Electronic Journal of Statistics
Vol. 9 (2015) 2324–2347
ISSN: 1935-7524
DOI: 10.1214/15-EJS1074

Statistical properties of convex clustering

Kean Ming Tan

Department of Biostatistics
University of Washington
Seattle, WA 98195, U.S.A.
e-mail: keanming@uw.edu

and

Daniela Witten

Departments of Statistics and Biostatistics
University of Washington
Seattle, WA 98195, U.S.A.
e-mail: dwitten@uw.edu

Abstract: In this manuscript, we study the statistical properties of *convex clustering*. We establish that convex clustering is closely related to single linkage hierarchical clustering and k-means clustering. In addition, we derive the range of the tuning parameter for convex clustering that yields a non-trivial solution. We also provide an unbiased estimator of the degrees of freedom, and provide a finite sample bound for the prediction error for convex clustering. We compare convex clustering to some traditional clustering methods in simulation studies.

Keywords and phrases: Degrees of freedom, fusion penalty, hierarchical clustering, k-means, prediction error, single linkage.

Received March 2015.

1. Introduction

Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a data matrix with n observations and p features. We assume for convenience that the rows of $\mathbf{X}$ are unique. The goal of clustering is to partition the n observations into K clusters, $D_1, \ldots, D_K$, based on some similarity measure. Traditional clustering methods such as hierarchical clustering, k-means clustering, and spectral clustering take a greedy approach (see, e.g., Hastie, Tibshirani and Friedman, 2009).

In recent years, several authors have proposed formulations for *convex clustering* (Pelckmans et al., 2005; Hocking et al., 2011; Lindsten, Ohlsson and Ljung, 2011; Chi and Lange, 2014a). Chi and Lange (2014a) proposed efficient algorithms for convex clustering. In addition, Radchenko and Mukherjee (2014) studied the theoretical properties of a closely related problem to convex clustering, and Zhu et al. (2014) studied the condition needed for convex clustering to recover the correct clusters.

Convex clustering of the rows, $\mathbf{X}_{1.}, \ldots, \mathbf{X}_{n.}$, of a data matrix $\mathbf{X}$ involves solving the convex optimization problem

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times p}}{\text{minimize}}\ \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{X}_{i.}-\mathbf{U}_{i.}\|_2^2+\lambda \mathrm{Q}_q(\mathbf{U}), \tag{1}$$

where $\mathrm{Q}_q(\mathbf{U}) = \sum_{i<i'}\|\mathbf{U}_{i.}-\mathbf{U}_{i'.}\|_q$ for $q \in \{1, 2, \infty\}$. The penalty $\mathrm{Q}_q(\mathbf{U})$ generalizes the fused lasso penalty proposed in Tibshirani et al. (2005), and encourages the rows of $\hat{\mathbf{U}}$, the solution to (1), to take on a small number of unique values. On the basis of $\hat{\mathbf{U}}$, we define the estimated clusters as follows.

Definition 1. The ith and i'th observations are estimated by convex clustering to belong to the same cluster if and only if $\hat{\mathbf{U}}_{i.} = \hat{\mathbf{U}}_{i'.}$.

The tuning parameter λ controls the number of unique rows of $\hat{\mathbf{U}}$, i.e., the number of estimated clusters. When $\lambda = 0$, $\hat{\mathbf{U}} = \mathbf{X}$, and so each observation belongs to its own cluster. As λ increases, the number of unique rows of $\hat{\mathbf{U}}$ will decrease. For sufficiently large λ, all rows of $\hat{\mathbf{U}}$ will be identical, and so all observations will be estimated to belong to a single cluster. Note that (1) is strictly convex, and therefore the solution $\hat{\mathbf{U}}$ is unique.

To simplify our analysis of convex clustering, we rewrite (1). Let $\mathbf{x} = \mathrm{vec}(\mathbf{X}) \in \mathbb{R}^{np}$ and let $\mathbf{u} = \mathrm{vec}(\mathbf{U}) \in \mathbb{R}^{np}$, where the $\mathrm{vec}(\cdot)$ operator is such that $x_{(i-1)p+j} = X_{ij}$ and $u_{(i-1)p+j} = U_{ij}$. Construct $\mathbf{D} \in \mathbb{R}^{[p\cdot\binom{n}{2}]\times np}$, and define the index set $\mathcal{C}(i, i')$ such that the $p \times np$ submatrix $\mathbf{D}_{\mathcal{C}(i,i')}$ satisfies $\mathbf{D}_{\mathcal{C}(i,i')}\mathbf{u} = \mathbf{U}_{i.} - \mathbf{U}_{i'.}$. Furthermore, for a vector $\mathbf{b} \in \mathbb{R}^{p\cdot\binom{n}{2}}$, we define

$$\mathrm{P}_q(\mathbf{b}) = \sum_{i<i'}\|\mathbf{b}_{\mathcal{C}(i,i')}\|_q. \tag{2}$$

Thus, we have $\mathrm{P}_q(\mathbf{Du}) = \sum_{i<i'}\|\mathbf{D}_{\mathcal{C}(i,i')}\mathbf{u}\|_q = \sum_{i<i'}\|\mathbf{U}_{i.}-\mathbf{U}_{i'.}\|_q = \mathrm{Q}_q(\mathbf{U})$. Problem (1) can be rewritten as

$$\underset{\mathbf{u}\in\mathbb{R}^{np}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2+\lambda \mathrm{P}_q(\mathbf{Du}). \tag{3}$$

When $q = 1$, (3) is an instance of the generalized lasso problem studied in Tibshirani and Taylor (2011). Let $\hat{\mathbf{u}}$ be the solution to (3). By Definition 1, the ith and i'th observations belong to the same cluster if and only if $\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}} = 0$. In what follows, we work with (3) instead of (1) for convenience.

Let $\mathbf{D}^\dagger \in \mathbb{R}^{np\times[p\cdot\binom{n}{2}]}$ be the Moore-Penrose pseudo-inverse of $\mathbf{D}$. We state some properties of $\mathbf{D}$ and $\mathbf{D}^\dagger$ that will prove useful in later sections.

Lemma 1. *The matrices* $\mathbf{D}$ *and* $\mathbf{D}^\dagger$ *have the following properties.*

(i) $\mathrm{rank}(\mathbf{D}) = p(n-1)$.
(ii) $\mathbf{D}^\dagger = \frac{1}{n}\mathbf{D}^T$.
(iii) $(\mathbf{D}^T\mathbf{D})^\dagger\mathbf{D}^T = \mathbf{D}^\dagger$ *and* $(\mathbf{D}\mathbf{D}^T)^\dagger\mathbf{D} = (\mathbf{D}^T)^\dagger$.

(iv) $\mathbf{D}(\mathbf{D}^T\mathbf{D})^\dagger\mathbf{D}^T = \frac{1}{n}\mathbf{D}\mathbf{D}^T$ *is a projection matrix onto the column space of* $\mathbf{D}$.

(v) Define $\Lambda_{\min}(\mathbf{D})$ *and* $\Lambda_{\max}(\mathbf{D})$ *as the minimum non-zero singular value and maximum singular value of the matrix* $\mathbf{D}$*, respectively. Then,* $\Lambda_{\min}(\mathbf{D}) = \Lambda_{\max}(\mathbf{D}) = \sqrt{n}$.

In this manuscript, we study the statistical properties of convex clustering. In Section 2, we study the dual problem of (3) and use it to establish that convex clustering is closely related to single linkage hierarchical clustering. In addition, we establish a connection between k-means clustering and convex clustering. In Section 3, we present some properties of convex clustering. More specifically, we characterize the range of the tuning parameter λ in (3) such that convex clustering yields a non-trivial solution. We also provide a finite sample bound for the prediction error, and an unbiased estimator of the degrees of freedom for convex clustering. In Section 4, we conduct numerical studies to evaluate the empirical performance of convex clustering relative to some existing proposals. We close with a discussion in Section 5.

2. Convex clustering, single linkage hierarchical clustering, and k-means clustering

In Section 2.1, we study the dual problem of convex clustering (3). Through its dual problem, we establish a connection between convex clustering and single linkage hierarchical clustering in Section 2.2. We then show that convex clustering is closely related to k-means clustering in Section 2.3.

2.1. Dual problem of convex clustering

We analyze convex clustering (3) by studying its dual problem. Let $s, q \in \{1, 2, \infty\}$ satisfy $\frac{1}{s} + \frac{1}{q} = 1$. For a vector $\mathbf{b} \in \mathbb{R}^{p\cdot\binom{n}{2}}$, let $\mathrm{P}_q^*(\mathbf{b})$ denote the dual norm of $\mathrm{P}_q(\mathbf{b})$, which takes the form

$$\mathrm{P}_q^*(\mathbf{b}) = \max_{i<i'} \|\mathbf{b}_{\mathcal{C}(i,i')}\|_s. \tag{4}$$

We refer the reader to Chapter 6 in Boyd and Vandenberghe (2004) for an overview of the concept of duality.

Lemma 2. *The dual problem of convex clustering (3) is*

$$\underset{\boldsymbol{\nu}\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}}{\text{minimize}} \frac{1}{2}\|\mathbf{x} - \mathbf{D}^T\boldsymbol{\nu}\|_2^2 \qquad \text{subject to } \mathrm{P}_q^*(\boldsymbol{\nu}) \le \lambda, \tag{5}$$

where $\boldsymbol{\nu} \in \mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}$ *is the dual variable. Furthermore, let* $\hat{\mathbf{u}}$ *and* $\hat{\boldsymbol{\nu}}$ *be the solutions to (3) and (5), respectively. Then,*

$$\mathbf{D}\hat{\mathbf{u}} = \mathbf{D}\mathbf{x} - \mathbf{D}\mathbf{D}^T\hat{\boldsymbol{\nu}}. \tag{6}$$

While (3) is strictly convex, its dual problem (5) is not strictly convex, since $\mathbf{D}$ is not of full rank by Lemma 1(i). Therefore, the solution $\hat{\boldsymbol{\nu}}$ to (5) is not unique. Lemma 1(iv) indicates that $\frac{1}{n}\mathbf{D}\mathbf{D}^T$ is a projection matrix onto the column space of $\mathbf{D}$. Thus, the solution $\mathbf{D}\hat{\mathbf{u}}$ in (6) can be interpreted as the difference between $\mathbf{D}\mathbf{x}$, the pairwise difference between rows of $\mathbf{X}$, and the projection of a dual variable onto the column space of $\mathbf{D}$.

We now consider a modification to the convex clustering problem (3). Recall from Definition 1 that the ith and i'th observations are in the same estimated cluster if $\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}} = \mathbf{0}$. This motivates us to estimate $\boldsymbol{\gamma} = \mathbf{D}\mathbf{u}$ directly by solving

$$\underset{\boldsymbol{\gamma}\in\mathbb{R}^{[p\cdot\binom{n}{2}]}}{\text{minimize}} \frac{1}{2}\|\mathbf{D}\mathbf{x} - \boldsymbol{\gamma}\|_2^2 + \lambda \mathrm{P}_q(\boldsymbol{\gamma}). \tag{7}$$

We establish a connection between (3) and (7) by studying the dual problem of (7).

Lemma 3. *The dual problem of (7) is*

$$\underset{\boldsymbol{\nu}'\in\mathbb{R}^{[p\cdot\binom{n}{2}]}}{\text{minimize}} \frac{1}{2}\|\mathbf{D}\mathbf{x} - \boldsymbol{\nu}'\|_2^2 \qquad \text{subject to } \mathrm{P}_q^*(\boldsymbol{\nu}') \le \lambda, \tag{8}$$

where $\boldsymbol{\nu}' \in \mathbb{R}^{[p\cdot\binom{n}{2}]}$ is the dual variable. Furthermore, let $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\nu}}'$ be the solutions to (7) and (8), respectively. Then,

$$\hat{\boldsymbol{\gamma}} = \mathbf{D}\mathbf{x} - \hat{\boldsymbol{\nu}}'. \tag{9}$$

Comparing (6) and (9), we see that the solutions to convex clustering (3) and the modified problem (7) are closely related. In particular, both $\mathbf{D}\hat{\mathbf{u}}$ in (6) and $\hat{\boldsymbol{\gamma}}$ in (9) involve taking the difference between $\mathbf{D}\mathbf{x}$ and some function of a dual variable that has $\mathrm{P}_q^*(\cdot)$ norm less than or equal to λ. The main difference is that in (6), the dual variable is projected into the column space of $\mathbf{D}$.

Problem (7) is quite simple, and in fact it amounts to a thresholding operation on $\mathbf{D}\mathbf{x}$ when $q = 1$ or $q = 2$, i.e., the solution $\hat{\boldsymbol{\gamma}}$ is obtained by performing soft thresholding on $\mathbf{D}\mathbf{x}$, or group soft thresholding on $\mathbf{D}_{\mathcal{C}(i,i')}\mathbf{x}$ for all $i < i'$, respectively (Bach et al., 2011). When $q = \infty$, an efficient algorithm was proposed by Duchi and Singer (2009).

2.2. Convex clustering and single linkage hierarchical clustering

In this section, we establish a connection between convex clustering and single linkage hierarchical clustering. Let $\hat{\boldsymbol{\gamma}}^q$ be the solution to (7) with $\mathrm{P}_q(\cdot)$ norm and let $s, q \in \{1, 2, \infty\}$ satisfy $\frac{1}{s} + \frac{1}{q} = 1$. Since (7) is separable in $\boldsymbol{\gamma}_{\mathcal{C}(i,i')}$ for all $i < i'$, by Lemma 2.1 in Haris, Witten and Simon (2015), it can be verified that

$$\hat{\boldsymbol{\gamma}}^q_{\mathcal{C}(i,i')} = \mathbf{0} \quad \text{if and only if} \quad \|\mathbf{X}_{i\cdot} - \mathbf{X}_{i'\cdot}\|_s \le \lambda. \tag{10}$$

It might be tempting to conclude that a pair of observations (i, i') belong to the same cluster if $\hat{\gamma}^q_{\mathcal{C}(i,i')} = \mathbf{0}$. However, by inspection of (10), it could happen that $\hat{\gamma}^q_{\mathcal{C}(i,i')} = \mathbf{0}$ and $\hat{\gamma}^q_{\mathcal{C}(i',i'')} = \mathbf{0}$, but $\hat{\gamma}^q_{\mathcal{C}(i,i'')} \neq \mathbf{0}$.

To overcome this problem, we define the $n \times n$ adjacency matrix $\mathbf{A}^q(\lambda)$ as

$$[\mathbf{A}^q(\lambda)]_{ii'} = \begin{cases} 1 & \text{if } i = i', \\ 1 & \text{if } \hat{\gamma}^q_{\mathcal{C}(i,i')} = \mathbf{0}, \\ 0 & \text{if } \hat{\gamma}^q_{\mathcal{C}(i,i')} \neq \mathbf{0}. \end{cases} \tag{11}$$

Subject to a rearrangement of the rows and columns, $\mathbf{A}^q(\lambda)$ is a block-diagonal matrix with some number of blocks, denoted as R. On the basis of $\mathbf{A}^q(\lambda)$, we define R estimated clusters: the indices of the observations in the rth cluster are the same as the indices of the observations in the rth block of $\mathbf{A}^q(\lambda)$.

We now present a lemma on the equivalence between single linkage hierarchical clustering and the clusters identified by (7) using (11). The lemma follows directly from the definition of single linkage clustering (see, for instance, Chapter 3.2 of Jain and Dubes, 1988).

Lemma 4. *Let $\hat{E}_1, \ldots, \hat{E}_R$ index the blocks within the adjacency matrix $\mathbf{A}_q(\lambda)$. Let s satisfy $\frac{1}{s} + \frac{1}{q} = 1$. Let $\hat{D}_1, \ldots, \hat{D}_K$ denote the clusters that result from performing single linkage hierarchical clustering on the dissimilarity matrix defined by the pairwise distance between the observations $\|\mathbf{X}_{i\cdot} - \mathbf{X}_{i'}\|_s$, and cutting the dendrogram at the height of $\lambda > 0$. Then $K = R$, and there exists a permutation $\pi : \{1, \ldots, K\} \to \{1, \ldots, K\}$ such that $D_k = E_{\pi(k)}$ for $k = 1, \ldots, K$.*

In other words, Lemma 4 implies that single linkage hierarchical clustering and (7) yield the same estimated clusters. Recalling the connection between (3) and (7) established in Section 2.1, this implies a close connection between convex clustering and single linkage hierarchical clustering.

2.3. Convex clustering and k-means clustering

We now establish a connection between convex clustering and k-means clustering. k-means clustering seeks to partition the n observations into K clusters by minimizing the within cluster sum of squares. That is, the clusters are given by the partition $\hat{D}_1, \ldots, \hat{D}_K$ of $\{1, \ldots, n\}$ that solves the optimization problem

$$\underset{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K \in \mathbb{R}^p, D_1, \ldots, D_K}{\text{minimize}} \sum_{k=1}^{K} \sum_{i \in D_k} \|\mathbf{X}_{i\cdot} - \boldsymbol{\mu}_k\|_2^2. \tag{12}$$

We consider convex clustering (1) with $q = 0$,

$$\underset{\mathbf{U} \in \mathbb{R}^{n \times p}}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{n} \|\mathbf{X}_{i\cdot} - \mathbf{U}_{i\cdot}\|_2^2 + \lambda \sum_{i < i'} \mathbb{I}(\mathbf{U}_{i\cdot} \neq \mathbf{U}_{i'\cdot}), \tag{13}$$

where $\mathbb{I}(\mathbf{U}_{i\cdot} \neq \mathbf{U}_{i'\cdot})$ is an indicator function that equals one if $\mathbf{U}_{i\cdot} \neq \mathbf{U}_{i'\cdot}$. Note that (13) is no longer a convex optimization problem.

We now establish a connection between (12) and (13). For a given value of λ, (13) is equivalent to

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times p},K,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\in\mathbb{R}^p,E_1,\ldots,E_K}{\text{minimize}} \frac{1}{2}\sum_{k=1}^{K}\sum_{i\in E_k}\|\mathbf{X}_{i\cdot}-\boldsymbol{\mu}_k\|_2^2+\lambda\sum_{i<i'}\sum_{k=1}^{K}\mathbb{I}(i\in E_k, i'\notin E_k), \tag{14}$$

subject to the constraint that $\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_K\}$ are the unique rows of $\mathbf{U}$ and $E_k = \{i : \mathbf{U}_{i\cdot} = \boldsymbol{\mu}_k\}$. Note that $\mathbb{I}(i \in E_k, i' \notin E_k)$ is an indicator function that equals to one if $i \in E_k$ and $i' \notin E_k$. Thus, we see from (12) and (14) that k-means clustering is equivalent to convex clustering with $q = 0$, up to a penalty term $\lambda\sum_{i<i'}\sum_{k=1}^{K}\mathbb{I}(i \in E_k, i' \notin E_k)$.

To interpret the penalty term, we consider the case when there are two clusters E_1 and E_2. The penalty term reduces to $\lambda|E_1|\cdot(n-|E_1|)$, where $|E_1|$ is the cardinality of the set E_1. The term $\lambda|E_1|\cdot(n-|E_1|)$ is minimized when $|E_1|$ is either 1 or $n-1$, encouraging one cluster taking only one observation. Thus, compared to k-means clustering, convex clustering with $q = 0$ has the undesirable behavior of producing clusters whose sizes are highly unbalanced.

3. Properties of convex clustering

We now study the properties of convex clustering (3) with $q \in \{1, 2\}$. In Section 3.1, we establish the range of the tuning parameter λ in (3) such that convex clustering yields a non-trivial solution with more than one cluster. We provide finite sample bounds for the prediction error of convex clustering in Section 3.2. Finally, we provide unbiased estimates of the degrees of freedom for convex clustering in Section 3.3.

3.1. Range of λ that yields non-trivial solution

In this section, we establish the range of the tuning parameter λ such that convex clustering (3) yields a solution with more than one cluster.

Lemma 5. *Let*

$$\lambda_{\text{upper}} := \begin{cases} \min\limits_{\boldsymbol{\omega}} \left\|\frac{1}{n}\mathbf{D}\mathbf{x}+\left(\mathbf{I}-\frac{1}{n}\mathbf{D}\mathbf{D}^T\right)\boldsymbol{\omega}\right\|_\infty & \text{for } q=1, \\ \min\limits_{\boldsymbol{\omega}} \left\{\max\limits_{i<i'}\left\{\left\|\left(\frac{1}{n}\mathbf{D}\mathbf{x}+\left(\mathbf{I}-\frac{1}{n}\mathbf{D}\mathbf{D}^T\right)\boldsymbol{\omega}\right)_{\mathcal{C}(i,i')}\right\|_2\right\}\right\} & \text{for } q=2. \end{cases} \tag{15}$$

Convex clustering (3) with $q = 1$ or $q = 2$ yields a non-trivial solution of more than one cluster if and only if $\lambda < \lambda_{\text{upper}}$.

By Lemma 5, we see that calculating λ_{upper} boils down to solving a convex optimization problem. This can be solved using a standard solver such as `CVX` in

MATLAB. In the absence of such a solver, a loose upper bound on λ_{upper} is given by $\|\frac{1}{n}\mathbf{D}\mathbf{x}\|_\infty$ for $q=1$, or $\max_{i<i'} \|\frac{1}{n}\mathbf{D}_{\mathcal{C}(i,i')}\mathbf{x}\|_2$ for $q=2$.

Therefore, to obtain the entire solution path of convex clustering, we need only consider values of λ that satisfy $\lambda \leq \lambda_{\text{upper}}$.

3.2. Bounds on prediction error

In this section, we assume the model $\mathbf{x} = \mathbf{u} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^{np}$ is a vector of independent sub-Gaussian noise terms with mean zero and variance σ^2, and $\mathbf{u}$ is an arbitrary np-dimensional mean vector. We refer the reader to pages 24-25 in Boucheron, Lugosi and Massart (2013) for the properties of sub-Gaussian random variables. We now provide finite sample bounds for the prediction error of convex clustering (3). Let λ be the tuning parameter in (3) and let $\lambda' = \frac{\lambda}{np}$.

Lemma 6. *Suppose that $\mathbf{x} = \mathbf{u} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^{np}$ and the elements of $\boldsymbol{\epsilon}$ are independent sub-Gaussian random variables with mean zero and variance σ^2. Let $\hat{\mathbf{u}}$ be the estimate obtained from (3) with $q=1$. If $\lambda' \geq 4\sigma\sqrt{\frac{\log\left(p\cdot\binom{n}{2}\right)}{n^3p^2}}$, then*

$$\frac{1}{2np}\|\hat{\mathbf{u}} - \mathbf{u}\|_2^2 \leq \frac{3\lambda'}{2}\|\mathbf{D}\mathbf{u}\|_1 + \sigma^2\left[\frac{1}{n} + \sqrt{\frac{\log(np)}{n^2p}}\right]$$

holds with probability at least $1 - \frac{2}{p\cdot\binom{n}{2}} - \exp\left\{-\min\left(c_1\log(np), c_2\sqrt{p\log(np)}\right)\right\}$, where c_1 and c_2 are positive constants appearing in Lemma 10.

We see from Lemma 6 that the average prediction error is bounded by the oracle quantity $\|\mathbf{D}\mathbf{u}\|_1$ and a second term that decays to zero as $n, p \to \infty$. Convex clustering with $q=1$ is prediction consistent only if $\lambda'\|\mathbf{D}\mathbf{u}\|_1 = o(1)$. We now provide a scenario for which $\lambda'\|\mathbf{D}\mathbf{u}\|_1 = o(1)$ holds.

Suppose that we are in the high-dimensional setting in which $p > n$ and the true underlying clusters differ only with respect to a fixed number of features (Witten and Tibshirani, 2010). Also, suppose that each element of $\mathbf{D}\mathbf{u}$ — that is, $U_{ij} - U_{i'j}$ — is of order $O(1)$. Therefore, $\|\mathbf{D}\mathbf{u}\|_1 = O(n^2)$, since by assumption only a fixed number of features have different means across clusters. Assume that $\sqrt{\frac{n\log\left(p\cdot\binom{n}{2}\right)}{p^2}} = o(1)$. Under these assumptions, convex clustering with $q=1$ is prediction consistent.

Next, we present a finite sample bound on the prediction error for convex clustering with $q=2$.

Lemma 7. *Suppose that $\mathbf{x} = \mathbf{u} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \in \mathbb{R}^{np}$ and the elements of $\boldsymbol{\epsilon}$ are independent sub-Gaussian random variables with mean zero and variance σ^2. Let $\hat{\mathbf{u}}$ be the estimate obtained from (3) with $q=2$. If $\lambda' \geq 4\sigma\sqrt{\frac{\log\left(p\cdot\binom{n}{2}\right)}{n^3p}}$, then*

$$\frac{1}{2np}\|\hat{\mathbf{u}} - \mathbf{u}\|_2^2 \leq \frac{3\lambda'}{2}\sum_{i<i'}\|\mathbf{D}_{\mathcal{C}(i,i')}\mathbf{u}\|_2 + \sigma^2\left[\frac{1}{n} + \sqrt{\frac{\log(np)}{n^2p}}\right]$$

holds with probability at least $1-\frac{2}{p\cdot\binom{n}{2}}-\exp\left\{-\min\left(c_1\log(np), c_2\sqrt{p\log(np)}\right)\right\}$, *where* c_1 *and* c_2 *are positive constants appearing in Lemma 10.*

Under the scenario described above, $\|\mathbf{D}_{\mathcal{C}(i,i')}\mathbf{u}\|_2 = O(1)$, and therefore $\sum_{i<i'}\|\mathbf{D}_{\mathcal{C}(i,i')}\mathbf{u}\|_2 = O(n^2)$. Convex clustering with $q=2$ is prediction consistent if $\sqrt{\frac{n\log\left(p\cdot\binom{n}{2}\right)}{p}} = o(1)$.

3.3. Degrees of freedom

Convex clustering recasts the clustering problem as a penalized regression problem, for which the notion of degrees of freedom is established (Efron, 1986). Under this framework, we provide an unbiased estimator of the degrees of freedom for clustering. Recall that $\hat{\mathbf{u}}$ is the solution to convex clustering (3). Suppose that $\mathrm{Var}(\mathbf{x}) = \sigma^2\mathbf{I}$. Then, the degrees of freedom for convex clustering is defined as $\frac{1}{\sigma^2}\sum_{\mathrm{j}=1}^{np}\mathrm{Cov}(\hat{u}_j, x_j)$ (see, e.g., Efron, 1986). An unbiased estimator of the degrees of freedom for convex clustering with $q=1$ follows directly from Theorem 3 in Tibshirani and Taylor (2012).

Lemma 8. *Assume that* $\mathbf{x}\sim\mathrm{MVN}(\mathbf{u},\sigma^2\mathbf{I})$, *and let* $\hat{\mathbf{u}}$ *be the solution to (3) with* $q=1$. *Furthermore, let* $\hat{\mathcal{B}}_1 = \{j : (\mathbf{D}\hat{\mathbf{u}})_j \neq 0\}$. *We define the matrix* $\mathbf{D}_{-\hat{\mathcal{B}}_1}$ *by removing the rows of* $\mathbf{D}$ *that correspond to* $\hat{\mathcal{B}}_1$. *Then*

$$\hat{\mathrm{df}}_1 = \mathrm{tr}\left(\mathbf{I}-\mathbf{D}^T_{-\hat{\mathcal{B}}_1}(\mathbf{D}_{-\hat{\mathcal{B}}_1}\mathbf{D}^T_{-\hat{\mathcal{B}}_1})^{\dagger}\mathbf{D}_{-\hat{\mathcal{B}}_1}\right) \tag{16}$$

is an unbiased estimator of the degrees of freedom of convex clustering with $q=1$.

The following corollary follows directly from Corollary 1 in Tibshirani and Taylor (2011).

Corollary 1. *Assume that* $\mathbf{x}\sim\mathrm{MVN}(\mathbf{u},\sigma^2\mathbf{I})$, *and let* $\hat{\mathbf{u}}$ *be the solution to (3) with* $q=1$. *The fit* $\hat{\mathbf{u}}$ *has degrees of freedom*

$$\mathrm{df}_1(\hat{\mathbf{u}}) = \mathrm{E}\left[\text{number of unique elements in } \hat{\mathbf{u}}\right].$$

There is an interesting interpretation of the degrees of freedom estimator for convex clustering with $q=1$. Suppose that there are K estimated clusters, and all elements of the estimated means corresponding to the K estimated clusters are unique. Then the degrees of freedom is Kp, the product of the number of estimated clusters and the number of features.

Next, we provide an unbiased estimator of the degrees of freedom for convex clustering with $q=2$.

Lemma 9. *Assume that* $\mathbf{x}\sim\mathrm{MVN}(\mathbf{u},\sigma^2\mathbf{I})$, *and let* $\hat{\mathbf{u}}$ *be the solution to (3) with* $q=2$. *Furthermore, let* $\hat{\mathcal{B}}_2 = \{(i,i') : \|\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\|_2 \neq 0\}$. *We define the matrix* $\mathbf{D}_{-\hat{\mathcal{B}}_2}$ *by removing rows of* $\mathbf{D}$ *that correspond to* $\hat{\mathcal{B}}_2$. *Let* $\mathbf{P}$ =

$\left(\mathbf{I} - \mathbf{D}_{-\hat{\mathcal{B}}_2}^T (\mathbf{D}_{-\hat{\mathcal{B}}_2} \mathbf{D}_{-\hat{\mathcal{B}}_2}^T)^\dagger \mathbf{D}_{-\hat{\mathcal{B}}_2}\right)$ *be the projection matrix onto the complement of the space spanned by the rows of* $\mathbf{D}_{-\hat{\mathcal{B}}_2}$. *Then*

$$\hat{\mathrm{df}}_2 = \mathrm{tr}\left(\left[\mathbf{I} + \lambda \mathbf{P} \sum_{(i,i') \in \hat{\mathcal{B}}_2} \left(\frac{\mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}}{\|\mathbf{D}_{\mathcal{C}(i,i')} \hat{\mathbf{u}}\|_2} - \frac{\mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')} \hat{\mathbf{u}} \hat{\mathbf{u}}^T \mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}}{\|\mathbf{D}_{\mathcal{C}(i,i')} \hat{\mathbf{u}}\|_2^3} \right) \right]^{-1} \mathbf{P}\right) \quad (17)$$

is an unbiased estimator of the degrees of freedom of convex clustering with $q = 2$.

When $\lambda = 0$, $\|\mathbf{D}_{\mathcal{C}(i,i')} \hat{\mathbf{u}}\|_2 \neq 0$ for all $i < i'$. Therefore, $\mathbf{P} = \mathbf{I} \in \mathbb{R}^{np \times np}$ and the degrees of freedom estimate is equal to $\mathrm{tr}(\mathbf{I}) = np$. When λ is sufficiently large that $\hat{\mathcal{B}}_2$ is an empty set, one can verify that $\mathbf{P} = \mathbf{I} - \mathbf{D}^T (\mathbf{D}\mathbf{D}^T)^\dagger \mathbf{D}$ is a projection matrix of rank p, using the fact that $\mathrm{rank}(\mathbf{D}) = p(n-1)$ from Lemma 1(i). Therefore $\hat{\mathrm{df}}_2 = \mathrm{tr}(\mathbf{P}) = p$.

We now assess the accuracy of the proposed unbiased estimators of the degrees of freedom. We simulate Gaussian clusters with $K = 2$ as described in Section 4.1 with $n = p = 20$ and $\sigma = 0.5$. We perform convex clustering with $q = 1$ and $q = 2$ across a fine grid of tuning parameters λ. For each λ, we compare the quantities (16) and (17) to

$$\frac{1}{\sigma^2} \sum_{j=1}^{np} (\hat{u}_j - u_j)(x_j - u_j), \quad (18)$$

which is an unbiased estimator of the true degrees of freedom, $\frac{1}{\sigma^2} \sum_{j=1}^{np} \mathrm{Cov}(\hat{u}_j, x_j)$, averaged over 500 data sets. In addition, we plot the pointwise intervals of the estimated degrees of freedom (mean $\pm$ 2 $\times$ standard deviation). Note that (18) cannot be computed in practice, since it requires knowledge of the unknown quantity $\mathbf{u}$. Results are displayed in Figure 1. We see that the estimated degrees of freedom are quite close to the true degrees of freedom.

4. Simulation studies

We compare convex clustering with $q = 1$ and $q = 2$ to the following proposals:

1. Single linkage hierarchical clustering with the dissimilarity matrix defined by the Euclidean distance between two observations.
2. The k-means clustering algorithm (Lloyd, 1982).
3. Average linkage hierarchical clustering with the dissimilarity matrix defined by the Euclidean distance between two observations.

We apply convex clustering (3) with $q = \{1, 2\}$ using the `R` package `cvxclustr` (Chi and Lange, 2014b). In order to obtain the entire solution path for convex

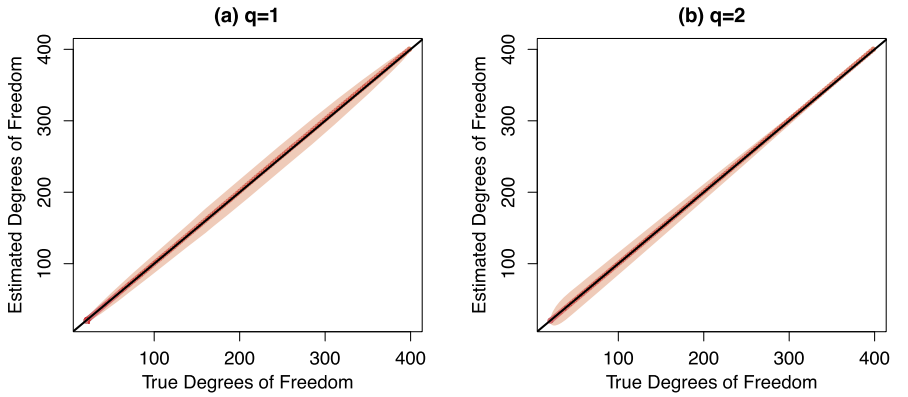

FIG 1. *We compare the true degrees of freedom of convex clustering (x-axis), given in (18), to the proposed unbiased estimators of the degrees of freedom (y-axis), given in Lemmas 8 and 9. Panels (a) and (b) contain the results for convex clustering with* $q = 1$ *and* $q = 2$*, respectively. The red line is the mean of the estimated degrees of freedom for convex clustering over 500 data sets, obtained by varying the tuning parameter* λ*. The shaded bands indicate the point-wise intervals of the estimated degrees of freedom (mean* $\pm$ *2* $\times$ *standard deviation), over 500 data sets. The black line indicates* $y = x$*.*

clustering, we use a fine grid of λ values for (3), in a range guided by Lemma 5. We apply the other methods by allowing the number of clusters to vary over a range from 1 to n clusters. To evaluate and quantify the performance of the different clustering methods, we use the Rand index (Rand, 1971). A high value of the Rand index indicates good agreement between the true and estimated clusters.

We consider two different types of clusters in our simulation studies: Gaussian clusters and non-convex clusters.

4.1. Gaussian clusters

We generate Gaussian clusters with $K = 2$ and $K = 3$ by randomly assigning each observation to a cluster with equal probability. For $K = 2$, we create the mean vectors $\boldsymbol{\mu}_1 = \mathbf{1}_p$ and $\boldsymbol{\mu}_2 = -\mathbf{1}_p$. For $K = 3$, we create the mean vectors $\boldsymbol{\mu}_1 = -3 \cdot \mathbf{1}_p$, $\boldsymbol{\mu}_2 = \mathbf{0}_p$, and $\boldsymbol{\mu}_3 = 3 \cdot \mathbf{1}_p$. We then generate the $n \times p$ data matrix $\mathbf{X}$ according to $\mathbf{X}_{i\cdot} \sim \text{MVN}(\boldsymbol{\mu}_k, \sigma^2\mathbf{I})$ for $i \in D_k$. We consider $n = p = 30$ and $\sigma = \{1, 2\}$. The Rand indices for $K = 2$ and $K = 3$, averaged over 200 data sets, are summarized in Figures 2 and 3, respectively.

Recall from Section 2.2 that there is a connection between convex clustering and single linkage clustering. However, we note that the two clustering methods are not equivalent. From Figure 2(a), we see that single linkage hierarchical clustering performs very similarly to convex clustering with $q = 2$ when the signal-to-noise ratio is high. However, from Figure 2(b), we see that single linkage hierarchical clustering outperforms convex clustering with $q = 2$ when the signal-to-noise ratio is low.

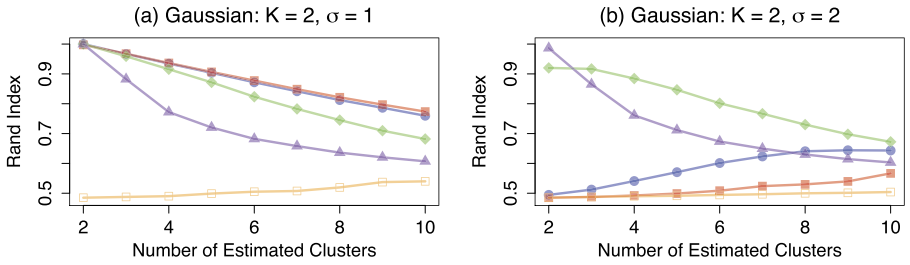

FIG 2. *Simulation results for Gaussian clusters with $K = 2$, $n = p = 30$, averaged over 200 data sets, for two noise levels $\sigma = \{1, 2\}$. Colored lines correspond to single linkage hierarchical -clustering (●—●—●), average linkage hierarchical clustering (◆—◆—◆), k-means clustering (1074i03), convex clustering with $q = 1$ (□—□—□), and convex clustering with $q = 2$ (●—●—●).*

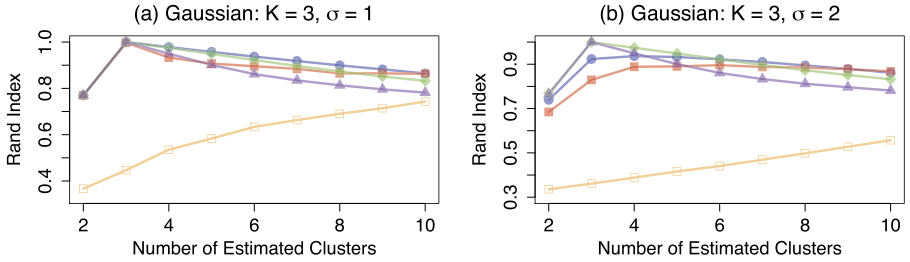

FIG 3. *Simulation results for Gaussian clusters with $K = 3$, $n = p = 30$, averaged over 200 data sets, for two noise levels $\sigma = \{1, 2\}$. Line types are as described in Figure 2.*

We also established a connection between convex clustering and k-means clustering in Section 2.3. From Figure 2(a), we see that k-means clustering and convex clustering with $q = 2$ perform similarly when two clusters are estimated and the signal-to-noise ratio is high. In this case, the first term in (14) can be made extremely small if the clusters are correctly estimated, and so both k-means and convex clustering yield the same (correct) cluster estimates. In contrast, when the signal-to-noise ratio is low, the first term in (14) is relatively large regardless of whether or not the clusters are correctly estimated, and so convex clustering focuses on minimizing the penalty term in (14). Therefore, when convex clustering with $q = 2$ estimates two clusters, one cluster is of size one and the other is of size $n - 1$, as discussed in Section 2.3. Figure 2(b) illustrates this phenomenon when both methods estimate two clusters: convex clustering with $q = 2$ has a Rand index of approximately 0.5 while k-means clustering has a Rand index of one.

All methods outperform convex clustering with $q = 1$. Moreover, k-means clustering and average linkage hierarchical clustering outperform single linkage hierarchical clustering and convex clustering when the signal-to-noise ratio is low. This suggests that the minimum signal needed for convex clustering to identify the correct clusters may be larger than that of average linkage hierarchical clustering and k-means clustering. We see similar results for the case when $K = 3$ in Figure 3.

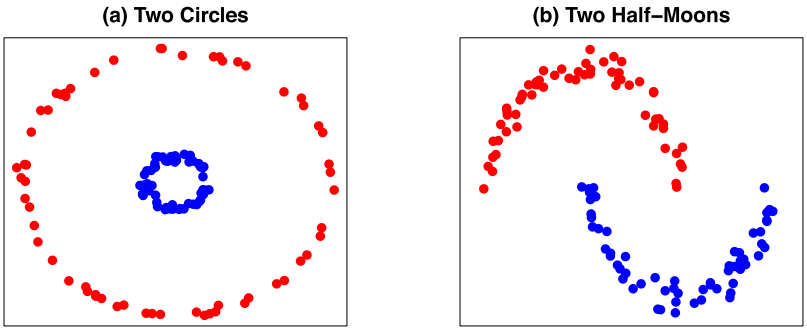

FIG 4. *Illustrations of two circles clusters and two half-moons clusters with* $n = 100$.

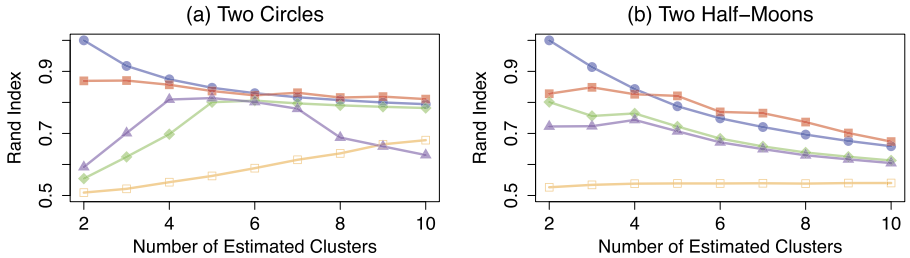

FIG 5. *Simulation results for the two circles and two half-moons clusters with* $n = 100$, *averaged over 200 data sets. Line types are as described in Figure 2.*

4.2. Non-convex clusters

We consider two types of non-convex clusters: two *circles clusters* (Ng, Jordan and Weiss, 2002) and two *half-moon clusters* (Hocking et al., 2011; Chi and Lange, 2014a). For two circles clusters, we generate 50 data points from each of the two circles that are centered at $(0, 0)$ with radiuses two and 10, respectively. We then add Gaussian random noise with mean zero and standard deviation 0.1 to each data point. For two half-moon clusters, we generate 50 data points from each of the two half-circles that are centered at $(0, 0)$ and $(30, 3)$ with radius 30, respectively. We then add Gaussian random noise with mean zero and standard deviation one to each data point. Illustrations of both types of clusters are given in Figure 4. The Rand indices for both types of clusters, averaged over 200 data sets, are summarized in Figure 5.

We see from Figure 5 that convex clustering with $q = 2$ and single linkage hierarchical clustering have similar performance, and that they outperform all of the other methods. Single linkage hierarchical clustering is able to identify non-convex clusters since it is an agglomerative algorithm that merges the closest pair of observations not yet belonging to the same cluster into one cluster. In contrast, average linkage hierarchical clustering and k-means clustering are known to perform poorly on identifying non-convex clusters (Ng, Jordan and Weiss, 2002; Hocking et al., 2011). Again, convex clustering with $q = 1$ has the worst performance.

TABLE 1
Simulation study to evaluate the performance of the extended BIC for tuning parameter selection for convex clustering with $q = 2$. Results are reported over 100 simulated data sets. We report the proportion of data sets for which the correct number of clusters was identified, and the average Rand index.

	$\text{eBIC}_{2,\gamma}$	Correct number of clusters	Rand index
Gaussian clusters, $K = 2$	$\gamma = 0$	0.94	0.9896
	$\gamma = 0.5$	0.98	0.9991
	$\gamma = 0.75$	0.99	0.9995
	$\gamma = 1$	0.99	0.9995
Gaussian clusters, $K = 3$	$\gamma = 0$	0.06	0.7616
	$\gamma = 0.5$	0.59	0.9681
	$\gamma = 0.75$	0.70	0.9768
	$\gamma = 1$	0.84	0.9873

4.3. Selection of the tuning parameter λ

Convex clustering (3) involves a tuning parameter λ, which determines the estimated number of clusters. Some authors have suggested a hold-out validation approach to select tuning parameters for clustering problems (see, for instance, Tan and Witten, 2014; Chi, Allen and Baraniuk, 2014). In this section, we present an alternative approach for selecting λ using the unbiased estimators of the degrees of freedom derived in Section 3.3.

The Bayesian Information Criterion (BIC) developed in Schwarz (1978) has been used extensively for model selection. However, it is known that the BIC does not perform well unless the number of observations is far larger than the number of parameters (Chen and Chen, 2008, 2012). For convex clustering (3), the number of observations is equal to the number of parameters. Thus, we consider the extended BIC (Chen and Chen, 2008, 2012), defined as

$$\text{eBIC}_{q,\gamma} = np \cdot \log\left(\frac{\text{RSS}_q}{np}\right) + \hat{\text{df}}_q \cdot \log(np) + 2\gamma \cdot \hat{\text{df}}_q \cdot \log(np), \tag{19}$$

where $\text{RSS}_q = \|\mathbf{x} - \hat{\mathbf{u}}_q\|_2^2$, $\hat{\mathbf{u}}_q$ is the convex clustering estimate for a given value of q and λ, $\gamma \in [0, 1]$, and $\hat{\text{df}}_q$ is given in Section 3.3. Note that we suppress the dependence of $\hat{\mathbf{u}}_q$ and $\hat{\text{df}}_q$ on λ for notational convenience. We see that when $\gamma = 0$, the extended BIC reduces to the classical BIC.

To evaluate the performance of the extended BIC in selecting the number of clusters, we generate Gaussian clusters with $K = 2$ and $K = 3$ as described in Section 4.1, with $n = p = 20$, and $\sigma = 0.5$. We perform convex clustering with $q = 2$ over a fine grid of λ, and select the value of λ for which the quantity $\text{eBIC}_{q,\gamma}$ is minimized. We consider $\gamma \in \{0, 0.5, 0.75, 1\}$. Table 1 reports the proportion of datasets for which the correct number of clusters was identified, as well as the average Rand index.

From Table 1, we see that the extended BIC is able to select the true number of clusters accurately for $K = 2$. When $K = 3$, the classical BIC ($\gamma = 0$) fails to select the true number of clusters. In contrast, the extended BIC with $\gamma = 1$ has the best performance.

5. Discussion

Convex clustering recasts the clustering problem into a penalized regression problem. By studying its dual problem, we show that there is a connection between convex clustering and single linkage hierarchical clustering. In addition, we establish a connection between convex clustering and k-means clustering. We also establish several statistical properties of convex clustering. Through some numerical studies, we illustrate that the performance of convex clustering may not be appealing relative to traditional clustering methods, especially when the signal-to-noise ratio is low.

Many authors have proposed a modification to the convex clustering problem (1),

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times p}}{\text{minimize}}\ \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{X}_{i\cdot}-\mathbf{U}_{i\cdot}\|_2^2+\lambda \mathrm{Q}_q(\mathbf{W},\mathbf{U}), \tag{20}$$

where $\mathbf{W}$ is an $n\times n$ symmetric matrix of positive weights, and $\mathrm{Q}_q(\mathbf{W},\mathbf{U})=\sum_{i<i'}W_{ii'}\|\mathbf{U}_{i\cdot}-\mathbf{U}_{i'\cdot}\|_q$ (Pelckmans et al., 2005; Hocking et al., 2011; Lindsten, Ohlsson and Ljung, 2011; Chi and Lange, 2014a). For instance, the weights can be defined as $W_{ii'}=\exp\left(-\phi\|\mathbf{X}_{i\cdot}-\mathbf{X}_{i'\cdot}\|_2^2\right)$ for some constant $\phi>0$. This yields better empirical performance than (1) (Hocking et al., 2011; Chi and Lange, 2014a). We leave an investigation of the properties of (20) to future work.

Appendix A: Proof of Lemmas 2–3

Proof of Lemma 2. We rewrite problem (3) as

$$\underset{\mathbf{u}\in\mathbb{R}^{np},\boldsymbol{\eta}_1\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2+\lambda \mathrm{P}_q(\boldsymbol{\eta}_1)\qquad \text{subject to } \boldsymbol{\eta}_1=\mathbf{D}\mathbf{u},$$

with the Lagrangian function

$$\mathcal{L}(\mathbf{u},\boldsymbol{\eta}_1,\boldsymbol{\nu})=\frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2+\lambda \mathrm{P}_q(\boldsymbol{\eta}_1)+\boldsymbol{\nu}^T(\mathbf{D}\mathbf{u}-\boldsymbol{\eta}_1), \tag{A-1}$$

where $\boldsymbol{\nu}\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}$ is the Lagrangian dual variable. In order to derive the dual problem, we need to minimize the Lagrangian function over the primal variables $\mathbf{u}$ and $\boldsymbol{\eta}_1$. Recall from (4) that $\mathrm{P}_q^*(\cdot)$ is the dual norm of $\mathrm{P}_q(\cdot)$. It can be shown that

$$\inf_{\boldsymbol{\eta}_1\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}}\mathcal{L}(\mathbf{u},\boldsymbol{\eta}_1,\boldsymbol{\nu})=\begin{cases}\frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2+\boldsymbol{\nu}^T\mathbf{D}\mathbf{u} & \text{if}\quad \mathrm{P}_q^*(\boldsymbol{\nu})\le\lambda,\\ -\infty & \text{otherwise,}\end{cases}$$

and

$$\inf_{\boldsymbol{\eta}_1\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]},\mathbf{u}\in\mathbb{R}^{np}}\mathcal{L}(\mathbf{u},\boldsymbol{\eta}_1,\boldsymbol{\nu})=\begin{cases}-\frac{1}{2}\|\mathbf{x}-\mathbf{D}^T\boldsymbol{\nu}\|_2^2+\frac{1}{2}\|\mathbf{x}\|_2^2 & \text{if}\quad \mathrm{P}_q^*(\boldsymbol{\nu})\le\lambda.\\ -\infty & \text{otherwise.}\end{cases}$$

Therefore, the dual problem for (3) is

$$\underset{\boldsymbol{\nu}\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{x}-\mathbf{D}^T\boldsymbol{\nu}\|_2^2 \qquad \text{subject to } \mathrm{P}_q^*(\boldsymbol{\nu})\le\lambda. \tag{A-2}$$

We now establish an explicit relationship between the solution to convex clustering and its dual problem. Differentiating the Lagrangian function (A-1) with respect to $\mathbf{u}$ and setting it equal to zero, we obtain

$$\hat{\mathbf{u}}=\mathbf{x}-\mathbf{D}^T\hat{\boldsymbol{\nu}},$$

where $\hat{\boldsymbol{\nu}}$ is the solution to the dual problem, which satisfies $\mathrm{P}_q^*(\hat{\boldsymbol{\nu}})\le\lambda$ by (A-2). Multiplying both sides by $\mathbf{D}$, we obtain the relationship (6). □

Proof of Lemma 3. We rewrite (7) as

$$\underset{\boldsymbol{\gamma}\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]},\boldsymbol{\eta}_2\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{D}\mathbf{x}-\boldsymbol{\gamma}\|_2^2+\lambda\mathrm{P}_q(\boldsymbol{\eta}_2) \qquad \text{subject to } \boldsymbol{\eta}_2=\boldsymbol{\gamma},$$

with the Lagrangian function

$$\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\eta}_2,\boldsymbol{\nu}')=\frac{1}{2}\|\mathbf{D}\mathbf{x}-\boldsymbol{\gamma}\|_2^2+\lambda\mathrm{P}_q(\boldsymbol{\eta}_2)+(\boldsymbol{\nu}')^T(\boldsymbol{\gamma}-\boldsymbol{\eta}_2), \tag{A-3}$$

where $\boldsymbol{\nu}'\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}$ is the Lagrangian dual variable. In order to derive the dual problem, we minimize the Lagrangian function over the primal variables $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}_2$. It can be shown that

$$\inf_{\boldsymbol{\eta}_2\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}}\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\eta}_2,\boldsymbol{\nu}')=\begin{cases}\frac{1}{2}\|\mathbf{D}\mathbf{x}-\boldsymbol{\gamma}\|_2^2+(\boldsymbol{\nu}')^T\boldsymbol{\gamma} & \text{if}\quad \mathrm{P}_q^*(\boldsymbol{\nu}')\le\lambda,\\ -\infty & \text{otherwise,}\end{cases}$$

and

$$\inf_{\boldsymbol{\eta}_2\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]},\boldsymbol{\gamma}\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}}\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\eta}_2,\boldsymbol{\nu}')=\begin{cases}-\frac{1}{2}\|\mathbf{D}\mathbf{x}-\boldsymbol{\nu}'\|_2^2+\frac{1}{2}\|\mathbf{D}\mathbf{x}\|_2^2 & \text{if}\quad \mathrm{P}_q^*(\boldsymbol{\nu}')\le\lambda.\\ -\infty & \text{otherwise.}\end{cases}$$

Therefore, the dual problem for (7) is

$$\underset{\boldsymbol{\nu}'\in\mathbb{R}^{\left[p\cdot\binom{n}{2}\right]}}{\text{minimize}}\ \frac{1}{2}\|\mathbf{D}\mathbf{x}-\boldsymbol{\nu}'\|_2^2 \qquad \text{subject to } \mathrm{P}_q^*(\boldsymbol{\nu}')\le\lambda. \tag{A-4}$$

We now establish an explicit relationship between the solution to (7) and its dual problem. Differentiating the Lagrangian function (A-3) with respect to $\boldsymbol{\gamma}$ and setting it equal to zero, we obtain

$$\hat{\boldsymbol{\gamma}}=\mathbf{D}\mathbf{x}-\hat{\boldsymbol{\nu}}',$$

where $\hat{\boldsymbol{\nu}}'$ is the solution to the dual problem, which we know from (A-4) satisfies $\mathrm{P}_q^*(\hat{\boldsymbol{\nu}}')\le\lambda$. □

Appendix B: Proof of Lemma 5

Proof of Lemma 5. Since $\mathbf{D}$ is not of full rank by Lemma 1(i), the solution to (5) in the absence of constraint is not unique, and takes the form

$$\begin{aligned}\hat{\boldsymbol{\nu}} &= (\mathbf{D}\mathbf{D}^T)^\dagger \mathbf{D}\mathbf{x} + (\mathbf{I} - \mathbf{D}(\mathbf{D}^T\mathbf{D})^\dagger \mathbf{D}^T)\boldsymbol{\omega} \\ &= (\mathbf{D}^T)^\dagger \mathbf{x} + (\mathbf{I} - \mathbf{D}\mathbf{D}^\dagger)\boldsymbol{\omega} \\ &= \frac{1}{n}\mathbf{D}\mathbf{x} + (\mathbf{I} - \frac{1}{n}\mathbf{D}\mathbf{D}^T)\boldsymbol{\omega},\end{aligned} \tag{B-1}$$

for $\boldsymbol{\omega} \in \mathbb{R}^{[p\cdot\binom{n}{2}]}$. The second equality follows from Lemma 1(iii) and the last equality follows from Lemma 1(ii).

Let $\hat{\mathbf{u}}$ be the solution to (3). Substituting $\hat{\boldsymbol{\nu}}$ given in (B-1) into (6), we obtain

$$\begin{aligned}\mathbf{D}\hat{\mathbf{u}} &= \mathbf{D}\mathbf{x} - \mathbf{D}\mathbf{D}^T\hat{\boldsymbol{\nu}} \\ &= \mathbf{D}\mathbf{x} - \frac{1}{n}\mathbf{D}\mathbf{D}^T\mathbf{D}\mathbf{x} - \mathbf{D}\mathbf{D}^T\boldsymbol{\omega} + \frac{1}{n}\mathbf{D}\mathbf{D}^T\mathbf{D}\mathbf{D}^T\boldsymbol{\omega} \\ &= \mathbf{D}\mathbf{x} - \mathbf{D}\mathbf{x} - \mathbf{D}\mathbf{D}^T\boldsymbol{\omega} + \mathbf{D}\mathbf{D}^T\boldsymbol{\omega} \\ &= \mathbf{0}.\end{aligned}$$

Recall from Definition 1 that all observations are estimated to belong to the same cluster if $\mathbf{D}\hat{\mathbf{u}} = \mathbf{0}$. For any $\hat{\boldsymbol{\nu}}$ in (B-1), picking $\lambda = \mathrm{P}_q^*(\hat{\boldsymbol{\nu}})$ guarantees that the constraint on the dual problem (5) is inactive, and therefore that convex clustering has a trivial solution of $\mathbf{D}\hat{\mathbf{u}} = \mathbf{0}$.

Since $\hat{\boldsymbol{\nu}}$ is not unique, $\mathrm{P}_q^*(\hat{\boldsymbol{\nu}})$ is not unique. In order to obtain the smallest tuning parameter λ such that $\mathbf{D}\hat{\mathbf{u}} = \mathbf{0}$, we take

$$\lambda_{\text{upper}} := \min_{\boldsymbol{\omega}\in\mathbb{R}^{[p\cdot\binom{n}{2}]}} \mathrm{P}_q^*\left(\frac{1}{n}\mathbf{D}\mathbf{x} + \left(\mathbf{I} - \frac{1}{n}\mathbf{D}\mathbf{D}^T\right)\boldsymbol{\omega}\right).$$

Any tuning parameter $\lambda \geq \lambda_{\text{upper}}$ results in an estimate for which all observations belong to a single cluster. The proof is completed by recalling the definition of the dual norm $\mathrm{P}_q^*(\cdot)$ in (4). □

Appendix C: Proof of Lemmas 6–7

To prove Lemmas 6 and 7, we need a lemma on the tail bound for quadratic forms of independent sub-Gaussian random variables.

Lemma 10. *(Hanson and Wright, 1971) Let $\mathbf{z}$ be a vector of independent sub-Gaussian random variables with mean zero and variance σ^2. Let $\mathbf{M}$ be a symmetric matrix. Then, there exists some constants $c_1, c_2 > 0$ such that for any $t > 0$,*

$$\Pr\left(\mathbf{z}^T\mathbf{M}\mathbf{z} \geq t + \sigma^2\mathrm{tr}(\mathbf{M})\right) \leq \exp\left\{-\min\left(\frac{c_1 t^2}{\sigma^4\|\mathbf{M}\|_{\mathrm{F}}}, \frac{c_2 t}{\sigma^2\|\mathbf{M}\|_{\mathrm{sp}}}\right)\right\},$$

where $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{\mathrm{sp}}$ are the Frobenius norm and spectral norm, respectively.

In order to simplify our analysis, we start by reformulating (3) as in Liu, Yuan and Ye (2013). Let $\mathbf{D} = \mathbf{A}\Lambda\mathbf{V}_\beta^T$ be the *singular value decomposition* of $\mathbf{D}$, where $\mathbf{A} \in \mathbb{R}^{[p\cdot\binom{n}{2}]\times p(n-1)}$, $\Lambda \in \mathbb{R}^{p(n-1)\times p(n-1)}$, and $\mathbf{V}_\beta \in \mathbb{R}^{np\times p(n-1)}$. Construct $\mathbf{V}_\alpha \in \mathbb{R}^{np\times p}$ such that $\mathbf{V} = [\mathbf{V}_\alpha, \mathbf{V}_\beta] \in \mathbb{R}^{np\times np}$ is an orthogonal matrix, that is, $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$. Note that $\mathbf{V}_\alpha^T\mathbf{V}_\beta = 0$.

Let $\boldsymbol{\beta} = \mathbf{V}_\beta^T\mathbf{u} \in \mathbb{R}^{p(n-1)}$ and $\boldsymbol{\alpha} = \mathbf{V}_\alpha^T\mathbf{u} \in \mathbb{R}^p$. Also, let $\lambda' = \frac{\lambda}{np}$. Optimization problem (3) then becomes

$$\underset{\boldsymbol{\alpha}\in\mathbb{R}^p,\boldsymbol{\beta}\in\mathbb{R}^{p(n-1)}}{\text{minimize}} \ \frac{1}{2np}\|\mathbf{x} - \mathbf{V}_\alpha\boldsymbol{\alpha} - \mathbf{V}_\beta\boldsymbol{\beta}\|^2 + \lambda' \mathrm{P}_q(\mathbf{Z}\boldsymbol{\beta}), \tag{C-1}$$

where $\mathbf{Z} = \mathbf{A}\Lambda \in \mathbb{R}^{[p\cdot\binom{n}{2}]\times p(n-1)}$. Note that $\operatorname{rank}(\mathbf{Z}) = p(n-1)$ and therefore, there exists a pseudo-inverse $\mathbf{Z}^\dagger \in \mathbb{R}^{p(n-1)\times[p\cdot\binom{n}{2}]}$ such that $\mathbf{Z}^\dagger\mathbf{Z} = \mathbf{I}$. Recall from Section 1 that the set $\mathcal{C}(i,i')$ contains the row indices of $\mathbf{D}$ such that $\mathbf{D}_{\mathcal{C}(i,i')}\mathbf{u} = \mathbf{U}_{i\cdot} - \mathbf{U}_{i'\cdot}$. Let the submatrices $\mathbf{Z}_{\mathcal{C}(i,i')}$ and $\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}$ denote the rows of $\mathbf{Z}$ and the columns of $\mathbf{Z}^\dagger$, respectively, corresponding to the indices in the set $\mathcal{C}(i,i')$. By Lemma 1(v),

$$\begin{aligned}\Lambda_{\min}(\mathbf{Z}) = \Lambda_{\min}(\mathbf{D}) &= \frac{1}{\Lambda_{\max}(\mathbf{Z}^\dagger)} = \sqrt{n}\\ \Lambda_{\max}(\mathbf{Z}) = \Lambda_{\max}(\mathbf{D}) &= \frac{1}{\Lambda_{\min}(\mathbf{Z}^\dagger)} = \sqrt{n}.\end{aligned} \tag{C-2}$$

Let $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ denote the solution to (C-1).

Proof of Lemma 6. We establish a finite sample bound for the prediction error of convex clustering with $q = 1$ by analyzing (C-1). First, note that $\hat{\mathbf{u}} = \mathbf{V}_\alpha\hat{\boldsymbol{\alpha}} + \mathbf{V}_\beta\hat{\boldsymbol{\beta}}$ and $\mathbf{u} = \mathbf{V}_\alpha\boldsymbol{\alpha} + \mathbf{V}_\beta\boldsymbol{\beta}$. Thus, $\frac{1}{2np}\|\hat{\mathbf{u}} - \mathbf{u}\|^2 = \frac{1}{2np}\|\mathbf{V}_\alpha(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \mathbf{V}_\beta(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2$. Recall from (2) that $\mathrm{P}_1(\mathbf{Z}\boldsymbol{\beta}) = \|\mathbf{Z}\boldsymbol{\beta}\|_1$. By the definition of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, we have

$$\frac{1}{2np}\|\mathbf{x} - (\mathbf{V}_\alpha\hat{\boldsymbol{\alpha}} + \mathbf{V}_\beta\hat{\boldsymbol{\beta}})\|^2 + \lambda'\|\mathbf{Z}\hat{\boldsymbol{\beta}}\|_1 \le \frac{1}{2np}\|\mathbf{x} - (\mathbf{V}_\alpha\boldsymbol{\alpha} + \mathbf{V}_\beta\boldsymbol{\beta})\|^2 + \lambda'\|\mathbf{Z}\boldsymbol{\beta}\|_1,$$

implying

$$\frac{1}{2np}\|\mathbf{V}_\alpha(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \mathbf{V}_\beta(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 + \lambda'\|\mathbf{Z}\hat{\boldsymbol{\beta}}\|_1 \le \frac{1}{np}\mathrm{G}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) + \lambda'\|\mathbf{Z}\boldsymbol{\beta}\|_1, \tag{C-3}$$

where $\mathrm{G}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}) = \boldsymbol{\epsilon}^T\left[\mathbf{V}_\alpha(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \mathbf{V}_\beta(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right]$. Recall that $\mathbf{V}_\alpha^T\mathbf{V}_\alpha = \mathbf{I}$ and $\mathbf{V}_\alpha^T\mathbf{V}_\beta = \mathbf{0}$. By the optimality condition of (C-1),

$$\begin{aligned}\hat{\boldsymbol{\alpha}} &= \mathbf{V}_\alpha^T(\mathbf{x} - \mathbf{V}_\beta\hat{\boldsymbol{\beta}})\\ &= \mathbf{V}_\alpha^T\left(\mathbf{V}_\alpha\boldsymbol{\alpha} + \mathbf{V}_\beta\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{V}_\beta\hat{\boldsymbol{\beta}}\right)\\ &= \boldsymbol{\alpha} + \mathbf{V}_\alpha^T\boldsymbol{\epsilon}.\end{aligned}$$

Therefore, substituting $\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} = \mathbf{V}_\alpha^T \boldsymbol{\epsilon}$ into $\mathrm{G}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$, we obtain

$$\begin{aligned}
\frac{1}{np}\left|\mathrm{G}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})\right| &= \frac{1}{np}\left|\boldsymbol{\epsilon}^T\left[\mathbf{V}_\alpha(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) + \mathbf{V}_\beta(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right]\right| \\
&= \frac{1}{np}\left|\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T\mathbf{V}_\beta(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right| \\
&\leq \frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} + \frac{1}{np}\left|\boldsymbol{\epsilon}^T\mathbf{V}_\beta(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right| \\
&= \frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} + \frac{1}{np}\left|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right| \\
&\leq \frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} + \frac{1}{np}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger\|_\infty\|\mathbf{Z}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|_1.
\end{aligned}$$

We now establish bounds for $\frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon}$ and $\frac{1}{np}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger\|_\infty$ that hold with high probability.

Bound for $\frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon}$:

First, note that $\mathbf{V}_\alpha\mathbf{V}_\alpha^T$ is a projection matrix of rank p, and therefore $\|\mathbf{V}_\alpha\mathbf{V}_\alpha^T\|_{\mathrm{sp}} = 1$ and $\|\mathbf{V}_\alpha\mathbf{V}_\alpha^T\|_{\mathrm{F}} = p$. By Lemma 10 and taking $\mathbf{z} = \boldsymbol{\epsilon}$ and $\mathbf{M} = \mathbf{V}_\alpha\mathbf{V}_\alpha^T$, we have that

$$\Pr\left(\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} \geq t + \sigma^2 p\right) \leq \exp\left\{-\min\left(\frac{c_1 t^2}{\sigma^4 p}, \frac{c_2 t}{\sigma^2}\right)\right\},$$

where c_1 and c_2 are constants in Lemma 10. Picking $t = \sigma^2\sqrt{p\log(np)}$, we have

$$\begin{aligned}
&\Pr\left(\frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} \geq \sigma^2\left[\frac{1}{n} + \sqrt{\frac{\log(np)}{n^2 p}}\right]\right) \\
&\quad\leq \exp\left\{-\min\left(c_1\log(np), c_2\sqrt{p\log(np)}\right)\right\}. \qquad \text{(C-4)}
\end{aligned}$$

Bound for $\frac{1}{np}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger\|_\infty$:

Let e_j be a vector of length $p\cdot\binom{n}{2}$ with a one in the jth entry and zeroes in the remaining entries. Let $v_j = e_j^T(\mathbf{Z}^\dagger)^T\mathbf{V}_\beta^T\boldsymbol{\epsilon}$. Using the fact that $\Lambda_{\max}(\mathbf{V}_\beta) = 1$ and $\Lambda_{\max}(\mathbf{Z}^\dagger) = \frac{1}{\sqrt{n}}$ (C-2), we know that each v_j is a sub-Gaussian random variable with zero mean and variance at most $\frac{\sigma^2}{n}$. Therefore, by the union bound,

$$\Pr\left(\max_j |v_j| \geq z\right) \leq p\cdot\binom{n}{2}\cdot\Pr\left(|v_j| \geq z\right) \leq 2p\cdot\binom{n}{2}\exp\left(-\frac{nz^2}{2\sigma^2}\right).$$

Picking $z = 2\sigma\sqrt{\frac{\log\left(p\cdot\binom{n}{2}\right)}{n}}$, we obtain

$$\Pr\left(\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger\|_\infty \geq 2\sigma\sqrt{\frac{\log\left(p\cdot\binom{n}{2}\right)}{n}}\right) \leq \frac{2}{p\cdot\binom{n}{2}}. \qquad \text{(C-5)}$$

Combining the two upper bounds: Setting $\lambda' > 4\sigma\sqrt{\frac{\log\left(p\cdot\binom{n}{2}\right)}{n^3p^2}}$ and combining the results from (C-4) and (C-5), we obtain

$$\frac{1}{np}\mathrm{G}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) \leq \sigma^2\left[\frac{1}{n}+\sqrt{\frac{\log(np)}{n^2p}}\right]+\frac{\lambda'}{2}\|\mathbf{Z}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_1 \tag{C-6}$$

with probability at least $1-\frac{2}{p\cdot\binom{n}{2}}-\exp\left\{-\min\left(c_1\log(np), c_2\sqrt{p\log(np)}\right)\right\}$. Substituting (C-6) into (C-3), we obtain

$$\begin{aligned}&\frac{1}{2np}\|\mathbf{V}_\alpha(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})+\mathbf{V}_\beta(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2+\lambda'\|\mathbf{Z}\hat{\boldsymbol{\beta}}\|_1\\ &\leq \sigma^2\left[\frac{1}{n}+\sqrt{\frac{\log(np)}{n^2p}}\right]+\frac{\lambda'}{2}\|\mathbf{Z}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_1+\lambda'\|\mathbf{Z}\boldsymbol{\beta}\|_1.\end{aligned}$$

We get Lemma 6 by an application of the triangle inequality and by rearranging the terms. □

Proof of Lemma 7. We establish a finite sample bound for the prediction error of convex clustering with $q=2$ by analyzing (C-1). Recall from (2) that $\mathrm{P}_2(\mathbf{Z}\boldsymbol{\beta}) = \sum_{i<i'}\|\mathbf{Z}_{\mathcal{C}(i,i')}\boldsymbol{\beta}\|_2$. By the definition of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, we have

$$\begin{aligned}&\frac{1}{2np}\|\mathbf{x}-(\mathbf{V}_\alpha\hat{\boldsymbol{\alpha}}+\mathbf{V}_\beta\hat{\boldsymbol{\beta}})\|^2+\lambda'\sum_{i<i'}\|\mathbf{Z}_{\mathcal{C}(i,i')}\hat{\boldsymbol{\beta}}\|_2\\ &\quad\leq \frac{1}{2np}\|\mathbf{x}-(\mathbf{V}_\alpha\boldsymbol{\alpha}+\mathbf{V}_\beta\boldsymbol{\beta})\|^2+\lambda'\sum_{i<i'}\|\mathbf{Z}_{\mathcal{C}(i,i')}\boldsymbol{\beta}\|_2,\end{aligned}$$

implying

$$\begin{aligned}&\frac{1}{2np}\|\mathbf{V}_\alpha(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})+\mathbf{V}_\beta(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2+\lambda'\sum_{i<i'}\|\mathbf{Z}_{\mathcal{C}(i,i')}\hat{\boldsymbol{\beta}}\|_2\\ &\quad\leq \frac{1}{np}\mathrm{G}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})+\lambda'\sum_{i<i'}\|\mathbf{Z}_{\mathcal{C}(i,i')}\boldsymbol{\beta}\|_2,\end{aligned} \tag{C-7}$$

where $\mathrm{G}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) = \boldsymbol{\epsilon}^T\left[\mathbf{V}_\alpha(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})+\mathbf{V}_\beta(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right]$. Again, by the optimality condition of (C-1), we have that $\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha} = \mathbf{V}_\alpha^T\boldsymbol{\epsilon}$. Substituting this into $\frac{1}{np}\mathrm{G}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})$, we obtain

$$\begin{aligned}\frac{1}{np}\left|\mathrm{G}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}})\right| &= \frac{1}{np}\left|\boldsymbol{\epsilon}^T\left[\mathbf{V}_\alpha(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})+\mathbf{V}_\beta(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right]\right|\\ &= \frac{1}{np}\left|\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon}+\boldsymbol{\epsilon}^T\mathbf{V}_\beta(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right|\\ &\leq \frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon}+\frac{1}{np}\left|\boldsymbol{\epsilon}^T\mathbf{V}_\beta(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right|\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} + \frac{1}{np}\left|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger\mathbf{Z}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\right| \\
&= \frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} + \frac{1}{np}\left|\sum_{i<i'}(\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')})(\mathbf{Z}_{\mathcal{C}(i,i')}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\right| \\
&\le \frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} + \frac{1}{np}\sum_{i<i'}\left|(\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')})(\mathbf{Z}_{\mathcal{C}(i,i')}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}))\right| \\
&\le \frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} + \frac{1}{np}\sum_{i<i'}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}\|_2\|\mathbf{Z}_{\mathcal{C}(i,i')}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_2 \\
&\le \frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} + \frac{1}{np}\cdot\max_{i<i'}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}\|_2\sum_{i<i'}\|\mathbf{Z}_{\mathcal{C}(i,i')}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_2,
\end{aligned}
$$

where the second inequality follows from an application of the triangle inequality and the third inequality from an application of the Cauchy-Schwarz inequality. We now establish bounds for $\frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon}$ and $\frac{1}{np}\cdot\max_{i<i'}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}\|_2$ that hold with large probability.

Bound for $\frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon}$:

This is established in the proof of Lemma 6 in (C-4), i.e.,

$$
\Pr\left(\frac{1}{np}\boldsymbol{\epsilon}^T\mathbf{V}_\alpha\mathbf{V}_\alpha^T\boldsymbol{\epsilon} \ge \sigma^2\left[\frac{1}{n}+\sqrt{\frac{\log(np)}{n^2p}}\right]\right) \le \frac{1}{np}.
$$

Bound for $\frac{1}{np}\cdot\max_{i<i'}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}\|_2$:

First, note that there are p indices in each set $\mathcal{C}(i,i')$. Therefore,

$$
\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}\|_2 \le \sqrt{p}\cdot\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}\|_\infty.
$$

Note that

$$
\frac{1}{np}\cdot\max_{i<i'}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}\|_2 \le \sqrt{\frac{1}{n^2p}}\cdot\max_{i<i'}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}\|_\infty = \sqrt{\frac{1}{n^2p}}\cdot\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger\|_\infty. \tag{C-8}
$$

Therefore, using (C-8),

$$
\begin{aligned}
&\Pr\left(\frac{1}{np}\cdot\max_{i<i'}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}\|_2 \ge 2\sigma\sqrt{\frac{\log\left(p\cdot\binom{n}{2}\right)}{n^3p}}\right) \\
&\le \Pr\left(\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger\|_\infty \ge 2\sigma\sqrt{\frac{\log\left(p\cdot\binom{n}{2}\right)}{n}}\right) \qquad \text{(C-9)} \\
&\le \frac{2}{p\cdot\binom{n}{2}},
\end{aligned}
$$

where the last inequality follows from (C-5) in the proof of Lemma 6.

Therefore, for $\lambda' > 4\sigma\sqrt{\frac{\log\left(p\cdot\binom{n}{2}\right)}{n^3 p}}$, we have $\frac{\lambda'}{2} < \frac{1}{np}\cdot\max_{i<i'}\|\boldsymbol{\epsilon}^T\mathbf{V}_\beta\mathbf{Z}^\dagger_{\mathcal{C}(i,i')}\|_2$ with probability at most $\frac{2}{p\cdot\binom{n}{2}}$. Combining the results from (C-4) and (C-9), we have that

$$\frac{1}{np}\mathrm{G}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}}) \le \sigma^2\left[\frac{1}{n}+\sqrt{\frac{\log(np)}{n^2p}}\right]+\frac{\lambda'}{2}\sum_{i<i'}\|\mathbf{Z}_{\mathcal{C}(i,i')}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_2 \tag{C-10}$$

with probability at least $1-\frac{2}{p\cdot\binom{n}{2}}-\exp\left\{-\min\left(c_1\log(np), c_2\sqrt{p\log(np)}\right)\right\}$. Substituting (C-10) into (C-7), we obtain

$$\begin{aligned}&\frac{1}{2np}\|\mathbf{V}_\alpha(\hat{\boldsymbol{\alpha}}-\boldsymbol{\alpha})+\mathbf{V}_\beta(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|^2+\lambda'\sum_{i<i'}\|\mathbf{Z}_{\mathcal{C}(i,i')}\hat{\boldsymbol{\beta}}\|_2\\ &\le\sigma^2\left[\frac{1}{n}+\sqrt{\frac{\log(np)}{n^2p}}\right]+\frac{\lambda'}{2}\sum_{i<i'}\|\mathbf{Z}_{\mathcal{C}(i,i')}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\|_2+\lambda'\sum_{i<i'}\|\mathbf{Z}_{\mathcal{C}(i,i')}\boldsymbol{\beta}\|_2.\end{aligned}$$

We get Lemma 7 by an application of the triangle inequality and by rearranging the terms. □

Appendix D: Proof of Lemma 9

Proof of Lemma 9. Directly from the dual problem (5), $\mathbf{D}^T\hat{\boldsymbol{\nu}}$ is the projection of $\mathbf{x}$ onto the convex set $K=\left\{\mathbf{D}^T\boldsymbol{\nu}:\mathrm{P}_2^*(\boldsymbol{\nu})\le\lambda\right\}$. Using the primal-dual relationship $\hat{\mathbf{u}}=\mathbf{x}-\mathbf{D}^T\hat{\boldsymbol{\nu}}$, we see that $\hat{\mathbf{u}}$ is the residual from projecting $\mathbf{x}$ onto the convex set K. By Lemma 1 of Tibshirani and Taylor (2012), $\hat{\mathbf{u}}$ is continuous and almost differentiable with respect to $\mathbf{x}$. Therefore, by Stein's formula, the degrees of freedom can be characterized as $\mathrm{E}\left[\mathrm{tr}\left(\frac{\partial\hat{\mathbf{u}}}{\partial\mathbf{x}}\right)\right]$.

Recall that $\mathbf{D}_{\mathcal{C}(i,i')}$ denotes the rows of $\mathbf{D}$ corresponding to the indices in the set $\mathcal{C}(i,i')$. Let $\hat{\mathcal{B}}_2=\{(i,i'):\|\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\|_2\neq 0\}$. By the optimality condition of (3) with $q=2$, we obtain

$$(\mathbf{x}-\hat{\mathbf{u}})=\lambda\sum_{i<i'}\mathbf{D}^T_{\mathcal{C}(i,i')}g_{\mathcal{C}(i,i')}, \tag{D-1}$$

where

$$g_{\mathcal{C}(i,i')}=\begin{cases}\frac{\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}}{\|\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\|_2} & \text{if } (i,i')\in\hat{\mathcal{B}}_2.\\ \in\{\Gamma:\|\Gamma\|_2\le 1\} & \text{if } (i,i')\notin\hat{\mathcal{B}}_2.\end{cases}$$

We define the matrix $\mathbf{D}_{-\hat{\mathcal{B}}_2}$ by removing the rows of $\mathbf{D}$ that correspond to elements in $\hat{\mathcal{B}}_2$. Let $\mathbf{P}=\left(\mathbf{I}-\mathbf{D}^T_{-\hat{\mathcal{B}}_2}(\mathbf{D}_{-\hat{\mathcal{B}}_2}\mathbf{D}^T_{-\hat{\mathcal{B}}_2})^\dagger\mathbf{D}_{-\hat{\mathcal{B}}_2}\right)$ be the projection matrix onto the complement of the space spanned by the rows of $\mathbf{D}_{-\hat{\mathcal{B}}_2}$.

By the definition of $\mathbf{D}_{-\hat{\mathcal{B}}_2}$, we obtain $\mathbf{D}_{-\hat{\mathcal{B}}_2}\hat{\mathbf{u}} = \mathbf{0}$. Therefore, $\mathbf{P}\hat{\mathbf{u}} = \hat{\mathbf{u}}$. Multiplying $\mathbf{P}$ onto both sides of (D-1), we obtain

$$\begin{aligned}\mathbf{P}\mathbf{x} - \hat{\mathbf{u}} &= \lambda \mathbf{P} \sum_{i<i'} \mathbf{D}_{\mathcal{C}(i,i')}^T g_{\mathcal{C}(i,i')} \\ &= \lambda \mathbf{P} \sum_{(i,i')\in\hat{\mathcal{B}}_2} \frac{\mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}}{\|\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\|_2},\end{aligned} \tag{D-2}$$

where the second equality follows from the fact that $\mathbf{P}\mathbf{D}_{\mathcal{C}(i,i')}^T = \mathbf{0}$ for any $(i,i') \notin \hat{\mathcal{B}}_2$.

Vaiter et al. (2014) showed that there exists a neighborhood around almost every $\mathbf{x}$ such that the set $\hat{\mathcal{B}}_2$ is locally constant with respect to $\mathbf{x}$. Therefore, the derivative of (D-2) with respect to $\mathbf{x}$ is

$$\mathbf{P} - \frac{\partial \hat{\mathbf{u}}}{\partial \mathbf{x}} = \lambda \mathbf{P} \sum_{(i,i')\in\hat{\mathcal{B}}_2} \left(\frac{\mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}}{\|\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\|_2} - \frac{\mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\hat{\mathbf{u}}^T \mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}}{\|\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\|_2^3} \right) \frac{\partial \hat{\mathbf{u}}}{\partial \mathbf{x}}, \tag{D-3}$$

using the fact that for any matrix $\mathbf{A}$ with $\|\mathbf{A}\mathbf{v}\|_2 \neq 0$, $\frac{\partial}{\partial \mathbf{v}} \frac{\mathbf{A}^T\mathbf{A}\mathbf{v}}{\|\mathbf{A}\mathbf{v}\|_2} = \frac{\mathbf{A}^T\mathbf{A}}{\|\mathbf{A}\mathbf{v}\|_2} - \frac{\mathbf{A}^T\mathbf{A}\mathbf{v}\mathbf{v}^T\mathbf{A}^T\mathbf{A}}{\|\mathbf{A}\mathbf{v}\|_2^3}$.

Solving (D-3) for $\frac{\partial \hat{\mathbf{u}}}{\partial \mathbf{x}}$, we have

$$\begin{aligned}\frac{\partial \hat{\mathbf{u}}}{\partial \mathbf{x}} = \Bigg[\mathbf{I} + \lambda \mathbf{P} \sum_{(i,i')\in\hat{\mathcal{B}}_2} \Bigg(&\frac{\mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}}{\|\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\|_2} \\ &- \frac{\mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\hat{\mathbf{u}}^T \mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}}{\|\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\|_2^3} \Bigg) \Bigg]^{-1} \mathbf{P}.\end{aligned} \tag{D-4}$$

Therefore, an unbiased estimator of the degrees of freedom is of the form

$$\begin{aligned}\operatorname{tr}\left(\frac{\partial \hat{\mathbf{u}}}{\partial \mathbf{x}}\right) = \operatorname{tr}\Bigg(\Bigg[\mathbf{I} + \lambda \mathbf{P} \sum_{(i,i')\in\hat{\mathcal{B}}_2} \Bigg(&\frac{\mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}}{\|\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\|_2} \\ &- \frac{\mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\hat{\mathbf{u}}^T \mathbf{D}_{\mathcal{C}(i,i')}^T \mathbf{D}_{\mathcal{C}(i,i')}}{\|\mathbf{D}_{\mathcal{C}(i,i')}\hat{\mathbf{u}}\|_2^3} \Bigg) \Bigg]^{-1} \mathbf{P} \Bigg).\end{aligned}$$

□

Acknowledgments

We thank Ashley Petersen, Ali Shojaie, and Noah Simon for helpful conversations on earlier drafts of this manuscript. We thank the editor and two reviewers for helpful comments that improved the quality of this manuscript. D. W. was partially supported by a Sloan Research Fellowship, NIH Grant DP5OD009145, and NSF CAREER DMS-1252624.

References

Bach, F., Jenatton, R., Mairal, J. and Obozinski, G. (2011). Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning* 19–53.

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: a Nonasymptotic Theory of Independence*. OUP Oxford.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge university press. MR2061575

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771. MR2443189

Chen, J. and Chen, Z. (2012). Extended BIC for small-n-large-P sparse GLM. *Statistica Sinica* **22** 555. MR2954352

Chi, E. C., Allen, G. I. and Baraniuk, R. G. (2014). Convex biclustering. *arXiv preprint arXiv:1408.0856*.

Chi, E. and Lange, K. (2014a). Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*. in press.

Chi, E. and Lange, K. (2014b). cvxclustr: Splitting methods for convex clustering, `http://cran.r-project.org/web/packages/cvxclustr.` `R package version 1.1.1.`

Duchi, J. and Singer, Y. (2009). Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research* **10** 2899–2934. MR2579916

Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association* **81** 461–470. MR0845884

Hanson, D. L. and Wright, F. T. (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics* **42** 1079–1083. MR0279864

Haris, A., Witten, D. and Simon, N. (2015). Convex modeling of interactions with strong heredity. *Journal of Computational and Graphical Statistics*. in press.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*. Springer Verlag, New York. MR2722294

Hocking, T. D., Joulin, A., Bach, F., Vert, J.-P. et al. (2011). Clusterpath: an algorithm for clustering using convex fusion penalties. In *28th International Conference on Machine Learning*.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall. MR0999135

Lindsten, F., Ohlsson, H. and Ljung, L. (2011). Clustering using sum-of-norms regularization: with application to particle filter output computation. In *Statistical Signal Processing Workshop (SSP)* 201–204. IEEE.

Liu, J., Yuan, L. and Ye, J. (2013). Guaranteed sparse recovery under linear transformation. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)* 91–99.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory* **28** 129–137. MR0651807

Ng, A. Y., Jordan, M. I. and Weiss, Y. (2002). On spectral clustering: analysis and an algorithm. *Advances in Neural Information Processing Systems*.

Pelckmans, K., De Brabanter, J., Suykens, J. and De Moor, B. (2005). Convex clustering shrinkage. In *PASCAL Workshop on Statistics and Optimization of Clustering Workshop*.

Radchenko, P. and Mukherjee, G. (2014). Consistent clustering using ℓ_1 fusion penalty. *arXiv preprint arXiv:1412.0753*.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* **66** 846–850.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464. MR0468014

Tan, K. M. and Witten, D. M. (2014). Sparse biclustering of transposable data. *Journal of Computational and Graphical Statistics* **23** 985–1008. MR3270707

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics* **39** 1335–1371. MR2850205

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics* **40** 1198–1232. MR2985948

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 91–108. MR2136641

Vaiter, S., Deledalle, C.-A., Peyré, G., Fadili, J. M. and Dossal, C. (2014). The degrees of freedom of partly smooth regularizers. *arXiv preprint arXiv:1404.5557*. MR3281282

Witten, D. M. and Tibshirani, R. (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association* **105** 713–726. MR2724855

Zhu, C., Xu, H., Leng, C. and Yan, S. (2014). Convex optimization procedure for clustering: theoretical revisit. In *Advances in Neural Information Processing Systems*.