# PENSE: a Robust Penalized Estimator

Gabriela Cohen Freue

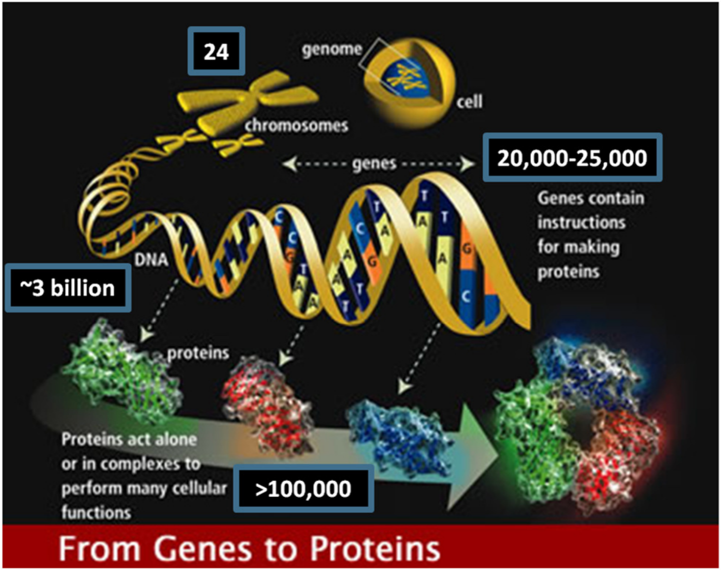
(joint with Dr. Salibian-Barrera, Kepplinger, and Dr. Smucler)

Department of Statistics, University of British Columbia

Statistical and Computational Challenges in High-Throughput Genomics with Application to Precision Medicine

November 7th, 2018


***When biology speaks, we listen.***




http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml


***... but it may get too loud, too noisy...***

- Many problem in high dimensional biology can be analyzed using linear regression

$$y_i = \mu + \mathbf{x}_i^t \boldsymbol{\beta} + \varepsilon_i, \quad \text{for} \quad i = 1, \ldots, n$$

where $\mathbf{x}_i \in \mathcal{R}^p$ are standardized; $y_i \in \mathcal{R}$ is centered; $\mu \in \mathcal{R}$; and $\boldsymbol{\beta} \in \mathcal{R}^p$

- Many problem in high dimensional biology can be analyzed using linear regression

$$y_i = \mu + \mathbf{x}_i^t \boldsymbol{\beta} + \varepsilon_i, \quad \text{for} \quad i = 1, \ldots, n$$

where $\mathbf{x}_i \in \mathcal{R}^p$ are standardized; $y_i \in \mathcal{R}$ is centered; $\mu \in \mathcal{R}$; and $\boldsymbol{\beta} \in \mathcal{R}^p$

- In -omics studies $p >> n$, and not all $p$ covariates are equally relevant

- Many problem in high dimensional biology can be analyzed using linear regression

$$y_i = \mu + \mathbf{x}_i^t \boldsymbol{\beta} + \varepsilon_i, \quad \text{for} \quad i = 1, \ldots, n$$

where $\mathbf{x}_i \in \mathcal{R}^p$ are standardized; $y_i \in \mathcal{R}$ is centered; $\mu \in \mathcal{R}$; and $\boldsymbol{\beta} \in \mathcal{R}^p$

- In -omics studies $p >> n$, and not all $p$ covariates are equally relevant

- Among the $p$ covariates available, many may be highly correlated, e.g., many genes from a common pathway
  - Do we need to listen to the whole rock band? or can we just listen to the singer?

**Which covariates should be included in the model ...**

**Which covariates should be included in the model ...**

***Variable Selection***

**Which covariates should be included in the model ...**

***Variable Selection***

**... to optimize the prediction of the response?**

***focused on Prediction Performance***

**Which covariates should be included in the model ...**

***Variable Selection***

**... to optimize the prediction of the response?**

***focused on Prediction Performance***

**... in a complex high dimensional setting**

**Select coefficients in a continuous way by adding a bound to their size**

# Regularization

**Select coefficients in a continuous way by adding a bound to their size**

For example,

**LASSO**: least absolute shrinkage and selection operator (Tibshirani, *JRSS*, 1996)

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) \;=\; \arg\min_{\mu,\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \mu - \mathbf{x}_i^t \boldsymbol{\beta}\right)^2$$

subject to

$$\|\boldsymbol{\beta}\|_1 \;\leq\; C \;\text{ for some } C > 0$$

# Penalized Estimators

More general, one can define

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) \;=\; \arg\min_{\mu, \boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \mu - \mathbf{x}_i^t \boldsymbol{\beta}\right)^2 \;+\; \lambda P(\boldsymbol{\beta}) \right\}$$

where $P$ is a penalty function and $\lambda$ controls the level of penalization.

**For example:**

**LASSO**: (Tibshirani, *JRSS*, 1996)

- $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$

# Penalized Estimators

More general, one can define

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) = \arg\min_{\mu, \boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left(y_i - \mu - \mathbf{x}_i^t \boldsymbol{\beta}\right)^2 + \lambda P(\boldsymbol{\beta}) \right\}$$

where $P$ is a penalty function and $\lambda$ controls the level of penalization.

**For example:**

**LASSO**: (Tibshirani, *JRSS*, 1996)

- $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$

**Other examples:**

- **Ridge**: (Hoerl and Kennard, *Technometrics*, 1970)
  - $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$
- **Bridge**: (Frank and Friedman, *Technometrics*, 1993)
  - $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_q^q$

**Other examples:**

- **Ridge**: (Hoerl and Kennard, *Technometrics*, 1970)
  - $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$
- **Bridge**: (Frank and Friedman, *Technometrics*, 1993)
  - $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_q^q$

**Limitations**

- Ridge and Bridge do not give sparse solutions.
- If $p > n$, LASSO can select at most $n$ variables out of $p$ candidates (Efron et al., *Annals of Statistics*, 2004).
- If there is a group of highly correlated variables, LASSO tends to select only one covariate from the group.

# Elastic Net Penalty

Zou and Hastie (*JRSS*, 2005) proposed

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) = \underset{\mu, \boldsymbol{\beta}}{\arg\min} \left\{ \sum_{i=1}^{n} (y_i - \mu - \mathbf{x}_i \boldsymbol{\beta})^2 + \lambda \left( \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right) \right\}$$

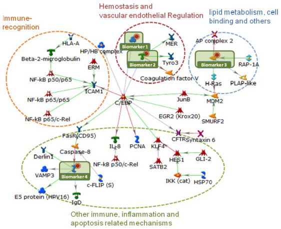
# Elastic Net Penalty

Zou and Hastie (*JRSS*, 2005) proposed

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) = \arg\min_{\mu, \boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mu - \mathbf{x}_i \boldsymbol{\beta})^2 + \lambda \left( \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right) \right\}$$

- EN combines the selection property of the $L_1$ penalty of LASSO with the smooth shrinkage of the $L_2$ penalty of Ridge
- EN can select at more variables than observations
- It preserves groups of highly correlated variables

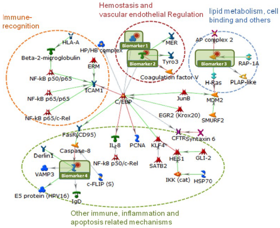
# We propose...




Penalized Elastic Net

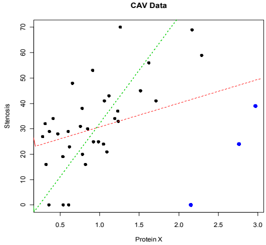
PENSE

# We propose...




Penalized Elastic Net

PENSE




S-estimator (Rousseeuw Yohai,1984)

PENSE

# Are regularized estimators robust?

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) = \arg\min_{\mu, \boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left( y_i - \mu - \mathbf{x}_i^t \boldsymbol{\beta} \right)^2 + \lambda P(\boldsymbol{\beta}) \right\}$$

**Regularized estimators are not necessarily robust!!**

- RLARS: Khan, Van Aelst and Zamar, *JASA* 2007
- S- and MM-Ridge: Maronna, *Technometrics*, 2011
- sparseLTS: Alfons, Croux, and Gelper, *Ann. Appl. Stat*, 2013
- MM-Bridge and MM-LASSO: Smucler and Yohai

# PENSE: Penalized Elastic Net S-Estimator

**Non-robust:**

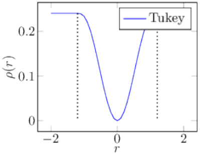
$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) = \arg\min_{\mu, \boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \left( y_i - \mu - \mathbf{x}_i^t \boldsymbol{\beta} \right)^2 + \lambda P(\boldsymbol{\beta}) \right\}$$

**Robust:**

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) = \operatorname*{argmin}_{\mu, \boldsymbol{\beta}} \left\{ n \hat{\sigma}(\mu, \boldsymbol{\beta})^2 + \lambda P(\boldsymbol{\beta}) \right\}$$

where

$$\hat{\sigma} : \frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{r_i}{\hat{\sigma}(r_i)} \right) = \delta,$$



# PENSE: Penalized Elastic Net Estimator

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) \;=\; \underset{\mu, \boldsymbol{\beta}}{\operatorname{argmin}} \; \left\{ n \hat{\sigma}(\mu, \boldsymbol{\beta})^2 \;+\; \lambda \left( \frac{1-\alpha}{2} \|\boldsymbol{\beta}\|_2^2 + \alpha \|\boldsymbol{\beta}\|_1 \right) \right\}$$

The generalized gradient of the penalized S loss is given by

$$\nabla_{(\mu,\boldsymbol{\beta})} \mathcal{L}(\mu, \boldsymbol{\beta}) = 2 \left[ -\frac{1}{n} \sum_{i=1}^{n} r_i(\mu, \boldsymbol{\beta}) w_i(\mu, \boldsymbol{\beta}) \begin{pmatrix} 1 \\ \boldsymbol{x}_i \end{pmatrix} + \frac{\lambda_S}{2} \begin{pmatrix} 0 \\ \nabla_{\boldsymbol{\beta}} P_\alpha(\boldsymbol{\beta}) \end{pmatrix} \right],$$

# IRWEN Algorithm

1. Given an initial $\hat{\mu}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(0)}$, compute the weights $w_i$
2. Solve an EN problem and get updated $\hat{\boldsymbol{\beta}}^{(0)}$ and corresponding $\hat{\mu}^{(0)}$
3. Iterate until convergence (or maximum number of steps)

# IRWEN Algorithm

1. Given an initial $\hat{\mu}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(0)}$, compute the weights $w_i$
2. Solve an EN problem and get updated $\hat{\boldsymbol{\beta}}^{(0)}$ and corresponding $\hat{\mu}^{(0)}$
3. Iterate until convergence (or maximum number of steps)

**Caution**

- Robust objective functions are often non-convex and exhibit multiple optima

# IRWEN Algorithm

1. Given an initial $\hat{\mu}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(0)}$, compute the weights $w_i$
2. Solve an EN problem and get updated $\hat{\boldsymbol{\beta}}^{(0)}$ and corresponding $\hat{\mu}^{(0)}$
3. Iterate until convergence (or maximum number of steps)

### Caution

- Robust objective functions are often non-convex and exhibit multiple optima
- Most optimization algorithms only find local optima

# IRWEN Algorithm

1. Given an initial $\hat{\mu}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(0)}$, compute the weights $w_i$
2. Solve an EN problem and get updated $\hat{\boldsymbol{\beta}}^{(0)}$ and corresponding $\hat{\mu}^{(0)}$
3. Iterate until convergence (or maximum number of steps)

**Caution**

- Robust objective functions are often non-convex and exhibit multiple optima
- Most optimization algorithms only find local optima
- Where we start determines what local optima is attained

# IRWEN Algorithm

1. Given an initial $\hat{\mu}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(0)}$, compute the weights $w_i$
2. Solve an EN problem and get updated $\hat{\boldsymbol{\beta}}^{(0)}$ and corresponding $\hat{\mu}^{(0)}$
3. Iterate until convergence (or maximum number of steps)

**Caution**

- Robust objective functions are often non-convex and exhibit multiple optima
- Most optimization algorithms only find local optima
- Where we start determines what local optima is attained
- Initial values play an important role

# IRWEN Algorithm

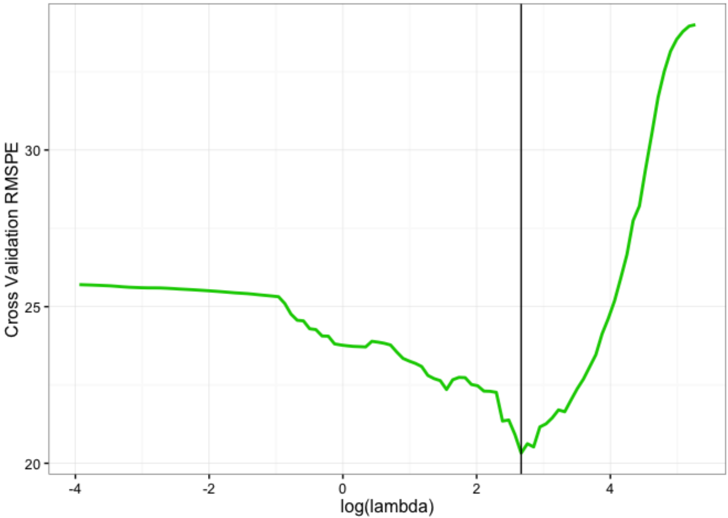
1. Given an initial $\hat{\mu}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(0)}$, compute the weights $w_i$
2. Solve an EN problem and get updated $\hat{\boldsymbol{\beta}}^{(0)}$ and corresponding $\hat{\mu}^{(0)}$
3. Iterate until convergence (or maximum number of steps)

**Caution**

- Robust objective functions are often non-convex and exhibit multiple optima
- Most optimization algorithms only find local optima
- Where we start determines what local optima is attained
- Initial values play an important role

**Topic for another talk...**

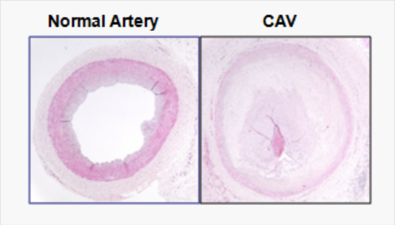
# Choosing Lambda: another talk!



# Proteomics Case Study

Proteomics Biomarker Study of Cardiac Allograft Vasculopathy

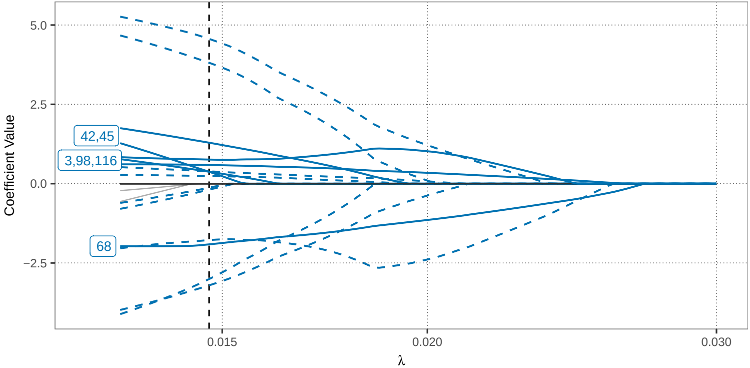
- Biomarkers in Transplantation: enrolled patients who received a heart transplant at St. Paul's Hospital, BC
- Around one year after transplantation, some patients presented signs of coronary artery narrowing




- BiT measured (81) protein levels in plasma 37 plasma samples
- **Goal**: identify potential biomarkers of CAV

# Potential Biomarkers

PENSE(M) can be uses select the most relevant proteins is plasma to predict CAV



# Potential Biomarkers

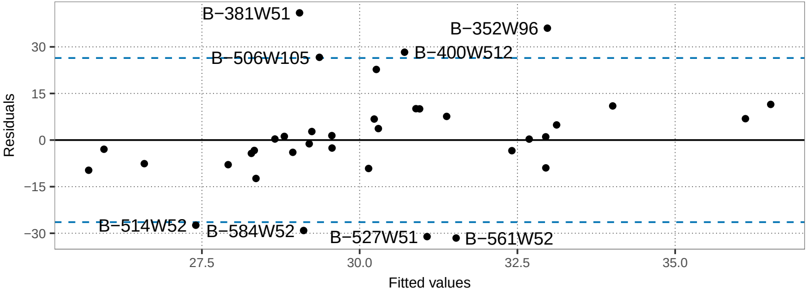
Identified by PENSEM ($\alpha = 0.6$)

| Protein ID | Gene Symbol | Protein Name |
|---|---|---|
| 3 | C4B/C4A | Complement C4-B/C4-A |
| 20 | C7 | Complement component C7 |
| 42 | APOE | Apolipoprotein E |
| 45 | AMBP | Protein AMBP |
| 64 | CFI | Complement factor I |
| 68 | SHBG | Sex hormone-binding globulin |
| 103 | C1QC | Complement C1q subunit C |
| 116 | APOC2 | Apolipoprotein C-II |
| 139 | HBD | Hemoglobin subunit delta |
| 161 | SEPP1 | Selenoprotein P |
| 298 | HBA2;HBA1;HBZ | Hemoglobin subunit alpha/zeta |

Some of these were previously associated with CAV (Lin*, Cohen Freue*, et al., 2013)

# Outlying Patients

PENSE(M) can be uses to flag outlying patients



# Validation

- Independent set of 52 patients collected by BiT in the second phase of their study.

# Validation

- Independent set of 52 patients collected by BiT in the second phase of their study.
- A subset of 6 proteins were analyzed by a Multiple Reaction Monitoring assay (Cohen Freue and Borchers, 2012; Domanski et al., 2012).

# Validation

- Independent set of 52 patients collected by BiT in the second phase of their study.
- A subset of 6 proteins were analyzed by a Multiple Reaction Monitoring assay (Cohen Freue and Borchers, 2012; Domanski et al., 2012).
- We built a proteomic score using an MM-estimator based on the resulting data.

# Validation

- Independent set of 52 patients collected by BiT in the second phase of their study.
- A subset of 6 proteins were analyzed by a Multiple Reaction Monitoring assay (Cohen Freue and Borchers, 2012; Domanski et al., 2012).
- We built a proteomic score using an MM-estimator based on the resulting data.
- Among these test samples, 12 were flagged as outlying and removed from the prediction set.

# Validation

- Independent set of 52 patients collected by BiT in the second phase of their study.
- A subset of 6 proteins were analyzed by a Multiple Reaction Monitoring assay (Cohen Freue and Borchers, 2012; Domanski et al., 2012).
- We built a proteomic score using an MM-estimator based on the resulting data.
- Among these test samples, 12 were flagged as outlying and removed from the prediction set.
- For the remaining 40 test patients, the stenosis was predicted by our model.

# Validation

- Independent set of 52 patients collected by BiT in the second phase of their study.
- A subset of 6 proteins were analyzed by a Multiple Reaction Monitoring assay (Cohen Freue and Borchers, 2012; Domanski et al., 2012).
- We built a proteomic score using an MM-estimator based on the resulting data.
- Among these test samples, 12 were flagged as outlying and removed from the prediction set.
- For the remaining 40 test patients, the stenosis was predicted by our model.
- A classification based on the predicted stenosis yielded an AUC of 0.85.

# Conclusions

- We propose robust penalized S- and MM-estimators using an EN penalty

# Conclusions

- We propose robust penalized S- and MM-estimators using an EN penalty
- LASSO and Ridge penalties are particular cases of the EN penalty

# Conclusions

- We propose robust penalized S- and MM-estimators using an EN penalty
- LASSO and Ridge penalties are particular cases of the EN penalty
- We propose an efficient algorithm to compute PENSE and PENSEM (‘pense‘ available in CRAN)

# Conclusions

- We propose robust penalized S- and MM-estimators using an EN penalty
- LASSO and Ridge penalties are particular cases of the EN penalty
- We propose an efficient algorithm to compute PENSE and PENSEM (‘pense‘ available in CRAN)
- PENSE(M) have strong robustness properties: can have a $50\%$ BDP

# Conclusions

- We propose robust penalized S- and MM-estimators using an EN penalty
- LASSO and Ridge penalties are particular cases of the EN penalty
- We propose an efficient algorithm to compute PENSE and PENSEM (‘pense‘ available in CRAN)
- PENSE(M) have strong robustness properties: can have a $50\%$ BDP
- Under some conditions, PENSE(M) are consistent

# Conclusions

- We propose robust penalized S- and MM-estimators using an EN penalty
- LASSO and Ridge penalties are particular cases of the EN penalty
- We propose an efficient algorithm to compute PENSE and PENSEM ('pense' available in CRAN)
- PENSE(M) have strong robustness properties: can have a $50\%$ BDP
- Under some conditions, PENSE(M) are consistent
- PENSE and PENSEM showed a competitive performance compared to other existing methods

# Conclusions

- We propose robust penalized S- and MM-estimators using an EN penalty
- LASSO and Ridge penalties are particular cases of the EN penalty
- We propose an efficient algorithm to compute PENSE and PENSEM (‘pense‘ available in CRAN)
- PENSE(M) have strong robustness properties: can have a $50\%$ BDP
- Under some conditions, PENSE(M) are consistent
- PENSE and PENSEM showed a competitive performance compared to other existing methods
- PENSE and PENSEM can be used to flag outlying patients

# Thank you!

**Acknowledgements**

- Coauthors: Dr. Matias Salibian-Barrera, David Kepplinger (PhD candidate), Ezequiel Smucler (PDF)
- NSERC grant and CFI computational infrastructure
- Data provided by the NCE CECR PRevention of Organ Failure (PROOF) Centre of Excellence

**https://gcohenfr.github.io**

# References


- Hoerl, A.E. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55-67.
- Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 267-288.
- Peña, D. and Yohai, V. (1999) A Fast Procedure for Outlier Diagnostics in Large Regression Problems. Journal of the American Statistical Association **94**, 434-445.
- Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics.
- Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32**, 406-499.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, **67**, 301-320.
- Friedman, J.; Hastie, T., and Tibshirani, R. (2010) Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, **33**,1.
- Maronna, R. A. (2011) Robust Ridge Regression for High-Dimensional Data. Technometrics **53**, 44-53.
- Smucler, E. and Yohai, V. (2017) Robust and sparse estimators for linear regression models. In: Computational Statistics & Data Analysis **111**, 116-130.
- Lin D*, Cohen Freue G*, et al. (2013) Plasma protein biosignatures for detection of cardiac allograft vasculopathy. The Journal of Heart and Lung Transplantation 7, 723-733. *Equal contributors.
- Gabriela V. Cohen Freue and Christoph H. Borchers. (2012) Multiple Reaction Monitoring (MRM), Principles and Application to Coronary Artery Disease. Circulation: Cardiovascular Genetics 5, 378.
- Domanski D, Percy AJ, Yang J, Chambers AG, Hill JS, Freue GV, Borchers CH. (2012) MRM-based Multiplexed Quantitation of 67 Putative Cardiovascular Disease Biomarkers in Human Plasma. Proteomics 12, 1222-1243.