Robust Score Matching

Richard Schwank
Technical University of Munich, Germany

Andrew McCormack
University of Alberta, Canada

Mathias Drton
Technical University of Munich and Munich Center for Machine Learning, Germany

Abstract

Proposed in Hyvärinen (2005), score matching is a parameter estimation procedure that does not require computation of distributional normalizing constants. In this work we utilize the geometric median of means to develop a robust score matching procedure that yields consistent parameter estimates in settings where the observed data has been contaminated. A special appeal of the proposed method is that it retains convexity in exponential family models. The new method is therefore particularly attractive for non-Gaussian, exponential family graphical models where evaluation of normalizing constants is intractable. Support recovery guarantees for such models when contamination is present are provided. Additionally, support recovery is studied in numerical experiments and on a precipitation dataset. We demonstrate that the proposed robust score matching estimator performs comparably to the standard score matching estimator when no contamination is present but greatly outperforms this estimator in a setting with contamination.

1 INTRODUCTION

Detecting and mitigating the influence of outliers or contaminated observations in multivariate data is a challenging task (Maronna et al., 2019), particularly in high-dimensional settings where there are many possible ways in which an observation can be deemed an outlier and where computational considerations play an important role (Diakonikolas and Kane, 2023). In this work we take up the problem of designing robust estimators of high-dimensional joint densities from exponential family models. The solution we propose is a robustified version of score matching, developed with the help of a carefully chosen multivariate median-of-means technique.

An exponential family consists of a collection of probability distributions that have densities of the form

$$p(\mathbf{x}|\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^\top \mathbf{t}(\mathbf{x}) - a(\boldsymbol{\theta}) + b(\mathbf{x})\right), \quad \mathbf{x} \in \mathcal{X}. \quad (1)$$

The parameter $\boldsymbol{\theta}$ ranges over the natural parameter space Ω, which is comprised of points where the integral $\int_{\mathcal{X}} \exp(\boldsymbol{\theta}^\top \mathbf{t}(\mathbf{x}) + b(\mathbf{x}))d\mathbf{x} = \exp(a(\boldsymbol{\theta}))$ is finite. Aside from very special cases like Gaussian models, estimating the parameter $\boldsymbol{\theta}$ in (1) by maximum likelihood is not feasible: in general models (Sun et al., 2015; Roy and Dunson, 2020) the normalizing constant $\exp(a(\boldsymbol{\theta}))$ does not have a closed-form expression and must be found by expensive numerical integration, a problem that is exacerbated by the fact that maximizing the likelihood typically requires iterative optimization procedures.

Score matching (SM), proposed by Hyvärinen (2005), is an estimation procedure that avoids the aforementioned shortcomings of maximum likelihood estimation in exponential families for continuous data. It does not require that the normalizing constant $\exp(a(\boldsymbol{\theta}))$ be known. Moreover, score matching amounts to minimizing a convex, quadratic loss function in $\boldsymbol{\theta}$, a task that is easily solved even when $\boldsymbol{\theta}$ is high-dimensional. One prominent application of score matching is estimation of non-Gaussian graphical models (Yu et al., 2019) that may be formed by structuring the sufficient statistics $\mathbf{t}(\mathbf{x})$ in (1) so that the vanishing of components of $\boldsymbol{\theta}$ correspond to conditional independence relations between variables (Lauritzen, 1996).

The central contribution of this work is to extend the score matching methodology to handle data that has been corrupted or contains outliers. Specifically, we propose using the geometric median of means (GMoM) (Minsker, 2015) to robustify the quadratic empirical loss function for exponential family score matching.

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

Crucially, by using the geometric median of means the robustified objective function remains convex. This property is in general not preserved by other multivariate medians, such as the componentwise median. The GMoM interpolates between the mean and geometric median of data points, with a block-size parameter determining the relative proximity of the GMoM to the mean and geometric median. By altering the block-size parameter we show how the proposed method can be tuned to handle different levels of corruption.

Our paper is structured as follows. Sections 2.1-2.2 review score matching and the geometric median of means, respectively. Section 3 details the proposed robust score matching procedure, provides theoretical results on robustness against contamination, and discusses hyperparameter tuning. Section 4 introduces an ℓ_1-regularized version of robust score matching for graphical models and presents a support recovery guarantee under corruption. Simulations in Section 5 illustrate the efficacy of robust score matching. This section concludes by applying the proposed procedure to a data set on precipitation in the Alps. Proofs of all theorems can be found in the appendix.

Notation: Scalars are denoted by α, x, etc., vectors by $\mathbf{x}$, $\boldsymbol{\theta}$, etc., and matrices by $\mathbf{X}, \boldsymbol{\Theta}$, etc. Subscripts X_{ij} and x_i indicate matrix and vector components, while the superscripts on $\mathbf{x}^{(i)}$ index different observations in a random sample. The gradient is denoted by $\nabla = (\partial_1, \dots, \partial_p)$ and ∂_{jj} denotes the second partial derivatives with respect to an argument x_j. Important vector norms are the Euclidean norm $\|\cdot\|_2$, the Manhattan norm $\|\cdot\|_1$, and the maximum norm $\|\cdot\|_\infty$. For a matrix $\mathbf{A}$, $|||\mathbf{A}|||_{\infty,\infty} = \max_{i=1,\dots,a} \sum_{j=1}^{b} |A_{ij}|$, and $\operatorname{tr}(\mathbf{A})$ and $\operatorname{diag}(\mathbf{A})$ are the trace and the diagonal part, respectively. Finally, $1\{b\}$ is the indicator function that equals 1 if the boolean b is true and zero otherwise.

2 PRELIMINARIES

2.1 Generalized score matching

In this section, we review the main aspects of score matching in the generalized form introduced by Yu et al. (2019, Sect. 2).

Suppose that n i.i.d. observations $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^m$ are sampled from a distribution in an r-dimensional exponential family $\mathcal{P} = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Omega \subseteq \mathbb{R}^r\}$ with densities of the form (1). In our use cases, the densities are considered with respect to Lebesgue measure on $\mathcal{X} = \mathbb{R}^m$ or $\mathcal{X} = [0, \infty)^m$. Write $\mathbf{X}$ for the $n \times m$ matrix of all observations. For practical consideration of general families $\mathcal{P}$, it is crucial to be able to obtain an estimate of the true parameter value $\boldsymbol{\theta}_0$ that does not require the computation of the normalizing constant $\exp(-a(\boldsymbol{\theta}))$. This may be achieved via the generalized score matching estimator, which minimizes an empirical loss function that approximates the loss function

$$J_h(\boldsymbol{\theta}) = \frac{1}{2} \int_{\mathcal{X}} \|\nabla_{\mathbf{x}} \log(p(\mathbf{x}|\boldsymbol{\theta})) \circ \mathbf{h}(\mathbf{x})^{1/2} - \nabla_{\mathbf{x}} \log(p(\mathbf{x}|\boldsymbol{\theta}_0)) \circ \mathbf{h}(\mathbf{x})^{1/2}\|_2^2 \, p(\mathbf{x}|\boldsymbol{\theta}_0) \mathrm{d}\mathbf{x}, \quad (2)$$

where gradient $\nabla_{\mathbf{x}}$ is taken with respect to the data $\mathbf{x}$, $\circ$ is the componentwise product of vectors, and $\mathbf{h}(\mathbf{x})^{1/2} = (h_1(\mathbf{x})^{1/2}, \dots, h_m(\mathbf{x})^{1/2})$ comprises the square roots of m non-negative weight functions. Nontrivial weighting is needed when the support $\mathcal{X}$ is constrained. The loss $J_h(\boldsymbol{\theta})$ in (2) is the expected weighted squared distance between the true score function and the score function $\nabla_{\mathbf{x}} \log(p(\mathbf{x}|\boldsymbol{\theta}))$ at $\boldsymbol{\theta}$. Under mild conditions on $\mathbf{h}$, the unique minimizer of (2) is $\boldsymbol{\theta}_0$ (Yu et al., 2019, Prop 2).

A key property of $J_h(\boldsymbol{\theta})$ is that after an integration by parts, $J_h(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}_0 \boldsymbol{\theta} - \mathbf{g}_0^\top \boldsymbol{\theta}$ up to constants not depending on $\boldsymbol{\theta}$, with $\boldsymbol{\Gamma}_0 = \mathrm{E}_{\boldsymbol{\theta}_0}[\boldsymbol{\Gamma}(\mathbf{x})]$ and $\mathbf{g}_0 = \mathrm{E}_{\boldsymbol{\theta}_0}[\mathbf{g}(\mathbf{x})]$ where

$$\boldsymbol{\Gamma}(\mathbf{x}) = \sum_{j=1}^{m} h_j(\mathbf{x}) \partial_j \mathbf{t}(\mathbf{x}) \partial_j \mathbf{t}(\mathbf{x})^\top \in \mathbb{R}^{r \times r}, \quad (3)$$

$$\mathbf{g}(\mathbf{x}) = -\sum_{j=1}^{m} \big(h_j(\mathbf{x}) \partial_j b(\mathbf{x}) \partial_j \mathbf{t}(\mathbf{x}) + h_j(\mathbf{x}) \partial_{jj} \mathbf{t}(\mathbf{x}) + \partial_j h_j(\mathbf{x}) \partial_j \mathbf{t}(\mathbf{x}) \big) \in \mathbb{R}^r. \quad (4)$$

The empirical loss $J_h(\boldsymbol{\theta}; \mathbf{X})$ replaces $\boldsymbol{\Gamma}_0$ by $\overline{\boldsymbol{\Gamma}}(\mathbf{X}) := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\Gamma}(\mathbf{x}^{(i)})$ and $\mathbf{g}_0$ by $\overline{\mathbf{g}}(\mathbf{X}) := \frac{1}{n}\sum_{i=1}^n \mathbf{g}(\mathbf{x}^{(i)})$. The resulting score matching estimator is

$$\hat{\boldsymbol{\theta}}(\mathbf{X}) := \operatorname*{argmin}_{\boldsymbol{\theta} \in \Omega} J_h(\boldsymbol{\theta}; \mathbf{X}) = \overline{\boldsymbol{\Gamma}}(\mathbf{X})^{-1} \overline{\mathbf{g}}(\mathbf{X}). \quad (5)$$

The score matching estimator $\hat{\boldsymbol{\theta}}$ depends on the weighting function $\mathbf{h}$ through $\boldsymbol{\Gamma}$ and $\mathbf{g}$. When the sample space $\mathcal{X}$ is $\mathbb{R}^m$ the constant weighting $\mathbf{h}(x) = (1, \dots, 1)$ can be used. When $\mathcal{X}$ has a boundary, a suitable choice of $\mathbf{h}$ can dampen boundary effects to ensure that the integration by parts argument is valid (Hyvärinen, 2007; Yu et al., 2019). Under mild regularity conditions, $\hat{\boldsymbol{\theta}}$ consistently estimates $\boldsymbol{\theta}_0$ as $n \to \infty$. In high-dimensional scenarios where $r > n$, the matrix $\overline{\boldsymbol{\Gamma}}(\mathbf{X})$ is not invertible and the score matching estimator does not exist. To handle such settings, Yu et al. (2019) modify the objective function $J_h(\boldsymbol{\theta}; \mathbf{X})$ in (5) by adding positive offsets to the diagonal entries of $\overline{\boldsymbol{\Gamma}}(\mathbf{X})$.

Example 2.1 (Square Root Graphical Model)**.** *Consider non-negative data with $\mathcal{X} = [0, \infty)^m$. The square root graphical model is parametrized by a pair $\boldsymbol{\theta} = (\boldsymbol{\Theta}, \boldsymbol{\eta})$ with $\boldsymbol{\Theta} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{\eta} \in \mathbb{R}^m$, and has*

densities

$$p(\mathbf{x}|\boldsymbol{\theta}) = \exp\Bigg(-\sum_{i=1}^{m} \Theta_{ii} x_i + \sum_{1\le i<j\le m} 2\Theta_{ij} x_i^{1/2} x_j^{1/2} + \sum_{i=1}^{m} 2\eta_i x_i^{1/2} - a(\boldsymbol{\theta}) \Bigg) \quad (6)$$

with respect to Lebesgue measure (Inouye et al., 2016). This is an example of a pairwise interaction model (Yu et al., 2016) where the interaction parameter Θ_{ij} represents the degree of dependence between x_i and x_j conditionally on all of the other components. In particular, if $\Theta_{ij} = 0$ then x_i is conditionally independent of x_j given $\{x_k : k \neq i, j\}$. In the formalism of graphical models, these independencies can be expressed in an undirected graph (Maathuis et al., 2019). The normalizing constant in this model is intractable, making score matching an attractive approach.

2.2 Geometric median of means and robustness

Recent literature has popularized the (univariate) median-of-means (MoM) as a robust mean estimator (Devroye et al., 2016; Laforgue et al., 2021). The geometric median of means (GMoM) is a multivariate generalization of the MoM (Minsker, 2015).

Definition 2.2. *The GMoM of $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ with K (a divisor of n) blocks is defined as*

$$\mathrm{GMoM}_K\Big(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\Big) := \mathrm{GMed}\Big(\hat{\boldsymbol{\mu}}^{(1)}, \dots, \hat{\boldsymbol{\mu}}^{(K)}\Big),$$

where $\hat{\boldsymbol{\mu}}^{(j)}$ is the sample mean of $\mathbf{x}^{((j-1)K+1)}, \dots, \mathbf{x}^{(jK)}$ and GMed *denotes the* geometric median *defined as*

$$\mathrm{GMed}\Big(\hat{\boldsymbol{\mu}}^{(1)}, \dots, \hat{\boldsymbol{\mu}}^{(K)}\Big) = \operatorname*{argmin}_{\mathbf{m}\in\mathbb{R}^p} \sum_{i=1}^{K} \|\hat{\boldsymbol{\mu}}^{(i)} - \mathbf{m}\|. \quad (7)$$

When $p = 1$, the GMoM reduces to the MoM, because the geometric median of real numbers equals the standard median. In this case, the MoM partitions observations into K blocks, computes the sample mean within each block, and then aggregates the block means by taking a sample median. Hence, the MoM is an interpolation between the sample mean ($K = 1$) and the sample median ($K = n$). For intermediate values of K, the MoM inherits robustness properties of the geometric median while also being an approximately unbiased estimator of the population mean. If the block-sizes of the MoM are increasing, asymptotically the MoM is a consistent estimator of the population mean (Minsker, 2019, Sect. 2.5). An advantage of the MoM over the mean is that if the moment generating function of the population distribution does not exist the sample mean will concentrate around the mean at a polynomial rate, whereas the MoM achieves sub-Gaussian concentration when second moments exist; see Lugosi and Mendelson (2019) for an in-depth discussion.

If the ambient dimension p is larger than one, the GMoM inherits concentration properties and robustness against outliers from its univariate counterpart, as shown in Section 3. We consider outliers originating from the contamination of entire observations, also referred to as *rowwise* corruption. See Section 3.1 for details. A basic quantity to assess robustness against rowwise corruption is the *breakdown point* (Lopuhaä and Rousseeuw, 1991). It is the minimal proportion of observations that, if tampered with arbitrarily, can force the estimator to diverge to infinity.

In principle, one could generalize the MoM to higher dimensions using any multivariate median concept (see the survey of Small, 1990) instead of the geometric median. However, subleties arise for our later application in robust score matching as we seek robust estimates of a collection of many means that feature in a loss that ought to admit a well-defined minimizer. For this reason, a candidate median concept for robust score matching should satisfy the following properties:

- **(R1)** The median should be a convex combination of its arguments. This is to ensure that the median of positive semidefinite matrices is again positive semidefinite, which is needed for applying the GMoM to $\boldsymbol{\Gamma}(\mathbf{x})$ (Section 3).
- **(R2)** Computation should be feasible in high dimensions. This is because the number of parameters in a graphical model scales quadratically with the number of nodes.
- **(R3)** The median should have a high breakdown point against rowwise contamination.

Many high-dimensional estimation problems can be addressed surprisingly well by seemingly simple coordinate-wise procedures. However, our argument against a componentwise median is that it fails to satisfy **(R1)**. In practice, this entails (robustly) estimated loss functions that end up being unbounded below, with no associated score matching estimator. In contrast, the geometric median $\mathbf{m}$, if it does not equal one of its arguments $\hat{\boldsymbol{\mu}}^{(1)}, \dots, \hat{\boldsymbol{\mu}}^{(K)}$, can be rewritten as

$$\mathbf{m} = \frac{1}{\sum_{i=1}^{K} 1/\|\mathbf{m} - \hat{\boldsymbol{\mu}}^{(i)}\|_2} \sum_{i=1}^{K} \frac{\hat{\boldsymbol{\mu}}^{(i)}}{\|\mathbf{m} - \hat{\boldsymbol{\mu}}^{(i)}\|_2} \quad (8)$$

by setting gradient with respect to $\mathbf{m}$ in (7) to zero. The GMed thus satisfies **(R1)**.

Regarding the computational requirement **(R2)**, note that while there is not a closed-form solution for the geometric median, equation (8) immediately suggests a fixed point algorithm called *Weiszfeld's algorithm*. The computational complexity of a single iteration step is only $\mathcal{O}(pK)$. In contrast, other well-known multivariate medians often have exponential complexity in the ambient dimension p; see Ronkainen et al. (2003) for the *Oja median* and Liu et al. (2019) for the *Tukey median*. Convergence of Weiszfeld's algorithm is guaranteed under slight modifications that prevent getting stuck on the input vectors (Vardi and Zhang, 2001).

Lastly, the geometric median satisfies **(R3)** as its breakdown point is $\lfloor (K+1)/2 \rfloor / K$ (Lopuhaä and Rousseeuw, 1991), the same as that of the univariate median. In contrast, the breakdown point of the Oja median tends to zero with the sample size n (Niinimaa et al., 1990). It is between $1/3$ and $1/(p+1)$ for the Tukey median (Donoho and Gasko, 1992).

While our approach uses the GMoM to obtain a robust aggregate, we would like to mention that alternative frameworks for this problem exist, e.g. distributionally robust optimization Blanchet et al. (2024); Kuhn et al. (2024).

3 ROBUST SCORE MATCHING FOR CONTAMINATED DATA

This section introduces a generalization of the score matching estimator $\hat{\boldsymbol{\theta}}$ from (5) and investigates its robustness against contamination.

3.1 Contamination assumptions

In the classical Tukey-Huber contamination model (Maronna et al., 2019, Sect. 2.2), the observed vector $\mathbf{y} \in \mathbb{R}^p$ equals $\mathbf{y} = (\mathbf{I} - \mathbf{B})\mathbf{x} + \mathbf{B}\mathbf{z}$, where $\mathbf{x}$ is the uncorrupted observation, $\mathbf{z}$ is a random contamination vector, $\mathbf{I}$ is the $p \times p$ identity matrix, and $\mathbf{B}$ is a diagonal matrix, either being $\mathbf{I}$ with probability $\varepsilon > 0$ or the zero-matrix otherwise. In a data frame where rows are observations, under the Tukey-Huber model any row is either corrupted or not, and thus this is a form of *rowwise* corruption.

In this paper, we consider rowwise contamination, however, we do not require that the contamination occurs at random like in the Tukey-Huber model. Instead, we assume that a proportion ε of the rows could have been altered arbitrarily. This includes *adversarial* contamination by an intelligent attacker; see, e.g., Bhatt et al. (2022). We note that yet other forms of corruption could be considered in future work; compare, e.g., the *cellwise* contamination treated by Alqallaf et al., 2009.

3.2 A robust estimator based on the GMoM

The classical score matching estimator $\hat{\boldsymbol{\theta}}$ from (5) minimizes $\frac{1}{2}\boldsymbol{\theta}^\top \overline{\boldsymbol{\Gamma}}(\mathbf{X})\boldsymbol{\theta} - \boldsymbol{\theta}^\top \overline{\mathbf{g}}(\mathbf{X})$. We propose to replace $\overline{\boldsymbol{\Gamma}}(\mathbf{X})$ and $\overline{\mathbf{g}}(\mathbf{X})$ with a more robust version using the GMoM. In symbols, we set

$$\begin{aligned} \hat{\boldsymbol{\Gamma}}_K(\mathbf{X}) &:= \mathrm{GMoM}_K\Big(\boldsymbol{\Gamma}(\mathbf{x}^{(1)}), \dots, \boldsymbol{\Gamma}(\mathbf{x}^{(n)}) \Big), \\ \hat{\mathbf{g}}_K(\mathbf{X}) &:= \mathrm{GMoM}_K\Big(\mathbf{g}(\mathbf{x}^{(1)}), \dots, \mathbf{g}(\mathbf{x}^{(n)}) \Big). \end{aligned} \tag{9}$$

Note that when applying the GMoM each $\boldsymbol{\Gamma}(\mathbf{x}^{(i)}) \in \mathbb{R}^{r \times r}$ is interpreted as a vector in $\mathbb{R}^{r^2}$. When the parameter K for the number of blocks equals one, $\hat{\boldsymbol{\Gamma}}_K(\mathbf{X})$ and $\hat{\mathbf{g}}_K(\mathbf{X})$ reduce to $\overline{\boldsymbol{\Gamma}}(\mathbf{X})$ and $\overline{\mathbf{g}}(\mathbf{X})$. For $K > 1$, the equal weights $1/n$ are replaced by non-negative weights that sum to one in which block means that contain outliers are downweighted, as shown in (8). We then propose the estimator

$$\hat{\boldsymbol{\theta}}(K) := \operatorname*{argmin}_{\boldsymbol{\theta} \in \Omega} \tfrac{1}{2}\boldsymbol{\theta}^\top \hat{\boldsymbol{\Gamma}}_K(\mathbf{X})\boldsymbol{\theta} - \boldsymbol{\theta}^\top \hat{\mathbf{g}}_K(\mathbf{X}), \tag{10}$$

which exists uniquely if and only if $\hat{\boldsymbol{\Gamma}}_K(\mathbf{X})$ is positive definite, in which case $\hat{\boldsymbol{\theta}}(K) = \hat{\boldsymbol{\Gamma}}_K^{-1}(\mathbf{X})\hat{\mathbf{g}}_K(\mathbf{X})$. In the classical problem (5), $\overline{\boldsymbol{\Gamma}}(\mathbf{X})$ looks very similar to a sample covariance matrix, which would be almost surely positive definite when the sample size n exceeds the ambient dimension m (Eaton and Perlman, 1973). Similarly, a sufficient sample size guarantees the positive definiteness of $\hat{\boldsymbol{\Gamma}}_K(\mathbf{X})$ under mild regularity conditions on the sufficient statistic $\mathbf{t}$, as detailed in the appendix. It is this guarantee of positive definiteness that stems from the use of the geometric median over conceptually and computationally simpler methods like the componentwise median.

Remark 3.1. *Barp et al. (2019) analyze minimum Stein discrepancy estimation, including so-called* diffusion score matching *(DSM) as a special case. They demonstrate DSM's robustness against contamination for suitable diffusion functions, which generalize the* $\mathbf{h}$*-functions considered in this paper. This approach could complement the robust aggregation of* $(\boldsymbol{\Gamma}(\mathbf{x}^{(i)}))_i$ *and* $(\mathbf{g}(\mathbf{x}^{(i)}))_i$ *via the GMoM.*

3.3 A first robustness guarantee under contamination

We now show that the robust score matching estimator from (10) consistently estimates the true parameter $\boldsymbol{\theta}_0$ even if a part of the observations are contaminated as described in Section 3.1. We begin by deriving a concentration result of the GMoM around the population mean under corruption, which is also useful when we consider sparse graphical models in the next section.

Theorem 3.2. *Let* $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ *be independent samples from a p-dimensional distribution with mean*

$\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. Fix a confidence level of $0 < \delta \leq 1$. We allow for up to $(\lfloor 17 \cdot \log(1/\delta) \rfloor + 1)\tau$ samples to be arbitrarily corrupted, where $0 \leq \tau < 1/2$. There exists functions $k(\tau) = \mathcal{O}(1/(\frac{1}{2} - \tau)^2)$ and $c(\tau) = \mathcal{O}(1/(\frac{1}{2} - \tau)^{2.5})$ as $\tau \to \frac{1}{2}$ such that when the number of blocks K defined as $K = K(\delta, \tau) := \lfloor k(\tau) \cdot \log(1/\delta) \rfloor + 1$ satisfies $K \leq n/2$, it holds that

$$\mathrm{P}\Bigg(\| \operatorname{GMoM}_K\Big(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \Big) - \boldsymbol{\mu} \|_2 > c(\tau) \sqrt{\log\left(\frac{4}{(1-\tau)^2} \frac{1}{\delta} \right) \frac{\operatorname{tr}(\boldsymbol{\Sigma})}{n}} \Bigg) \leq \delta. \quad (11)$$

To interpret Theorem 3.2, it is helpful to consider the complementary statement of (11). It reads that with probability at least $1 - \delta$, the GMoM approximates $\boldsymbol{\mu}$ correctly up to some bound $B(n, \delta, \tau)$. To illustrate how this can be used, assume that the number of corrupt samples n_c grows with n but is $o(n)$. For some fixed τ_0, one can set $\delta := \exp(-(\lceil n_c/\tau_0 \rceil - 1)/17)$ to satisfy the assumptions of the theorem, resulting in $K = o(n)$. By the assumption on n_c, both δ and $B(n, \delta, \tau_0)$ are $o(1)$ as $n \to \infty$. Thus, the GMoM converges in probability to $\boldsymbol{\mu}$ as $n \to \infty$ in the considered setting:

Corollary 3.3. *Let $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ be independent samples from an exponential family distribution with parameter $\boldsymbol{\theta}_0 \in \mathbb{R}^r$. Assume the data generating model satisfies the mild regularity assumptions listed in the appendix. Allow for up to n_c samples to be arbitrarily corrupted, where $n_c = o(n)$. Then, there exists a sequence $K = K(n_c)$ such that the robust score matching estimator $\hat{\boldsymbol{\theta}}(K(n_c))$ from (10) converges to $\boldsymbol{\theta}_0$ in probability as $n \to \infty$.*

The concentration statement (11) is based on the work of Minsker (2015, Cor. 4.1 & Rem. 3.1). The theorem shares traits with results in the literature: a logarithmic relation between $1/\delta$ and K (Lugosi and Mendelson, 2019) and between $1/\delta$ and the corruption τ (Laforgue et al., 2021). The range of the corruption parameter τ between 0 and $1/2$ reflects the high breakdown point of $1/2$ of the geometric median, cf. (**R3**). Concretely, the assumptions of the theorem ensure that at most $\tau K < K/2$ block means are corrupted, as detailed in the proof. For the Tukey median for instance, we would expect $\tau < 1/(p+1)$.

3.4 Choice of number of blocks K

Choosing the number of blocks K is a trade-off between robustness, bias and variance. As the number of blocks increases, the GMoM becomes more robust since the breakdown point of the GMoM is equal to $\lfloor \frac{1}{2}(K+1) \rfloor / n$, which grows with K. The effect that increasing K has on the variance is problem dependent. For Gaussian location estimation, the maximal choice $K = n$ has higher asymptotic variance than the mean $K = 1$ as shown in Brown (1983). In heavy tailed scenarios on the other hand, the GMoM has relatively light tails as shown in Theorem 3.2, which indicates that it can have lower variance than the sample mean. The bias of the GMoM also depends on the problem. The GMoM is an unbiased location estimator for any K when the underlying distribution is centrally symmetric, i.e., when $\mathbf{x} - \mathrm{E}[\mathbf{x}]$ and $\mathrm{E}[\mathbf{x}] - \mathbf{x}$ have the same distribution (Serfling, 2006). In general however, the geometric median is a biased estimator for the population mean, making the GMoM biased as well. Still, the GMoM typically has small bias for small K, as the central limit theorem implies that the block means are approximately Gaussian and thus centrally symmetric. For large K the bias will generally be larger as the GMoM approaches the geometric median at $K = n$.

A general choice of K when a proportion ε of samples is corrupted should comfortably exceed the breakdown point for robustness, but not be too large in order to avoid bias. This reasoning is supported by a simulation study in the appendix. Since the breakdown point is exceeded when $K \geq 2\varepsilon$, we propose $K := 4\varepsilon n$ as a compromise. This choice works well empirically as shown in Section 5. If $4\varepsilon n$ is not an integer, is smaller than one, or is greater than n, K is chosen to be the nearest admissible integer.

4 APPLICATION TO HIGH-DIMENSIONAL GRAPHICAL MODELING

As an application of special interest, we consider a general pairwise interaction model given by

$$p(x|\boldsymbol{\Theta}, \boldsymbol{\eta}) := \exp\Bigg(- \sum_{1 \leq i \leq j \leq m} \Theta_{ij} t_{ij}(x_i, x_j) - \sum_{i=1}^{m} \eta_i t_i(x_i) - a(\boldsymbol{\Theta}, \boldsymbol{\eta}) \Bigg), \quad \mathbf{x} \in \mathcal{X}, \quad (12)$$

where the domain $\mathcal{X}$ can be $\mathbb{R}^m$ or have boundaries like in Example 2.1. Let $\mathbf{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\}$ be an i.i.d. sample from (12). Score matching for pairwise interaction models simplifies structurally when the dummy variables $\Theta_{ji} := \Theta_{ij}$ for $j < i$ are introduced; for example, $\boldsymbol{\Gamma}(\mathbf{x})$ is block-diagonal (Yu et al., 2019). To apply the theory from Yu et al. (2019), this section assumes that $\boldsymbol{\Gamma}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ are derived for the (extended) square parameter matrix $\boldsymbol{\Theta} = (\Theta_{ij})$. We abbreviate the pair of $\boldsymbol{\Theta}$ together with the parameter vector $\boldsymbol{\eta}$ by a single r-dimensional parameter $\boldsymbol{\theta}$.

Motivated by applications such as gene regulatory networks, we focus on the case that the dimension m is

large, most Θ_{ij} are zero (Oh and Deasy, 2014), and the sample size n is smaller than the dimension m (Chu et al., 2009). To incorporate the sparsity assumption, we include an ℓ_1-regularization penalty in the objective function. For $n < m$, we follow Yu et al. (2019) and include a diagonal multiplier that ensures that the Gram matrix is positive definite. We thus propose the following estimator:

Definition 4.1. *Using $\hat{\mathbf{\Gamma}}_K$ and $\hat{\mathbf{g}}_K$ from (9), define for $\beta, \lambda > 0$ with $\hat{\mathbf{\Gamma}}_{K;\beta} := \hat{\mathbf{\Gamma}}_K + \beta \cdot \operatorname{diag}(\hat{\mathbf{\Gamma}}_K)$,*

$$\hat{\boldsymbol{\theta}}(K, \beta, \lambda) := \operatorname*{argmin}_{\boldsymbol{\theta} \in \Omega} \frac{1}{2} \boldsymbol{\theta}^\mathsf{T} \hat{\mathbf{\Gamma}}_{K;\beta} \boldsymbol{\theta} + \hat{\mathbf{g}}_K^\mathsf{T} \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_1. \quad (13)$$

Computationally, this estimator is attractive. First, $\mathbf{\Gamma}(\mathbf{x}^{(i)})$ and $\mathbf{g}(\mathbf{x}^{(i)})$ are assembled for $i = 1, \ldots, n$. This needs $\mathcal{O}(nmr^2)$ operations under the assumption that $\partial_i T$ is evaluated in constant time. Since the number of parameters r is $\mathcal{O}(m^2)$, the complexity is $\mathcal{O}(nm^5)$, although factoring in the block-diagonal structure of $\mathbf{\Gamma}(\mathbf{x})$ reduces this to $\mathcal{O}(nm^4)$ and even $\mathcal{O}(nm^3)$ in very symmetric models like the square root graphical model (Yu et al., 2019). Next, the GMoMs $\hat{\mathbf{\Gamma}}_K(\mathbf{X})$ and $\hat{\mathbf{g}}_K(\mathbf{X})$ are computed iteratively, where each iteration requires $\mathcal{O}(r^2 K) = \mathcal{O}(m^4 K)$ operations (only $\mathcal{O}(m^3 K)$ when factoring in the block structure of $\mathbf{\Gamma}(\mathbf{x})$). The actual ℓ_1-regularized optimization problem can be solved by iterative methods like coordinate descent (Friedman et al., 2007a).

To treat the estimator from (13) theoretically, we introduce the following notation and definitions:

Definition 4.2. *Let $\boldsymbol{\theta}_0 = (\mathbf{\Theta}_0, \boldsymbol{\eta}_0)$ be the unknown true parameter and $\mathbf{\Gamma}_0 := \mathrm{E}_{\boldsymbol{\theta}_0}[\mathbf{\Gamma}(\mathbf{x})]$. Define*

$$d_{\boldsymbol{\theta}_0} := \max_{j=1,\ldots,m} \big(\#\{i : (\mathbf{\Theta}_0)_{ij} \neq 0\} + 1\{(\boldsymbol{\eta}_0)_j \neq 0\} \big).$$

Let $c_{\boldsymbol{\theta}_0} := |||\mathbf{\Theta}_0|||_{\infty,\infty}$. Write $S(\boldsymbol{\theta}) := \{i : \boldsymbol{\theta}_i \neq 0\}$ for the support of a parameter vector. Abbreviate $S_0 := S(\boldsymbol{\theta}_0)$. Further, if $\mathbf{\Gamma}_{0,S_0 S_0}$ is invertible, set

$$c_{\mathbf{\Gamma}_0} := \left| \left| \left| (\mathbf{\Gamma}_{0,S_0 S_0})^{-1} \right| \right| \right|_{\infty,\infty}$$
$$I_{S_0} := \left| \left| \left| \mathbf{\Gamma}_{0,S_0^c S_0} (\mathbf{\Gamma}_{0,S_0 S_0}^{-1}) \right| \right| \right|_{\infty,\infty}.$$

Finally, we say $\mathbf{\Gamma}_0$ satisfies the irrepresentability condition *with incoherence parameter $\alpha \in (0, 1]$ and edge set S_0, if $I_{S_0} \leq (1 - \alpha)$.*

The following theorem shows concentration of $\hat{\boldsymbol{\theta}}(K, \beta, \lambda)$ around the true parameter $\boldsymbol{\theta}_0$ with high probability even under contamination, extending the work of Yu et al. (2019) and Lin et al. (2016).

Theorem 4.3. *Let $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \in \mathbb{R}^m$ be i.i.d. samples from a pairwise interaction model with parameter $\boldsymbol{\theta}_0$. Assume that $\mathbf{\Gamma}_0$ satisfies the irrepresentability condition with parameter α and edge set S_0. Further, suppose $\mathbf{\Sigma}_{\mathbf{\Gamma}_0} := \operatorname{Var}_{\boldsymbol{\theta}_0}(\mathbf{\Gamma}(\mathbf{x}))$ and $\mathbf{\Sigma}_{\mathbf{g}_0} := \operatorname{Var}_{\boldsymbol{\theta}_0}(\mathbf{g}(\mathbf{x}))$ exist with $\operatorname{tr}(\mathbf{\Sigma}_{\mathbf{\Gamma}_0}) > 0$.*

Fix a confidence level $0 < \delta \leq 1$. We allow up to $n_c := \tau(\lfloor 17 \cdot \log(1/\delta) \rfloor + 1)$ samples being arbitrarily corrupted, with $0 \leq \tau < 1/2$. Let $K = K(\delta, \tau)$ and $c(\tau)$ be as in Theorem 3.2. Let

$$0 \leq \beta \leq \frac{1}{1 + (\|\mathbf{\Gamma}_0\|_2 / \sqrt{2 \operatorname{tr}(\mathbf{\Sigma}_{\mathbf{\Gamma}_0})}) \sqrt{n/K}}. \quad (14)$$

Finally, with constants and notation from Definition 4.2, if

$$n > \left(\frac{24 d_{\boldsymbol{\theta}_0} c_{\mathbf{\Gamma}_0} c(\tau)}{\alpha} \right)^2 \log\left(\frac{4}{(1-\tau)^2} \frac{1}{\delta} \right) \operatorname{tr}(\mathbf{\Sigma}_{\mathbf{\Gamma}_0}), \quad (15)$$

$$\lambda > \frac{6 c(\tau)(2 - \alpha)}{\alpha} \sqrt{\log\left(\frac{4}{(1-\tau)^2} \frac{1}{\delta} \right) \frac{1}{n}} \cdot \max\left(2 c_{\boldsymbol{\theta}_0} \sqrt{\operatorname{tr}(\mathbf{\Sigma}_{\mathbf{\Gamma}_0})}, \sqrt{\operatorname{tr}(\mathbf{\Sigma}_{g_0})} \right), \quad (16)$$

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

$$\|\hat{\boldsymbol{\theta}}(K, \beta, \lambda) - \boldsymbol{\theta}_0\|_\infty \leq \frac{c_{\mathbf{\Gamma}_0}}{2 - \alpha} \lambda. \quad (17)$$

Theorem 4.3 guarantees that with high probability the maximal difference between the estimated model parameters $\hat{\boldsymbol{\theta}}(K, \beta, \lambda)$ and the true parameter $\boldsymbol{\theta}_0$ is small, and that any non-zero interaction in the model induced by $\hat{\boldsymbol{\theta}}(K, \beta, \lambda)$ is also present in the true model. To illustrate the implications of the theorem, like in Section 3.3, consider $n_c = o(n)$, $\tau := \tau_0$ and $\delta := \exp(-(\lceil n_c/\tau_0 \rceil - 1)/17)$. Then, $K = o(n)$ and $\beta = o(1)$ as $n \to \infty$. Since $\log(4/(1 - \tau_0)^2 \cdot 1/\delta) = o(n)$, the requirement (15) is satisfied for large n, and the lower bound in (16) allows a choice $\lambda = o(1)$. By (17), $\hat{\boldsymbol{\theta}}(K, \beta, \lambda)$ converges to $\boldsymbol{\theta}_0$ in probability as $n \to \infty$.

Theorem 4.3 reads similarly to the theorems in (Yu et al., 2019, Sect. 6). However, since Theorem 4.3 does not assume an underlying Gaussian distribution, that is possibly truncated, the bounds on β, n and λ depend on $\mathbf{\Gamma}_0$ and $\mathbf{g}_0$ explicitly.

5 NUMERICAL EXPERIMENTS

5.1 Simulation study

We apply the estimator $\hat{\boldsymbol{\theta}}(K, \beta, \lambda)$ from Definition 4.1 to simulated data from square root graphical models, under scenarios that include rowwise corruption. The 'standard' regularized score matching estimator, corresponding to $K = 1$ block, serves as a baseline. Performance is judged by how well the zero structure of $\mathbf{\Theta}$

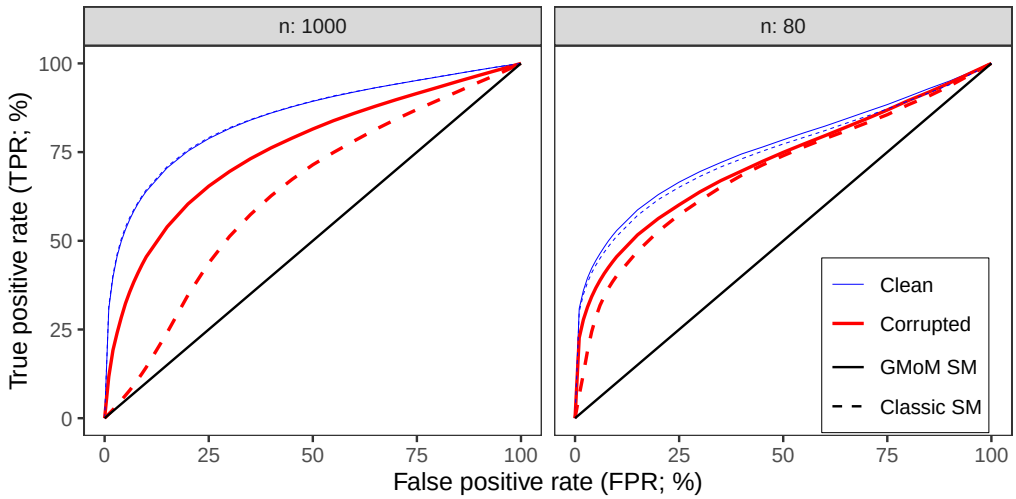

Figure 1: ROC curves for support recovery in the square root model. The pointwise uncertainty in TPR is at most ±0.75% based on 100 (left) and 500 simulations (right).

is learned, assessed through receiver operator characteristic (ROC) curves in Figure 1. The simulations run within a few hours on a personal laptop. Experiments on Gaussian data, one of the classic graphical models classes, lead to similar results and are contained in the appendix.

Data-generating model: As in Yu et al. (2019), we considered a $m = 100$ dimensional model with either $n = 1000$ or $n = 80$ samples, the latter acting as a high-dimensional scenario. Additional choices for n are considered in the appendix. The interaction matrix $\boldsymbol{\Theta}$ was determined by first selecting a graph on m vertices uniformly from the set of all graphs with κ edges (one of the variants of the Erdős–Rényi graph model), and then drawing the edge strength Θ_{ij} from $\pm\text{Unif}(0.5, 1)$. The ratio κ/n was kept constant and set to $1/2$ to have an average node degree of $m/10$ for $n = 1000$. Each ROC curve reports average ROCs (Fawcett, 2006) for 10 randomly chosen $\boldsymbol{\Theta}$. The location-like parameter $\boldsymbol{\eta}$ was randomly drawn from $\{0, 0.5, -0.5\}$.

Contamination details: For the simulations with contamination, 5% of data rows were replaced by independent Pareto draws. The Pareto scale parameter was set to the respective column mean and the shape parameter to 1, which ensures that most corrupted values are similar to the uncorrupted values and, due to the heavy tails of the Pareto, a small portion are strong outliers. Results under different contamination settings are reported in the appendix.

Hyperparameter tuning: The number of blocks was set to $K := 4 \cdot 0.05\, n = n/5$ as discussed in Section 3.4. In simulations with 5% contamination, $\hat{\boldsymbol{\theta}}(K, \beta, \lambda)$ is thus adapted to the actual corruption amount, while in the uncorrupted case it represents a conservative block size choice. The baseline from Yu et al. (2019) represents the opposite: it has $K = 1$ by definition, making it adapted to the uncontaminated simulations, but underestimating the corruption amount otherwise. The diagonal multiplier β was set to 0 for $n = 1000$, and for $n = 80$ to the upper bound in Theorem 4.3 with $\boldsymbol{\Gamma}_0$ being estimated from uncorrupted data (yielding $\beta \approx 0.01$). The upper bound was chosen since Yu et al. (2019) experimentally found this aided support recovery. The regularization parameter λ was varied to obtain a ROC curve. The weights are set to $\mathbf{h}(\mathbf{x}) := \mathbf{x}^{3/2}$, a choice found to be favorable in (Yu et al., 2019).

Interpretation: Figure 1 shows that the GMoM procedure is on par with the baseline in terms of support recovery on uncorrupted data and outperforms the baseline on contaminated data. A pointwise 95% bootstrap confidence band around the curves has a maximal width of ±0.75%, implying that this conclusion is statistically significant. For the high-dimensional experiment $n = 80$, the effect is less pronounced and the difference under contamination only significant for the lower end of the FPR spectrum. This is due to the small sample size and increased sparsity, making the system already very noisy even without corruption.

5.2 Data on precipitation across the Alps

Consider the task of learning how precipitation at different weather stations in central Europe is related. We model the dependence between monthly total precipitation at $m = 30$ stations using the European Climate Assessment & Dataset (ECA&D) (Tank et al., 2002) from www.ecad.eu. The records of the stations share

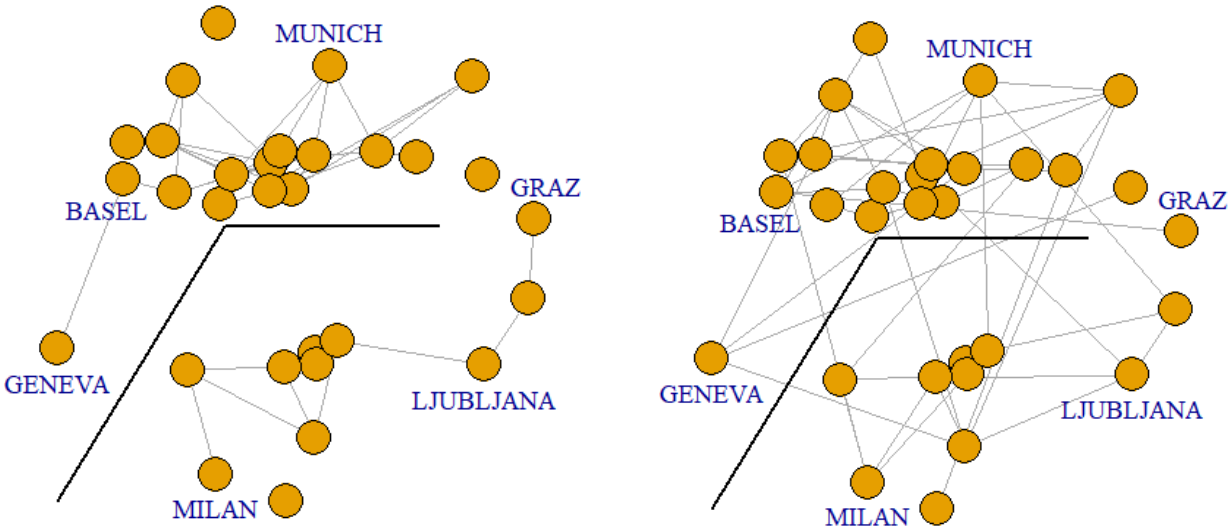

Figure 2: Precipitation dependencies (grey lines) learned from uncorrupted (left) and contaminated data (right). Black lines sketch the location of some of the highest mountains of the Alps.

a span of 87 years. Data from November, January and March was used to obtain similar precipitation distributions, which are roughly independent due to the one-month gap. This leads to $n = 3 \cdot 87 = 261$ samples. The data clearly is non-Gaussian, for example due to positivity. Instead, inspired by the marginal distributions, the square root graphical model is chosen. To learn the model, λ is tuned such that the graph contains 45 edges to get an average node degree of 3, which roughly equals the number of geographical neighbors of the average station. The diagonal multiplier β is not needed since $n \gg m$. The graph learned by the baseline score matching approach ($K = 1$) is shown on the left of Figure 2.

While there is no direct ground truth information on the precipitation dependence, the graph from the uncorrupted data has the notable feature that it only connects stations within the same geographical neighborhood. Moreover, no edge crosses the Alps mountains sketched by black lines in Figure 2. This is expected since high mountains act as a barrier for clouds. According to the latter physical consideration, edges crossing the Alps are false discoveries. The corresponding false discovery rate, termed Alps-FDR, is used to judge a learned precipitation network. No edges crossing the Alps yields the optimal 0% Alps-FDR.

Now, as our experiment, we alter the data through random contamination as in Section 5.1. A graph learned under 5% contamination using the non-robust the baseline ($K = 1$) is shown on the right of Figure 2. It is noticeably more noisy and connects stations further away. Its Alps-FDR is 19%, meaning that roughly every fifth connection is considered unrealistic. In a Monte Carlo simulation with 100 respective runs, different proportions ε of the sample were contaminated and the Alps-FDR of the baseline $K = 1$ compared with $K = 4\varepsilon n$. Additionally, a Gaussian graphical model (GGM), arguably the most studied graphical model, was fit to the data, knowing that the Gaussianity assumption was violated. Results are reported in Table 1 with 95% bootstrap confidence intervals. It is evident that the GMoM version with $K = 4\varepsilon n$ has the best Alps-FDR in every corruption scenario. For comparison, chosing a graph uniformly at random from all graphs with 45 edges has an Alps-FDR of roughly 42%.

Table 1: The Alps-FDR for different corruption proportions computed from 100 Monte Carlo runs.

Corr. (%)	GMoM SM	Classic SM	GGM
1	2.4 ± 0.5	4.2 ± 0.8	12.8 ± 1.6
5	7.0 ± 0.8	13.5 ± 1.0	26.9 ± 1.7
10	8.1 ± 0.8	17.8 ± 1.0	31.7 ± 1.3
20	10.5 ± 1.0	22.4 ± 1.1	38.7 ± 1.4

6 CONCLUSION AND FUTURE WORK

This paper introduces a robust score matching estimator that utilizes the geometric median of means to circumvent existence issues that result from more naive robustification approaches. Theoretical guarantees and empirical evidence demonstrate our estimator's ability to recover the dependence structure of a pairwise interaction model, even when a portion of the observations

is contaminated. In the presented numerical experiments on uncorrupted data, the dependence recovery was on par with that of the classical regularized score matching estimator from Yu et al. (2019).

An interesting topic for future work is to further examine the optimal choice of the number of blocks K. Evidently, there is a trade-off between bias and variance inherent to score matching; especially for asymmetric, heavy tailed distributions. Neither bias nor variance of the geometric median of means in this scenario seems to be well understood. How the trade-off is influenced by contamination, possibly also in different forms such as cellwise contamination, is another open problem. Additionally, the concentration guarantee in Theorem 4.3 could likely be improved if one were able to refine the interplay between $\|\cdot\|_2$ from the concentration of the geometric median and $\|\cdot\|_1$ from the ℓ_1-regularization.

Acknowledgements

We acknowledge the data providers in the ECA&D project (Tank et al., 2002). Data and metadata available at https://www.ecad.eu. This work is supported by the DAAD programme Konrad Zuse Schools of Excellence in Artificial Intelligence, sponsored by the Federal Ministry of Education and Research. Further, this work has been funded by the German Federal Ministry of Education and Research and the Bavarian State Ministry for Science and the Arts. The authors of this work take full responsibility for its content. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 883818).

References

F. Alqallaf, S. Van Aelst, V. J. Yohai, and R. H. Zamar. Propagation of outliers in multivariate data. *Ann. Statist.*, 37(1):311–331, 2009.

A. Barp, F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey. Minimum stein discrepancy estimators. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

S. Bhatt, G. Fang, P. Li, and G. Samorodnitsky. Minimax m-estimation under adversarial contamination. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1906–1924. PMLR, 17–23 Jul 2022.

J. Blanchet, J. Li, S. Lin, and X. Zhang. Distributionally robust optimization and robust statistics, 2024. URL https://arxiv.org/abs/2401.14655.

B. M. Brown. Statistical uses of the spatial median. *J. Roy. Statist. Soc. Ser. B*, 45(1):25–30, 1983.

J.-h. Chu, S. T. Weiss, V. J. Carey, and B. A. Raby. A graphical model approach for inferring large-scale networks integrating gene expression and genetic polymorphism. *BMC Systems Biology*, 3:1–9, 2009.

L. Devroye, M. Lerasle, G. Lugosi, and R. I. Oliveira. Sub-Gaussian mean estimators. *Ann. Statist.*, 44(6): 2695–2725, 2016.

I. Diakonikolas and D. M. Kane. *Algorithmic High-Dimensional Robust Statistics*. Cambridge University Press, 2023.

D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on halfspace depth and projected outlyingness. *Ann. Statist.*, 20(4):1803–1827, 1992.

M. L. Eaton and M. D. Perlman. The non-singularity of generalized sample covariance matrices. *Ann. Statist.*, 1:710–717, 1973.

T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007a.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 12 2007b.

A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6: 695–709, 2005.

A. Hyvärinen. Some extensions of score matching. *Comput. Statist. Data Anal.*, 51(5):2499–2512, 2007.

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*, volume 48 of *Proceedings of Machine Learning Research*, pages 2445–2453, New York, New York, USA, 2016. PMLR.

D. Kuhn, S. Shafiee, and W. Wiesemann. Distributionally robust optimization, 2024. URL https://arxiv.org/abs/2411.02549.

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*, volume 139 of *Proceedings of Machine Learning Research*, pages 5937–5947. PMLR, 18–24 Jul 2021.

S. L. Lauritzen. *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996.

L. Lin, M. Drton, and A. Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10(1):806–854, 2016.

X. Liu, K. Mosler, and P. Mozharovskyi. Fast computation of Tukey trimmed regions and median in dimension $p > 2$. *J. Comput. Graph. Statist.*, 28(3): 682–697, 2019.

P.-L. Loh and X. L. Tan. High-dimensional robust precision matrix estimation: Cellwise corruption under ϵ-contamination. *Electronic Journal of Statistics*, 12 (1):1429 – 1467, 2018.

H. P. Lopuhaä and P. J. Rousseeuw. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *Ann. Statist.*, 19(1): 229–248, 1991.

G. Lugosi and S. Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5): 1145–1190, Aug. 2019.

M. Maathuis, M. Drton, S. Lauritzen, and M. Wainwright, editors. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019.

R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, second edition, 2019.

S. Minsker. Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21(4):2308–2335, 2015.

S. Minsker. Distributed statistical estimation and rates of convergence in normal approximation. *Electron. J. Stat.*, 13(2):5213–5252, 2019.

A. Niinimaa, H. Oja, and M. Tableman. The finite-sample breakdown point of the Oja bivariate median and of the corresponding half-samples version. *Statist. Probab. Lett.*, 10(4):325–328, 1990.

J. H. Oh and J. O. Deasy. Inference of radio-responsive gene regulatory networks using the graphical lasso algorithm. *BMC Bioinformatics*, 15:1–8, 2014.

T. Ronkainen, H. Oja, and P. Orponen. Computation of the multivariate Oja median. In *Developments in robust statistics (Vorau, 2001)*, pages 344–359. Physica, Heidelberg, 2003.

A. Roy and D. B. Dunson. Nonparametric graphical model for counts. *Journal of Machine Learning Research*, 21(229):1–21, 2020.

R. J. Serfling. Multivariate symmetry and asymmetry. *Encyclopedia of Statistical Sciences*, 8:5338–5345, 2006.

C. G. Small. A survey of multidimensional medians. *International Statistical Review / Revue Internationale de Statistique*, 58(3):263–277, 1990.

S. Sun, M. Kolar, and J. Xu. Learning structured densities via infinite dimensional exponential families. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

K. Tank, A.M.G., and Coauthors. Daily dataset of 20th-century surface air temperature and precipitation series for the european climate assessment. *International Journal of Climatology*, 22:1441–1453, 2002.

Y. Vardi and C.-H. Zhang. A modified Weiszfeld algorithm for the Fermat-Weber location problem. *Math. Program.*, 90(3):559–566, 2001.

M. Yu, M. Kolar, and V. Gupta. Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

S. Yu, M. Drton, and A. Shojaie. Generalized score matching for non-negative data. *J. Mach. Learn. Res.*, 20(1):2779–2848, 2019.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes (https://gitlab.lrz.de/robust_score_matching/reproducibility)

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes (https://gitlab.lrz.de/robust_score_matching/reproducibility)
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Yes

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

   (a) Citations of the creator If your work uses existing assets. Yes

   (b) The license information of the assets, if applicable. Yes

   (c) New assets either in the supplemental material or as a URL, if applicable. Not Applicable

   (d) Information about consent from data providers/curators. Not Applicable

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. Not Applicable

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

7 APPENDIX

7.1 Regularity conditions to ensure the positive definiteness of $\mathbf{\Gamma}$

For the almost sure positive definiteness of the score matching design matrix $\mathbf{\Gamma}(x) \in \mathbb{R}^{r\times r}$, we require two assumptions on the sufficient statistic $\mathbf{t}\colon \mathbb{R}^m \to \mathbb{R}^r$:

(A1) Without loss of generality, assume that t_i is not constant for any $i \in \{1,\dots,r\}$

(A2) With $R_j := \{i \in \{1,\dots,r\} \mid \partial_j t_i \neq 0\}$ and $d_j := |R_j|$ for $j \in \{1,\dots,m\}$, define the function $\mathbf{v}^{(j)}\colon \mathbb{R}^m \to \mathbb{R}^{d_j}, \mathbf{v}^{(j)}(\mathbf{x}) := \sqrt{h_j(\mathbf{x})}\cdot \partial_j \mathbf{t}|^{R_j}(\mathbf{x})$. We assume that for any proper linear subspace L of $\mathbb{R}^{d_j}$, the pre-image $\left(\mathbf{v}^{(j)}\right)^{-1}(L)$ is a Lebesgue null set in $\mathbb{R}^m$.

Example 7.1. *In the square root graphical model with $\boldsymbol{\eta}$ known, it holds that $d_j = m$ and up to permutations of the components, $\mathbf{v}^{(j)}(\mathbf{x}) = -\sqrt{h_j(x_j)/x_j}\cdot\sqrt{\mathbf{x}}$. If $h > 0$ is invertible and sufficiently smooth, this is a diffeomorphism (its inverse equals $h^{-1}(y_j^2)(\mathbf{e}^{(j)} + (\mathbf{1}-\mathbf{e}^{(j)})(\mathbf{y}/y_j)^2)$ with the j-th Euclidean basis vector $\mathbf{e}^{(j)}$ and the all-one vector $\mathbf{1}$) and thus null sets, in particular proper linear subspaces, are mapped to null sets by the change of variables theorem for Lebesgue's measure.*

Lemma 7.2. *Let $\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)} \in \mathbb{R}^m$ be i.i.d. according to an exponential family satisfying (A1) and (A2). Further, let (c_{ij}) be variables on the same probability space that are positive almost surely. Assume $n \geq \max_{j=1,\dots,m} d_j$. Then,*

$$\mathbf{M} := \sum_{j=1}^{m}\sum_{i=1}^{n} c_{ij} h_j(\mathbf{x}^{(i)})\partial_j\mathbf{t}(\mathbf{x}^{(i)})\partial_j\mathbf{t}(\mathbf{x}^{(i)})^\mathsf{T}$$

is positive definite almost surely.

Proof. Let $j \in \{1,\dots,m\}$. We show that $(\mathbf{v}^{(j)}(\mathbf{x}^{(i)}))_{i=1,\dots,d_j}$ are independent almost surely. Note that this collection of vectors requires $n \geq d_j$. We show that the probability of linear dependence is zero:

$$\mathrm{P}\Big(\mathbf{v}^{(j)}(\mathbf{x}^{(i)}) \in \operatorname{span}(\mathbf{v}^{(j)}(\mathbf{x}^{(2)}),\dots,\mathbf{v}^{(j)}(\mathbf{x}^{(d_j)}))\Big) = \mathrm{E}\Big[\mathrm{P}\Big(\mathbf{x}^{(i)} \in (\mathbf{v}^{(j)})^{-1}\Big(\operatorname{span}(\mathbf{v}^{(j)}(\mathbf{x}^{(2)}),\dots,\mathbf{v}^{(j)}(\mathbf{x}^{(d_j)}))\Big) \,\Big|\, (\mathbf{v}^{(j)}(\mathbf{x}^{(i)}))_{i=2,\dots,d_j}\Big)\Big] \overset{(A2)}{=} \mathrm{E}[0] = 0.$$

Define the $d_j \times d_j$ matrices $\mathbf{M}^{(j)}(d) := \sum_{i=1}^{d} c_{ij}\mathbf{v}^{(j)}(\mathbf{x}^{(i)})\mathbf{v}^{(j)}(\mathbf{x}^{(i)})^\mathsf{T}$. The independence result implies that $\mathbf{M}^{(j)}(d_j)$ has full rank almost surely. Otherwise, there would be $\mathbf{v} \neq 0$ in its kernel by the rank theorem. This would imply

$$\mathbf{M}^{(j)}(d_j)\mathbf{v} = \sum_{i=1}^{d_j} c_{ij}\mathbf{v}^{(j)}(\mathbf{x}^{(i)})\mathbf{v}^{(j)}(\mathbf{x}^{(i)})^\mathsf{T}\mathbf{v} = \sum_{i=1}^{d_j}\left(c_{ij}\mathbf{v}^{(j)}(\mathbf{x}^{(i)})^\mathsf{T}\mathbf{v}\right)\mathbf{v}^{(j)}(\mathbf{x}^{(i)}) = \mathbf{0},$$

a contradiction to $(\mathbf{v}^{(j)}(\mathbf{x}^{(i)}))_{i=1,\dots,d_j}$ being independent almost surely. Since $\mathbf{M}^{(j)}(d_j)$ is positive semidefinite due to the structure of its summands (recall $c_{ij}, h_j > 0$), it follows that $\mathbf{M}^{(j)}(d_j)$ is positive definite. Also, $\mathbf{M}^{(j)}(n)$ is positive definite since only more positive semidefinite terms are added.

To show the statement of this lemma, first note that $\mathbf{M}$ is positive semidefinite since $\sum_{i=1}^{n} c_{ij}h_j(\mathbf{x}^{(i)})\partial_j\mathbf{t}(\mathbf{x}^{(i)})\partial_j\mathbf{t}(\mathbf{x}^{(i)})^\mathsf{T} \in \mathbb{R}^{r\times r}$ are positive semidefinite. Assume $\mathbf{M}$ had a nontrivial vector $\mathbf{v} \in \mathbb{R}^r$ in its kernel. By positive semidefiniteness of the summands, $\mathbf{M}\mathbf{v} = \mathbf{0}$ implies $\sum_{i=1}^{n} c_{ij}h_j(\mathbf{x}^{(i)})\partial_j\mathbf{t}(\mathbf{x}^{(i)})\partial_j\mathbf{t}(\mathbf{x}^{(i)})^\mathsf{T}\mathbf{v} = 0$ for all j. Since $\partial_j t_i = 0$ for all $i \in \{1,\dots,r\}\setminus R_j$ by definition of R_j, this is equivalent to $\mathbf{M}^{(j)}(n)\cdot\mathbf{v}_{R_j} = \mathbf{0}$. By the previous result on $\mathbf{M}^{(j)}(n)$, it follows that $\mathbf{v}_{R_j} = \mathbf{0}$. Assumption (A1) guarantees that $\{1,\dots,r\} = \bigcup_{j=1,\dots,m} R_j$, which implies $\mathbf{v} = \mathbf{0}$, a contradiction. □

A direct consequence of Lemma 7.2 is that $\overline{\mathbf{\Gamma}}(\mathbf{X})$ is positive definite almost surely (choose $c_i := 1/n$).

To see that the same holds for the GMoM version $\hat{\mathbf{\Gamma}}_K(\mathbf{X})$, consider the following equation from the paper, which holds when the geometric median does not equal one of its arguments:

$$\mathbf{m} := \mathrm{GMed}\Big(\hat{\boldsymbol{\mu}}^{(1)}, \dots, \hat{\boldsymbol{\mu}}^{(K)} \Big) = \frac{1}{\sum_{i=1}^K 1/\|\mathbf{m} - \hat{\boldsymbol{\mu}}^{(i)}\|_2} \sum_{i=1}^K \frac{\hat{\boldsymbol{\mu}}^{(i)}}{\|\mathbf{m} - \hat{\boldsymbol{\mu}}^{(i)}\|_2}.$$

To apply this to $\hat{\mathbf{\Gamma}}_K(\mathbf{X})$, define

$$\hat{\boldsymbol{\mu}}^{(k)} := \frac{1}{n/K} \sum_{i=(k-1)K+1}^{kK} \sum_{j=1}^{m} h_j(\mathbf{x}^{(i)}) \partial_j \mathbf{t}(\mathbf{x}^{(i)}) \partial_j \mathbf{t}(\mathbf{x}^{(i)})^\top,$$

such that $\hat{\mathbf{\Gamma}}_K(\mathbf{X}) = \mathrm{GMed}\Big(\hat{\boldsymbol{\mu}}^{(1)}, \dots, \hat{\boldsymbol{\mu}}^{(K)} \Big)$. Lemma 7.2 guarantees positive definiteness of $\hat{\mathbf{\Gamma}}_K(\mathbf{X})$ with

$$c_i := \frac{K}{n} \left(\sum_{k=1}^K \frac{\|\hat{\mathbf{\Gamma}}_K(\mathbf{X}) - \hat{\boldsymbol{\mu}}^{(k_i)}\|_2}{\|\hat{\mathbf{\Gamma}}_K(\mathbf{X}) - \hat{\boldsymbol{\mu}}^{(k)}\|_2} \right)^{-1},$$

where k_i is the block index that $i \in \{1, \dots, n\}$ belongs to.

7.2 Proof of Theorem 3.1

Theorem 3.1. *Let $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ be independent samples from a p-dimensional distribution with mean $\boldsymbol{\mu}$ and variance $\mathbf{\Sigma}$. Fix a confidence level of $0 < \delta \leq 1$. We allow for up to $n_c := (\lfloor 17 \cdot \log(1/\delta) \rfloor + 1)\tau$ samples to be arbitrarily corrupted, where $0 \leq \tau < 1/2$. Then, there exist functions $k(\tau) = \mathcal{O}(1/(\frac{1}{2} - \tau)^2)$ and $c(\tau) = \mathcal{O}(1/(\frac{1}{2} - \tau)^{2.5})$ as $\tau \to \frac{1}{2}$ such that when the number of blocks K defined as $K = K(\delta, \tau) := \lfloor k(\tau) \cdot \log(1/\delta) \rfloor + 1$ satisfies $K \leq n/2$, it holds that*

$$\mathrm{P}\left(\| \mathrm{GMoM}_K\Big(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)} \Big) - \boldsymbol{\mu} \|_2 > c(\tau) \sqrt{\log\left(\frac{4}{(1-\tau)^2} \frac{1}{\delta} \right) \frac{\mathrm{tr}(\mathbf{\Sigma})}{n}} \right) \leq \delta. \tag{18}$$

This section proves the above theorem. Let

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

We base the proof on the following robustness result on the geometric median of independent estimators from Minsker (2015, Remark 3.1.a). Set $C_\alpha := (1-\alpha)/\sqrt{1-2\alpha}$ for $0 < \alpha < 1/2$.

Lemma 7.3 (Minsker, 2015)**.** *Let $\boldsymbol{\mu} \in \mathbb{R}^p$, and let $\hat{\boldsymbol{\mu}}_1, \dots, \hat{\boldsymbol{\mu}}_k \in \mathbb{R}^p$ be a collection of independent estimators of $\boldsymbol{\mu}$. Let the hyperparameters $0 < \alpha < 1/2$, $0 < p < \alpha$ and $\varepsilon > 0$ be such that*

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

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

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

The function $k(\tau)$ from the theorem statement can be set to

$$k(\tau) := \frac{1}{(1-\tau)\psi\left(\frac{(1/2-\tau)^2}{1-\tau}, \frac{1}{2}\left(\frac{1}{2} - \tau\right)^2 \right)}$$

such that K is given by

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

The second function $c(\tau)$ from the theorem statement can be set to

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

It follows that $k(\tau) = \mathcal{O}(1/(\frac{1}{2} - \tau)^2)$ and $c(\tau) = \mathcal{O}(1/(\frac{1}{2} - \tau)^{2.5})$ since $\log(1-x) = \mathcal{O}(x)$ as $x \to 0$.

We can now prove the theorem for $k(\tau)$ and $c(\tau)$ given above.

Proof. To simplify notation, let

$$\hat{\boldsymbol{\mu}} := \mathrm{GMoM}_K\left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(K \cdot \lfloor n/K \rfloor)}\right) = \mathrm{GMed}(\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_K).$$

For theoretical simplicity, we prove the theorem for $\hat{\boldsymbol{\mu}}$ with K blocks of equal block size $\lfloor n/K \rfloor$

The main step of this proof is applying Lemma 7.3 to the block means $\hat{\boldsymbol{\mu}}_1, \ldots, \hat{\boldsymbol{\mu}}_K$. We start by fixing α, p and ε in the Lemma. Consider the following choices that depend on the corruption parameter τ:

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

It remains to verify that these choices can satisfy the conditions in Lemma 7.3. 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 (since $17 \leq 1/\psi(1/4, 1/8)$). 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 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

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

$$\frac{1}{\lfloor n/K \rfloor^2} \sum_{i=1}^{\lfloor n/K \rfloor} \mathrm{E}\Big[\, (\mathbf{x}^{(i)} - \boldsymbol{\mu})^T (\mathbf{x}^{(i)} - \boldsymbol{\mu}) \,\Big] = \frac{\mathrm{E}\left[\, \|\mathbf{x} - \boldsymbol{\mu}\|_2^2 \,\right]}{\lfloor n/K \rfloor} \leq \frac{2K}{n} \operatorname{tr}(\boldsymbol{\Sigma}).$$

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

For the second condition, check that

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

By Lemma 7.3, we have established

$$\mathrm{P}\big(\, \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 > C_{\alpha(\tau)} \varepsilon(\tau) \,\big) \overset{7.3}{\leq} e^{-K(1-\tau)\psi\left(\frac{\alpha(\tau)-\tau}{1-\tau}, p(\tau)\right)}. \tag{19}$$

We start by simplifying the exponent in the right hand side of (19) 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 (19) by

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

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

$$C_\alpha\varepsilon = C_\alpha\sqrt{\frac{2K\,\mathrm{tr}(\boldsymbol{\Sigma})}{np}} = \frac{C_\alpha\sqrt{2}}{\sqrt{p}\sqrt{(1-\tau)\psi\left(\frac{\alpha-\tau}{1-\tau},p\right)}}\cdot\sqrt{K\cdot(1-\tau)\psi\left(\frac{\alpha-\tau}{1-\tau},p\right)}\cdot\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{n}} \leq$$
$$c(\tau)\sqrt{\left(\frac{\log(1/\delta)}{(1-\tau)\psi\left(\frac{2p}{1-\tau},p\right)}+1\right)\cdot(1-\tau)\psi\left(\frac{2p}{1-\tau},p\right)}\cdot\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{n}} =$$
$$c(\tau)\sqrt{\log(1/\delta)+(1-\tau)\psi\left(\frac{2p}{1-\tau},p\right)}\cdot\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{n}} \overset{\text{First term of } \psi \text{ negative}}{\leq}$$
$$c(\tau)\sqrt{\log(1/\delta)+(1-\tau)\frac{2p}{1-\tau}\log\left(\frac{2p}{(1-\tau)p}\right)}\cdot\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{n}} \overset{p\leq 1}{\leq}$$
$$c(\tau)\sqrt{\log(1/\delta)+2\log\left(\frac{2}{1-\tau}\right)}\cdot\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{n}} = c(\tau)\sqrt{\log\left(\frac{4}{(1-\tau)^2\cdot\delta}\right)}\cdot\sqrt{\frac{\mathrm{tr}(\boldsymbol{\Sigma})}{n}}.$$

□

7.3 Proof of Corollary 3.2

Corollary 3.2. *Let $\mathbf{x}^{(1)},\dots,\mathbf{x}^{(n)} \in \mathbb{R}^p$ be independent samples from an exponential family with parameter $\boldsymbol{\theta}_0 \in \mathbb{R}^r$. Assume that the following regularity conditions are met:*

1) *The conditions from Proposition 2 in Yu et al. (2019) hold, i.e. the exponential family satisfies the basic requirements for score matching and, if its support is restricted, the dampening function $\mathbf{h}$ satisfies regularity assumptions.*

2) *The exponential family satisfies (A1) and (A2) from Section 7.1 of this supplement.*

3) *$\boldsymbol{\Gamma}(\mathbf{x})$ and $\mathbf{g}(\mathbf{x})$ have finite second moments when $\mathbf{x}$ comes from the distribution indexed by $\boldsymbol{\theta}_0$*

Allow for up to n_c samples to be arbitrarily corrupted, where $n_c = o(n)$. Then, there exists a sequence $K = K(n_c)$ such that the robust score matching estimator $\hat{\boldsymbol{\theta}}(K(n_c))$ converges against $\boldsymbol{\theta}_0$ in probability when $n \to \infty$.

Proof. Assume without loss of generality that $n_c \to \infty$ as $n \to \infty$. Should the true number of corrupt samples be bounded by $M < \infty$, it certainly holds true that at most $n_c := M + \log(n)$ samples have been corrupted. Hence, the theorem assumptions are also satisfied for this larger n_c, which diverges as $n \to \infty$.

Fix some $0 < \tau_0 < 1/2$ independent of n. Setting $\delta(n_c) := \exp(-(\lceil n_c/\tau_0 \rceil - 1)/17)$, it holds that

$$(\lfloor 17 \cdot \log(1/\delta(n_c)) \rfloor + 1)\tau_0 = (\lfloor \lceil n_c/\tau_0 \rceil - 1 \rfloor + 1)\tau_0 = \lceil n_c/\tau_0 \rceil \tau_0 \geq n_c,$$

so in other words $(\delta(n_c), n_c, \tau_0)$ satisfy the assumption of Theorem 3.1. We set $K(n_c) := \lfloor k(\tau_0) \log(1/\delta(n_c)) \rfloor + 1$ in line with Theorem 3.1. By our choice of $\delta(n_c)$, we have $K(n_c) = \lfloor \frac{k(\tau_0)}{17}(\lceil n_c/\tau_0 \rceil - 1) \rfloor + 1 = o(n)$ by the assumption $n_c = o(n)$. Consequently, $K(n_c) \leq n/2$ for n large enough, which is the last assumption of Theorem 3.1 to check.

Denoting by $\mathbf{X}$ the $n \times p$ matrix having $\mathbf{x}^{(i)}$ as the i-th row and

$$B(n, \delta, \mathbf{\Sigma}) := c(\tau_0) \sqrt{\log\left(\frac{4}{(1-\tau_0)^2}\frac{1}{\delta}\right)} \frac{\operatorname{tr}(\mathbf{\Sigma})}{n},$$

Theorem 3.1 guarantees that

$$\mathrm{P}\Big(\|\hat{\mathbf{\Gamma}}_{K(n_c)}(\mathbf{X}) - \mathbf{\Gamma}_0\| > B(n, \delta(n_c), \Sigma_{\mathbf{\Gamma}}) \Big) \leq \delta(n_c), \quad \mathrm{P}\big(\|\hat{\mathbf{g}}_{K(n_c)}(\mathbf{X}) - \mathbf{g}_0\| > B(n, \delta(n_c), \Sigma_{\mathbf{g}}) \big) \leq \delta(n_c), \quad (20)$$

where $\mathbf{\Sigma}_{\mathbf{\Gamma}}$ denotes the variance of $\mathbf{\Gamma}(\mathbf{x})$ and $\mathbf{\Sigma}_{\mathbf{g}}$ that of $\mathbf{g}(\mathbf{x})$ when $\mathbf{x}$ is distributed according to $\boldsymbol{\theta}_0$.

Since $\log(1/\delta(n_c)) = (\lceil n_c/\tau_0 \rceil - 1)/17 = o(n)$, we have that $B(n, \delta(n_c), \Sigma_{\mathbf{\Gamma}})$ and $B(n, \delta(n_c), \Sigma_{\mathbf{g}})$ converge to zero as $n \to \infty$. Also, since we assumed without loss of generality that $n_c \to \infty$ as $n \to \infty$, we have that $\delta(n_c) \to 0$ when $n \to \infty$. These observations together with (20) imply that

$$\hat{\mathbf{\Gamma}}_{K(n_c)}(\mathbf{X}) \xrightarrow{P} \mathbf{\Gamma}_0, \qquad \hat{\mathbf{g}}_{K(n_c)}(\mathbf{X}) \xrightarrow{P} \mathbf{g}_0 \quad (21)$$

in probability as $n \to \infty$. As matrix inversion and multiplication are continuous, we have that

$$\hat{\boldsymbol{\theta}}(K(n_c)) := \operatorname*{argmin}_{\boldsymbol{\theta} \in \Omega} \tfrac{1}{2}\boldsymbol{\theta}^\top \hat{\mathbf{\Gamma}}_{K(n_c)}(\mathbf{X})\boldsymbol{\theta} - \boldsymbol{\theta}^\top \hat{\mathbf{g}}_{K(n_c)}(\mathbf{X}) \overset{(*1)}{=} \hat{\mathbf{\Gamma}}_{K(n_c)}(\mathbf{X})^{-1}\hat{\mathbf{g}}_{K(n_c)}(\mathbf{X}) \xrightarrow{P} \mathbf{\Gamma}_0^{-1}\mathbf{g}_0 \overset{(*2)}{=} \boldsymbol{\theta}_0,$$

where the minimizer in equation $(*1)$ exists almost surely and is given by the matrix inversion formula because of assumption 2) in the corollary statement above, and equation $(*2)$ holds because of assumption 1) in the corollary statement. □

7.4 Choice of K

To understand what choice for the number of blocks K in the GMoM leads to the best mean squared error (MSE) under contamination, two simulation studies are conducted. In the first, we estimate a Gaussian mean vector and a covariance matrix from contaminated data. Conceptually, this simulation corresponds to estimating $\mathbf{\Gamma}$ and $\boldsymbol{g}$ individually. Figure 3 displays the results. In view of the simulation results, we propose $K := 4\varepsilon n$ as a heuristic.

In the second simulation, we use the robust score matching estimator $\hat{\boldsymbol{\theta}}(K)$ from section 3.2 of the main paper to estimate the parameters of a square root graphical model from contaminated data. This simulation sheds some light on how to tune the number of blocks under contamination if one cares about the downstream accuracy of a score matching estimator that incorporates $\mathbf{\Gamma}$ and $\boldsymbol{g}$ estimated through a GMoM procedure. Figure 4 validates the heuristic $K := 4\varepsilon n$ proposed earlier.

7.4.1 Gaussian mean and covariance estimation

Two estimation problems are considered, involving a 10-dimensional Gaussian distribution with mean $\boldsymbol{\mu} = 0$ and randomly fixed covariance matrix $\mathbf{\Sigma}$. The first problem is finding the population mean $\boldsymbol{\mu}$, which is a classic problem of general interest. From a sample $\mathbf{x}^{(1)}, \ldots \mathbf{x}^{(n)}$ with $n = 100$, the mean is estimated simply by

$$\hat{\boldsymbol{\mu}} := \mathrm{GMoM}_K\Big(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\Big).$$

The second problem is estimating the Gaussian covariance matrix $\mathbf{\Sigma}$, a problem that is structurally similar to estimating $\mathbf{\Gamma}_0$ in score matching, especially for pairwise interaction models. From a sample $\mathbf{x}^{(1)}, \ldots \mathbf{x}^{(100)}$ from $N(\mathbf{0}, \mathbf{\Sigma})$, the covariance is estimated by

$$\hat{\mathbf{\Sigma}} := \mathrm{GMoM}_K\Big(\mathbf{x}^{(1)} \cdot \mathbf{x}^{(1)^\top}, \ldots, \mathbf{x}^{(n)} \cdot \mathbf{x}^{(n)^\top}\Big).$$

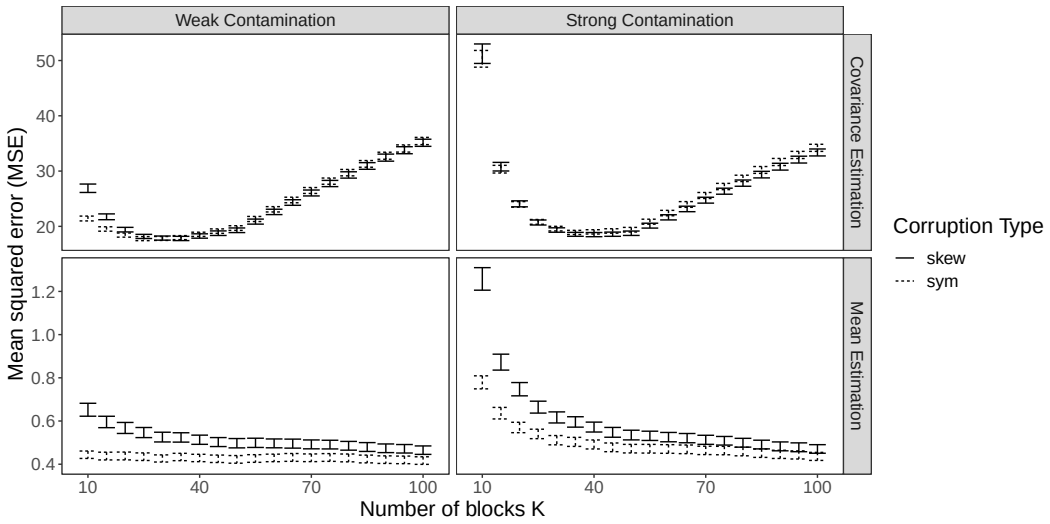

Figure 3: MSE versus number of blocks K for Gaussian covariance (top) and mean (bottom) estimation. 5% of observations were corrupted at varying intensity (weak (left) versus strong (right)) and using different types of corrupting distributions (skewed versus symmetric).

Both problems together cover a range of distributional properties. When estimating a Gaussian mean, the underlying distribution is symmetric and has light tails. Conversely, when estimating the Gaussian covariance, the underlying Wishart distribution is not symmetric and has heavier tails than the Gaussian.

Four contamination scenarios were considered, combining two levels of corruption intensity with two types of corrupting distributions. Each scenario corrupted 5 observations $\mathbf{x}^{(i)}$ of the sample $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(100)}$ with independent draws from a corrupting distribution. On the corrupted sample, $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ were computed. The corruption intensity was varied by setting α to 2 or 10 in the following distributions. The first corrupting distribution was Gaussian with mean zero and covariance $\alpha \cdot \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_{10}^2)$, where $\hat{\sigma}_i$ was estimated from $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(100)}$. This is labeled as *sym* in Figure 3. The second corrupting distribution was a Pareto with independent components $(P_1, \ldots, P_{10})$. Each Pareto P_i had its scale parameter chosen such that its 3/4-th quantile equalled α times the 3/4-th quantile of $\mathbf{x}_i^{(1)}, \ldots, \mathbf{x}_i^{(100)}$, and the distribution was shifted such that the lower Pareto cutoff agreed with the population mean of 0. This is labeled as *skew* in Figure 3.

Each simulation output is comprised of the empirical error $\|\hat{\boldsymbol{\mu}} - \mathbf{0}\|_2^2$ and $\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\|_2^2$ for the two estimation problems and four contamination scenarios, respectively, where the number of blocks K was ranged from the breakdown point of $K = 10$ to the geometric median $K = 100$. Aggregates over 1000 simulations are reported in Figure 3.

Interpretation: Increased corruption strength degrades the MSE for $K \lesssim 40$, but has little effect on MSE for large K. It seems that once a comfortable distance from the breakdown point is reached, the corruption is irrelevant. The MSE curves share similar shapes per corruption type, showing that the GMoM reacts similarly to different corrupting distributions. The number of blocks K resulting in optimal MSE is $K = n = 100$ for mean estimation and $K \approx 30$ for covariance estimation. The different optimum can be explained by the fact that $\hat{\boldsymbol{\Sigma}}$ with $K = n$ is biased for $\boldsymbol{\Sigma}$, which degrades the MSE for large K in covariance estimation, while $\hat{\boldsymbol{\mu}}$ is unbiased.

Take away: A choice for K optimizing the MSE of the GmoM when a proportion ε of samples is contaminated should get some distance to the breakdown point of $2\varepsilon n$, however not be to large in order to avoid the bias witnessed in covariance estimation. We propose $K := 4\varepsilon n$.

7.4.2 Estimating a square root graphical model with score matching

Data generation: As an example of an exponential family of interest, we consider the square root graphical model introduced in the main paper. We consider a $m = 5$ dimensional model with an interaction matrix $\boldsymbol{\Theta}$ and coefficient vector $\boldsymbol{\eta}$ being randomly determined. From this model, $n = 1000$ samples were created. Of these, 5% and 10% respectively were contaminated by draws from a Pareto distribution, with parameters chosen such that

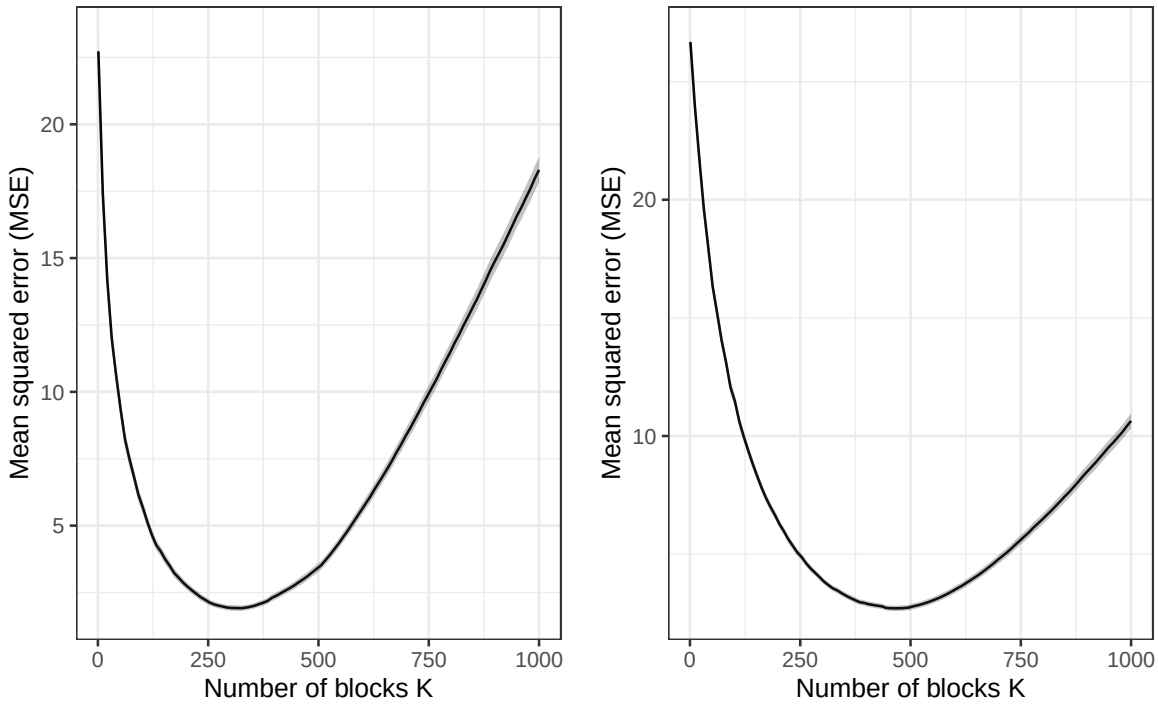

Figure 4: MSE versus number of blocks K in the robust score matching procedure $\hat{\boldsymbol{\theta}}(K)$ for estimating the parameters of a square root graphical model under 5% contamination (left) and 10% contamination (right).

most contaminated data points remained close to the mean of the uncorrupted data, while some became serious outliers.

Estimation: Estimates $\hat{\boldsymbol{\Theta}}$ and $\hat{\boldsymbol{\eta}}$ were obtained using the robust score matching estimator $\hat{\boldsymbol{\theta}}(K)$ from section 3.2 of the main paper, where the number of blocks K was varied on a grid from 1 to n, corresponding to the mean and geometric median on the extremes of the spectrum.

Simulation target: For each contaminated data set and choice of K, the squared error $\text{SE} := \|\hat{\boldsymbol{\Theta}}-\boldsymbol{\Theta}\|_2^2+\|\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}\|_2^2$ was computed. Averages and uncertainty estimates over $N = 100$ independent Monte Carlo repetitions are presented in Figure 4.

Take away: In Figure 4, the number of blocks K minimizing the MSE increases as the number of contaminated samples increases (left (5% contamination): $K_{\text{opt}} \approx 320$, right (10% contamination): $K_{\text{opt}} \approx 460$). This agrees with the intuition that a higher number of blocks makes the GMoM tolerate a higher number of outliers. The proposed choice $K := 4\varepsilon n$ corresponds to $K = 200$ and $K = 400$ respectively. Both fall in area of low MSE under corruption respectively and are thus suitable choices of the block-size. Still, they don't quite optimize the MSE, highlighting the need for further research into how to optimally tune the geometric median of means under contamination.

Also note that compared to the simulations in Figure 3, the MSE curves of the score matching estimator resemble that of Gaussian covariance estimation more closely than that of Gaussian mean estimation. This is not surprising, given that the distributions of $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ need not be centrally symmetric.

7.5 Proof of Theorem 4.3

We begin with a lemma that extends Theorem 3.1 by allowing for a diagonal multiplier. Set $\mathbf{b} := \beta \cdot \text{vec}\,(\mathbf{I}_m)$ with the vectorized $m \times m$ identity matrix $\mathbf{I}_m$ to obtain the diagonal multiplier β as it is used in the paper.

Lemma 7.4. *Let $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)} \in \mathbb{R}^p$ be independent samples from a p-dimensional distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Fix a confidence level $\delta \in (0, 1]$. We allow for up to $\tau(\lfloor 17 \log(1/\delta) \rfloor + 1)$ 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 3.1. Further, let $c(\tau)$ as in Theorem 3.1.

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

$$\|\mathbf{b}\|_\infty \leq \frac{1}{1 + (\|\boldsymbol{\mu}\|_2 / \sqrt{2\operatorname{tr}(\mathbf{\Sigma})})\sqrt{n/K}}.$$

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

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

where $\circ$ *denotes elementwise multiplication.*

Proof. To simplify notation, let

$$\hat{\boldsymbol{\mu}} := \mathrm{GMoM}_K\left(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\right), \qquad t := c(\tau)\sqrt{\log\left(\frac{4}{(1-\tau)^2}\frac{1}{\delta}\right)\frac{\operatorname{tr}(\mathbf{\Sigma})}{n}}.$$

We show the implication

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 \leq t \implies \|\mathbf{b} \circ \hat{\boldsymbol{\mu}}\|_2 \leq t. \tag{22}$$

If the left hand side of (22) holds, we find (recall $t > 0$ since $\operatorname{tr}(\mathbf{\Sigma}) > 0$)

$$\|\hat{\boldsymbol{\mu}}\|_2 \leq \|\boldsymbol{\mu}\|_2 + t = t\left(\frac{\|\boldsymbol{\mu}\|_2}{t} + 1\right) \iff \frac{1}{1 + \|\boldsymbol{\mu}\|_2/t}\|\hat{\boldsymbol{\mu}}\|_2 \leq t. \tag{23}$$

We can use (23) for the right hand side of (22). Recalling the definitions of $\alpha(\tau), p(\tau)$ and $\varepsilon(\tau)$ from the proof of Theorem 3.1 as well as the fact that the end of said proof can be rephrased as $C_{\alpha(\tau)}\varepsilon(\tau) \leq t$, we find

$$\begin{aligned}
\|\mathbf{b} \circ \hat{\boldsymbol{\mu}}\|_2 \leq \frac{1}{1 + (\|\boldsymbol{\mu}\|_2 / \sqrt{2\operatorname{tr}(\mathbf{\Sigma})})\sqrt{n/K}}\|\hat{\boldsymbol{\mu}}\|_2 &\overset{\sqrt{p(\tau)} \leq 1}{\leq} \\
\frac{1}{1 + \sqrt{p(\tau)}(\|\boldsymbol{\mu}\|_2 / \sqrt{2\operatorname{tr}(\mathbf{\Sigma})})\sqrt{n/K}}\|\hat{\boldsymbol{\mu}}\|_2 &\overset{C_{\alpha(\tau)} \geq 1}{\leq} \\
\frac{1}{1 + \|\boldsymbol{\mu}\|_2/(C_{\alpha(\tau)}\varepsilon(\tau))}\|\hat{\boldsymbol{\mu}}\|_2 \leq \frac{1}{1 + \|\boldsymbol{\mu}\|_2/t}\|\hat{\boldsymbol{\mu}}\|_2 &\overset{(23)}{\leq} t.
\end{aligned}$$

This bound proves (22) which allows us to deduce the inclusion of events

$$\{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu} + \mathbf{b} \circ \hat{\boldsymbol{\mu}}\|_2 > 2t\} \subset \{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 + \|\mathbf{b} \circ \hat{\boldsymbol{\mu}}\|_2 > 2t\} \overset{(22)}{\subset} \{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 > t\}.$$

Hence, by inclusion of events $\{\|\cdot\|_\infty \geq 2t\} \subset \{\|\cdot\|_2 \geq 2t\}$ and Theorem 3.1

$$\mathrm{P}(\,\|(1+\mathbf{b}) \circ \hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_\infty > 2t\,) \leq \mathrm{P}(\,\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|_2 > t\,) \overset{\text{Thm 3.1}}{\leq} \delta.$$

□

Our strategy is now to apply the following theorem from Yu et al. (2019):

Theorem 7.5 (Yu et al.)**.** *Suppose,* $\mathbf{\Gamma}_{0,S_0S_0}$ *is invertible and satisfies the irrepresentability condition with incoherence parameter* α. *Assume*

$$\| \left(\hat{\mathbf{\Gamma}}_K + \beta \cdot \operatorname{diag}(\hat{\mathbf{\Gamma}}_K)\right) - \mathbf{\Gamma}_0\|_\infty < \varepsilon_1, \qquad \|\hat{\mathbf{g}}_K - \mathbf{g}_0\|_\infty < \varepsilon_2,$$

and $d_{\boldsymbol{\theta}_0}\varepsilon_1 \leq \alpha/(6c_{\mathbf{\Gamma}_0})$. *If*

$$\lambda > \frac{3(2-\alpha)}{\alpha}\max(c_{\boldsymbol{\theta}_0}\varepsilon_1, \varepsilon_2),$$

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

$$\|\hat{\boldsymbol{\theta}}(K, \beta, \lambda) - \boldsymbol{\theta}_0\|_\infty \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda.$$

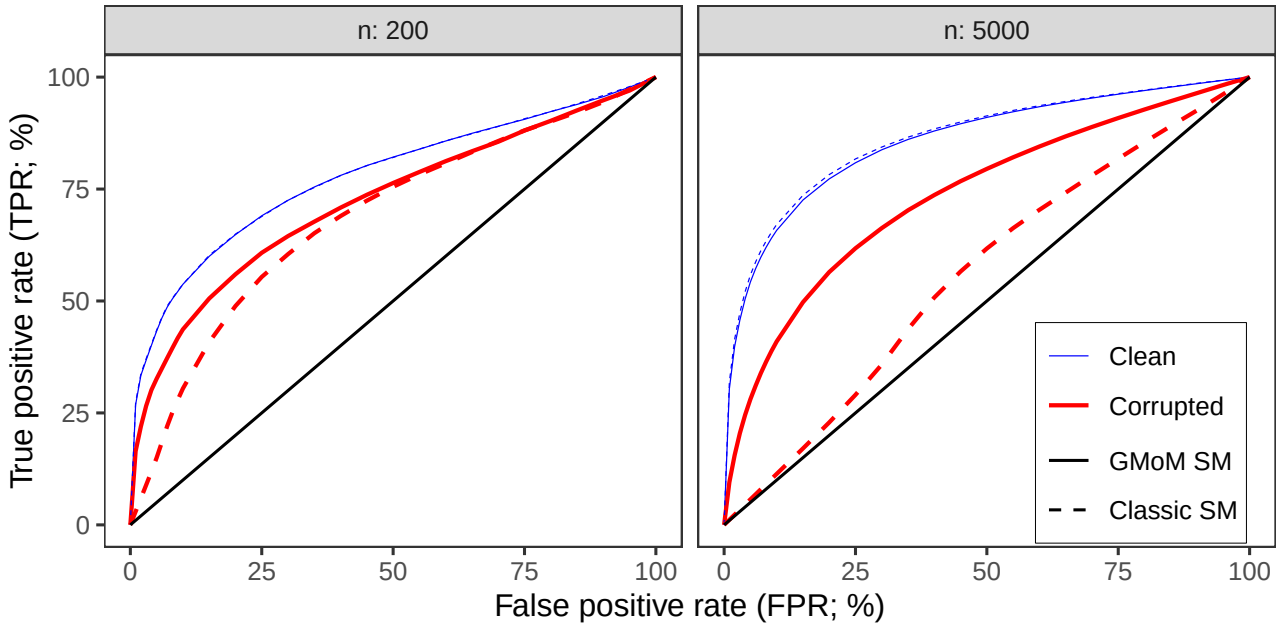

Figure 5: ROC curves for support recovery in the square root model. The experimental setup is the same as for Figure 1 of the main paper; the difference is the values for the sample size n.

Combining Lemma 7.4 with Theorem 7.5, we can prove Theorem 4.3:

Proof. Define

$$\varepsilon_1 := 4c(\tau)\sqrt{\log\left(\frac{4}{(1-\tau)^2}\frac{1}{\delta}\right)\frac{\operatorname{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\Gamma}_0})}{n}}, \qquad \varepsilon_2 := 2c(\tau)\sqrt{\log\left(\frac{4}{(1-\tau)^2}\frac{1}{\delta}\right)\frac{\operatorname{tr}(\boldsymbol{\Sigma}_{g_0})}{n}}.$$

Treating $\hat{\boldsymbol{\Gamma}}_K + \beta\,\mathrm{diag}(\hat{\boldsymbol{\Gamma}}_K)$ by Lemma 7.4 and $\hat{g}_K$ by Theorem 3.1 (together with the inclusion of events $\{\|\cdot\|_\infty > \text{const}\} \subset \{\|\cdot\|_2 > \text{const}\}$), applying a union bound yields that with probability at least $1 - 2\delta$

$$\|\hat{\boldsymbol{\Gamma}}_K + \beta\,\mathrm{diag}(\hat{\boldsymbol{\Gamma}}_K)\|_\infty \le \varepsilon_1/2 < \varepsilon_1, \qquad \|\hat{\mathbf{g}}_K - \mathbf{g}_0\|_\infty \le \varepsilon_2/2 < \varepsilon_2.$$

Furthermore, the growth condition on n ensures that

$$d_{\boldsymbol{\theta}_0}\varepsilon_1 \le \alpha/(6c_{\boldsymbol{\Gamma}_0})$$

and, by construction,

$$\lambda > 3(2-\alpha)\max(c_{\boldsymbol{\theta}_0}\varepsilon_1, \varepsilon_2)/\alpha.$$

The claim thus follows from Theorem 7.5. □

7.6 Additional simulations

7.6.1 Additional choices for the sample size n in Section 5.1

Results of the experiment described in Section 5.1 of the main paper for $n = 200$ and $n = 5000$ are reported in Figure 5. The figure supports the conclusions from Section 5.1. The classic score matching procedure is almost non-informative in the experiment with $n = 5000$ under contamination, highlighting the improvement the GMoM can have on robustness.

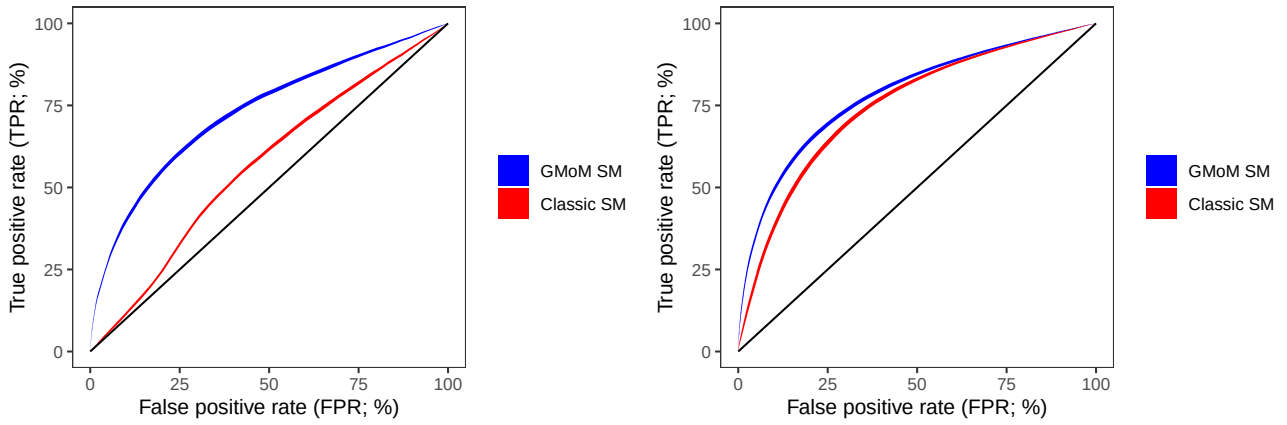

Figure 6: ROC curves for support recovery in the square root model under contamination only. 95% confidence bands are shown. Left: 10% Pareto contamination; Right: 5% contamination with Gaussian data of different dependence structure.

7.6.2 Additional contamination scenarios

Figure 6 shows how the experiments in Section 5.1 of the main paper are affected by changes to the contamination scenario. Concretely, we consider the experiment with $n = 1000$.

For the left hand side of Figure 6, the contamination percentage was increased from 5% to 10%, while the contamination distribution was maintained to be Pareto. As expected, when there is more contaminated samples, the GMoM version of score matching outperforms the classic version even more clearly.

For the right hand side of Figure 6, the contamination percentage was set to 5% again, but the contaminating samples were drawn from a Gaussian graphical model. The dependence network of the contaminating distribution was chosen at random independently of the network underlying the uncontaminated sample. To make the contaminated samples not blatantly inconsistent with the true model, their absolute value was taken such that the support constraint of the square root model is satisfied. As the ROC curves show, the GMoM version outperforms the classic version significantly in this contamination setting, albeit the absolute difference is relatively small. In a way, it is surprising the GMoM has a significantly better ROC curve at all, given that it downweighs based on magnitude and not on semantics.

7.6.3 Results for Gaussian graphical models

In this section, we apply regularized score matching (Classic SM) and our extension using the GMoM (GMoM SM) to simulated data from the familiar class of Gaussian graphical models (GGMs). We compare their support recovery performance in terms of ROC to that of GLASSO (Friedman et al., 2007b), a widely adopted tool for estimating sparse GGMs.

Again, we consider a scenario where $\varepsilon = 5\%$ of the observations are contaminated. Here, contaminated observations are replaced with draws from a Gaussian with iid components, having as variance 10 times the maximum component variance of the uncontaminated model. To make for a fair comparison, we provide GLASSO with a robust covariance estimate in the contaminated case. Specifically, we use the MAD-Spearman combination theoretically treated for GLASSO in (Loh and Tan, 2018).

Data was generated from a $m = 100$ dimensional Gaussian graphical model on an Erdős–Rényi graph with 100 edges. In the spirit of the experiments in section 4.1 of (Lin et al., 2016), we consider the borderline high-dimensional scenario $n = m = 100$ samples. The number of blocks for the GMoM SM was set to $4\varepsilon n = 20$, thus again being conservative on uncorrupted data and being adapted to the contamination amount ε on corrupted data. The diagonal multiplier was not needed since $n \geq m$. The penalty parameter λ was varied to cover the entire ROC space. No dampening function $\mathbf{h}$ is needed for the Gaussian, as the domain is unrestricted.

ROC curves based on 100 independent Monte Carlo simulations are displayed in Figure 7. On uncorrupted data, all three methods have practically identical ROC curves, which is in line with the findings from Lin et al. (2016).

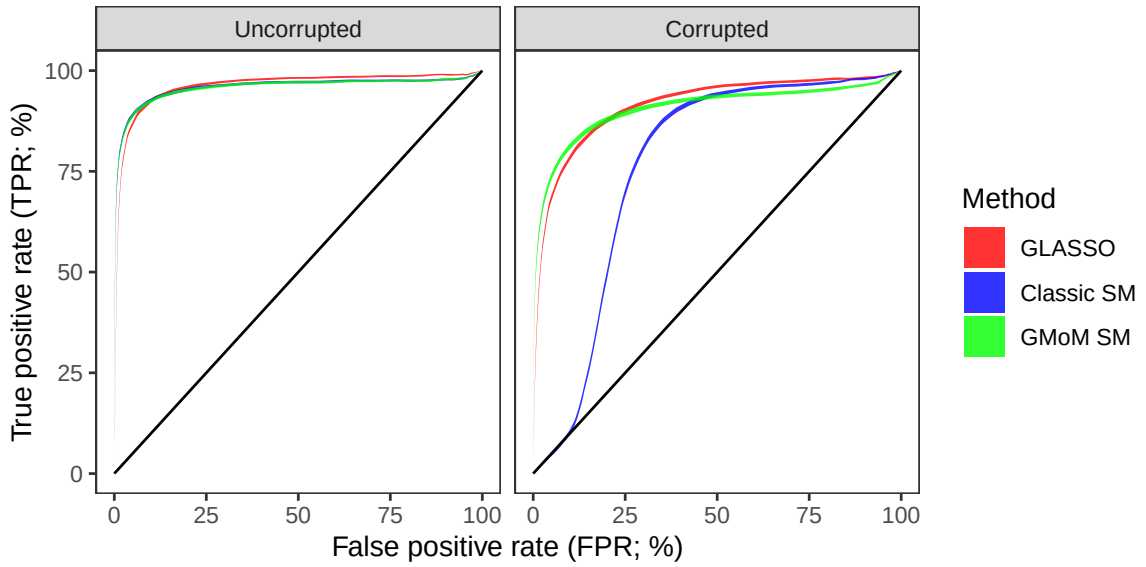

Figure 7: ROC curves for support recovery in the Gaussian graphical model. Right: 5% of observations have been contaminated. Line width of ROC curves shows a 95% confidence band.

On corrupted data, classic SM performs worse than the two robust methods, in line with the experiments from the main paper. The robust GLASSO and GMoM SM have very similar ROC curves, with GLASSO performing a bit better for high FPRs and GMoM SM a bit better for low FPRs. To conclude, the experiment shows that GMoM SM is a strong contender for estimating the support of sparse graphical models, especially when a part of the observations has been contaminated.