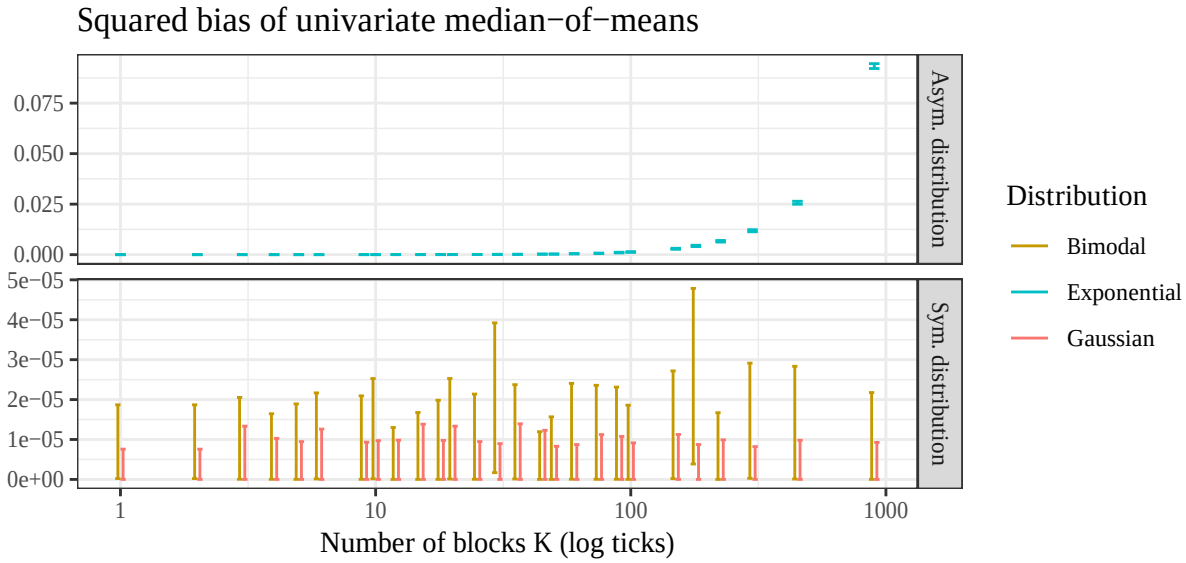


Figure 4.4 For the univariate estimators from fig. 4.2, bootstrap confidence intervals for the squared bias are shown. While there is no significant bias for the symmetric distributions, the exponential distribution displays a strictly increasing bias with K.

Squared bias Even if the pairwise interaction model has a symmetric distribution, we cannot expect the distributions of Γ or g to be symmetric. Consequently, we have to expect a bias when employing median-like procedures to estimate the mean.

Again, we start with the *univariate* median-of-means and simulate the squared bias for the distributions from fig. 4.2. The result is shown in fig. 4.4. For the asymmetric distribution (i.e. the exponential distribution), the squared bias of the median-of-means increases strictly with K. This is to be expected, since K interpolates between the mean (unbiased) and the median (biased). When the underlying distribution is symmetric, we see no significant bias.

When we repeat the same experiment in dimension $p = 10$ (as in fig. 4.3), the qualitative picture for the squared bias doesn't change. The results are reported in fig. 4.5.

One detail we can observe is the reduced bias phenomenon described in section 3.2. Since the random vector for the exponential distribution contains iid components, we would expect the squared bias to scale like p. However, for example the squared bias of the geometric median equals roughly 0.06 for $p = 10$ and is thus actually *lower* than the squared bias of roughly 0.09 for the univariate median.

Applying the results from section 3.2, we expect the component squared bias to be of order $1/p^2$. When the component squared biases get summed up to the vector squared bias, this becomes $1/p$. We don't quite observe a factor of 10 between the two squared biases though. This could be because $p = 10$ is too small for the asymptotic result and because the univariate median behaves slightly differently to the geometric median (e.g. taking derivatives of the expected loss like in section 3.2 wouldn't work).

Mean squared error (MSE) As the MSE is the sum of variance and squared bias, the optimal choice in terms of MSE will depend on the magnitude of variance and squared bias relative to each other. This relative magnitude depends on the sample size n: While the variance asymptotically decreases with rate $1/n$, the squared bias roughly stays constant in n. More precisely, set $K(n) := \varepsilon\, n$ for some fixed $0 < \varepsilon < 1$ to compare concrete estimators $\hat{\mu}^{\varepsilon}_{K(n)}$ only depending on n. For the choice of $K(n)$, each block contains a fixed number of samples $1/\varepsilon$, hence the block means have a population median m_ε independent of n. The central limit theorem for the geometric median implies $\hat{\mu}^{\varepsilon}_{K(n)} \approx N(m_\varepsilon, \Sigma_\varepsilon/n)$. Thus, the squared bias of $\hat{\mu}^{\varepsilon}_{K(n)}$ approximately equals $(m_\varepsilon - \mu)^2$, which is independent of n, and the variance of $\hat{\mu}^{\varepsilon}_{K(n)}$ is roughly Σ_ε/n, which decreases at rate $1/n$.

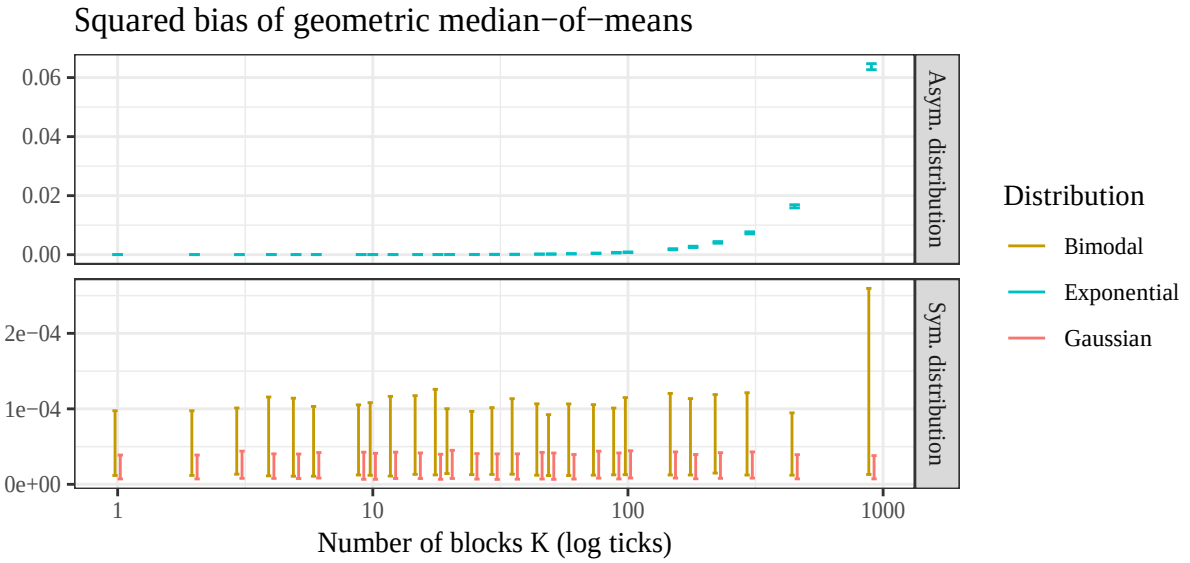


Figure 4.5 For the multivariate estimators from fig. 4.3, bootstrap confidence intervals for the squared bias are shown. The qualitative picture is very similar to the one in fig. 4.4, yet notice that the squared bias for the geometric median ($K = n$) in the exponential case is actually *lower* than in the univariate setting. This can be explained by the results from section 3.2.

Consequently, the MSE will be *bias-dominated* for large n. What happens in the opposite case of small n depends on the initial relative magnitude between variance and squared bias. We will see an example from a pairwise interaction model where the MSE is *variance dominated* for small n.

We conjecture that the MSE of the geometric median-of-means is often variance-dominated in *high-dimensional* scenarios, i.e. when the sample size n is not too large relative to the dimension p. As described in the preceding discussion on squared bias, we can expect the squared bias to be of order $1/p$ when components aren't too dependent (the setting from section 3.2). On the other hand, the trace of the variance usually grows with order p. In the scenario of section 3.2, the trace of variance thus surpasses the squared bias for large p.

In this scenario, the choice of K is hence mostly determined by the *variance* when n is *small* relative to p, and determined by the *bias* when n is large relative to p.

Consider the task of Gaussian covariance estimation, i.e. to estimate Γ in the Gaussian graphical model (GGM) from example 2.2.1. The block means are Wishart-distributed, which fits best to the exponential examples in fig. 4.3 and fig. 4.5. For the exponential distribution, we observed

- optimal variance for $K = n$ in fig. 4.3, and
- a monotonically increasing bias with K in fig. 4.5.

When the sample size n is small relative to the dimension p, we expect variance dominance and therefore $K = n$ to be the optimal choice. When n is large relative to p, we expect bias dominance and thus $K = 1$ to be the best choice.

This intuition is confirmed by an experiment. A positive definite 10×10 covariance matrix was randomly sampled and re-estimated using the geometric median-of-means with the three sample sizes $n = 12$, $n = 105$ and $n = 900$ (roughly powers of 10 but with more divisors). The results of a Monte Carlo simulation are presented in fig. 4.6. The optimal block sizes are $K = n$ for $n = 12$, the second-to-max choice $K = 35$ for $n = 105$, and for $n = 900$, the mean ($K = 1$) and most $K \leq 150$ are tied or very close to being tied in terms of MSE.

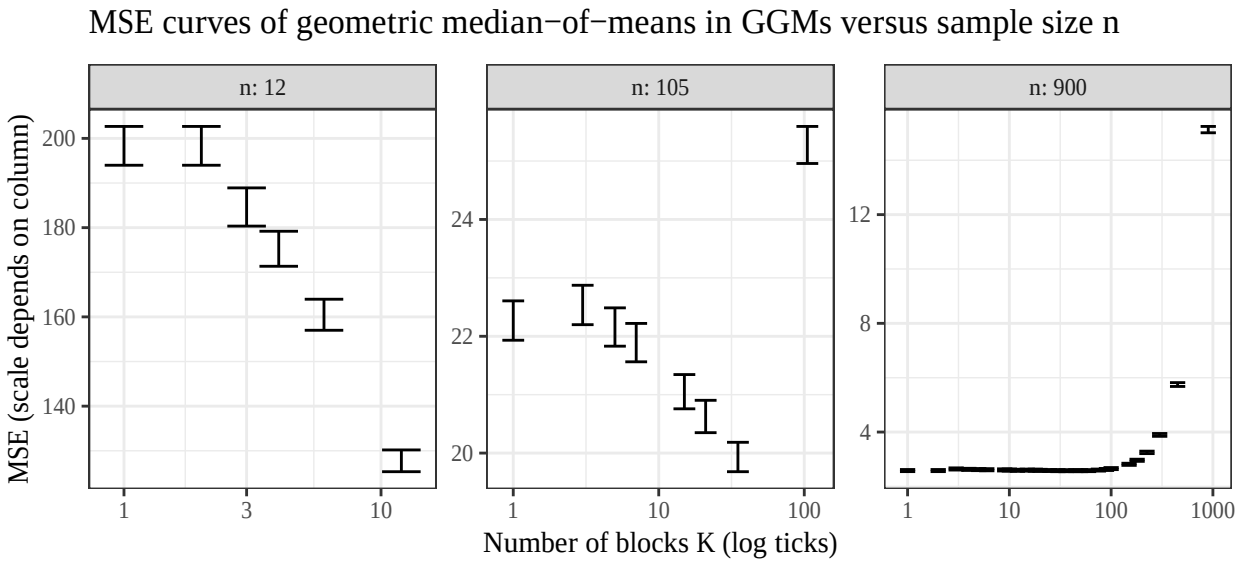


Figure 4.6 The geometric median-of-means was used to estimate a random 10×10 Gaussian covariance matrix from n samples. For $n \in \{12, 105, 900\}$ (chosen roughly as powers of 10 but with more divisors), the number of blocks K ranged through all divisors of n. Reported in black are bootstrap confidence intervals for the MSE based on 10^3 Monte Carlo repetitions. While the geometric median ($K = n$) performs best for $n = 12$, it performs the worst for $n = 900$. This comes from the changing of the relative magnitude of bias and variance, as discussed in section 4.4.1.

Remark 4.4.2. *In this experiment, it's not quite clear what dimension p the sample size n should be compared to in order to designate the problem high-dimensional. On the one hand, the estimation problem is 100-dimensional, so naturally $p = 100$. On the other hand, there are only 10 (not even fully) independent sources of randomness, so maybe a lower p would be more appropriate. This is an open question.*

To conclude, what number of blocks K should be chosen when estimating Γ from a Gaussian graphical model?

In a high-dimensional setting, it can be attractive to chose a large K in order to obtain optimal MSE. Further, a large K improves the breakdown point. And unless $K = n$ (the geometric median), the estimator even has a low breakdown probability beyond the breakdown point if corruption location is random (see section 3.1).

If on the other hand the sample size n is rather large, very large K perform poorly because of the dominating bias. In this case, it would be better to choose $K \approx n/10$ to ensure some symmetrization by the block means to limit the bias. In theory one could choose K even lower, but one loses robustness, the bias doesn't reduce much further and variance-wise, it's best to avoid the "jump" reported at the end of the variance discussion.

Remark 4.4.3. *For the univariate exponential distribution and an odd number of blocks K, one can compute the variance and bias of the median-of-means* exactly. *Since the block means are Gamma-distributed and the median is an order statistic, the problem reduces to moments of Gamma order statistics for which explicit recursive formulas are known (see [Gup60]). Routines for this approach are contained in the appended code files. However, due to the combinatorics involved the computation is only numerically stable for small n (roughly 50).*

4.4.2 Regularization parameter λ

One can consider extensions to the Bayesian information criterion (BIC) for choosing the L1 regularization parameter λ.

In the classical BIC, having n observations and models with a different number of parameters to choose from, one picks the number of parameters k minimizing

$$\mathrm{BIC}(k) = -2l_n(\hat{\theta}_k) + k\log(n),$$

where l_n is the log likelihood of the n observations and $\hat{\theta}_k$ the maximum likelihood estimator for k parameters. The rationale is to find a balance between high likelihood and the number of parameters.

There are two problems with the BIC in the setting of score matching for graphical models.

First, the classical BIC was found to select overly complex graphical models, especially in high-dimensional cases. This is because in its derivation, the BIC places a uniform prior on all possible models (see chapter 2 of [CC08]). Particularly on many vertices, the number of sparse graphs is greatly outmatched by the number of denser graphs. Extended BICs remedy this imbalance with additional dimension-dependent terms ([BD15; CC08] and others).

Second, the log likelihood requires the intractable normalizing constant. In [YDS19], this was addressed by exchanging the log likelihood with the (unregularized) score matching loss.

The extended BIC from [YDS19] in the setting of this thesis reads

$$\mathrm{eBIC}(\lambda) = -n \cdot (\hat{\theta})^{\mathsf{T}}\hat{\Gamma}_K\hat{\theta} - 2n\hat{g}_K^{\mathsf{T}}\hat{\theta} + \#S_\lambda \cdot \log(n) + 2\log\left(\binom{\binom{m}{2}}{\#S_\lambda}\right),$$

where $\hat{\theta} = \hat{\theta}(K, \beta, \lambda)$, again n is the sample size, $\hat{\Gamma}_K$ and $\hat{g}_K$ come from eq. (4.1), $S_\lambda = \{(i,j) : \mathrm{Mat}(\hat{\theta})_{ij} \neq 0, i < j\}$ is the support of $\hat{\theta}$ interpreted as an interaction matrix, and m is the dimension of the pairwise interaction model. As in [YDS19], we don't include the diagonal multiplier in the substitute for the log likelihood. Background to this extended BIC is given in [LDS16], [YDS19] and the references therein. The version from [YDS19] is closest to the one in [BD15].

It's an open problem to show consistency results for this extended BIC, especially under corruption. Lacking theoretical insights, we judge the performance of estimators in section 4.5 by receiver operating characteristic curves, which don't require to fix λ.

4.4.3 Diagonal multiplier β

As mentioned in section 4.1, it's typically not necessary to pick a positive diagonal multiplier β when the problem isn't high-dimensional. When one still chooses $\beta > 0$, one risks a directional bias as shown in section 4.3. On the other hand, this bias was experimentally found to aid support recovery in [YDS19].

We set a bound on the diagonal multiplier in theorem 4.2.2 to ensure sufficient concentration. Its disadvantage is that the bound depends on the (unknown) Γ_0. For the numerical experiment in section 4.5.3, estimates of Γ_0 are plugged into the bound and the middle between zero and this bound is chosen as the diagonal multiplier. This is to balance between the larger-is-better from [YDS19] and the directional bias from section 4.3.

As an alternative, [YDS19] derives a data-free bound for the diagonal multiplier based on the (truncated) normal distribution, which also is the default in the `genscore` package ([YLG23]). It's important to only employ this rule in high-dimensional settings since it would have evaluated to 0.67 in the example of fig. 4.1, where it would have severely biased the interactions.

4.5 Numerical experiments

We study the performance of $\hat{\theta}(K, \beta, \lambda)$ from definition 4.1.1 for different pairwise interaction models. A special emphasis is placed on performance on rowwise *corrupted* samples.

$\hat{\theta}$ was computed using coordinate descent. Some details on the sub-problem for each coordinate are given in example 2.2.1 (b). The implementation can be found in the appended code files.

The interaction matrix Θ was determined such that in expectation 70% of the interactions would be zero. Following [YDS19], the remaining entries were randomly sampled between 0.5 and 1 with random sign. The diagonal was also randomly sampled and raised until the smallest eigenvalue was 0.1.

4.5.1 Gaussian graphical models

We begin with *the* classical graphical model, the Gaussian graphical model from example 2.2.1. We set the dimension $m := 20$ and choose an interaction matrix $\Theta \in \mathbb{R}^{20\times 20}$ as described in the introduction of section 4.5.

In an experiment, $n = 200$ independent samples were drawn from a Gaussian distribution with mean zero and precision matrix Θ. On a grid of 20 values for λ resulting in very sparse to very dense graphs, the classical regularized score matching estimator $\hat{\theta}(K = 1, \beta = 0, \lambda)$ based on the sample mean from [LDS16] was compared to $\hat{\theta}(K = 40, \beta = 0, \lambda)$, a version based on the geometric median-of-means with $K = 40$ blocks.

The number of blocks K was chosen with fig. 4.6 and the findings from section 4.4.1 in mind: choose K rather large to profit from the MSE optimum in the middle column of fig. 4.6, but not too large to avoid biasing.

Since $n >> m$, we chose the diagonal multiplier β to be zero as discussed in 4.1.

In a second step, 19 out of the $n = 200$ samples were corrupted randomly (the largest corruption amount below the breakdown point for $K = 40$). They were replaced by draws from an independent m-dimensional Gaussian whose variance was roughly 10 times as high as the original distribution. Afterwards, as before $\hat{\theta}(K = 1, \beta = 0, \lambda)$ and $\hat{\theta}(K = 40, \beta = 0, \lambda)$ were computed. The experiment was repeated 1000 times and the results reported in fig. 4.7.

The estimator based on the geometric median-of-means performs favorably in both the corrupted and uncorrupted setting. Explicitly, its performance is on par with the previous estimator based on the mean from [LDS16] in the uncorrupted setting, and it outperforms the previous estimator under corruption.

We support this visual conclusion from fig. 4.7 by computing the paired difference in AUC (area under curve) between $\hat{\theta}(K = 1, \beta = 0, \lambda)$ and $\hat{\theta}(K = 40, \beta = 0, \lambda)$. The difference in AUC is neglible in the uncorrupted case (0.0011 ± 0.0005), yet is quite pronounced in favor of $K = 40$ under corruption (0.16 ± 0.004). The uncertainty estimates are 95% bootstrap confidence intervals.

Remark 4.5.1. *Figure 4.7 judges $\hat{\theta}$ from a graphical model perspective, i.e. by whether or not the correct edges were identified. From an estimation perspective, it would also be interesting to judge the distance between $\hat{\theta}$ and the true interaction matrix Θ_0 in some norm. The appended computer code generates this information, however the results are qualitatively very similar to fig. 4.7, which is why they aren't reported separately.*

Finally, we come back to the natural idea of using the componentwise median in the median-of-means. Briefly discussed in section 2.4.1, the componentwise median wasn't pursued further because it doesn't preserve positive definiteness. Still, it's computationally attractive and very robust. Consequently, we simply try it out here without theoretical analysis.

Again, $n = 200$ samples were drawn from a multivariate Gaussian with mean zero and precision matrix Θ_0. This time, the previously used estimator $\hat{\theta}(K = 40, \beta = 0, \lambda)$ based on the geometric median-of-means was compared with $\hat{\theta}_{\text{comp}}(K = 40, \beta = 0, \lambda)$, an analogous version of $\hat{\theta}$ employing the componentwise median-of-means, if the optimization problem to definition 4.1.1 admits a minimizer. Only in 5.5% of 3000 Monte Carlo runs was the matrix $\hat{\Gamma}_K$ from the componentwise median-of-means actually positive definite, i.e. only in 5.5% of the runs could $\hat{\theta}_{\text{comp}}$ be computed.

In a second step, the same corruption procedure as earlier in this section was applied. Among the 5.5% of runs where $\hat{\theta}_{\text{comp}}$ existed, in 72% of cases did $\hat{\theta}_{\text{comp}}$ exist on the corrupted data. This may seem surprisingly

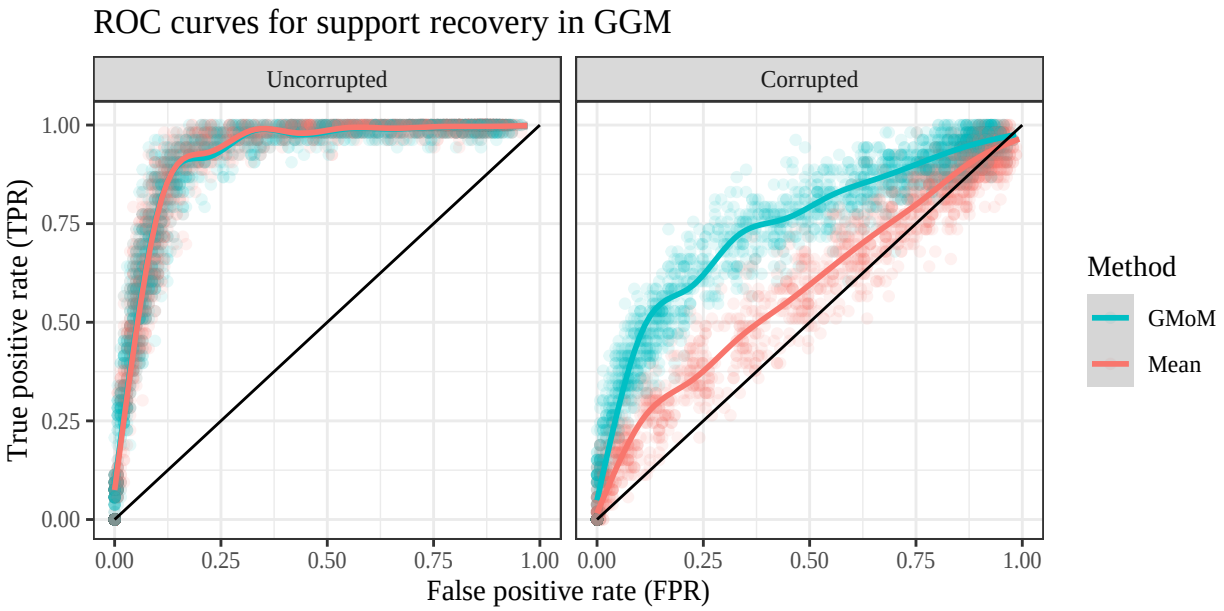


Figure 4.7 Receiver operating characteristic (ROC) curves for how well the zero structure of a Gaussian precision matrix is estimated from simulated data. The two competing versions of $\hat{\theta}(K, \beta, \lambda)$ are $K = 1$ (Mean) and $K = 40$ (Geometric median-of-means (GMoM)). Both ROC curves are on par if the data is uncorrupted, yet the version with GMoM outperforms the mean version when a portion of the samples is corrupted. This figure reports the results of 10^3 Monte Carlo runs. The colored lines represent smoothed aggregate ROC curves, while the transparent dots are points in ROC space originating from different choices of λ in each run. Only a random fraction of these dots is plotted to simplify the visual. Details on the simulation setup and corruption procedure are given in section 4.5.1.

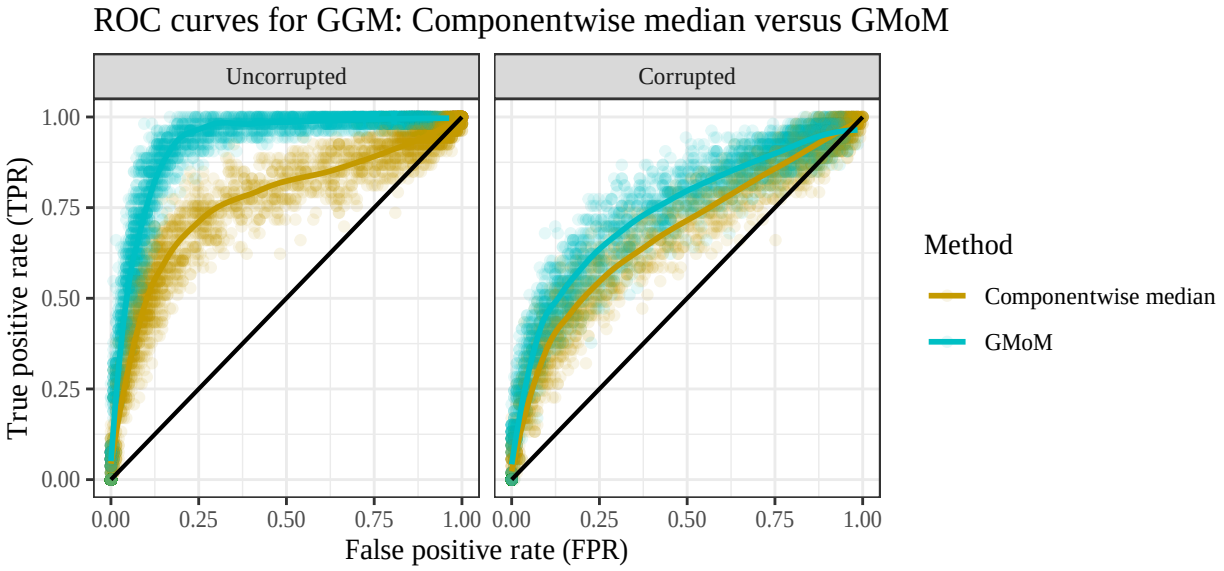


Figure 4.8 The same experiment from fig. 4.7, this time testing a version of $\hat{\theta}$ where the geometric median-of-means (GMoM) is replaced by the componentwise median from section 2.4.1. A Monte Carlo simulation with 3000 runs is performed, however only in 165 of the runs did the componentwise median-of-means result in a positive definite $\hat{\Gamma}_K$ for which the optimization problem from definition 4.1.1 admits a minimizer. The support recovery performance in these 165 runs is reported in the left column. For the right column, data in these 165 runs was corrupted as in fig. 4.7. In 119 cases the the estimator based on the componentwise median-of-means existed. We see that $\hat{\theta}$ based on the GMoM outperforms the mean version in both corruption scenarios.

high, but ironically the heavy-tailed corruption gets squared for the diagonal of $\hat{\Gamma}_K$, i.e. the block means tend to have larger positive diagonals, which carries over through the componentwise median and explains the relatively high proportion of positive definite $\hat{\Gamma}_K$ under corruption.

The support recovery performance of $\hat{\theta}_{\text{comp}}$ (when it existed) is detailed in fig. 4.8.

We conclude from fig. 4.8 that $\hat{\theta}_{\text{comp}}$ is outperformed by $\hat{\theta}$ in support recovery with the gap being smaller in the corrupted case. The analysis in section 4.4.1 suggests that the bias of the univariate median for asymmetric distributions could play a role.

The componentwise median-of-means thus won't be pursued further here.

4.5.2 Square root graphical models

We also investigate how $\hat{\theta}$ performs for the practically relevant square root graphical model from example 2.1.3 and example 2.2.1. Since we include a location-like parameter η and since the blocks of the Γ matrix are no longer equal, the optimization is more expensive. We consider the dimension $m = 10$ and choose the interaction matrix $\Theta_0 \in \mathbb{R}^{10\times 10}$ as described in section 4.5. The parameter η_0 is randomly sampled from a standard normal distribution.

Similar to before, $n = 200$ samples were drawn from a square root graphical model with parameters (Θ_0, η_0). On a grid of 10 values for the regularization parameter λ, the classical estimator $\hat{\theta}(K = 1, \beta = 0, \lambda)$ was compared to $\hat{\theta}(K = 40, \beta = 0, \lambda)$. The function h to smooth at the boundaries (see example 2.2.1) was chosen as $h(x) := x^{3/2}$, the optimal choice found in [YDS19]. For the reasoning behind the choices of β and $K = 40$, see section 4.5.1.

In a second step, again 19 out of the 200 data rows were corrupted. They were replaced by iid draws from a Pareto distribution whose threshold was set to the grand mean of all (clean) data points. The shape parameter was set to one. This way, most of the corrupted points would not stand out, yet the very heavy tails would produce some outliers. The experiment was repeated 200 times and the results are reported in fig. 4.9.

As in section 4.5.1, the results in fig. 4.9 suggest that both the choices $K = 1$ (Mean) and $K = 40$ (GMoM) perform equally well on uncorrupted data, while the estimator involving the GMoM outperforms the mean version under corruption. As in section 4.5.1, we analyze the difference in AUC. Here, there is no significant difference in AUC in the uncorrupted case, and under corruption the difference is 0.24 ± 0.02 in favor of the GMoM procedure.

Remark 4.5.2. *Several choices for the smoothing function h in the square root graphical model are compared in section 7.3.1 of [YDS19]. It's investigated whether truncating h away from the boundary is beneficial. While the optimal candidate $h(x) = x^{3/2}$ also used here performed best without truncation, it's worth noting that truncation naturally improves robustness, since outliers can no longer be amplified through h or its derivative.*

4.5.3 High-dimensional Gaussian graphical model

To conclude the experiments, we consider a truely high-dimensional example where the sample size $n = 40$ is exceeded by the number of nodes $m = 60$. As discussed in section 4.4.3, we require a diagonal multiplier $\beta > 0$.

Similar to before, $n = 40$ samples are drawn from a multivariate Gaussian distribution with mean zero and a randomly chosen interaction matrix $\Theta_0 \in \mathbb{R}^{60\times 60}$ as described in section 4.5. On a grid of 15 values for the regularization parameter λ, the estimator $\hat{\theta}(K = 1, \beta, \lambda)$ was compared to $\hat{\theta}(K = 10, \beta, \lambda)$. The diagonal multiplier β was chosen as discussed in section 4.4.3. The same rationale of choosing K as in section 4.5.1 was applied.

In a second step, 4 out of the 40 samples (again, the corruption amount just below the breakdown point) were corrupted through the same corruption model as in section 4.5.1, and both $\hat{\theta}(K = 1, \beta, \lambda)$ and

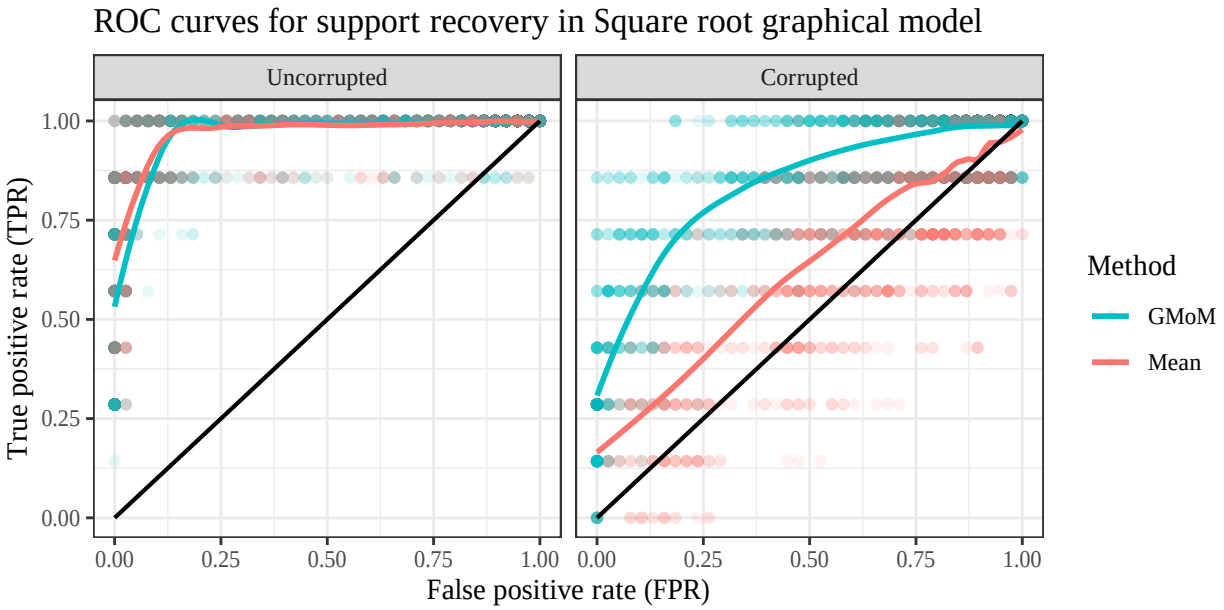


Figure 4.9 A repeat of the experiment in fig. 4.7 for the square root graphical model with 200 Monte Carlo runs. We draw the same conclusion that choosing $K = 1$ (Mean) versus $K = 40$ (geometric median-of-means (GMoM)) in $\hat{\theta}$ has little effect on uncorrupted data, while under corruption the estimator based on the GMoM outperforms the mean version. Details on the simulation and corruption procedure are given in section 4.5.2. The randomly determined underlying graph on $m = 10$ nodes had 7 edges, which is why only 8 different values for the true positive rate are possible. In contrast to fig. 4.7, all transparent points in ROC space could be shown without overloading the plot.

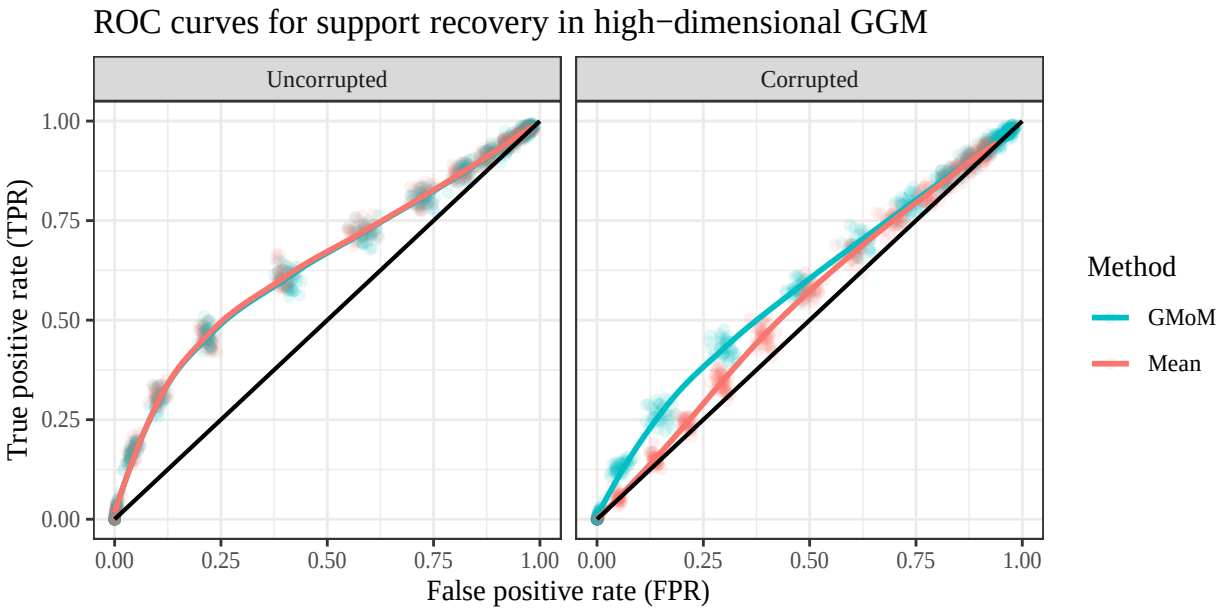


Figure 4.10 A repeat of fig. 4.7 for the high-dimensional case $n < m$. Again, both methods are on par in the uncorrupted case and the method based on the geometric median-of-means outperforms the mean version under corruption. Better coverage of the FPR range would have required a denser grid of regularization parameters λ, which would have increased computation time in addition to the scaling in m. For details, see section 4.5.3.

$\hat{\theta}(K = 10, \beta, \lambda)$ were re-computed. The experiment was repeated 200 times and the results are shown in fig. 4.10.

Visually, our previous conclusion is upheld that both methods perform equally well on uncorrupted data, while the method with $K = 10$ based on the geometric median-of-means has the upper hand under corruption. The difference in AUC is neglible without corruption (0.003 ± 0.001) and an order of magnitude higher under corruption (0.039 ± 0.003) in favour of the geometric median-of-means estimator.

5 Discussion and future work

This thesis presented a score matching approach involving the geometric median-of-means. The simulations in section 4.5 demonstrate promising performance, especially under corruption. It would have been interesting to include other robust estimation procedures in the comparison, but this is out of scope for the thesis. Comparison would have been immediate for the Gaussian graphical model, as there is extensive literature on robust (sparse) precision matrix estimation (see [ÖC15] and references therein). For a general pairwise interaction model, especially on restricted domains, it's less clear what kind of methods make for a good comparison.

Theoretical guarantees for the new estimator are presented in section 4.2. However these don't fully live up to similar guarantees in the literature. Concretely, note that theorem 15 in [YDS19] applies to $\hat{\theta}(K = 1, \beta, \lambda)$, i.e. the special case that the geometric median-of-means reduces to the mean. The difference in the scaling of λ and n between said theorem and theorem 4.2.5 is that theorem 4.2.5 additionally includes $\mathrm{tr}(\Sigma_{\Gamma_0})$ and $\mathrm{tr}(\Sigma_{g_0})$. Since e.g. the diagonal of Γ_0 has m^2 entries, the term $\mathrm{tr}(\Sigma_{\Gamma_0})$ introduces a strong scaling with the number of nodes m.

There are two explanations for this discrepancy. First, theorem 15 in [YDS19] applies to the Gaussian distributions and as discussed in remark 3.3.3, the concentration result from theorem 3.3.2 cannot quite deliver multivariate gaussian concentration. Second, theorem 4.2.2 requires a deviation bound from $\|\cdot\|_\infty$ on Γ_0 and g_0, while the geometric median controls $\|\cdot\|_2$ as in theorem 3.3.2. This gap is bridged by bounding the former norm by the latter in lemma 4.2.4 and theorem 4.2.5, which introduces inefficiency.

Closing this thought, controlling $\hat{\Gamma}_K - \Gamma_0$ in the Frobenius norm (as opposed to e.g. a matrix norm like the spectral norm) could seem peculiar in the first place. However, note that theorem 4.2.2 from [LDS16] also works with the vector maximum norm as opposed to the matrix norm.

The diagonal multiplier β was introduced following [YDS19] in section 4.1 to guarantee that the score matching optimization problem is always well posed. As shown in section 4.3 and particularly in example 4.3.2, this multiplier can bias parameter estimates in a certain direction. It consequently would be desirable to study alternatives to the diagonal multiplier in future work.

Finally, it should be stressed that theory and experiments in this thesis considered the corruption of entire observations (rowwise corruption). Alternatives like cell-wise corruption are harder to study. Very recently, it was shown in [RR24] that a large class of robust estimators including the geometric median can only roughly achieve a cell-wise breakdown point of $1/p$, where p is the ambient dimension. This suggests that there are better robust estimators than the geometric median-of-means under such corruption.

Staying with corruption assumptions, the theoretical analysis in this thesis discarded corrupted bins entirely and only made use of completely uncorrupted bins. This has the advantage that results hold universally without a corruption model and that they even cover adversarial corruption. On the other hand, results can be too pessimistic. Although the breakdown probability from section 3.1 loosened some worst-case assumptions, it too considers the estimation failed once enough blocks are corrupted.

6 Conclusion

A robust estimator for sparse pairwise interaction models was constructed using the median-of-means principle. Robustness against corruption of a few observations was established theoretically and verified in simulations.

The geometric median was chosen for the median-of-means, since it fit the problem-specific requirements best. Surprisingly, the geometric median has another advantage. Usually, tuning the number of blocks in a median-of-means procedure is a tradeoff between robustness from the median and unbiasedness from the block means. Evidence was presented that bias is less of an issue for the geometric median when the ambient dimension is large and components relatively independent. This insight was applied to tuning the number of blocks of the geometric median-of-means in terms of MSE.

A Appendix

The following two plots illustrate how the geometric median changes when a part of the sample is corrupted arbitrarily. The resulting regions are more rich than in the univariate case, but remain bounded since we corrupt below the breakdown point.

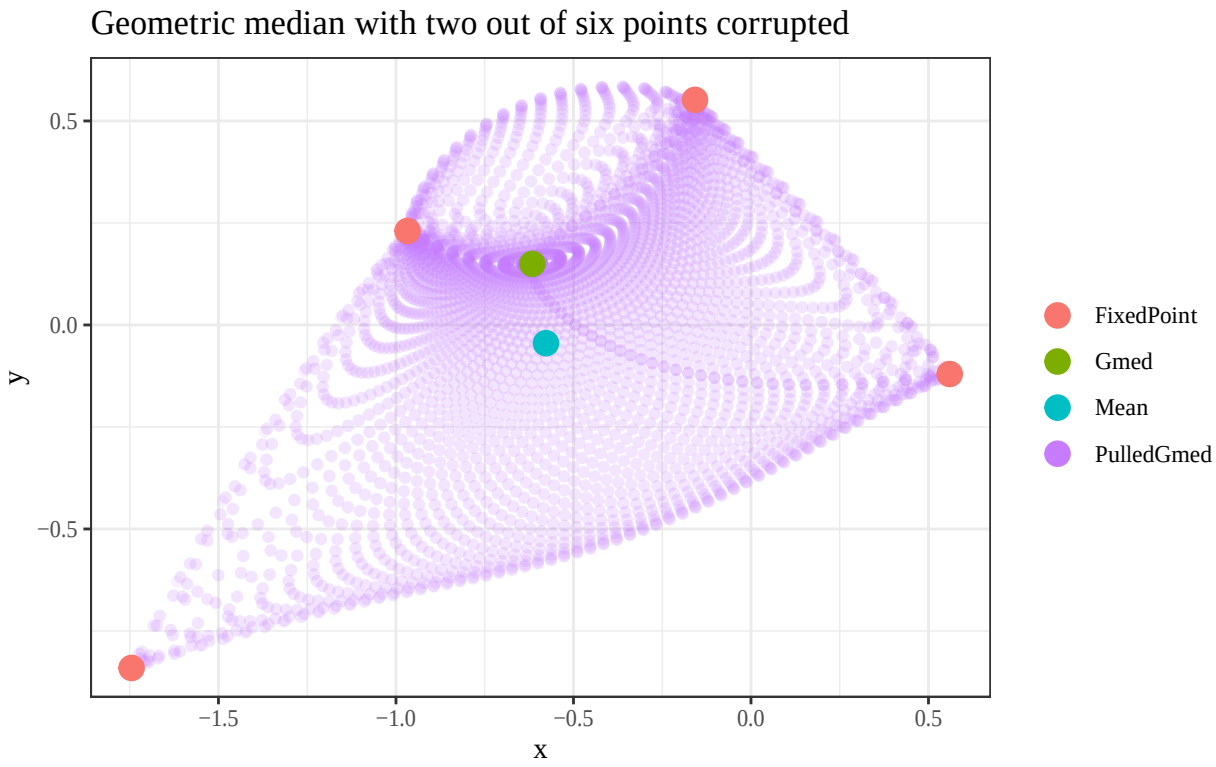


Figure A.1 This purple shape shows the resulting geometric median, when two out of six points are corrupted arbitrarily (i.e. four points remain fixed). The shape remains bounded as two corruptions is below breakdown point of three. However, the geometric median can still vary considerably. There definitely is some structure, but it's an open problem to theoretically derive the purple shape - and as shown in fig. A.2, the picture is much less clear when more points are involved. A more detailed explanation of the experimental setup: four red points are placed in the plane and remain fixed in the coming experiment. Their mean (blue) and geometric median (green) are shown as dots. Two points are added and the geometric median of all six points computed. The two new points are moved to infinity in fixed directions until the geometric median changes little, when it is plotted as a purple point. All combinations of the two corruption directions are iterated through and yield the purple shape.

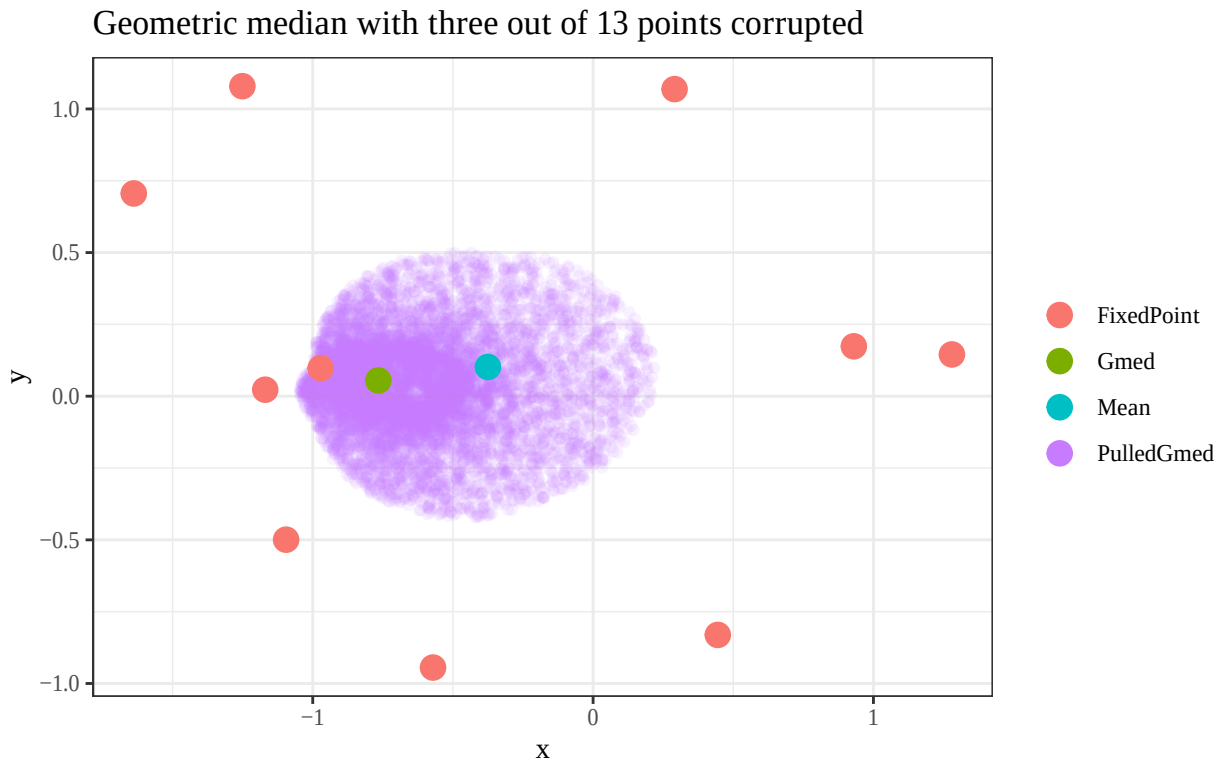


Figure A.2 A repeat of fig. A.1 with more points: ten points remain fixed and three corrupt points are added. This time, the three corruption directions are sampled randomly, which is why the shape seems to be less structured on the inside. We see that the purple shape seems more detached from the fixed points when compared to fig. A.1.

Bibliography

[Alq+09] F. Alqallaf et al. "Propagation of outliers in multivariate data". In: *The Annals of Statistics* 37.1 (Feb. 2009). ISSN: 0090-5364.

[AE06] H. Amann and J. Escher. *Analysis II*. 2., corr. Birkhäuser (Basel), 2006. ISBN: 9783764371050.

[BM23] C. Barber and P. Mozharovskyi. *TukeyRegion: Tukey Region and Median*. R package version 0.1.6.3. 2023.

[BD15] R. F. Barber and M. Drton. "High-dimensional Ising model selection with Bayesian information criteria". In: *Electronic Journal of Statistics* 9.1 (Jan. 2015). ISSN: 1935-7524.

[Bee15] R. A. Beeler. *How to Count: An Introduction to Combinatorics and Its Applications*. Springer International Publishing, 2015. ISBN: 9783319138442.

[Bro83] B. M. Brown. "Statistical Uses of the Spatial Median". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 45.1 (Sept. 1983), pp. 25–30. ISSN: 2517-6161.

[CC08] J. Chen and Z. Chen. "Extended Bayesian information criteria for model selection with large model spaces". In: *Biometrika* 95.3 (Sept. 2008), pp. 759–771. ISSN: 1464-3510.

[Dev+16] L. Devroye et al. "Sub-Gaussian mean estimators". In: *The Annals of Statistics* 44.6 (Dec. 2016). ISSN: 0090-5364.

[DG92] D. L. Donoho and M. Gasko. "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness". In: *The Annals of Statistics* 20.4 (Dec. 1992). ISSN: 0090-5364.

[DM17] M. Drton and M. H. Maathuis. "Structure Learning in Graphical Modeling". In: *Annual Review of Statistics and Its Application* 4.1 (Mar. 2017), pp. 365–393. ISSN: 2326-831X.

[Gup60] S. S. Gupta. "Order Statistics from the Gamma Distribution". In: *Technometrics* 2.2 (May 1960), pp. 243–262. ISSN: 1537-2723.

[Gup71] S. D. Gupta. "Nonsingularity of the Sample Covariance Matrix". In: *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)* 33.4 (1971), pp. 475–478. ISSN: 0581572X.

[HTW15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, May 2015. ISBN: 9780429171581.

[Hyv05] A. Hyvärinen. "Estimation of Non-Normalized Statistical Models by Score Matching". In: *Journal of Machine Learning Research* 6.24 (2005), pp. 695–709.

[IRD16] D. Inouye, P. Ravikumar, and I. Dhillon. "Square Root Graphical Models: Multivariate Generalizations of Univariate Exponential Families that Permit Positive Dependencies". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by M. F. Balcan and K. Q. Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 2016, pp. 2445–2453.

[LSC21] P. Laforgue, G. Staerman, and S. Clémençon. "Generalization Bounds in the Presence of Outliers: a Median-of-Means Study". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 5937–5947.

[Lau04] S. L. Lauritzen. *Graphical models*. Reprinted 2004 with corrections. Oxford statistical science series 17. Oxford [u.a.]: Clarendon Press, 2004. 298 pp. ISBN: 9780198522195.

[LDS16] L. Lin, M. Drton, and A. Shojaie. "Estimation of high-dimensional graphical models using regularized score matching". In: *Electronic Journal of Statistics* 10.1 (2016).

[LMM19] X. Liu, K. Mosler, and P. Mozharovskyi. "Fast Computation of Tukey Trimmed Regions and Median in Dimension p > 2". In: *Journal of Computational and Graphical Statistics* 28.3 (2019), pp. 682–697.

[LR91] H. P. Lopuhaa and P. J. Rousseeuw. "Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices". In: *The Annals of Statistics* 19.1 (1991), pp. 229–248. ISSN: 00905364.

[LM16] G. Lugosi and S. Mendelson. "Risk minimization by median-of-means tournaments". In: (Aug. 2016). arXiv: 1608.00757 [math.ST].

[LM19] G. Lugosi and S. Mendelson. "Mean Estimation and Regression Under Heavy-Tailed Distributions: A Survey". In: *Foundations of Computational Mathematics* 19.5 (Aug. 2019), pp. 1145–1190. ISSN: 1615-3383.

[Mar+19] R. A. Maronna et al., eds. *Robust statistics. Theory and methods*. Second Edition. Wiley series in probability and statistics. Chichester: Wiley, 2019. 1430 pp. ISBN: 9781119214663.

[Mat21] T. Mathieu. "M-estimation and Median of Means applied to statistical learning". Theses. Université Paris-Saclay, Jan. 2021.

[MD87] P. Milasevic and G. R. Ducharme. "Uniqueness of the Spatial Median". In: *The Annals of Statistics* 15.3 (Sept. 1987). ISSN: 0090-5364.

[Min19] S. Minsker. "Distributed statistical estimation and rates of convergence in normal approximation". In: *Electronic Journal of Statistics* 13.2 (Jan. 2019). ISSN: 1935-7524.

[Min15] S. Minsker. "Geometric median and robust estimation in Banach spaces". In: *Bernoulli* 21.4 (Nov. 2015). ISSN: 1350-7265.

[MS23] S. Minsker and N. Strawn. "The Geometric Median and Applications to Robust Mean Estimation". In: (July 2023). arXiv: 2307.03111 [math.ST].

[MNO10] J. Möttönen, K. Nordhausen, and H. Oja. "Asymptotic theory of the spatial median". In: *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková*. Institute of Mathematical Statistics, 2010, pp. 182–193.

[ÖC15] V. Öllerer and C. Croux. "Robust high-dimensional precision matrix estimation". In: *Modern nonparametric, robust and multivariate methods* (2015), pp. 325–350.

[Pol84] D. Pollard. *Convergence of Stochastic Processes*. Springer Series in Statistics Ser. New York, NY: Springer New York, 1984. 1228 pp. ISBN: 9781461252542.

[RR24] J. Raymaekers and P. J. Rousseeuw. "Challenges of cellwise outliers". In: *Econometrics and Statistics* (Feb. 2024). ISSN: 2452-3062.

[Sma90] C. G. Small. "A Survey of Multidimensional Medians". In: *International Statistical Review / Revue Internationale de Statistique* 58.3 (Dec. 1990), p. 263. ISSN: 0306-7734.

[Tib96] R. Tibshirani. "Regression Shrinkage and Selection Via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (Jan. 1996), pp. 267–288. ISSN: 2517-6161.

[Wai19] M. J. Wainwright. *High-dimensional statistics. A non-asymptotic viewpoint*. Cambridge series in statistical and probabilistic mathematics 48. Cambridge: Cambridge University Press, 2019. 1552 pp. ISBN: 9781108627771.

[YDS19] S. Yu, M. Drton, and A. Shojaie. "Generalized Score Matching for Non-Negative Data". In: *Journal of Machine Learning Research* 20.76 (2019), pp. 1–70.

[YLG23] S. Yu, L. Lin, and W. Gilks. *genscore: Generalized Score Matching Estimators*. R package version 1.0.2.1. 2023.