SUPPLEMENTARY MATERIAL

Proof of Lemma 2.5. Without loss of generality, assume h_i is the origin. Let $K \subset \mathbf{R}^k$ be the smallest circular cone with axis

$$a = \frac{1}{\alpha_i} P_{\mathbf{conv}(\mathrm{ext}(P)\setminus h_i)}(h_i)$$

that contains P. Since K is a cone, it also contains $T_P(h_i) = \mathrm{cone}(P)$. Thus $K^\circ \subset N_P(h_i)$ and $\omega(K^\circ) \leq \omega_i$. Figure 1 gives an illustrates the described geometry in a simple case.

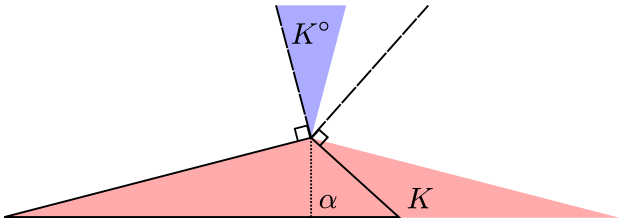


Figure 1: Example of the cone K from the proof of lemma 2.5. The normal cone is outlined by the dashed black line and the code K and its polar are shaded.

Further, K° is a circular cone with axis $-a$. By Lemma 2.1, the radius of the spherical cap $K^\circ \cap \mathbf{S}^{k-1}$ is at most

$$r(\omega_i) = 2(2\omega)^{1/(k-1)}. \tag{0.1}$$

Thus the angle of K° is at most $\arccos\left(1 - \frac{1}{2}r(\omega_i)^2\right)$ and the angle of K is at least

$$\frac{\pi}{2} - \arccos\left(1 - \frac{1}{2}r(\omega_i)^2\right) = \arcsin\left(1 - \frac{1}{2}r(\omega_i)^2\right).$$

To obtain a bound on the simplicial constant $\alpha_i = \alpha_P(h_i)$, we study 2-dimensional slices of K and P:

$$K \cap \mathbf{span}\,(a, \hat{n}) \text{ and } P \cap \mathbf{span}\,(a, \hat{n}) \text{ for any } \hat{n} \perp a.$$

Given a slice of P along the direction $\hat{n}$, the simplicial constant is given by

$$\alpha_i = \frac{r_{\hat{n}}}{\tan(\theta_{\hat{n}})}$$

for some radius $r_{\hat{n}}$ and some angle $\theta_{\hat{n}} \in \left[0, \frac{\pi}{2}\right)$. Since K is the smallest circular cone (with axis a) that contains P, the angle for K is equal to $\theta_{\hat{n}}$ for some slice. Further, $P \subset K$ so $r_{\hat{n}}$ is at most the diameter of the "base" of the pyramid that is the convex hull of the neighbors of h_i. Mathematically, the base is the set

$$B_P(h_i) = \mathbf{conv}\left(\{u \in \mathrm{ext}(P) \setminus h_i \mid u \text{ shares a face with } h_i\}\right).$$

Thus

$$\alpha_i \leq \frac{\mathbf{diam}\,(B_P(h_i))}{\tan\left(\arcsin\left(1 - \frac{1}{2}r(\omega_i)^2\right)\right)} \leq \mathbf{diam}\,(B_P(h_i)) \frac{r(\omega_i)\sqrt{1 - \frac{1}{4}r(\omega_i)^2}}{1 - \frac{1}{2}r(\omega_i)^2}.$$

□

Proof of Lemma 2.7. The proof is similar to the proof of Lemma 2.5. Assume (w.l.o.g.) h_i is at the origin. Let $K \subset \mathbf{R}^k$ be the largest circular cone that sits in $T_P(h_i)$, and let $a \in \mathbf{S}^{k-1}$ be its axis. Figure 2 gives an illustrates the described geometry in a simple case.

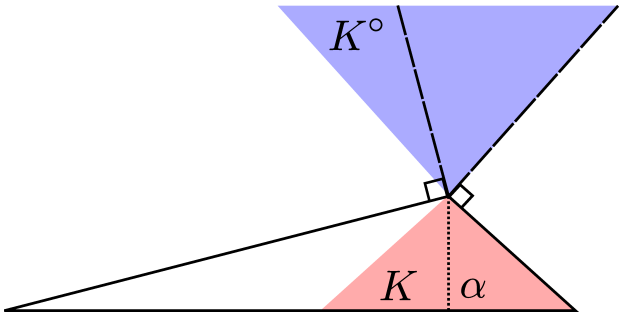


Figure 2: Example of the cone K from the proof of lemma 2.7. The normal cone is outlined by the dashed black line and the code K and its are shaded.

Let

$$B = T_P(h_i) \cap \{x \in \mathbf{R}^k \mid a^T(x - P_{\mathbf{conv}(\mathrm{ext}(P)\setminus h_i)}(h_i)) = 0\}.$$

Consider the 2-dimensional slices of K and P given by

$$K \cap \mathbf{span}\,(a, \hat{n}) \text{ and } P \cap \mathbf{span}\,(a, \hat{n}) \text{ for any } \hat{n} \perp a.$$

Given a slice of P along the direction $\hat{n}$, a bound on the simplicial constant is

$$\alpha_i \geq \frac{r_{\hat{n}}}{\tan(\theta_{\hat{n}})}$$

for some radius $r_{\hat{n}}$ and some angle $\theta_{\hat{n}} \in \left[0, \frac{\pi}{2}\right)$. Since K is the largest circular cone that sits in $T_P(h_i)$, the angle of K is equal to $\theta_{\hat{n}}$ for some slice. Further, $r_{\hat{n}}$ is well defined and its value depends only on the geometry of B. We let $r_{\min}$ be the smallest possible value of $r_{\hat{n}}$. Thus the angle of K is at least $\arctan\left(\frac{r_{\min}}{\alpha_i}\right)$. Since K° is a circular cone with axis $-a$, the angle of K° is at most

$$\frac{\pi}{2} - \arctan\left(\frac{r_{\min}}{\alpha_i}\right) = \operatorname{arccot}\left(\frac{r_{\min}}{\alpha_i}\right).$$

An elementary trigonometric calculation shows the height of the spherical cap $\mathbf{S}^{k-1} \cap K^\circ$ is at least

$$\cos\left(\operatorname{arccot}\left(\frac{r_{\min}}{\alpha_i}\right)\right) = \frac{r_{\min}}{\sqrt{\alpha_i^2 + (r_{\min})^2}}$$

By Lemma 2.2, the solid angle of K° is at most

$$\left(\frac{\sqrt{\alpha_i^2 + (r_{\min})^2}}{2r_{\min}}\right)^k,$$

since $K \subset T_P(h_i)$, $N_P(h_i) \subset K^\circ$ and $\omega_i \leq \omega(K^\circ)$. □

Proof of Lemma 2.10. Given a closed convex cone $K \subset \mathbf{R}^n$, a point $x \in \mathbf{R}^n$ has a unique orthogonal decomposition into $P_K(x) + P_{K^\circ}(x)$. To show $P_{K_2^n}\left(P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)\right)$ is the projection of x onto $K_2^n \cap \mathbf{R}_+^n$, it suffices to check

1. $P_{K_2^n}\left(P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)\right) \in K_2^n \cap \mathbf{R}_+^n$
2. $x - P_{K_2^n}\left(P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)\right) \in \left(K_2^n \cap \mathbf{R}_+^n\right)^\circ = \mathbf{conv}\left(-K_2^n \cap -\mathbf{R}_+^n\right)$
3. $P_{K_2^n}\left(P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)\right) \perp x - P_{K_2^n}\left(P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)\right)$

for any point $x \in \mathbf{R}^n$. To begin, we decompose x into its projection onto $\mathbf{R}_+^{n-1} \times \mathbf{R}$ and $\left(\mathbf{R}_+^{n-1} \times \mathbf{R}\right)^\circ$:

$$x = P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x) + P_{\left(\mathbf{R}_+^{n-1}\times\mathbf{R}\right)^\circ}(x).$$

We further decompose $P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)$ into its projection onto K_2^n and $K^\circ = -K_2^n$:

$$P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x) = P_{K_2^n}\left(P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)\right) + P_{-K_2^n}\left(P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)\right).$$

The projection onto K_2^n preserves the zero pattern of $P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)$. Thus a point $x \in \mathbf{R}^n$ admits the decomposition

$$x = P_{K_2^n}\left(P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)\right) + P_{-K_2^n}\left(P_{\mathbf{R}_+^{n-1}\times\mathbf{R}}(x)\right) + P_{\left(\mathbf{R}_+^{n-1}\times\mathbf{R}\right)^\circ}(x),$$

where the three parts are mutually orthogonal. Given this decomposition, it is easy to check 1, 2, and 3. □

1 Hyperspectral image example

Based on the origins of PPI, we demonstrate the use of random projections for finding important pixels in a hyperspectral image. We used a hyperspectral image of the National Mall in Washington, DC Landgrebe (2003)[1]. The image is 1280×307 pixels in size, contains 191 bands and is displayed in Figure 3. We utilize algorithm 1 to find the important pixels in the image. Intuitively, we should find pixels that represent pure versions of each class of objects, *e.g.*, trees, roofs, roads, etc., in the image. We then use these important pixels to broadly classify the remaining pixels in the image as each type of object. The assumption that predicate such a process is that in the image there appear to be a few key, or dominant, object classes. Figure 4 shows the relative frequency with which each selected extreme point is chosen. Here, we observe that there are roughly 10-15 key pixels identified by the algorithm.

[1] Available on the web at `engineering.purdue.edu/~biehl/MultiSpec/hyperspectral.html`

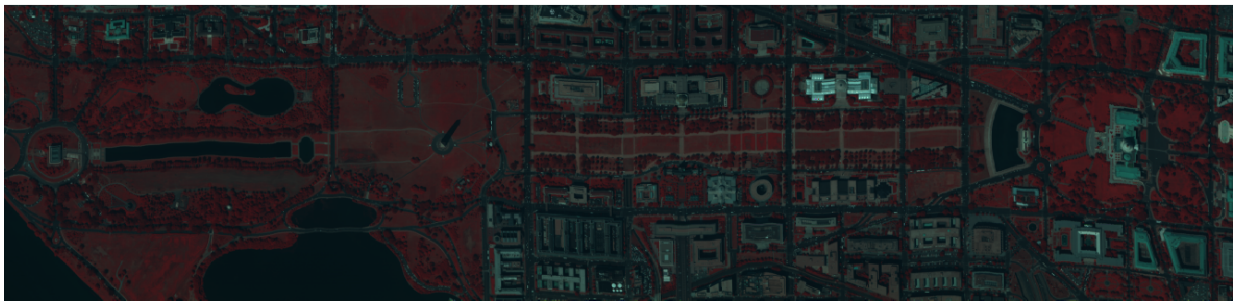


Figure 3: Hyperspectral image of the National Mall in Washington, DC. The RGB values of the image are set by choosing, for each color, a single color band.

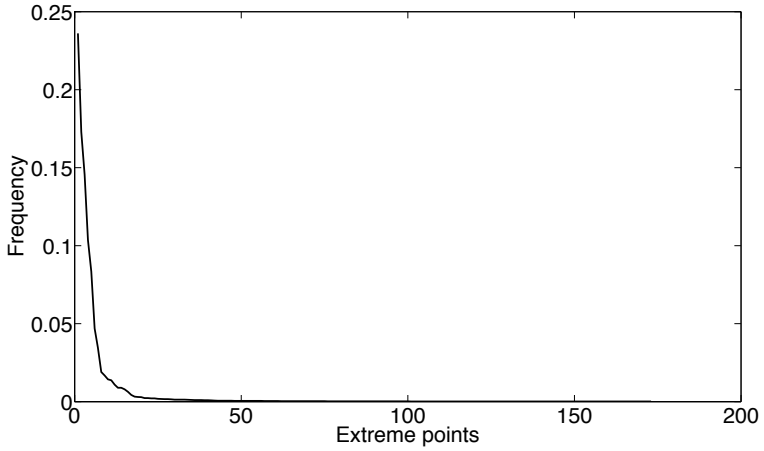
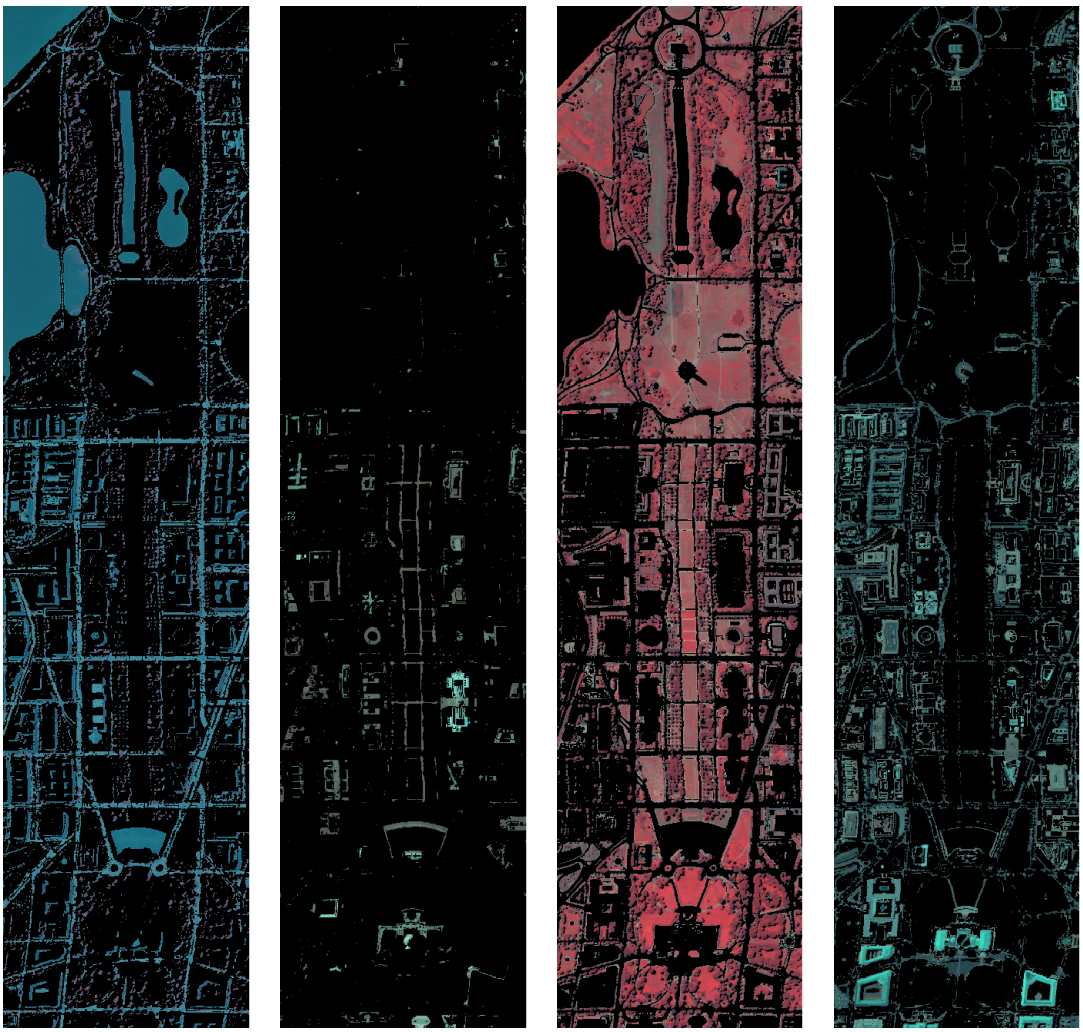


Figure 4: Relative frequency that each extreme point is selected via algorithm 1 using 5,000 random projections.

We now partially decompose the image using the most interpretable of the 11 most frequently selected extreme points. There are some pixels in the image that may be considered outliers, and because they are distinct from the remaining pixels they will be selected a lot. In fact, these points correspond to very pointy extreme points. These points are, in fact, important as they represent objects unlike the remainder of the image. In this situation, one example is that there appears to be a bright red light on the roof of the National Gallery of Art; such an object does not appear elsewhere in the image. However, for presentation purposes we stick to the important pixels that represent large sections of the image.

To broadly classify the image, we selected four pixels that appear to represent key features. We classify the remaining pixels by simply asking which representative pixel, of the 11 most selected, their spectrum looks most similar to in the ℓ_2 sense. Figure 5 shows the pixels classified into 4 categories. In each image the pixels that are classified as such are left colored and the remaining pixels are colored black. In fact, the images corresponding to the other seven pixels represent very little of the image.



(a) (b) (c) (d)

Figure 5: Classification of pixels into (a) roads and water, (b) concrete, (c) trees and grass, and (d) roofs. The contrast has been exaggerated for presentation.

2 Hereditary breast cancer dataset

We adopt the approach of Brunet et al. (2004) to discover "metagenes" from gene expression data with NMF. Given a dataset consisting of the expression levels of d genes in n samples, we seek to represent the expression pattern of the samples in terms of conical combinations of a small number of metagenes. The data is usually represented by an expression matrix $X \in \mathbf{R}^{d \times n}$. In most studies, $d \gg n$. Thus expression matrices are usually "tall and skinny." Mathematically, we seek an approximate factorization of the expression matrix $X = UV^T$ in terms of non-negative factors $U \in \mathbf{R}^{d \times k}$ and $V \in \mathbf{R}^{n \times k}$: $X \sim UV^T$. The columns of U are metagenes, and the rows of V are the coefficients of the conical combinations.

The hereditary breast cancer dataset collected by Hedenfalk et al. (2001) consists of the expression levels of 3226 genes on 22 samples from breast cancer patients. The patients

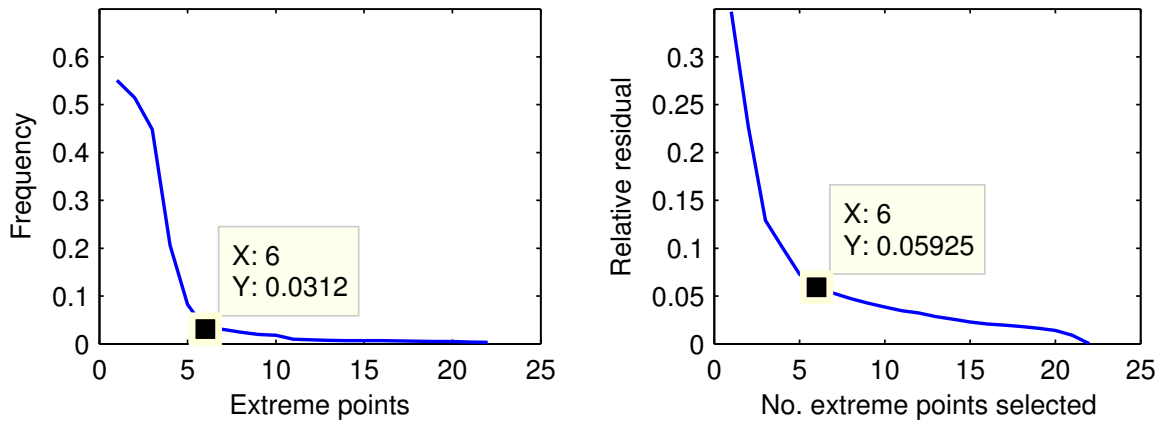


Figure 6: Scree plot (left) of how often each extreme point is found and plot of the relative residual $\left\|X - UV^T\right\|_F / \|X\|_F$ (right) versus how many extreme points are selected. There is a noticeable "elbow" at 6 on both plots.

consist of three groups: 7 patients with a BRCA1 mutation, 8 samples with a BRCA2 mutation, and 7 additional patients with sporadic (either estrogen-receptor-negative, aggressive cancers or estrogen-receptor-positive, less aggressive) cancers. The dataset is available at `http://www.expression.washington.edu/publications/kayee/bma/`. We exponentiate the data to make the log-expression levels non-negative.

We normalize the expression profiles (columns of X) and look for extreme points with the proto-algorithm. Figure 6 shows a scree plot of how often each extreme point is found by the proto-algorithm. Figure 6 also shows a plot of the relative residual versus how many extreme points are selected. The extreme points were selected by keeping the points found most often. On both plots, we notice an "elbow" at 6. This suggests the expression matrix is nearly-separable and has non-negative rank 6. Biologically, this means the expression pattern is mostly explained by the expression pattern of 6 metagenes.

We also selected metagenes by sparse regression (2.7). To compute the regularization path of (2.7), we implemented a solver on top of TFOCS by Becker et al. (2011). Figure 7 shows a coefficient plot and a spy plot of the regularization path. Although the sparse regression approach accounts for correlation among metagenes, the effect is negligible for the beginning (large regularization parameter) of the regularization path. Figure 8 shows the first 4 metagenes selected by the group lasso approach and by the greedy approach are the same and the sixth metagene selected by the greedy approach is the seventh to enter the regularization path.

References

Becker, S. R., Candès, E. J., and Grant, M. C. (2011), "Templates for convex cone problems with applications to sparse signal recovery," *Mathematical Programming Computation*, 3, 165–218.

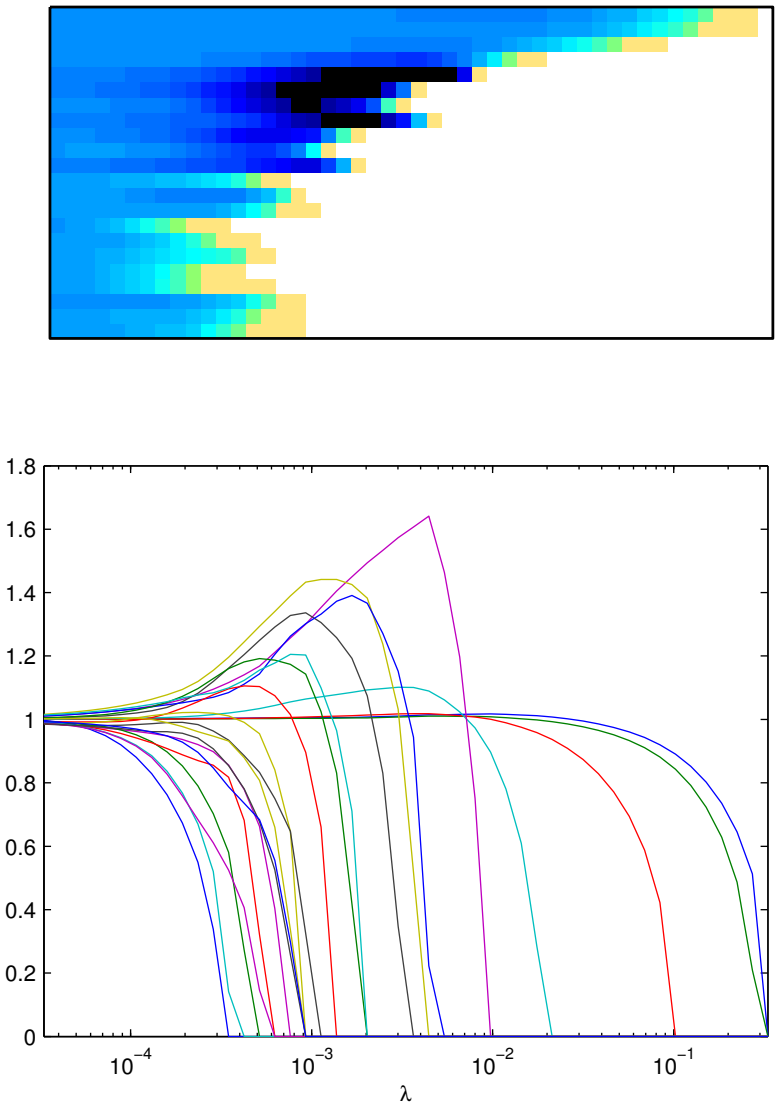


Figure 7: A spy plot (top) and a profile (bottom) of the group lasso path ($\|v_i\|_2$, $i = 1, \ldots, 22$ versus the regularization parameter λ). The rows (of pixels) in the spy plot and the lines in the profile correspond to groups. In the spy plot, lighter pixels correspond to small coefficients, while darker pixels correspond to large coefficients.

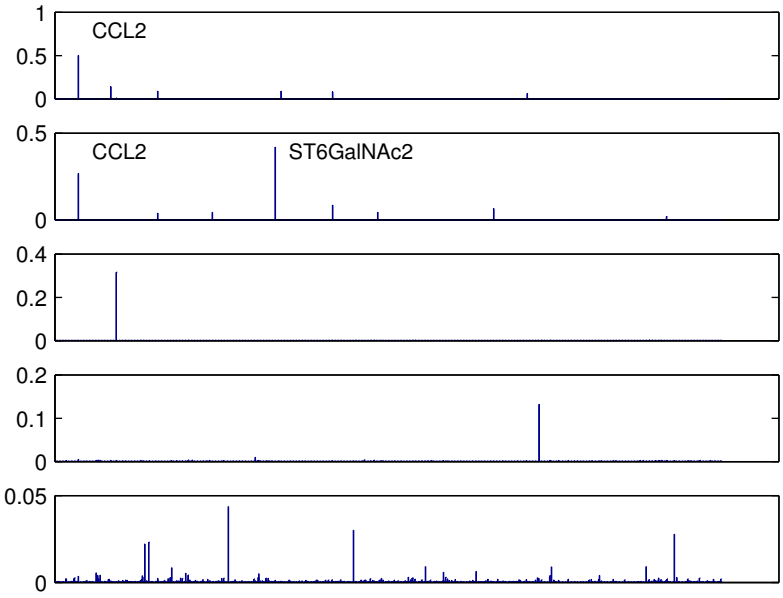


Figure 8: The first five metagenes selected by both the greedy and the group lasso approachs. The first two metagenes show high expression levels of an inflammatory chemokine CCL2 (also called MCP-1). Soria and Ben-Baruch (2008) showed levated CCL2 expression is associated with advanced disease course and with progression in breast cancers. This is consistent with the fact that 12 (of 22) samples in the study were (histologically) graded and all showed moderate to poor-differentiation (grades 6 to 9 on a scale of 1 to 9), an indication of advanced disease progression. The second metagene also shows high expression levels of ST6GalNAc2. Recently, Murugaesu et al. (2014) showed the enzyme encoded by ST6GalNAc2 is a metastasis suppressor in breast cancers. Unfortunately, the study only included patients with primary cancers, so the data cannot support the association between high expression of ST6GalNAc2 and lower incidence of metastasis.

Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P. (2004), "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences*, 101, 4164–4169.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P., Borg, k., and Trent, J. (2001), "Gene-expression profiles in hereditary breast cancer," *New England Journal of Medicine*, 344, 539–548, pMID: 11207349.

Landgrebe, D. A. (2003), *Signal theory methods in multispectral remote sensing*, John Wiley & Sons.

Murugaesu, N., Iravani, M., van Weverwijk, A., Ivetic, A., Johnson, D. A., Antonopoulos, A., Fearns, A., Jamal-Hanjani, M., Sims, D., Fenwick, K., et al. (2014), "An in vivo functional screen identifies ST6GalNAc2 sialyltransferase as a breast cancer metastasis suppressor," *Cancer discovery*, 4, 304–317.

Soria, G. and Ben-Baruch, A. (2008), "The inflammatory chemokines CCL2 and CCL5 in breast cancer," *Cancer letters*, 267, 271–285.