Characterization and estimation of high dimensional sparse regression parameters under linear inequality constraints

Neha Agarwala[1], Arkaprava Roy [2] and Anindya Roy[3]

[1] Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland

[2] Department of Biostatistics, University of Florida

[3] Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, USA

Abstract

Modern statistical problems often involve such linear inequality constraints on model parameters. Ignoring natural parameter constraints usually results in less efficient statistical procedures. To this end, we define a notion of 'sparsity' for such restricted sets using lower-dimensional features. We allow our framework to be flexible so that the number of restrictions may be higher than the number of parameters. One such situation arise in estimation of monotone curve using a non parametric approach e.g. splines. We show that the proposed notion of sparsity agrees with the usual notion of sparsity in the unrestricted case and proves the validity of the proposed definition as a measure of sparsity. The proposed sparsity measure also allows us to generalize popular priors for sparse vector estimation to the constrained case.

Key words: sparsity, convex polyhedral cone, high dimension, adjacency graph, spike-and-slab prior, continuous shrinkage prior.

1 Introduction

In this chapter, we consider Bayesian estimation of possibly high dimensional parameter that are known to be restricted to a pointed closed convex polyhedral cone. We develop everything in the backdrop of normal mean estimation problem where the mean vector is constrained to a convex polyhedral cone but the concepts and the prior probability distributions developed here generalize easily to other models. Often, in constrained problems, the restricted models have to be embedded in higher dimensional models where the parameter space is unrestricted or at least more amenable to standard estimation methods. Thus, model complexity can be high in constrained problems even if the dimension of observations is not. In such situation some form of low dimensional formulation of the problem is required for making statistical inference possible without demanding a large sample size. The embedding to a higher dimensional space provides a parameterization of the model. For successful inference over a 'low dimensional' set of parameters the embedding needs to be an identifiable parameterization over that set. This property of the embedding is not guaranteed. We look at the restriction of parameters to a pointed full-dimensional closed convex cone defined by a set of linear inequalities

$$\mathbb{C} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{\mu} \geq 0\} \tag{1}$$

where $\boldsymbol{A}$ is some fixed $m \times n$ matrix. Since the cone is the intersection of finitely many half-spaces, it is a polyhedral cone. We consider the natural embedding of the cone using its minimal set of generators and consider its restriction to lower dimensional faces of the cone. We show that ascribing sparsity on the parameters of the embedding is not sufficient to have identifiable representation of the lower dimensional parameter vectors.

The main contribution of this chapter is an identifiable parameterization of vectors lying in lower dimensional subsets of the cone described in terms of the minimal generators representations. We define such vectors lying on the lower dimensional faces as 'sparse' vector because the notion of sparsity agrees with the usual notion of sparsity when the cone is an orthant. Then using the proposed definition of sparsity we defined flexible prior distributions that are either fully or nearly fully supported on the set of 'sparse' vectors and allows one to carry on Bayesian inference under sparsity and conic constraints.

There are many motivating applications where 'sparse' signals for constrained parameters arise and thus estimation of these parameters with these restrictions is desired. Some examples of that are popular in economics are estimation of cumulative distribution function (CDF), demand curve estimation, portfolio optimization and trend detection in econometrics. However, constrained parameter inference are not limited to business and economics, for example, dose-determination in treatments, signal detection in radar processing, and shape-constrained inference in non-parametric statistics.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)' \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$ where the parameter of interest $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)' \in \mathbb{C}$. We assume that $\mathbb{C}$ has non-zero interior volume with respect to the n dimensional Lebesgue measure. We consider a general framework where $\boldsymbol{\mu}$ is constrained to a proper polyhedral cone $\boldsymbol{A\mu} \geq \mathbf{0}$. A proper polyhedral cone is a closed convex full polyhedral cone that is pointed. A pointed cone is one that does not contain any non-trivial subspace and it is full or full-dimensional if the dual cone is pointed. We assume the cone is pointed (acute) and irreducible, i.e. the $m \times n$ $(m \geq n)$ matrix describing the linear inequalities, $\boldsymbol{A}$, is full column rank, and the rows are conically independent in the sense that there are no non-negative linear combinations, other than the trivial combination, of the rows that gives the zero vector.

The importance of linear inequality constraints in the practice of statistics is two fold. First, linear constraints arise extensively in shape restricted inference including, but not limited to, monotonicity, concavity or convexity. Such restrictions can be imposed directly on the mean function parameter or they can be modelled non-parametrically to obtain a flexible and smooth estimate. For example, if our goal is to fit a function, f to the data $(x_1, y_1), \ldots, (x_n, y_n)$, so that

$$y_i = f(x_i) + \epsilon_i$$

where f is assumed to have some restrictions and $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\text{cov}(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}$ then assuming a parametric approach, the mean function f is the parameter $\boldsymbol{\mu}$ with $A\boldsymbol{\mu} \geq \mathbf{0}$.

Second, the linear inequalities constraint framework can be used to extend estimation of $\boldsymbol{\mu}$ in the non-negative orthant when the covariance matrix is a general positive definite matrix $\boldsymbol{\Sigma}$. Consider the model $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \sigma^2 \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is completely known. A standard approach to dealing with general $\boldsymbol{\Sigma}$ matrix is to transform the observations to $\boldsymbol{z} = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{y}$ so that $\boldsymbol{z}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$ where $\boldsymbol{\theta} = \Sigma^{-1/2} \boldsymbol{\mu}$. However, the transformed mean $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\mu}$ need not remain in the positive orthant unless $\boldsymbol{\Sigma}$ is such that the square root $\boldsymbol{\Sigma}^{-1/2}$ is a *positive operator*, i.e. a matrix that leaves the cone unchanged. In the case of the positive orthant that would mean

$\boldsymbol{\Sigma}^{-1/2}$ is a non-negative matrix. e.g. $\boldsymbol{\Sigma}$ is an M-matrix with an inverse that admits a positive square-root. Hence one could reduce the problem of estimating $\boldsymbol{\mu}$ where $\boldsymbol{\mu} \geq \mathbf{0}$ to estimating $\boldsymbol{\theta}$ where $\boldsymbol{\Sigma}^{1/2}\boldsymbol{\theta} \geq \mathbf{0}$.

Of course, one could combine these two problems and consider the bigger problem of linear inequality constraints for a general $\boldsymbol{\Sigma}$. For $\boldsymbol{y}|\boldsymbol{\mu} \sim N(\boldsymbol{\mu}, \sigma^2\boldsymbol{\Sigma})$ with $\boldsymbol{A}\boldsymbol{\mu} \geq \mathbf{0}$. The problem can be transformed by taking $\boldsymbol{z} = \Sigma^{-1/2}\boldsymbol{y}$ so that $\boldsymbol{z}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \sigma^2\boldsymbol{I})$ where $\boldsymbol{\theta} = \Sigma^{-1/2}\boldsymbol{\mu}$. Hence, the estimation of $\boldsymbol{\mu}$ where $A\boldsymbol{\mu} \geq \mathbf{0}$ is condensed to estimating $\boldsymbol{\theta}$ where $\boldsymbol{A}\Sigma^{1/2}\boldsymbol{\theta} \geq \mathbf{0}$.

One way of estimating such a parameter is to first obtain an unrestricted estimate of the parameter and then truncate it so that the estimate lies in the constrained parameter space. Intuitively, the performance of the estimator is expected to be much better if such constraint conditions are incorporated in the estimation process. Hence the idea here is to incorporate the linear inequality restrictions into the model and in the inferential procedures.

From a frequentist estimation point of view this is a standard 'cone projection' problem of finding $\boldsymbol{\mu} \in \mathbb{C}$ such that it minimizes $||\boldsymbol{y} - \boldsymbol{\mu}||^2$. The cone projection problem is a special case of quadratic programming which involves finding $\boldsymbol{\theta}$ such that it minimizes $\boldsymbol{\theta}^T\boldsymbol{Q}\boldsymbol{\theta} - 2c^T\boldsymbol{\theta}$ over $\mathbb{C}$. When $\boldsymbol{Q}$ is positive definite, the objective function has a unique minimum and the solution reduces to finding the projection of a general Euclidean vector to the convex cone [16, 18]. Several algorithms have been studied in the literature to address the cone projection problem by Dykstra (1983), Karmarkar (1984), Fraser and Massam (1989) among others [7, 8, 9, 10, 13, 14, 15, 20]. A detailed account of the numerical stability and computational cost of the projection algorithms has been studied by Dimiccoli (2016) [6]. Constrained estimation of normal mean restricted to convex cones has been discussed in detail in Sen and Silvapulle (2001) [21]. Polyhedral cone constraints or equivalently linear inequalities arise extensively in shape restricted inference. There are many papers on estimation of regression function under shape restrictions which are special cases of the conic restriction problem. In the Bayesian set up, Danaher *et al.* (2012) provided an example of Bayesian estimation of normal mean when the mean is constrained to a convex polytope [4].

As mentioned, one of the most interesting question that naturally arises in the context of closed convex polyhedral cone restrictions is how to specify sparsity in constrained spaces such as $\mathbb{C}$. We provide a novel characterization of "sparse" parameters restricted to a polyhedral cone in Section 2. The notion of 'sparsity' defined here conforms with the general definition in the unrestricted case or in the case of the orthant. In Section 3 we define priors where bulk of the support is on the sparse vectors. Such priors would facilitate sparse signal extraction under general convex polyhedral cone restrictions. Finally, results from some of the examples are discussed in section 5.

2 Sparsity on Closed Convex Polyhedral Cones

To begin with, we provide some background on the geometry of cone and produce examples of three dimensional cone to understand and set the ideas. For any cone $\mathbb{C}$, let us denote its dimension by $\dim(\mathbb{C})$. A polyhedral cone is formed by the intersection of finitely many half spaces that contain the origin, i.e. for a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we define

$$\mathbb{C} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{\mu} \geq 0\} \tag{2}$$

to be a polyhedral cone with $\dim(\mathbb{C}) = n$. The halfspace representation of the cone containing the origin is called the facet representation or H-representation and the matrix A forming the set of linear inequalities is called the representation matrix. The face of a cone is a lower dimensional feature formed by the intersection of the cone with a supporting hyperplane. In particular, we focus on vertex, extreme ray and facet that are faces of a cone, each lying in different dimension. A vertex is a face of dimension 0, an extreme ray is a face of dimension 1 and a facet is a face of dimension $\dim(\mathbb{C}) - 1$.

We use the primal-dual representation of the cone to define 'sparsity'. Using Minkowski's theorem, a polyhedral cone (2) can also be represented using a finite set of vectors called generators or extreme rays. That is, for any $\boldsymbol{A}_{m\times n}$, there exists a generating matrix $\boldsymbol{\Delta}_{n\times d}$ such that

$$\mathbb{C} \quad = \quad \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{\mu} = \boldsymbol{\Delta}\boldsymbol{b} = \sum_{j=1}^{d} b_j \boldsymbol{\delta}_j,\ b_j \geq 0\} \tag{3}$$

where the columns $\boldsymbol{\delta}_j$ are the generators of the cone. This representation of a polyhedral cone is called the vertex representation or V-representation. The converse of the Minkowski's theorem is the Weyl's theorem for a polyhedral cone which states the existence of a representation matrix given a generating matrix. The generators are called *minimal* if they are conically independent, i.e. there is no positive linear combination of the generators that equals the origin vector. For the rest of this chapter, we assume $\boldsymbol{A}$ to be a *irreducible* matrix meaning that rows of $\boldsymbol{A}$ are conically independent. If $\boldsymbol{A}$ is full row rank, then it is irreducible. We also assume that the $rank(A) = n$. The resulting cone is called an *acute* cone and the set of extreme rays is its minimal generating system. In that case, d is the minimal number of extreme rays forming the skeleton of the cone.

Remark 1. *The parameterization of a cone $\mathbb{C}$ in terms of $\boldsymbol{b}$ in its vertex representation is not a proper parameterization in the sense for each vector $\boldsymbol{\mu} \in \mathbb{C}$ there could be multiple $\boldsymbol{b}$ such that $\boldsymbol{\Delta b} = \boldsymbol{\mu}$ even when the cone is irreducible and acute. Thus, the vector $\boldsymbol{b}$ is not generally identifiable from the vector $\boldsymbol{\mu}$. Only when $m = n = d$ and the cone is irreducible and acute, in which case $\Delta = \boldsymbol{A}^{-1}$ is non-singular and the parameterization is a bijection between the cone and the non-negative orthant.*

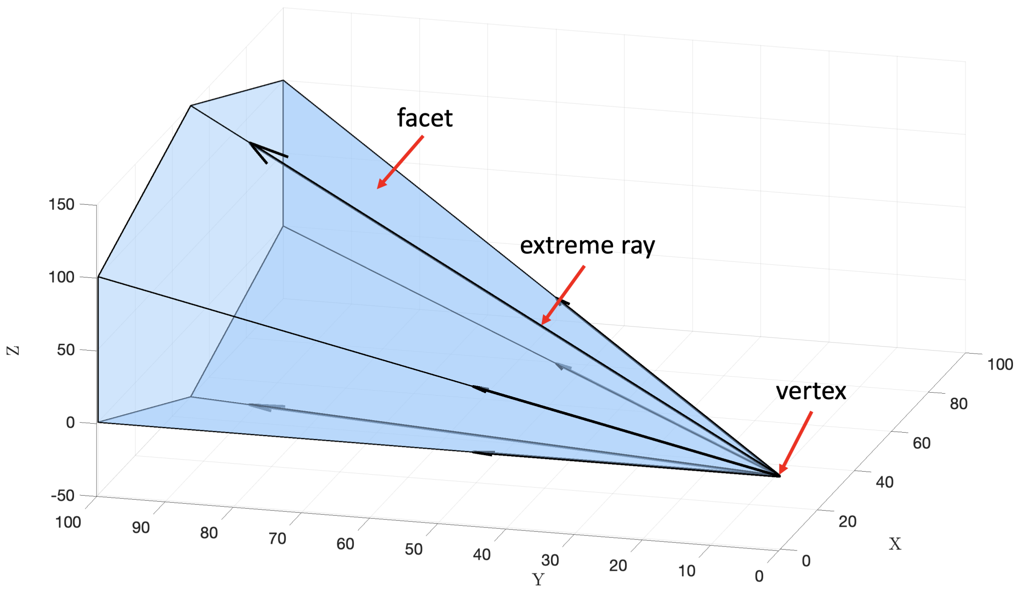

Figure 1: An example of polyhedral cone in $\mathbb{R}^3$ with $m = 6$ homogeneous linear inequalities and 6 extreme rays.

Figure (1) shows an example of a polyhedral cone in $\mathbb{R}^3$ i.e. $n = 3$ formed by $m = 6$ homogeneous linear inequalities. There are $m = 6$ hyperplanes intersecting with the cone and hence the number of facets is 6. Also, it turns out the number of extreme rays in $\mathbb{R}^3$ is equal to the number of facets. However, it is not true in general and d can be substantially larger than m which leads us to the next part.

Since there are two descriptions of a polyhedral cone, the pair $(\boldsymbol{A}, \boldsymbol{\Delta})$ is said be the Double description (DD) pair [19]. Switching between the two descriptions is called the representation conversion problem. Given the facet representation, the problem of finding the set of minimal extreme rays is called the extreme ray enumeration problem. Similarly, finding the irreducible representation from the vertex representation is called facet enumeration problem. When A is full row rank, the extreme rays $\boldsymbol{\delta}_j$'s are given by the columns of $\boldsymbol{\Delta} = \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{A}^T)^{-1}$ and $d = m$ [18]. When $\boldsymbol{A}$ is not full row rank, the number of extreme rays may be substantially larger than m. In that case, the extreme rays of the cone can be obtained using proposition 1 from Meyer (1999) [17].

There have been many variations and modifications of the Double Description (DD) method to move back and forth between the two representations, right from the primitive DD method to standard DD method [3, 12, 19, 23]. We use the R package "rcdd" by K. Fakuda, a R interface for cddlib which is a C-implementation of the DD method of Motzkin et al. [11, 12].

H-representation

$$x + \frac{3}{2}y - \frac{1}{2}z \geq 0$$
$$\frac{1}{2}x + \frac{3}{2}y - z \geq 0$$
$$-\frac{1}{2}x + \frac{3}{2}y - z \geq 0$$
$$-x + \frac{3}{2}y - \frac{1}{2}z \geq 0$$
$$-x + \frac{3}{2}y + \frac{1}{2}z \geq 0$$
$$-\frac{1}{2}x + \frac{3}{2}y + z \geq 0$$
$$\frac{1}{2}x + \frac{3}{2}y + z \geq 0$$
$$x + \frac{3}{2}y + \frac{1}{2}z \geq 0$$

V-representation

$$\begin{pmatrix} -1 & 0 & 1.5 & 1 & 0 & 1 & -1 & -1.5 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1.5 & 0 & 1 & -1.5 & -1 & -1 & 0 \end{pmatrix}_b$$

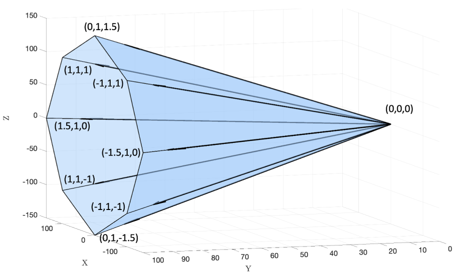

Figure 2: An illustration of the H-representation (left) and V-representation (center) for a irreducible polyhedral cone (right) in $\mathbb{R}^3$ with $n = 3, m = 8, d = 8$.

The methodology proposed here depends on the idea of describing points on the boundary of the cone or describing a points with proximity to the boundary of the cone when the point is in the interior. To this end, we need to use the adjacency graph with the list of adjacent extreme rays of the cone.

Definition 1. *For an acute cone $\mathbb{C} = \{\boldsymbol{\mu} : \boldsymbol{\mu} = \boldsymbol{\Delta b}\}$, two extreme rays $\boldsymbol{\delta}_i$ and $\boldsymbol{\delta}_j$ are adjacent if the minimal face containing both rays does not contain any other extreme rays of the cone.*

Two well-known tests for verifying the adjacency of extreme rays of a cone are the algebraic test and combinatorial test [19]. Given the adjacency relation one can define the *adjacency graph* of the cone. Let $\{\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_d\}$ correspond to a set of nodes in $V = \{1, \dots, d\}$. Then the edge set E is defined through adjacency. i.e. each pair of adjacent extreme rays, i and j correspond to a edge in the graph network. The edge set E can be written as the union of the edge set for each node. Suppose $E = \{E_1, E_2, \dots, E_d\}$ where E_i denote the set of adjacent extreme rays corresponding to the $\boldsymbol{\delta}_i$. Then $G = (V, E)$ forms an undirected graph. The degree of a node of a graph is the number of edges that are incident to the node. We denote the degree of the i^{th} node by $deg(\boldsymbol{\delta}_i)$. Then $|E_i| = deg(\boldsymbol{\delta}_i) + 1$.

To illustrate the geometry of polyhedral cones in 3D, consider the following example with $n = 3, m = 8, d = 8$ from Figure 2. The corresponding adjacency graph is shown in Figure 3. For instance, $\boldsymbol{\delta}_1$ is an extreme ray, which is is adjacent to $\boldsymbol{\delta}_2$ and $\boldsymbol{\delta}_8$. Hence in the corresponding adjacency graph, node 1 is connected to node 2 and node 8. In this case, each extreme ray is connected to two other extreme rays. So $deg(\boldsymbol{\delta}_i) = 2$ and $|E_i| = 3 \ \ \forall i$.

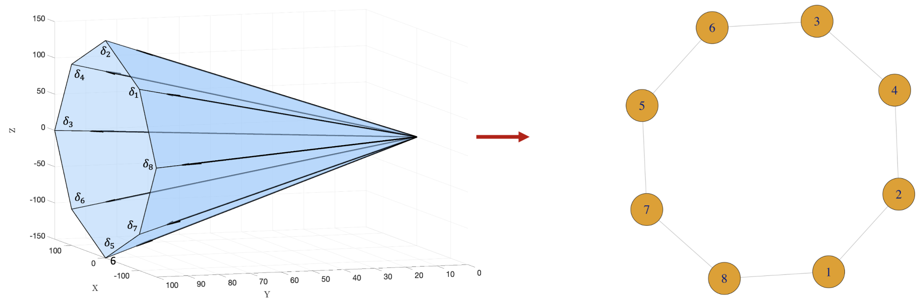

Figure 3: The graph network for the cone from Figure 2.

For high dimension with $n > 3$, the adjacency graph can become quite complicated with varying degree. A simple example for a polyhedral cone in $\mathbb{R}^4$ with $m = 7, d = 8$ is illustrated below with degree varying between 3 and 4.

When $m = n$, is the number of minimal generators is the same as the dimension, and

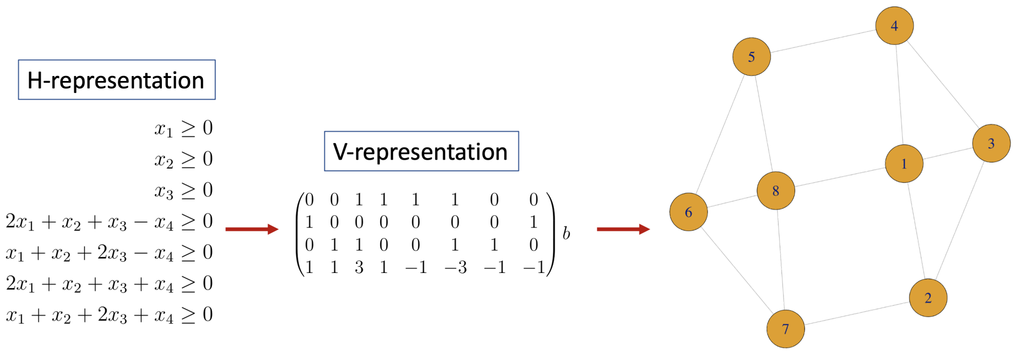

Figure 4: An illustration of the H-representation (left) and V-representation (center) for a irreducible polyhedral cone in $\mathbb{R}^4$ with its adjacency graph (right).

the adjacency graph is a *complete* graph. We will use the adjacency network to describe the notion of sparsity as well as the proposed priors.

When there are no restrictions, a sparse vector is a vector that has a large number of zeros (or for a weaker notion of sparsity the vector has a large number of entries that are negligible). For non-negative orthant, the same definition applies, except the non-zero entries are required to be positive. Thus, the sparse vectors are the one which lie on (or close to) one of the lower dimensional faces of the orthant. Following the description of sparsity in the orthant, we define a sparse vector to be any vector lying on or near a lower dimensional face. Since any $\boldsymbol{x} \in \mathcal{K}$ can be represented as $\boldsymbol{x} = \Delta \boldsymbol{b}$, extrinsic 'sparsity' can be defined as $\boldsymbol{x}$ being specified by smaller number of lower dimensional features. In other words, $\boldsymbol{x}$ is sparse when $\boldsymbol{b}$ is a sparse vector. The idea is to map the vector $\boldsymbol{x}$ in $\mathbb{R}^n$ to a non-negative orthant in $\mathbb{R}^d$, use the definition of sparsity in the orthant and then use the inverse map to lift the notion of sparsity back to the polyhedral cone. The dimension d in which the vector $\boldsymbol{x}$ is being embedded is either equal to or larger than the original dimension n.

For a non-negative orthant, the canonical vectors are the minimal generators and the usual definition of sparsity is that the vector can be written as a conic combination of a few of the full set of generators. Such vectors will lie on the boundary of the orthant, on a lower dimensional face of the orthant to be precise. It seems natural to use a similar definition of sparsity in the general case, i.e. vectors that lie on lower dimensional faces of the cone. The minimal two dimensional faces are the conic hull of pair of adjacent generator. Thus, to restrict the vector to the lower dimensional faces one can work with adjacent generator. However, simply generating a vector as a conic combination of a set of adjacent rays is not enough to guarantee that the vector lies on a lower dimensional face. It seems that the notion sparsity is more nuanced. For the vector to occupy a lower dimensional face, the sets of generators must form a *clique* or in other words the sub-adjacency graph corresponding to the set of generators used to define a sparse vector must be complete. This will ensure that the notion of sparsity is an identifiable notion in the sense that a sparse vector cannot have a non-sparse representation.

Recall that a clique, W, of an undirected graph $G = (V, E)$ is a subset of vertices, $W \subseteq V$ such that every two distinct nodes are connected by an edge. That is, a clique of a graph is an induced subgraph that is complete. A maximum clique of a graph, G, is a clique w such that $w \bigcup \{v\}$ in not a clique for any $v \in V \backslash w$. Then we have the following definition of a 'sparse' vector in a closed convex polyhedral cone.

Definition 2. *Let $\mathbb{C} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{\Delta b}\}$ be the vertex representation of a closed convex polyhedral cone $\mathbb{C}$ where the columns of $\boldsymbol{\Delta}$ is a set of d minimal generators of $\mathbb{C}$. Let $G = (V, E)$ be the adjacency graph of $\mathbb{C}$ where $V = \{1, \ldots, d\}$ and $E = \{E_1, \ldots, E_d\}$. Then $\boldsymbol{\mu} = \boldsymbol{\Delta} \in \mathbb{C}$ is sparse iff the subgraph corresponding to $i : b_i > 0$ is a clique.*

The following result proves that the above definition is 'proper' in the sense for a sparse vector there cannot be a non-sparse representation.

Theorem 1. *Suppose $\boldsymbol{\mu} \in \mathbb{C}$ has a vertex representation $\boldsymbol{\mu} = \boldsymbol{\Delta b}$ such that the set of nodes $\mathcal{I} = \{i : b_i > 0\}$ forms a clique. Then in any vertex representation of $\boldsymbol{\mu} = \boldsymbol{\Delta\beta}$ we have $\beta_i = 0$ for all $i \in \{1, \ldots, d\} \backslash \mathcal{I}$.*

Proof. We will use method of induction to prove the result. From the definition of adjacency, the result is obvious true when the size of the clique is $k = 2$. Now suppose it is true a positive integer $k > 2$. Let $\boldsymbol{\mu} = \sum_{i=1}^{k+1} b_i \boldsymbol{\delta}_i$ be a vertex representation of a vector $\boldsymbol{\mu}$ where without loss of generality we assume that the nodes $\{1, \ldots, k+1\}$ form a clique. Suppose there is another representation of $\boldsymbol{\mu}$ as

$$\boldsymbol{\mu} = \sum_{i=1}^{k+1} \beta_i \boldsymbol{\delta}_i + \sum_{i=k+2}^{d} b_i \boldsymbol{\delta}_i.$$

Then $\mathbf{0} = \sum_{i=1}^{k+1}(\beta_i - b_i)\boldsymbol{\delta}_i + \sum_{i=k+2}^{d} \beta_i \boldsymbol{\delta}_i$. Consider three cases.

case1: $(\beta_i - b_i) \geq 0, \forall i$ In this case a nonnegative linear combination of the columns of $\boldsymbol{\Delta}$ is zero which contradicts minimality of the generators.

case2: $(\beta_i - b_i) < 0$, for some i. Let $\mathcal{J} = \{i : (\beta_i - b_i) < 0\}$. Then

$$\boldsymbol{x} = \sum_{i \in \mathcal{J}} (b_i - \beta_i)\boldsymbol{\delta}_i = \sum_{i \in \{1,\dots,k+1\} \setminus \mathcal{J}} (\beta_i - b_i)\boldsymbol{\delta}_i + \sum_{i=k+2}^{d} b_i \boldsymbol{\delta}_i.$$

Thus, the vector $\boldsymbol{x}$ has two representations one of which is based on a clique since any sub-clique of a clique is also a clique. Since $|\mathcal{J}| \leq k$, this contradicts the assumption unless $\beta_i = b_i$ for $i = 1, \dots, k+1$ and $\beta_i = 0$ for $i = (k+1), \dots, d$. This completes the proof. $\square$

3 Sparse Priors for Closed Convex Polyhedral cone

To define probability measures on the cone that is supported mostly on lower dimensional sets, one could simply specify any sparse prior that are used in the unrestricted case as a prior on $\boldsymbol{b}$ in the vertex represetntaiton and invoke a prior on $\boldsymbol{\mu}$. Such a prior indeed works as a sparse prior on the cone provided the adjacency graph is a complete graph, as in the case of the positive orthant.

Thus, for the case when $d = n$, and hence the adjacency graph is a complete graph one could use popular sparse priors such as the continuous shrinkage priors like Horseshoe priors [2] or spike-and-slab priors like the Strawderman-Berger prior [1], where the continuous part is taken to be a density on the first orthant such as product of normal densities truncated to the positive half. Specifically, one could define priors on $\boldsymbol{b}$ as the Horseshoe prior

$$\begin{aligned} b_i|\tau, \lambda_i &\sim N(0, \tau^2\lambda_i^2)_+, \\ \lambda_i &\sim C(0,1)_+, \\ \tau|\sigma &\sim C(0,\sigma)_+, \\ \pi(\sigma) &\propto \tfrac{1}{\sigma} \end{aligned} \tag{4}$$

or as the Strawderman-Berger prior

$$
\begin{aligned}
\pi(b_i) &= (p\delta_o + (1-p)\ N(0, \tau^2\lambda_i^2)_+) \\
\pi(\lambda_i) &\propto \lambda_i(1+\lambda_i^2)^{\frac{3}{2}}, \\
p &\sim \text{Unif}(0,1), \\
\tau|\sigma &\sim C(\sigma,\sigma)\ 1(\tau \geq \sigma), \\
\pi(\sigma) &\propto \frac{1}{\sigma}
\end{aligned} \tag{5}
$$

One could also specify other priors such as Bayesian lasso [22] like prior on $\boldsymbol{b}$. However, when the adjacency graph is not complete, simply demanding that the vector $\boldsymbol{b}$ is sparse does not ensure that the resulting $\boldsymbol{\mu}$ vector is near a lower dimensional face. To guarantee sparsity, it is important to specify which of the components of $\boldsymbol{b}$ are zero. For instance, consider the above example of the 3D cone with eight extreme rays and suppose only b_4 and b_8 are the only positive entries in $\boldsymbol{b}$. The resulting vector will lie on the 2D cone generated by the vectors $\boldsymbol{\delta}_4$ and $\boldsymbol{\delta}_8$. Points on this set can be far away from any of the faces and can have many equivalent dense representation (Remark 1) where none of the entries in $\boldsymbol{b}$ is zero or small. Hence, the vector will not be sparse according to the notion described above. Thus, general sparse prior on $\boldsymbol{b}$ may still put substantial mass in the dense interior of the cone.

It is evident from the definition of sparsity that one could simply restrict to the clique-lattice of the adjacency graph and work with the maximal cliques to define priors that will be supported only on sparse vectors. To define a probability measure that is supported on the sparse vectors and hence on the maximal cliques corresponding to the adjacency graph, an obvious choice would be to define a Markov Random Field (MRF), specifically the Gibbs distribution describing the clique probabilities and then conditional on the clique defining a prior on the entries of $\boldsymbol{b}$ within the clique. We briefly review the Markov-Gibbs equivalence in the context of an undirected graph. Suppose $\{X_v : v \in V\}$ be a stochastic process with X_v taking values in S_v. Suppose further the joint distribution of the variables is $Q\{\boldsymbol{x}\} = P\{X_v = x_v \text{ for } v \in V\}$ where $\boldsymbol{x} = (x_1, \ldots, x_d)$ and $x_i \in S_i$.

Definition 3. *The probability distribution Q is called a Gibbs distribution for the graph if it can be written in the form*

$$Q\{\boldsymbol{x}\} = \prod_{S \in W} \phi_S(\boldsymbol{x})$$

where W is the set of cliques for G and ϕ_S is a positive function (also referred to as clique potential function) that depends on $\boldsymbol{x}$ only through $\{x_v : v \in S\}$. The definition is equivalent if maximal cliques are used instead of just cliques.

An MRF is characterized by its local property (the Markovianity) whereas a Gibbs Random Field (GRF) is characterized by its global property (the Gibbs distribution). The Hammersley-Clifford theorem establishes the equivalence of these two types of properties. The theorem asserts that the process $\{X_v : v \in V\}$ is a Markov Random field if and only if the corresponding Q is a Gibbs distribution. The practical value of the theorem is that it provides a simple way to parametrize the joint probability by specifying the clique potential functions. In other words, the theorem tells us it suffices to search over Gibbs distribution.

Given a particular maximal clique, then define the sparsity of a vector in the usual sense by generating the vector using possibly sparse coefficients on the generators belonging to the clique. This procedure agrees with the usual method of selecting sparse vectors on the orthant or $\mathbb{R}^n$ where the generators are the canonical vectors and all the extreme rays together for the unique maximal clique.

Thus, specifically we recommend the following class of sparse prior on $\mathbb{C}$. Let $\mathcal{W}$ be the set of maximal cliques of the adjacency graph of $\mathbb{C}$.

$$\begin{aligned} \boldsymbol{b}|w &\sim \pi(\boldsymbol{b}_w) \\ w &\sim \pi_{\mathcal{W}}(w) \end{aligned} \tag{6}$$

where given a clique $w \in \mathcal{W}$, $\boldsymbol{b}_w$ is the subvector of $\boldsymbol{b}$ constructed with the entries of $\boldsymbol{b}$ where the indices belong to w, $\pi(\boldsymbol{b}_w)$ is a 'sparse' prior, such as the Horseshoe prior or the Strawderman-Berger prior, on $\boldsymbol{b}_w$ in appropriate dimension, and $\pi_{\mathcal{W}}(w)$ is an MRF on $\mathcal{W}$. The priors $\pi(\cdot)$ and $\pi_{\mathcal{W}}(\cdot)$ an have their own hyper-parameters and hyperpriors can be specified accordingly.

In order to have a prior that is fully supported but has most of the support on the sparse vectors one could add a mixture component including the full set of extreme rays

$$\begin{aligned} \boldsymbol{b}|\delta, w &\sim \delta\pi^0(\boldsymbol{b}) + (1-\delta)\pi(\boldsymbol{b}_w) \\ w|\delta &\sim \delta I(w = V) + (1-\delta)\pi_{\mathcal{W}}(w) \\ \delta &\sim Bernoulli(\phi) \end{aligned} \tag{7}$$

where $\pi^0(\cdot)$ is a sparse prior on the interior of the positive orthant, $\mathbb{R}^n_+$, and the Bernoulli parameter ϕ is either a pre-specified small probability or a prior can be specified on ϕ.

3.1 Prior with adjacency on $\boldsymbol{b}$

A 'weaker' notion of sparsity will be to allow for mass to be spread along the boundary of the cone instead of being only supported on the boundary. In high dimension, the probability for most fully supported measures on the entire cone will concentrate on or near the boundary and hence so will the posterior. However, how the prior is specified will have impact on the recovery rate of the sparse sets.

Instead of restricting to cliques, one could choose priors that are supported on a cone generated by a single adjacency set. While not guaranteed, such priors would emphasize vectors where most of the coefficients in $\boldsymbol{b}$ are small in any representation of the vector. Of course the idea of small or negligible coefficients has to be formalized but in general this would mean $b_j < \epsilon, j \notin E_i$ fora given adjacency set E_i and for some pre-specified small value $\epsilon > 0$. Unfortunately, even when only a few coefficients within an adjacency are set to positive values, the resulting vector may still have equivalent representations that are very dense. If the prior specified on the elements of $\boldsymbol{b}$ within an adjacency set is sufficiently sparse, with high prior probability the generated vectors would be near one of the boundary sets, i.e. the minimum distance of the point to the boundary will be small.

To this end we define 'weakly sparse' priors that are fully supported on a closed convex polyhedral cone $\mathbb{C}$ and with most or all of its mass supported on or near the boundary. To formally define this, let

$$S(\boldsymbol{\mu}) = \{\boldsymbol{b} \in \mathbb{R}^d : \boldsymbol{\mu} = \Delta\boldsymbol{b}, \boldsymbol{b} \geq 0\}. \tag{8}$$

Then we have the following definition for a weakly sparse vector.

Definition 4. *Let $\boldsymbol{\mu} \in \mathbb{C}$ where $\mathbb{C}$ is a closed convex polyhedral cone with vertex representation given by $\mathbb{C} = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{x} = \boldsymbol{\Delta b}$ for some $\boldsymbol{b} \in \mathbb{R}^d_+\}$. Then $\boldsymbol{\mu}$ is weakly sparse if $\exists \boldsymbol{b} \in S(\boldsymbol{\mu})$ such that $\{i : b_i > 0\}$ corresponds to an adjacent set of an extreme ray in the adjacency graph of $\mathbb{C}$ where $S(\boldsymbol{\mu})$ is defined in* (8).

We propose adjacency prior based generalization of the Horseshoe or Strawderman-Berger priors as

$$\begin{aligned} \pi_1, \ldots, \pi_d &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_d) \\ u &\sim \text{Multinomial}(1, \pi_1, \ldots, \pi_d) \\ \boldsymbol{b}|u &\sim \pi(\boldsymbol{b}_{E_u}) \end{aligned} \tag{9}$$

where $\pi(\boldsymbol{b}_{E_u})$ can be $\pi_{HS}(\boldsymbol{b}_{E_u})$ or $\pi_{SB}(\boldsymbol{b}_{E_u})$. This is different from using a prior like modified lasso such as fused lasso (Tibshirani and Saunders, 2005) type selection, since we select only 1 adjacency set to stay on the surface whereas in fused lasso several clusters maybe selected and hence the results vectors may have dense representations.

4 Numerical Results

4.1 Distribution of points in a 3D cone

Figure 5 shows the distribution on 10000 points drawn from the Horseshoe kind prior on polyhedral cone. While most points lie near the face of the cone including the vertex, there are still many points in the interior of the cone. The 2D contour has been plotted by considering equal volume of circular cones inside the polyhedral cone and then calculating the relative frequency of 10000 points. The points very close to the vertex are included in the outermost region since they are anyway sparse for being close to the vertex. From the 2D contour, it is clearer that there is a heavy positive mass in the interior most circle.

Figure 6 and 7 presents the points inside the 3D polyhedral cone and the reciprocal 2D

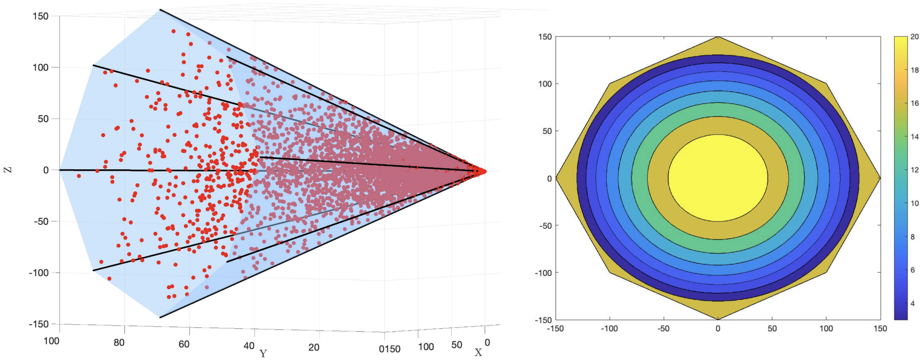

Figure 5: Plot showing points inside a 3D polyhedral cone by invoking a Horseshoe prior on $\boldsymbol{b}$ (left) and 2D contour of the cone showing the concentration of 10000 such points (right).

contour for Horseshoe prior on adjacent set and on maximal clique, respectively. The figures either show some positive mass and no positive mass in the interior of the cone for the two cases. All points are either closer to or lie exactly on the lower dimension features be it vertex, extreme rays or facet.

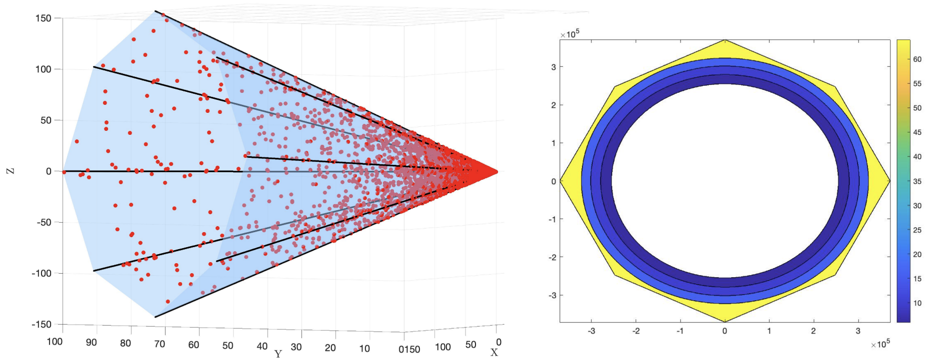

Figure 6: Plot showing points inside a 3D polyhedral cone by incorporating adjacency of extreme rays (left) and 2D contour of the cone showing the concentration of 10000 such points (right).

4.2 Max-min distance of points from facet

In this numerical study, we consider polyhedral cone in different dimensions and simulate $R = 100000$ points using the three different priors discussed in the previous section. For a fair comparison, for each of the adjacent set E_j chosen, we select randomly $|E_j|$ rays so that $\pi(u) \sim \frac{1}{\binom{d}{|E_j|}}$ and the rest of the prior specifications are same as in horseshoe prior with

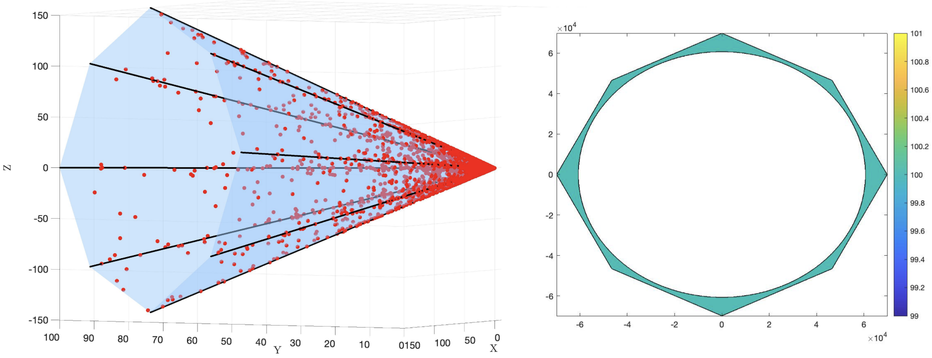

Figure 7: Plot showing points inside a 3D polyhedral cone by incorporating maximal cliques of extreme rays (left) and 2D contour of the cone showing the concentration of 10000 such points (right).

n	m	d	d_{ij}	d^r_{ij}	d^a_{ij}	d^c_{ij}
3	6	6	3.081	3.081	0.891	0
8	11	16	1.096	0.924	0.309	0
10	13	20	0.818	0.376	0.159	0

Table 1: Max-min distance of points from different priors

adjacency(9). We report the scaled max-min distance where the maximum is over R number of repetitions and minimum is considered with respect to the point's distance from the m facets. That is, we see which hyperplane of the cone it is closest to.

Let $d_{ij} = \max_{i=1:R} \min_{j=1:m} \text{distance}_{ij}$. Table 1 reports the max-min distance for Horseshoe prior d_{ij}, Horseshoe prior on a randomly selected set d^r_{ij}, Horseshoe prior with adjacency d^a_{ij} and Horseshoe prior on cliques d^c_{ij}.

5 Application

We discuss two examples in details. For the positive isotonic function estimation, we explain both the parametric and non-parametric approach. For the bell-shaped function, we show results for the parametric approach and additionally discuss how the non-parametric fit can be obtained.

5.1 Positive Isotonic Function

We consider the mean function $f(x) = \exp(x)$ over the interval $[-2, 2]$, a positive isotonic function so that A is $n \times n$ matrix with $A_{1,1} = 1$, $A_{i,i} = -1$, $A_{i,i+1} = 1$ for $i = 2, \dots, n$. Hence the Bayes estimator $\hat{\boldsymbol{\mu}}$ is obtained using MCMC by invoking a Horseshoe kind prior and Strawderman-Berger kind prior on $\boldsymbol{b}$ based on the model

$$\boldsymbol{\mu} = \Delta \boldsymbol{b}.$$

Figure 8 shows the plot of the estimators for the priors along with the MLE.

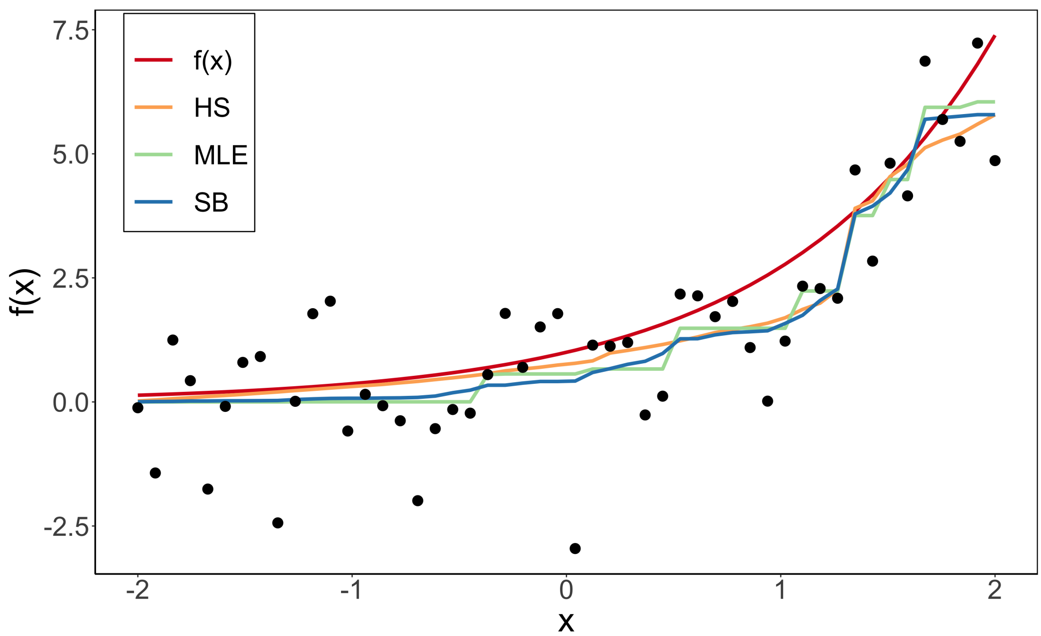

Figure 8: Bayes estimates for Horseshoe prior (HS), Strawderman-Berger prior (SB) and MLE for $n = 50$ points from $f(x) = \exp(x)$.

In the non-parametric approach, we model $f(x) = \boldsymbol{\Psi}(x)\boldsymbol{\beta}$ where $\boldsymbol{\Psi}(x)$ is the p dimensional basis function at x. This will produce a flexible and smooth estimate depending on the choice of p. To enforce the monotonicity of $f(x)$, we consider a set of fine grid points $t_1 < \cdots < t_m$ over the range of x and construct A such that the i^{th} row of A is the derivative of the basis functions $\boldsymbol{\Psi}'(x)$ evaluated at t_i. These constraints are then applied on the coefficients parameter such that $A\boldsymbol{\beta} \geq \mathbf{0}$ where A is a $m \times p$ matrix. Specifically, we consider cubic B-splines with no intercept and $k = 3$ equidistant internal knots so that $p = 6$ [5]. We consider

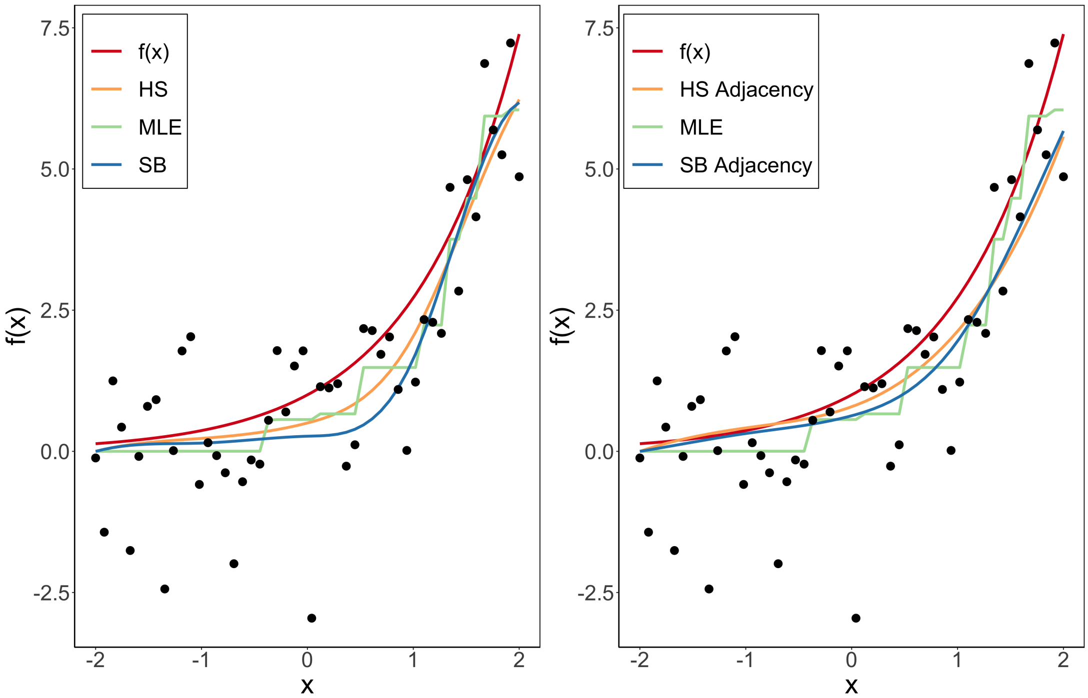

Figure 9: Bayes estimates using cubic B-spline with 3 internal knots for $n = 50$ points from $f(x) = \exp(x)$. Horseshoe prior (HS) and Strawderman-Berger prior (SB) (left) and Horseshoe prior (HS adjacency) and Strawderman-Berger prior with adjacency (SB adjacency) (right).

$m = 8$ equidistant grid points and since the number of constraints is greater than the number of parameters, the number of extreme rays $d = 18$ is greater than p. Similar to the parametric approach, $\hat{\boldsymbol{f}}$ is obtained using MCMC by invoking priors on $\boldsymbol{b}$ through the model

$$\boldsymbol{f} = \boldsymbol{\Psi}\boldsymbol{\beta} = \boldsymbol{\Psi}\Delta\boldsymbol{b} = \tilde{\Delta}\boldsymbol{b}.$$

Figure 9 presents the results from all four priors, the Horseshoe kind estimator and Strawderman Berger kind estimator as well as the priors incorporating adjacency. As expected, all four the estimators are smoother compared to ones obtained by parametric approach. For the priors incorporating adjacency, Figure 10 demonstrates the d estimates based solely on one of the d adjacency sets for Horseshoe kind prior (left) and for Strawderman-Berger kind prior (right). The final estimates for the priors invoking adjacency are an average of these d estimators presented in the right panel of Figure 9 since all these adjacent sets appear with almost equal frequency in the mcmc chains.

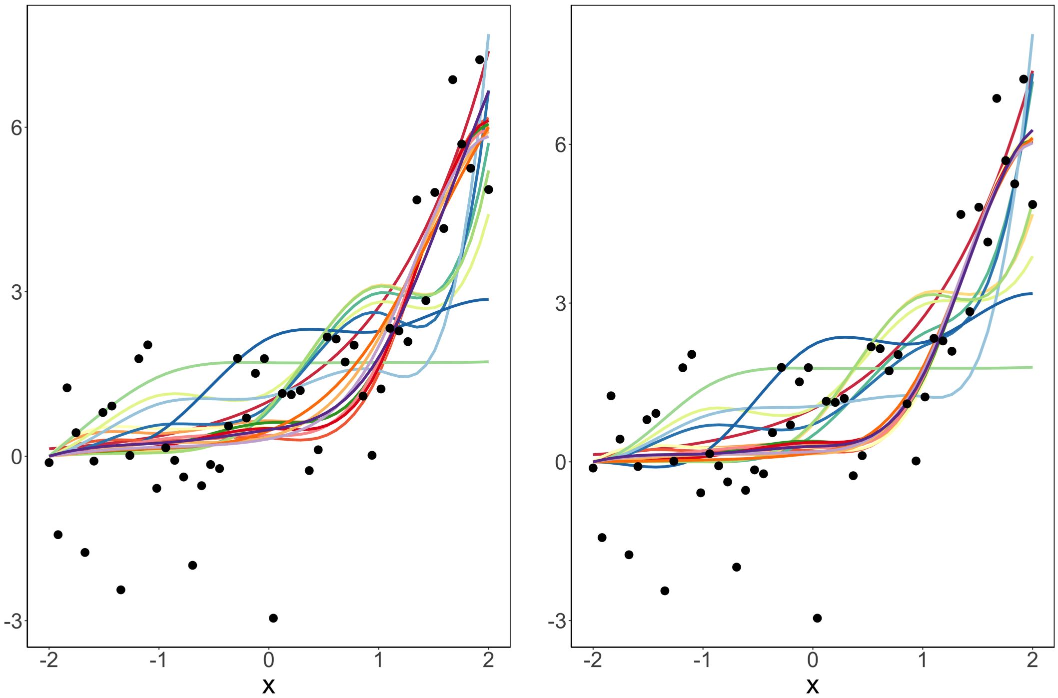

Figure 10: 18 estimates from the each of the $d = 18$ adjacency set for Horseshoe prior with adjacency (left) and Strawderman-Berger prior with adjacency (right) for $n = 50$ points from $f(x) = \exp(x)$.

5.2 Bell-shaped Function

In this example, we consider estimation of a symmetric bell-shaped curve. Given the inflection points k_1 and k_2, A is a $(n + 2) \times n$ matrix based on the constraints that the function is positive, increasing on the left, convex, concave, convex and then decreasing at the right. We consider the true mean function $f(x)$ to be a normal density scaled to have large values i.e. $f(x) = 50\ \frac{1}{\sigma}\phi(\frac{x}{\sigma})$ for $n = 40$ points over x in $[-2, 2]$. The estimated mean functions are obtained by invoking priors on $\boldsymbol{b}$ using the model $\boldsymbol{f} = \Delta\boldsymbol{b}$ where the number of extreme rays d becomes super large and is equal to 2551 for $n = 40$. The results are shown in Figure 11. Similar to the MLE, both simple Strawderman-Berger prior and the simple Horseshoe prior are piece-wise functions. Figure 12 provides 18 estimates out of $d = 2551$ estimates one for each of the d extreme sets for the priors incorporating adjacency. Since, each of these sets appear almost equally in the mcmc, we take an average of the 2551 estimates to obtain the final estimate for both the priors using adjacency.

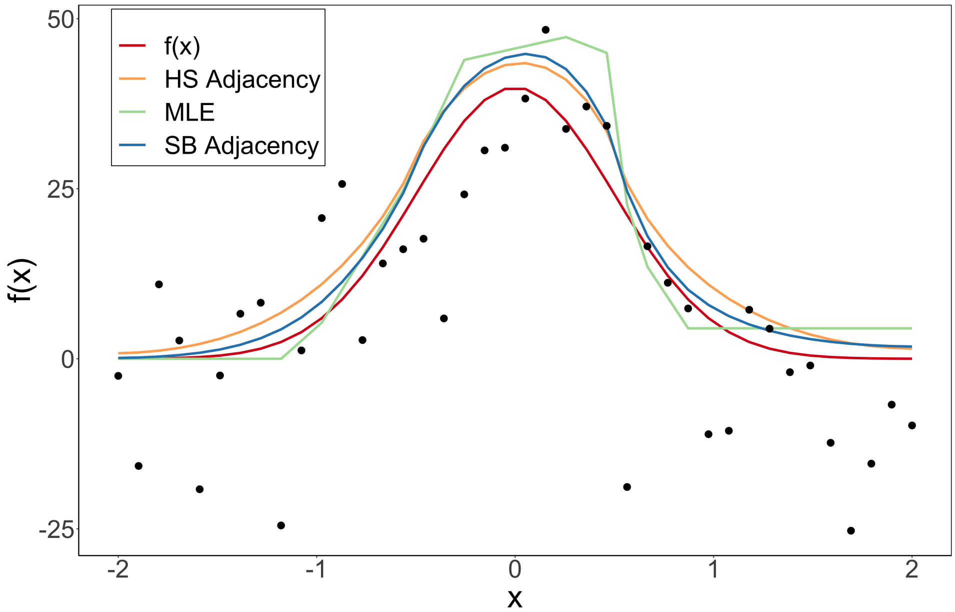

Figure 11: Bayes estimates for Horseshoe prior and Strawderman-Berger prior with adjacency for $n = 40$ points from $f(x) = 50\ \frac{1}{\sigma}\phi(\frac{x}{\sigma})$.

6 Discussion

In this chapter, we have have introduced new priors on high-dimensional closed convex cone where most of the mass is on lower dimensional sets on the boundary. The priors facilitate Bayesian estimation of constrained priors. While the motivating example is estimation of a constrained normal mean vector, the application of non-parametric estimation of shape-restricted functions show that the priors can easily applied to a regression model. In fact, it can be used for inference for any parameter vector with linear inequality constraints. For now, we have shown applications with inequality restrictions on the parameters but the notion of sparsity is related to having several of the inequalities reducing to equality in the true value of the parameter. While in the present set up these equality constraints are not necessarily binding, many examples where equality constraints are present as hard constraints in addition to inequality constraints can be also be incorporated in the proposed method. Another interesting application of our work is testing for $H_0 : \boldsymbol{A}\boldsymbol{\mu} = \mathbf{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\mu} \geq \mathbf{0}$ using Bayesian model comparison. When $\boldsymbol{A} = I$, the problem reduces to testing origin against non-negative orthant and the Likelihood Ratio Test is much easier to compute than for a general $\boldsymbol{A}$. The projection of the data vector to the cone will also lie on one of the lower dimensional faces and is the max-

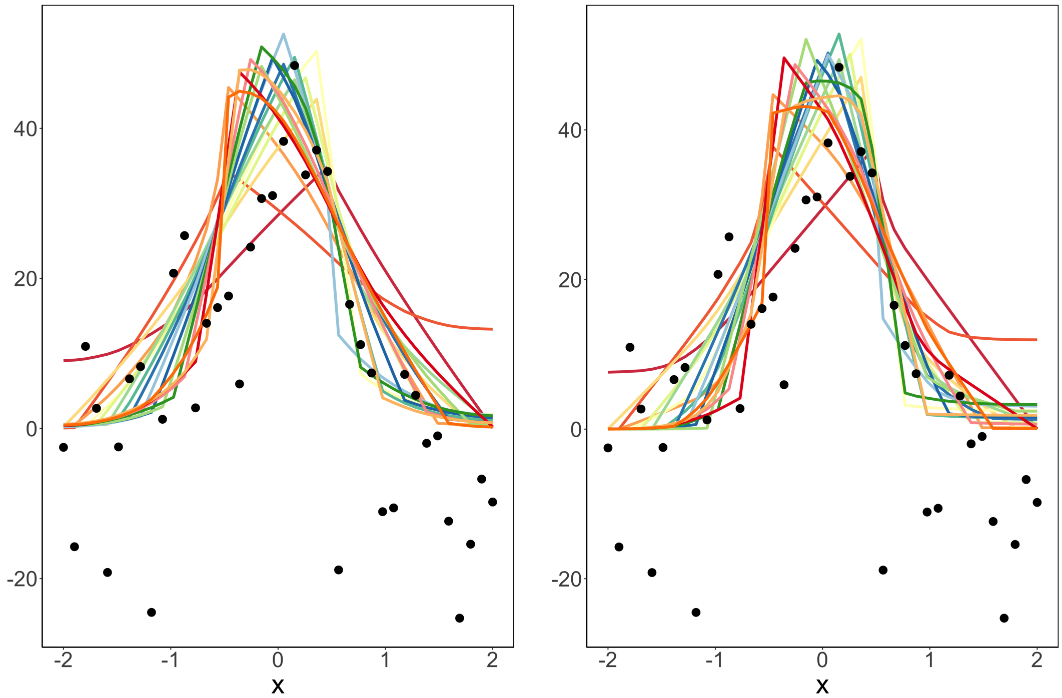

Figure 12: 18 out of $d = 2551$ estimates from the each of the 18 adjacency set for Horseshoe prior with adjacency (left) and Strawderman-Berger prior with adjacency (right) for $n = 40$ points from $f(x) = 50\ \frac{1}{\sigma}\phi(\frac{x}{\sigma})$.

imum likelihood estimator. In general the projection maybe hard to compute, but in principle the Bayesian posterior should concentrate around the Euclidean projection. Bayesian recovery results for the true clique and posterior concentration results need to be investigated.

References

[1] James O. Berger and William E. Strawderman. Choice of hierarchical priors: admissibility in estimation of normal means. *Ann. Statist.*, 24(3):931–951, June 1996. ISSN 0090-5364, 2168-8966. URL `https://projecteuclid.org/euclid.aos/1032526950`.

[2] Carlos M. Carvalho, Nicholas G. Polson, and James G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, June 2010. ISSN 0006-3444. URL `https://academic.oup.com/biomet/article/97/2/465/219397`.

[3] N. V. Chernikova. Algorithm for finding a general formula for the non-negative solutions of a system of linear inequalities. *USSR Computational Mathematics and Math-*

ematical Physics, 5(2):228–233, January 1965. ISSN 0041-5553. URL https://www.sciencedirect.com/science/article/pii/0041555365900455.

[4] Michelle R. Danaher, Anindya Roy, Zhen Chen, Sunni L. Mumford, and Enrique F. Schisterman. Minkowski–Weyl Priors for Models With Parameter Constraints: An Analysis of the BioCycle Study. *J Am Stat Assoc*, 107(500):1395–1409, 2012. ISSN 0162-1459. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4834988/.

[5] Carl de Boor. *A Practical Guide to Splines*. New York, Springer, 2001.

[6] Mariella Dimiccoli. Fundamentals of cone regression. *Statistics Surveys*, 10(none):53 – 99, 2016. URL https://doi.org/10.1214/16-SS114.

[7] A.M. Duguid. Studies in linear and non-linear programming, by k. j. arrow, l. hurwicz and h. uzawa. stanford university press, 1958. 229 pages. *Canadian Mathematical Bulletin*, 3 (3):196–198, 1960. doi: 10.1017/S0008439500025522.

[8] Richard L. Dykstra. An algorithm for restricted least squares regression. *Journal of the American Statistical Association*, 78(384):837–842, 1983. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1983.10477029.

[9] Yahya Fathi and Katta G. Murty. A critical index algorithm for nearest point problems on simplicial cones. December 1982. ISSN 0025-5610. URL http://deepblue.lib.umich.edu/handle/2027.42/47910. Accepted: 2006-09-11T19:32:29Z Publisher: Springer-Verlag; The Mathematical Programming Society, Inc.

[10] D. A. S. Fraser and Hélène Massam. A mixed primal-dual bases algorithm for regression under inequality constraints. application to concave regression. *Scandinavian Journal of Statistics*, 16:65–74, 2016.

[11] Komei Fukuda. cdd, cddplus and cddlib Homepage. https://people.inf.ethz.ch/fukudak/cdd_home/index.html. URL https://people.inf.ethz.ch/fukudak/cdd_home/index.html. Accessed: 2022-01-25.

[12] Komei Fukuda and Alain Prodon. Double description method revisited. In Michel Deza, Reinhardt Euler, and Ioannis Manoussakis, editors, *Combinatorics and Computer Science*,

Lecture Notes in Computer Science, pages 91–111. Springer, 1996. ISBN 978-3-540-70627-4.

[13] Magnus R. Hestenes. Multiplier and gradient methods. *J Optim Theory Appl*, 4(5):303–320, November 1969. ISSN 1573-2878. URL `https://doi.org/10.1007/BF00927673`.

[14] Clifford Hildreth. Point estimates of ordinates of concave functions. *Journal of the American Statistical Association*, 49(267):598–619, 1954. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1954.10483523`.

[15] N. Karmarkar. A new polynomial-time algorithm for linear programming. *Combinatorica*, 4(4):373–395, December 1984. ISSN 1439-6912. URL `https://doi.org/10.1007/BF02579150`.

[16] Xiyue Liao and Mary C. Meyer. coneproj: An R Package for the Primal or Dual Cone Projections with Routines for Constrained Regression. *Journal of Statistical Software*, 61:1–22, November 2014. ISSN 1548-7660. URL `https://doi.org/10.18637/jss.v061.i12`.

[17] Mary C. Meyer. An extension of the mixed primal–dual bases algorithm to the case of more constraints than dimensions. *Journal of Statistical Planning and Inference*, 81(1):13–31, October 1999. ISSN 0378-3758. URL `https://www.sciencedirect.com/science/article/pii/S0378375899000257`.

[18] Mary C. Meyer. A Simple New Algorithm for Quadratic Programming with Applications in Statistics. *Communications in Statistics - Simulation and Computation*, 42(5):1126–1139, May 2013. ISSN 0361-0918. URL `https://doi.org/10.1080/03610918.2012.659820`.

[19] T. S. Motzkin, H. Raiffa, G. L. Thompson, and R. M. Thrall. 3. The Double Description Method. In *3. The Double Description Method*, pages 51–74. Princeton University Press, March 2016. ISBN 978-1-4008-8197-0. doi: 10.1515/9781400881970-004. URL `https://www.degruyter.com/document/doi/10.1515/9781400881970-004/html`.

[20] M. J. D. Powell. Algorithms for nonlinear constraints that use lagrangian functions. *Mathematical Programming*, 14(1):224–248, December 1978. ISSN 1436-4646. URL `https://doi.org/10.1007/BF01588967`.

[21] Mervyn J. Silvapulle and Pranab Kumar Sen. *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. John Wiley & Sons, September 2011. ISBN 978-1-118-16563-8.

[22] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996. ISSN 0035-9246. URL `https://www.jstor.org/stable/2346178`.

[23] N. Yu. Zolotykh. New modification of the double description method for constructing the skeleton of a polyhedral cone. *Computational Mathematics and Mathematical Physics*, 52: 146–156, January 2012. ISSN 0965-5425. URL `https://ui.adsabs.harvard.edu/abs/2012CMMPh..52..146Z`.