Sparse Parameter Recovery from Aggregated Data : Supplement

Avradeep Bhowmik AVRADEEP.1@UTEXAS.EDU
Joydeep Ghosh GHOSH@ECE.UTEXAS.EDU
The University of Texas at Austin, TX, USA

Oluwasanmi Koyejo SANMI@ILLINOIS.EDU
Stanford University, CA & University of Illinois at Urbana Champaign, IL, USA

1. General Remarks on the Results

As mentioned in the main manuscript, existing analyses in the sparse sensing literature are inadequate for analysing the aggregated data case, and our guarantees are much stronger than what could be achieved by a naive analysis.

The most general setup of the problem under study can be written in the following form:

$$\begin{aligned} &\text{Estimate:} && \boldsymbol{\beta}_0 \\ &\text{Given:} && \widehat{\mathbf{M}}, \hat{\boldsymbol{\upsilon}} \\ &\text{where:} && \widehat{\mathbf{M}} = \mathbf{M} + \mathbf{E} \\ & && \hat{\boldsymbol{\upsilon}} = \mathbf{y} + \mathbf{s} \\ & && \mathbf{y} = \mathbf{M}\boldsymbol{\beta}_0 \end{aligned} \tag{1}$$

There are four variations of this problem that are of interest in our setup:

1. error in design matrix $\widehat{\mathbf{M}} = \mathbf{M} + \mathbf{E}$, without noise in observation vector $\mathbf{y}$ (that is, $\mathbf{s} = 0$)
2. noise in observations $\hat{\boldsymbol{\upsilon}} = \mathbf{y} + \mathbf{s}$, with exact design matrix $\mathbf{M}$ (that is, $\mathbf{E} = 0$)
3. design matrix error $\mathbf{E}$ and observation noise $\mathbf{s}$, where $\mathbf{E}$ and $\mathbf{s}$ are independent, $\mathbf{E} \perp\!\!\!\perp \mathbf{s}$
4. the aggregated data case (as we study in this work) which contains both design matrix error $\mathbf{E}$ and observation noise $\mathbf{s}$, and where $\mathbf{E}$ and $\mathbf{s}$ are linearly correlated

To our knowledge, all prior work in the literature (eg. [Herman & Strohmer 2010; Chi et al. 2011; Rosenbaum et al. 2013; Rudelson & Zhou 2015] among others) only concern themselves with cases 1, 2 and 3. Moreover, for papers that do deal with case 2 and 3, unless $\mathbf{s} = 0$ the existing analysis will be restricted to providing only *approximate* recovery guarantees. Thus, these methods do not apply directly to case 4, a setup that almost always arises in the context of data aggregation.

We focus our investigation on the aggregated data case, that is, case 4: where $\mathbf{E}$ and $\mathbf{s}$ are linearly correlated. First of all, the existing literature does not make it clear how linearly correlated noise affects sparse parameter recovery from standard methods (like the LASSO or basis pursuit), and if the parameter can be recoverable in such cases. Even ignoring the linear correlation in the noise model, naive application of existing techniques that involve bounding error magnitudes will only be able to provide approximate recovery guarantees (where the degree of ℓ_2-approximation would depend on $\|\boldsymbol{\beta}_0\|$).

The key observation that allows us to bypass all these limitations is the fact that while $\mathbf{E}$ and $\mathbf{s}$ are correlated, we have one more piece of the puzzle that can be used to augment the information in equation 1: the fact that not only are $\mathbf{E}$ and $\mathbf{s}$ linearly correlated, they are tied together via the true parameter $\boldsymbol{\beta}_0$ in the form of the expression $\mathbf{s} = \mathbf{E}\boldsymbol{\beta}_0$. This is an artefact of the natural structure that is generated by data aggregation in linear models.

This observation is key to bypassing the problems in parameter recovery outlined earlier. Indeed, we show that not only can we guarantee parameter recovery using standard compressed sensing algorithms, we can also guarantee *exact* parameter recovery, as we see in Theorem 3.1, and recovery upto arbitrarily accurate degree of estimation as we see in 3.2 and 3.3. These results, while seemingly intuitive after the fact, have not been shown in either the compressed sensing literature, or in the literature on ecological estimation dating back at least 60 years to [Goodman 1953], and to our knowledge, ours is the first work that examines and gives guarantees for the structured parameter recovery problem in the context of aggregated data.

Furthermore, as we mention in the manuscript, our analysis techniques generalise beyond the exact problem setup and estimation procedure that we present in this paper, and

can be easily extended to analyse sparse or approximately sparse parameter recovery from aggregated data in a wide variety of contexts (non-sparse $\boldsymbol{\beta}_0$, beyond sub-Gaussian assumptions, etc. see for example [Candes et al. 2006; Cai et al. 2009]) and using various kinds of estimators beyond the LASSO or basis pursuit (for example the Dantzig selector, Matrix Uncertainty-selector, etc., see [Candes & Tao 2007; Rosenbaum et al. 2013]). While the sample complexity required may vary a little from case to case, our main results, on exact parameter recovery or recovery to within any arbitrary degree of approximation, would remain the same.

2. Proofs of Main Results

Note that the analysis presented below is one out of many possible approaches. Slightly different bounds can be achieved using different methods of analysis, for example using the Bauer-Fike Theorem, Weyl's Inequality, Wielandt Hoffman theorem, etc. and the bounds derived below can be made tighter by making further assumptions on the distributions of covariates or noise terms, etc.

The main property that enables recovery of sparse parameters from an underdetermined linear system is the restricted isometry condition, also sometimes known as the Uniform Uncertainty Principle.

For the matrix $\boldsymbol{M} \in \mathbb{R}^{k\times d}$ and any set $T \subseteq \{1, 2, \cdots, d\}$, suppose $\boldsymbol{M}_T$ is the $k \times |T|$ matrix consisting of the columns of $\boldsymbol{M}$ corresponding to T. Then, the *s-restricted isometry constant* δ_s of the matrix $\boldsymbol{M}$ is defined as the smallest quantity such that the matrix $\boldsymbol{M}_T$ obeys

$$(1-\delta_s)\|c\|_2^2 \leq \|\boldsymbol{M}_T c\|_2^2 \leq (1+\delta_s)\|c\|_2^2$$

for every subset $T \subset \{1, 2, \cdots, d\}$ of size $|T| < s$ and all real $c \in \mathbb{R}^{|T|}$

As in the main manuscript, we assume that $\mathbf{M}$ satisfies the restricted isometry hypotheses for both exact recovery and noisy recovery. That is, there exists an s_0 such that the following conditions are satisfied with respect to the $2s_0$-restricted isometry constants δ_{2s_0} for $\mathbf{M}$ in the manner as defined below:

1. For exact recovery from noise-free measurements, we assume $\delta_{2s_0} < \Theta_0 = \frac{3}{4+\sqrt{6}} \approx 0.465$
2. For approximate recovery from noisy measurements, we assume $\delta_{2s_0} < \Theta_1 = \sqrt{2} - 1 \approx 0.414$

However, we do not know the true mean matrix $\mathbf{M}$, only the sample mean matrix $\widehat{\mathbf{M}}_n = \mathbf{M} + \mathbf{E}_n$, where $\mathbf{E}_n$ is the matrix of aggregation error owing to empirical estimation from a finite number of samples. We now show that when the true mean matrix $\mathbf{M}$ satisfies the restricted isometry conditions, given enough samples n so will the sample mean matrix $\widehat{\mathbf{M}}_n$ with high probability.

We first show the following result for the isometry constants for $\widehat{\mathbf{M}}_n = \mathbf{M}+\mathbf{E}_n$ in terms of the eigenvalues of $\mathbf{E}_n$.

Lemma 2.1. *Let δ_s be the s-restricted isometry constant for $\mathbf{M}$. Let $\sqrt{\lambda_n}$ denote the absolute value of the largest (in absolute value) singular value of $E_{n,T}$ for all subsets $T \subset \{1, 2, \cdots d\}$. Then, $\zeta_s = (\delta_s + \lambda_n + 2\sqrt{\lambda_n(1-\delta_s)})$ is such that for every subset $T \subset \{1, 2, \cdots, d\}$ of size $|T| < s$ and all real $c \in \mathbb{R}^{|T|}$*

$$(1-\zeta_s)\|c\|_2^2 \leq \|(\mathbf{M}_T + E_{n,T})c\|_2^2 \leq (1+\zeta_s)\|c\|_2^2 \quad (2)$$

Proof. For every subset $T \subset \{1, 2, \cdots, d\}$ and all real $c \in \mathbb{R}^{|T|}$ we have by triangle inequality,

$$\|(\mathbf{M}_T + E_{n,T})c\| \leq \|\mathbf{M}_T c\| + \|E_{n,T}c\| \leq (\sqrt{1+\delta_s} + \sqrt{\lambda_n})\|c\|$$

Also,

$$\begin{aligned}\sqrt{(1-\delta_s)}\|c\| &\leq \|\mathbf{M}_T c\| \\ &= \|\left((\mathbf{M}_T + E_{n,T}) - E_{n,T}\right)c\| \\ &\leq \|(\mathbf{M}_T + E_{n,T})c\| + \|E_{n,T}c\| \\ &\leq \|(\mathbf{M}_T + E_{n,T})c\| + \sqrt{\lambda_n}\|c\|\end{aligned}$$

Therefore, we have

$$\begin{aligned}(\sqrt{1-\delta_s} - \sqrt{\lambda_n})\|c\| &\leq \|(\mathbf{M}_T + E_{n,T})c\| \\ &\leq (\sqrt{1+\delta_s} + \sqrt{\lambda_n})\|c\|\end{aligned}$$

Assume[1] $\lambda_n < (1+\delta_s)$, and $\zeta_s = (\delta_s + \lambda_n + 2\sqrt{\lambda_n(1+\delta_s)}) < 1$, then we have

$$\sqrt{(1-\zeta_s)} \leq \sqrt{1-\delta_s} - \sqrt{\lambda_n}$$

and

$$\sqrt{(1+\zeta_s)} = \sqrt{1+\delta_s} + \sqrt{\lambda_n}$$

This completes the proof. □

We now bound the singular values of $E_{n,T}$.

Lemma 2.2. *Let $\sqrt{\lambda_n}$ denote the absolute value of the largest (in absolute value) singular value of $E_{n,T}$ for any $T \subset \{1, 2, 3, \cdots d\}$. Then*

$$\lambda_n \leq \|\mathbf{E}_n\|_F^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

[1] We shall prove later that with overwhelmingly high probability $\lambda_n < \left(\frac{(\Theta-\delta_s)^2}{9(1+\delta_s)}\right)$ where $\Theta < 1$. This subsumes both the assumptions stated here.

Proof. Let $\sqrt{\lambda_\tau^{(n,T)}}$ for $\tau = 1, 2, \cdots$ be absolute values of the non-zero singular values of $E_{n,T}$. Consider the singular value decomposition of $E_{n,T} = U\Lambda V^\top$. Then

$$\begin{aligned} \|E_{n,T}\|_F^2 &= Trace(E_{n,T}^\top E_{n,T}) = Trace(\Lambda^\top \Lambda) \\ &= \sum_\tau \lambda_\tau \geq \max_\tau \lambda_\tau^{(n,T)} \end{aligned}$$

Therefore, for every T we have

$$\max_\tau \lambda_\tau^{(n,T)} \leq \|E_{n,T}\|_F^2 \leq \|\mathbf{E}_n\|_F^2$$

Since $\lambda_n = \max_T \max_\tau \lambda_\tau^{(n,T)}$, we have the result. □

This is just one approach, similar results can also be obtained, for example, by bounding the eigenvalues using the Gershgorin Circle Theorem.

Finally we show that with high probability λ_n can be bounded.

Lemma 2.3. *Suppose each covariate has a sub-Gaussian distribution with parameter σ^2, that is, for each covariate $x_{j,i} \in \mathbf{x}_j = [x_{j,1}, x_{j,2} \cdots x_{j,d}]$ and each group $j \in \{1, 2, \cdots k\}$, we have for every $t \in \mathbb{R}$, the logarithm of the moment generating function is quadratically bounded*

$$lnE[e^{t(x_{j,i} - \mu_{j,i})}] < \frac{t^2\sigma^2}{2}$$

Then, for any positive $\theta > 0$, the probability $P(\lambda_n > \theta) < 2kd\, e^{-n\theta/2kd\sigma^2}$

Proof. Note that the $(j,i)^{th}$ element of the matrix $\mathbf{E}_n$ is the zero random variable $E_{n,(j,i)} = \frac{\sum_{m=1}^n (x_{j,i}^{(m)} - \mu_{ji})}{n}$, where $x_{j,i}^{(m)}$ is the m^{th} observation of the i^{th} covariate in the j^{th} group, and μ_{ji} is the mean of the i^{th} covariate in the j^{th} group.

Since each covariate has a sub-Gaussian distribution with parameter σ^2, we have, by Hoeffding's inequality for sub-Gaussian random variables, for any $\theta > 0$

$$\begin{aligned} P\left(|E_{n,(ij)}| > \sqrt{\theta}\right) &= P\left(|\frac{\sum_{m=1}^n (x_{j,i}^{(m)} - \mu_{ji})}{n}| > \sqrt{\theta}\right) \\ &< 2e^{-n\theta/2\sigma^2} \end{aligned}$$

Therefore, using Lemma 2.2, we have

$$\begin{aligned} P(\lambda_n > \theta) &\leq P(\|\mathbf{E}_n\|_F^2 > \theta) \\ &\leq \sum_{ij} P(E_{n,(ij)}^2 > \frac{\theta}{kd}) \\ &\leq \sum_{ij} 2e^{-n\theta/2kd\sigma^2} \\ &= 2kd\, e^{-n\theta/2kd\sigma^2} \end{aligned}$$

where the second inequality is by union bound and the third is due to Hoeffding's inequality. □

We are now in a position to prove the main results.

2.1. Proof of Theorem 3.1

Proof. We saw in Lemma 2.1 that is the s-restricted isometry constants for $\mathbf{M}$ are δ_s, then the corresponding s-restricted isometry constants for $\widehat{\mathbf{M}}_n$ are

$$\zeta_s = \delta_s + \lambda_n + 2\sqrt{\lambda_n(1+\delta_s)} < \delta_s + 3\sqrt{\lambda_n(1+\delta_s)}$$

for small enough λ_n

Let $\Theta_0 = \frac{3}{4+\sqrt{6}} \approx 0.465$. Suppose there exists an s_0 such that the isometry constant δ_{2s_0} for the true mean matrix $\mathbf{M}$ satisfy $\delta_{2s_0} < \Theta_0$. Using Theorem 2.1 [Foucart 2010], we can see that any κ_0 sparse $\boldsymbol{\beta}_0$ can be recovered from $\widehat{\mathbf{M}}_n$ if the corresponding isometry constants for $\widehat{\mathbf{M}}_n$ satisfy $\zeta_{2s_0} < \Theta_0$, that is

$$\begin{aligned} & \zeta_{2s_0} && < \Theta_0 \\ \equiv\ & \zeta_{2s_0} - \delta_{2s0} && < \Theta_0 - \delta_{2s_0} \\ \Leftarrow\ & 3\sqrt{\lambda_n(1+\delta_{2s_0})} && < \Theta_0 - \delta_{2s_0} \\ \equiv\ & \lambda_n && < \vartheta_{s_0} \end{aligned} \tag{3}$$

where

$$\vartheta_{s_0} = \left(\frac{(\Theta_0 - \delta_{2s_0})^2}{9(1+\delta_{2s_0})}\right)$$

All that is left to show is that the condition $\zeta_{2s_0} < \Theta_0$ is true with high probability. This is straightforward by using Lemma 2.3 and the results in equations (2) above. We have,

$$\begin{aligned} P(\zeta_{2s_0} < \Theta_0) &> P(\lambda_n < \vartheta_{s_0}) \\ &= 1 - P(\lambda_n > \vartheta_{s_0}) \\ &\geq 1 - e^{-C_0 n} \quad \text{by Lemma 2.3} \end{aligned}$$

where the constant C_0 is such that

$$C_0 = O\left(\frac{\vartheta_0}{kd\sigma^2}\right) = O\left(\frac{(\Theta_0 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$$

□

2.2. Proof of Theorem 3.2

Proof. Using Theorem 2.2 [Candes 2008], recovery of $\boldsymbol{\beta}_0$ within an $O(\xi)$ distance is possible if the restricted isometry constants for $\widehat{\mathbf{M}}_n$ satisfy $\zeta_{2s_0} < \Theta_1$ where $\Theta_1 = \sqrt{2} - 1 \approx 0.414$, and the error term $\boldsymbol{\epsilon}_n$ is bounded as $\|\boldsymbol{\epsilon}_n\|_2 < \xi$. For succinctness, we drop the subscript from the error term and denote $\boldsymbol{\epsilon}_n$ simply as $\boldsymbol{\epsilon}$.

The probability of the restricted isometry condition being violated for the sample means can be bounded in a manner similar to the proof of theorem 3.1 as

$$P(\zeta_{2s_0} > \Theta_1) \leq e^{-C_1 n}$$

where $C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$. The probability of the error being too large can be bounded in a similar fashion by using Hoeffding's inequality as

$$\begin{aligned} P(\|\boldsymbol{\epsilon}\|_2 > \xi) &= P(\sum_{j=1}^{k} \epsilon_j^2 > \xi^2) \\ &\leq \sum_{j=1}^{k} P(\epsilon_j^2 > \frac{\xi^2}{k}) \\ &= \sum_{j=1}^{k} P(|\epsilon_j| > \frac{\xi}{\sqrt{k}}) \\ &\leq \sum_{j=1}^{k} 2\, e^{-n\xi^2/2\rho^2 k} \\ &= 2k\, e^{-n\xi^2/2\rho^2 k} \end{aligned}$$

where the first inequality is by union bound and the second inequality is due to Hoeffding's inequality.

Therefore the probability of recovery within $O(\xi)$ is bounded below by

$$1 - P(\zeta_{2s_0} > \Theta_1) - P(\|\boldsymbol{\epsilon}\|_2 > \xi) = 1 - e^{-C_1 n} - e^{-C_2 n}$$

where $C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$ and $C_2 \sim O\left(\frac{\xi^2}{\rho^2 k}\right)$ □

As mentioned earlier, there are multiple other approaches for special cases and using alternative conditions for successful recovery of sparse or nearly sparse vectors from under-determined linear systems, see for instance [Candes & Tao 2007], [Candes & Plan 2011], [Cai et al. 2010b], [Cai et al. 2010a], [Cai et al. 2009], etc. The analysis with alternative assumptions follows along the same lines as that presented in this paper.

2.3. Proof of Theorem 3.3

Proof. Note that the observations where the target mean is estimated from aggregated data as $\hat{\boldsymbol{\upsilon}}_\Delta = \hat{\boldsymbol{\upsilon}}_n + h_\Delta$ can be considered noisy observations of the type $\widehat{\mathbf{M}}_n \boldsymbol{\beta}_0 = \boldsymbol{\upsilon}_\Delta - h_\Delta$. Therefore, using Theorem 2.2, recovery of $\boldsymbol{\beta}_0$ within an $O(\xi_\Delta)$ distance is possible if the restricted isometry constants for $\widehat{\mathbf{M}}_n$ satisfy $\zeta_{2s_0} < \Theta_1$ and the the error term h_Δ is bounded as $\|h_\Delta\|_2 < \xi_\Delta$. The probability of the restricted isometry hypothesis being violated is

$$P(\zeta_{2s_0} > \Theta_1) \leq e^{-C_1 n}$$

where $C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$. This part is exactly identical to the proof of Theorem 3.2.

The bound on the error in estimation of target means can be done in a deterministic manner as follows.

The mean estimation procedure from the histogram is exact if the targets in each bin are distributed symmetrically around the mid point of each bin. Note that since each target is at a maximum distance of $\frac{\Delta}{2}$ from the mid point of their corresponding bin, by setting every target to the mid point of the bin we incur at most an error of $\frac{\Delta}{2}$ for each target. Therefore, the maximum possible error in estimating the sample mean in each group is

$$|\hat{\nu}_n - \hat{\nu}_\Delta| < \frac{\Delta}{2}$$

And hence, the error term h_Δ is bounded in ℓ_2 as

$$\|h_\Delta\|_2 < \sqrt{k}\frac{\Delta}{2}$$

This is of course a loose bound which assumed a worst-case pathological condition. Better bounds on the recovery error can be obtained by appropriate regularity assumptions on the distribution of the targets. □

3. Higher Order Moments

Consider the τ^{th} order moments under a linear function

$$\rho_\tau = E[y^\tau] = E[(\mathbf{x}^\top \boldsymbol{\beta})^\tau], \quad \tau = 1, 2, 3, \cdots \tag{4}$$

If all moments of the covariates are known, that is, $\{E[\Pi_j x_j^{a_j}] : a_j \in \mathbb{Z}_+, \sum_j a_j = \tau\}$ is known, then the right hand side of (4) is a scalar valued (shifted) homogeneous polynomial function in $\boldsymbol{\beta}$ of degree τ. Therefore, (4) is essentially a set of multivariate polynomial equations in $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_d]$. First consider whether the problem is well-defined, that is, whether the system of equations (4) has a unique solution. There is a considerable amount of literature in computational algebraic geometry that deals with the determination of whether a system of multivariate polynomial equations has at least one solution or is inconsistent (using, for instance, techniques and results in [Adams & Loustaunau 1994; Ruiz 1985]). In our case, this question is moot since we assume that the data is generated according to a linear model and therefore, there exists at least one solution. Unfortunately, testing for uniqueness of solution is a much harder problem.

As a base case, consider only using the first two moments. This is a widely applicable case since for many commonly used distribution choices for $\mathcal{P}_x$ like Multivariate Gaussian, Poisson, etc. the first two moments completely characterise the entire distribution.

The equations (4) can now be written comprising of a set of linear and a set of quadratic equations. The linear system of equations involving first order moments from each population sub-group $j \in \{1, 2, \cdots k\}$ is as follows:

$$E[\mathbf{x}_{(j)}^\top \boldsymbol{\beta}] = E[y_{(j)}] \Leftrightarrow \boldsymbol{\mu}_j^\top \boldsymbol{\beta} = \nu_j \quad j = 1, 2, 3, \cdots k \quad (5)$$

Similarly, the set of quadratic equations involving second order moments from each population-subgroup $j \in \{1, 2, \cdots k\}$ can be written as follows:

$$E[\boldsymbol{\beta}^\top (\mathbf{x}\mathbf{x}_{(j)}^\top)\boldsymbol{\beta}] = E[y_{(j)}^2] \Leftrightarrow \boldsymbol{\beta}^\top \Sigma_j \boldsymbol{\beta} = \sigma_j^2 \quad j = 1, \cdots, k \quad (6)$$

where Σ_j and σ_j^2 are the covariance of $\mathbf{x}$ and variance of y corresponding to the j^{th} population subgroup.

Geometrically, (5) and (6) represent in terms of $\boldsymbol{\beta}$ a set of k hyperplanes and a set of k ellipsoids centred at the origin in $\mathbb{R}^d$ space. The problem has a unique solution if the set of hyperplanes and the set of ellipsoids have a single point of intersection.

Counting the number of points of intersection of polynomials in real space is a difficult problem in the general case. It is usually studied for the complex space $\mathbb{C}^d$ under the umbrella of enumerative geometry [Katz 2006]. As earlier, if $k \geq d$ and under the assumption that at least one solution exists (the system is consistent), the set of hyperplanes is sufficient to recover the true $\boldsymbol{\beta}_0$. We would ideally like to see if knowledge of second order moments can reduce the number of population subgroups k required for a unique solution, or aids the estimation process in any other way.

Let Σ be some covariance matrix and $U\Delta_S U^\top$ be its singular value decomposition, where U is an orthonormal matrix and $\Delta_S = diag(S)$ is a diagonal matrix of loadings $S = [s_1, s_2, \cdots s_d] \succeq \mathbf{0}$. Let $\sigma^2 \in \mathbb{R}_+$ be any positive real value. Then for a given $\boldsymbol{\beta}$ to satisfy the second order moment constraint

$$\boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta} = \sigma^2 \quad (7)$$

means that the ellipsoid Σ in $\mathbb{R}^d$ centred at the origin with axes defined by U and of size (S, σ^2) passes through $\boldsymbol{\beta}$.

We now show that in the general case, knowledge about second order moments do not help.

Proposition 3.1. *Suppose $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are two points in $\mathbb{R}^d$ such that the origin, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are not collinear. For any arbitrary $\sigma^2 > 0$ and any arbitrary choice of axes U, the set of loadings S for which both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ satisfy equation (7) with $\Sigma = U\Delta_S U^\top$ and $\Delta_S = diag(S)$ is given by the intersection of a $(d-2)$-dimensional vector space with the positive orthant.*

Before we prove this, let us unpack this result. The essential idea is that, barring non-degenerate cases like $S = \mathbf{0}$, and for $d > 2$, a $(d-2)$ dimensional vector space intersects the positive orthant in an infinite number of points, assuming they do intersect. Therefore, for any two points in $\mathbb{R}^d$, there exist an infinite number of ellipsoids for every given size σ^2 and axes U which passes through both the points.

The implications of the above result are the following. Suppose we place constraints on $\boldsymbol{\beta}$ to constrain it to some set $\mathcal{C}$. Then if $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are any two points in $\mathcal{C}$, we can easily find any number of arbitrary second order moment conditions that are satisfied by both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Therefore, estimation with information about second order moments from k groups for any $k < \infty$ cannot be guaranteed to be any better than estimation without second order moments in the general case.

Furthermore, since the result holds for arbitrary values of σ^2 and U, it also implies that many types of common assumptions like sparsity or norm constraints on $\boldsymbol{\beta}$, rank constraints on the covariances Σ_k, etc. are insufficient in general to make the parameter recovery problem well defined with second order moments alone. Similar results can potentially be obtained for higher order moments by noting that a set of higher order polynomial equations can be converted into polynomial equations of degree $\tau \leq 2$ by introducing auxiliary variables.

Proof. Let $\Sigma = U\Delta_S U^\top$ where U is a unitary matrix and $\Delta_S = diag(S) = diag(s_1, s_2, \cdots, s_d)$ is a diagonal matrix. Let $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d$ be any two arbitrary points. Take the projections of each $\boldsymbol{\beta}_i$ on the axes defined by the j^{th} column $\boldsymbol{u}_j$ of U for each j. Let $\lambda_{j,2} = (\boldsymbol{\beta}_1^\top \boldsymbol{u}_j)^2$ and $\lambda_{j,2} = (\boldsymbol{\beta}_2^\top \boldsymbol{u}_j)^2$ be the corresponding squared projections of the two points $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ on each axis $\boldsymbol{u}_j$ for $j = 1, 2, 3, \cdots d$.

Concatenate the projections into the matrix $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1; \boldsymbol{\Lambda}_2]^\top \in \mathbb{R}^{2\times d}$ where $\boldsymbol{\Lambda_1} = [\lambda_{1,1}, \lambda_{2,1}, \cdots \lambda_{d,1}]^\top$ and $\boldsymbol{\Lambda_2} = [\lambda_{1,2}, \lambda_{2,2}, \cdots \lambda_{d,2}]^\top$.

It is easy to verify that

$$\begin{aligned} \boldsymbol{\beta}_1^\top \Sigma \boldsymbol{\beta}_1 &= \boldsymbol{\Lambda}_1^\top S \\ \boldsymbol{\beta}_2^\top \Sigma \boldsymbol{\beta}_2 &= \boldsymbol{\Lambda}_2^\top S \end{aligned}$$

Therefore, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ will both satisfy the second moment equation (7) for any ellipsoid defined by $(\Sigma = U\Delta_S U^\top, \sigma^2)$ if

$$\boldsymbol{\Lambda}^\top S = [\sigma^2; \sigma^2] \quad (8)$$

$$S \succeq \mathbf{0} \quad (9)$$

In terms of S, this represents an intersection of a $d-2$ dimensional vector space $\boldsymbol{\Lambda}^\top S = [\sigma^2; \sigma^2]$ with the positive orthant $S \succeq \mathbf{0}$ which is satisfied by an infinite number of solutions in terms of S. □

Note that $\mathbf{\Lambda}^\top S = [\sigma^2; \sigma^2]$ is inconsistent if $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are collinear with the origin, that is, $\boldsymbol{\beta}_1 = \eta\boldsymbol{\beta}_2$ for some η with $|\eta| \neq 1$. If $\boldsymbol{\beta}_1 = \pm\boldsymbol{\beta}_2$, then if one satisfies the ellipsoid constraint, the other trivially satisfies it as well.

References

Adams, William W and Loustaunau, Philippe. *An introduction to Gröbner bases*, volume 3. American Mathematical Society Providence, 1994.

Cai, T Tony, Xu, Guangwu, and Zhang, Jun. On recovery of sparse signals via ℓ_1 minimization. *Information Theory, IEEE Transactions on*, 55(7):3388–3397, 2009.

Cai, T Tony, Wang, Lie, and Xu, Guangwu. Shifting inequality and recovery of sparse signals. *Signal Processing, IEEE Transactions on*, 58(3):1300–1308, 2010a.

Cai, Tony Tony, Wang, Lie, and Xu, Guangwu. Stable recovery of sparse signals and an oracle inequality. 2010b.

Candes, Emmanuel and Tao, Terence. The dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pp. 2313–2351, 2007.

Candes, Emmanuel J. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

Candes, Emmanuel J and Plan, Yaniv. A probabilistic and ripless theory of compressed sensing. *Information Theory, IEEE Transactions on*, 57(11):7235–7254, 2011.

Candes, Emmanuel J, Romberg, Justin K, and Tao, Terence. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

Chi, Yuejie, Scharf, Louis L, Pezeshki, Ali, and Calderbank, A Robert. Sensitivity to basis mismatch in compressed sensing. *Signal Processing, IEEE Transactions on*, 59(5):2182–2195, 2011.

Foucart, Simon. A note on guaranteed sparse recovery via ℓ_1-minimization. *Applied and Computational Harmonic Analysis*, 29(1):97–103, 2010.

Goodman, Leo A. Ecological regressions and behavior of individuals. *American Sociological Review*, 1953.

Herman, Matthew A and Strohmer, Thomas. General deviants: An analysis of perturbations in compressed sensing. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):342–349, 2010.

Katz, Sheldon. *Enumerative geometry and string theory*. American Mathematical Soc., 2006.

Rosenbaum, Mathieu, Tsybakov, Alexandre B, et al. Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pp. 276–290. Institute of Mathematical Statistics, 2013.

Rudelson, Mark and Zhou, Shuheng. High dimensional errors-in-variables models with dependent measurements. *arXiv preprint arXiv:1502.02355*, 2015.

Ruiz, Jesús M. On hilbert's 17th problem and real nullstellensatz for global analytic functions. *Mathematische Zeitschrift*, 190(3):447–454, 1985.