This document combines a number of threads for model selection in our problem (and other related optimization problems). First, we consider estimating the degrees of freedom. Then we give two results which can be used to select models.

1 Degrees of freedom for Gen-Gen Lasso

Consider a random vector $Y \in \mathbb{R}^n$ with distribution in the exponential family with independent components. Specifically, write the density of Y as

$$p(Y \mid \theta) = \left(\prod_{i=1}^n h(y_i)\right) \exp\left\{\sum_{i=1}^n w(y_i)\mu_i(\theta) - \psi_i(\theta)\right\}.$$

We will consider estimation of θ subject to the generalized lasso penalty. That is, our goal is to solve

$$\widehat{\theta}(y) = \operatorname*{argmin}_\theta \ell(y|\theta) + \lambda \left\| G\theta \right\|_1 = \operatorname*{argmin}_\theta \sum_{i=1}^n \psi_i(\theta) - w(y_i)\mu_i(\theta) + \lambda \left\| G\theta \right\|_1, \tag{1}$$

where $G \in \mathbb{R}^{q\times p}$ and $\theta \in \mathbb{R}^p$.

This generalizes many standard models. For example, ℓ_1-trend filtering has $p = n$, $\psi_i(x) = x_i^2/2$, $\mu_i(x) = x_i$, $w(x) = 2$, and G the first-order discrete difference operator. Variance estimation in our context has $\psi_i(x) = x_i$, $w(x) = y_i^2$ and $\mu_i(x) = e^{-x_i}$. Logistic loss also falls into this category. We use G rather than the more common choice of D to avoid confusion with the differential operator that we will require below.

Notation For a vector-valued function $f : \mathbb{R}^p \to \mathbb{R}^q$, we use the notation Df to denote the Jacobian matrix and ∇f to be the gradient.

We first consider the case that $\theta \in \mathbb{R}^n$ and that ψ_i and μ_i operate component-wise on θ. That is $p(y \mid \theta) = \prod_{i=1}^n p(y_i \mid \theta_i)$.

Theorem 1. *Suppose $p(y \mid \theta) = \prod_{i=1}^n p(y_i \mid \theta_i)$ and that $\widehat{\theta}$ is a solution of (1). Then, the divergence of $\widehat{\theta}(y)$ is given by*

$$\operatorname{tr}\left(D\widehat{\theta}(y)\right) = -P_{\mathcal{N}(G)}\left(P_{\mathcal{N}(G)} \operatorname{diag}\left(\frac{d^2}{d\theta_i^2}\ell\bigg|_{y_i,\widehat{\theta}_i}\right) P_{\mathcal{N}(G)}\right)^\dagger P_{\mathcal{N}(G)} \operatorname{diag}\left(\frac{d^2}{d\theta_i dy_i}\ell\bigg|_{y_i,\widehat{\theta}_i}\right),$$

where

$$P_{\mathcal{N}(G)} = I_n - G_S^\top (G_S^\top G_S)^\dagger G_S \tag{2}$$

is the projection onto the null-space of G_S, where $S = \{j \in [q] : G\widehat{\theta} = 0\}$, and the notation G_S means the rows of G whose indices are in S.

Special cases:

1. If we are interested in the natural exponential family (that is $\mu_i(\theta) = \theta_i$ and $w(y_i) = y_i$), then $\frac{d^2}{d\theta_i dy_i}\ell = -1$ and $\frac{d^2}{d\theta_i^2}\psi_i(\theta_i) = \mathrm{Var}[Y_i]$.
2. In particular, if $p(y_i \mid \theta_i) = \mathrm{N}(\theta_i, \sigma^2)$, then $\frac{d^2}{d\theta_i^2}\ell = \sigma^2$ and $\frac{d^2}{d\tau_i dy_i}\ell = -1$, so the divergence is σ^2 times the dimension of $\mathcal{N}(G_S)$ as shown in Tibshirani and Taylor [4].
3. For logistic loss, $\frac{d^2}{d\theta_i^2}\ell = e^{\theta_i}/(1 + e^{\theta_i})^2$ and $\frac{d^2}{d\tau_i dy_i}\ell = -1$.

4. For the case of variance estimation, we have $\frac{d^2}{d\theta_i^2}\ell = y_i^2 e^{-\theta}$ and $\frac{d^2}{d\theta_i dy_i}\ell = -e^{-\theta}$.

Now we generalize to the regression setting. Define $\theta_i = x_i^\top \beta$, $\beta \in \mathbb{R}^p$. And let

$$\widehat{\beta}(y) = \operatorname*{argmin}_{\beta} \sum_{i=1}^{n} \psi_i(x_i^\top \beta) - w(y_i)\mu_i(x_i^\top \beta) + \lambda \left\| G\beta \right\|_1 ,$$

Theorem 2. *The divergence of $\widehat{\theta}(y) := X\widehat{\beta}(y)$ is given by*

$$\operatorname{tr}\left(D\widehat{\theta}(y)\right) = -X_P \left(X_P^\top \operatorname{diag}\left(\frac{d^2}{d\theta_i^2}\ell \Big|_{y_i,\widehat{\theta}_i} \right) X_P \right)^\dagger X_P^\top \operatorname{diag}\left(\frac{d^2}{d\theta_i dy_i}\ell \Big|_{y_i,\widehat{\theta}_i} \right),$$

where

$$X_P = (I_n - G_S^\top (G_S^\top G_S)^\dagger G_S)X$$

is the projection of X onto the null-space of G_S.

Proof. This follows mainly from Theorem 2 in [5], though the conditions are non-trivial. To be expanded. □

2 Risk estimation

If $Y \sim \mathrm{N}(\theta, \sigma^2)$, a now common method of risk estimation makes use of Stein's Lemma.

Lemma 3 (Stein's Lemma)**.** *Assume $f(Y)$ is weakly differentiable with essentially bounded weak partial derivatives on $\mathbb{R}^n$, then*

$$\operatorname{tr}\operatorname{Cov}(Y, f(Y)) = \mathbb{E}\left[\langle Y,\ f(Y)\rangle\right] = \sigma^2 \mathbb{E}\left[\operatorname{tr} Df(Y)\Big|_y \right].$$

The utility of this result comes from examining the decomposition of the mean squared error of $f(Y)$ as an estimator of θ.

$$\begin{aligned}\mathbb{E}\left[\|\theta - f(Y)\|_2^2\right] &= \mathbb{E}\left[\|Y - f(Y)\|\right] - n\sigma^2 + 2\operatorname{tr}\operatorname{Cov}(Y, f(Y)) \\ &= \mathbb{E}\left[\|Y - f(Y)\|\right] - n\sigma^2 + 2\sigma^2 \mathbb{E}\left[\operatorname{tr} Df(y)\Big|_Y \right].\end{aligned}$$

This characterization motivates the definition of degrees-of-freedom for linear predictors (df $:= \frac{1}{\sigma^2}\operatorname{tr} Df(y)\big|_Y$) [2], where $f(y) = Hy$. Using Stein's Lemma, assuming σ^2 is known, we have Stein's Unbiased Risk Estimator

$$SURE_\theta = \|Y - f(Y)\| - n\sigma^2 + 2\sigma^2 \operatorname{tr} Df(y)\Big|_y .$$

Note that this is a risk for estimating the n-dimensional parameter θ.

The following result generalizes this idea to certain continuous exponential families.

Lemma 4 (Generalized Stein Lemma [3])**.** *Assume $f(Y)$ is weakly differentiable with essentially bounded weak partial derivatives on $\mathbb{R}^n$. Let Y be distributed according to a natural exponential family*

$$p(Y \mid \beta) = \left(\prod_{i=1}^{n} h(y_i) \right) \exp\left\{ \sum_{i=1}^{n} y_i \theta_i(\beta_i) - \psi(\beta_i) \right\}, \tag{3}$$

and assume that h is weakly differentiable and that θ_i is one-to-one. Then,

$$\mathbb{E}\left[\langle \theta(\beta),\ f(Y)\rangle\right] = -\mathbb{E}\left[\left\langle \frac{\nabla h(Y)}{h(Y)},\ f(Y)\right\rangle + \operatorname{tr} Df(y)\Big|_{Y}\right].$$

Note that $\nabla h(Y)$ here means the vector $[d/dy h(y)|_{y_i}]$ and $h(Y)$ means the vector $[h(y_i)]$.

Therefore we define the Generalized SURE [3] along the lines of the multivariate Gaussian case.

Lemma 5. *Assume h and ∇h are weakly differentiable, θ_i is one-to-one, $f(Y)$ is weakly differentiable with essentially bounded partial derivatives and that $p(Y \mid \beta)$ is as in (3). Then*

$$GSURE_\theta = \|f(Y)\|_2^2 + 2\left\langle \frac{\nabla h(Y)}{h(Y)},\ f(Y)\right\rangle + 2\operatorname{tr} Df(y)\Big|_{Y} + \frac{1}{h(Y)} \operatorname{tr} \frac{\partial^2 h(y)}{\partial y^2}\Big|_{Y}$$

is an unbiased estimator for the MSE of an estimator $f(Y)$ of θ: $\mathbb{E}\left[\|f(Y) - \theta(\beta)\|_2^2\right]$.

Proof. We have

$$\mathbb{E}\left[\|f(Y) - \theta(\beta)\|_2^2\right] = \mathbb{E}\left[\|f(Y)\|_2^2\right] + \mathbb{E}\left[\|\theta\|_2^2\right] - 2\mathbb{E}\left[\langle \theta(\beta),\ f(Y)\rangle\right].$$

Now, the first term is a function of the data only, and to the last term, we simply apply Lemma 4. For the second term,

$$\begin{aligned}
\mathbb{E}\left[\|\theta\|_2^2\right] = \mathbb{E}\left[\langle \theta,\ \theta\rangle\right] &= -\mathbb{E}\left[\left\langle \frac{\nabla h(Y)}{h(Y)},\ \theta\right\rangle\right] \\
&= \mathbb{E}\left[\left\langle \frac{\nabla h(Y)}{h(Y)},\ \frac{\nabla h(Y)}{h(Y)}\right\rangle\right] + \mathbb{E}\left[\operatorname{tr} \frac{\partial}{\partial y} \frac{\nabla h(y)}{h(y)}\Big|_{Y}\right] \\
&= \mathbb{E}\left[\frac{\|\nabla h(Y)\|_2^2}{h(Y)^2}\right] + \mathbb{E}\left[\operatorname{tr} \frac{\|\nabla h(Y)\|_2^2 + h(Y)\partial^2/\partial y^2 h(y)\big|_{Y}}{h(Y)^2}\right] \\
&= \mathbb{E}\left[\frac{1}{h(Y)} \operatorname{tr} \frac{\partial^2 h(y)}{\partial y^2}\Big|_{Y}\right],
\end{aligned}$$

by applying Lemma 4 twice along with the quotient rule. □

Theorem 6. *Let $Y_i \sim N(0, e^{x_i})$. Consider estimating x by solving*

$$\min_h \sum_{i=1}^n x_i + y_i^2 e^{-x_i} + \lambda \|Gx\|_1.$$

Then an unbiased estimator of $\mathbb{E}\left[\|\frac{1}{2e^x} - \frac{1}{2e^{\widehat{x}}}\|\right]_2^2$ is given by

$$\left\|\frac{1}{2e^{\widehat{x}}}\right\|_2^2 - \frac{1}{2}\left\langle \frac{1}{y^2},\ \frac{1}{e^{\widehat{x}}}\right\rangle + 2P_{\mathcal{N}(G)}\left(P_{\mathcal{N}(G)} \operatorname{diag}\left(y^2 e^{-\widehat{x}}\right) P_{\mathcal{N}(G)}\right)^{\dagger} P_{\mathcal{N}(G)} \operatorname{diag}\left(e^{-\widehat{x}}\right) + \frac{3}{4y^4},$$

where $P_{\mathcal{N}(G)}$ is as in (2).

Proof. Define $z_i := y_i^2$ and $\theta_i = -\frac{1}{2e^{x_i}}$. Then

$$\begin{aligned}
p(z_i \mid x_i) &= \frac{1}{\sqrt{2\pi z_i e^{x_i}}} \exp\left\{-\frac{1}{2}\frac{z_i}{e^{x_i}}\right\} \mathbf{1}_{(0,\infty)}(z_i) \\
&= \frac{1}{\sqrt{\pi z}} \mathbf{1}_{(0,\infty)}(z_i) \exp\left\{z_i \cdot \left(-\frac{1}{2e^{x_i}}\right) - \left(-\frac{1}{2}\log\left(\frac{1}{2e^{x_i}}\right)\right)\right\} \\
&= \frac{1}{\sqrt{\pi z}} \mathbf{1}_{(0,\infty)}(z_i) \exp\left\{z_i \cdot \theta_i - \psi(\theta_i)\right\},
\end{aligned}$$

which is a natural exponential family with $\psi(t) = -\frac{1}{2}\log(-t)$ and $h(t) = \frac{1}{\sqrt{\pi t}}\mathbf{1}_{(0,\infty)}(t)$. Therefore, $d/dt \log h(t) = -1/(2t)$ and $(d^2/dt^2 h(t))/h(t) = 3/(4t^2)$. Finally, we have $\frac{d^2}{d\theta_i^2}\ell = y_i^2 e^{-\theta}$ and $\frac{d^2}{d\theta_i dy_i}\ell = -e^{-\theta}$. The result follows from Theorem 1 and Lemma 5.

□

An alternate model selection technique is to examine the Kullback-Leibler Divergence between the density under $\theta = \widehat{\theta}(Y)$ and that under $\theta = \theta$: $\mathbb{E}\left[KL(p(Y \mid \widehat{\theta}(Y)) \| p(Y \mid \theta))\right]$. For exponential families, we have

$$\mathbb{E}\left[KL(\widehat{\theta}(Y)\|\theta)\right] = \mathbb{E}\left[KL(p(Y \mid \widehat{\theta}(Y)) \| p(Y \mid \theta))\right] = \left\langle \widehat{\theta}(Y) - \theta,\ \nabla\psi(\widehat{\theta}(Y)) \right\rangle + \psi(\theta) - \psi(\widehat{\theta}(Y)).$$

An application of Lemma 4 thus provides an unbiased estimator of this quantity [1].

Lemma 7. *Assume h and ∇h are weakly differentiable, θ_i is one-to-one, $f(Y)$ is weakly differentiable with essentially bounded partial derivatives and that $p(Y \mid \beta)$ is as in (3). Then*

$$SUKLS = \left\langle \widehat{\theta} + \frac{\nabla h(Y)}{h(Y)},\ \nabla\psi(\widehat{\theta}(Y)) \right\rangle + \operatorname{tr} Df(y)\Big|_Y - \psi(\widehat{\theta}(Y))$$

is unbiased for $\mathbb{E}\left[KL(\widehat{\theta}(Y)\|\theta)\right] - \psi(\theta)$.

Corollary 8. *Let $Y_i \sim N(0, e^{x_i})$. Consider estimating x by solving*

$$\min_h \sum_{i=1}^n x_i + y_i^2 e^{-x_i} + \lambda \left\|Gx\right\|_1.$$

Then an unbiased estimator of $\mathbb{E}\left[KL(\widehat{x}\|x\right] - \psi(\theta)$ is given by

$$-\frac{n}{2} - \left\langle \frac{1}{2y^2},\ e^{\widehat{x}} \right\rangle + -\frac{1}{2}\log\left(2e^{\widehat{x}_i}\right) + P_{\mathcal{N}(G)}\left(P_{\mathcal{N}(G)}\operatorname{diag}\left(y^2 e^{-\widehat{x}}\right) P_{\mathcal{N}(G)}\right)^\dagger P_{\mathcal{N}(G)}\operatorname{diag}\left(e^{-\widehat{x}}\right).$$

Proof. As $\widehat{\theta}_i = -\frac{1}{2e^{\widehat{x}_i}}$,

$$\psi(\widehat{\theta}_i) = -\frac{1}{2}\log\left(\frac{1}{2e^{\widehat{x}_i}}\right)$$

and $d/d\theta\psi(\theta) = -\frac{1}{2\theta}$. Thus

$$\begin{aligned} SUKLS &= \left\langle -\frac{1}{2e^{\widehat{x}}} - \frac{1}{2y^2},\ e^{\widehat{x}} \right\rangle + \operatorname{tr} Df(y)\Big|_Y + \frac{1}{2}\log\left(\frac{1}{2e^{\widehat{x}_i}}\right) \\ &= -\frac{n}{2} - \left\langle \frac{1}{2y^2},\ e^{\widehat{x}} \right\rangle + \operatorname{tr} Df(y)\Big|_Y - \frac{1}{2}\log\left(2e^{\widehat{x}_i}\right). \end{aligned}$$

NOTE: I think some signs are wrong here. □

References

[1] Deledalle, C.-A. (2017), "Estimation of Kullback-Leibler losses for noisy recovery problems within the exponential family," *Electronic Journal of Statistics*, **11**, 3141—3164.

[2] Efron, B. (1986), "How biased is the apparent error rate of a prediction rule?" *Journal of the American Statistical Association*, **81**(394), 461–470.

[3] Eldar, Y. C. (2009), "Generalized SURE for exponential families: Applications to regularization," *IEEE Transactions on Signal Processing*, **57**, 471–481.

[4] Tibshirani, R. J., and Taylor, J. (2012), "Degrees of freedom in lasso problems," *Annals of Statistics*, **40**, 1198–1232, arXiv:1111.0653.

[5] Vaiter, S., Deledalle, C., Fadili, J., Peyré, G., and Dossal, C. (2017), "The degrees of freedom of partly smooth regularizers," *Annals of the Institute of Statistical Mathematics*, **69**, 791–832.