A note on the generalized degrees of freedom under the $L_1$ loss function

By: Xiaoli Gao, Yixin Fang


Made available courtesy of Elsevier: [http://dx.doi.org/10.1016/j.jspi.2010.07.006](http://dx.doi.org/10.1016/j.jspi.2010.07.006)

***© Elsevier. Reprinted with permission. No further reproduction is authorized without written permission from Elsevier. This version of the document is not the version of record. Figures and/or pictures may be missing from this format of the document.***

This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nc-nd/4.0/).

**Abstract:**

Generalized degrees of freedom measure the complexity of a modeling procedure; a modeling procedure is a combination of model selection and model fitting. In this manuscript, we consider two definitions of generalized degrees of freedom for a modeling procedure under the $L_1$ loss function, and investigate the connections between those two definitions. We also propose the extended Akaike information criterion, the adaptive model selection, and the extended generalized cross-validation under the $L_1$ loss function. Finally, we extend the results to M-estimation.

**Keywords:** Adaptive model selection | Covariance penalty | Degrees of freedom | Generalized cross-validation | Least absolute deviations | Modeling procedure

**Article:**

1. Introduction

In the literature of model selection, modeling procedures are usually assessed according to prediction errors. Based on training data $y=(y_1, \ldots, y_T)'$, a modeling procedure is constructed and an estimator $\hat{\mu}$ is produced, and then the prediction error $L(y_0, \hat{\mu})$ measures how well the modeling procedure predicts the future data $y_0=(y_{1,0}, \ldots, y_{T,0})'$. Here $L$ is a loss function and $y_0$ is independently generated from the same mechanism that generated $y$. See for example, Efron (1986, 2004).

It is important to notice that a modeling procedure is different from a model fitting. As pointed out by Ye (1998), a modeling procedure $M$ is a combination process of both model selection and model fitting mapped by
The degrees of freedom (DF) measure the complexity of a model fitted a priori. Ye (1998) introduced the concept of generalized degrees of freedom (GDF) to measure the complexity of a modeling procedure. In the frame of the $L_2$ loss function and Normal assumption, Ye (1998) interpreted the GDF as “the sum of sensitivity of each fitted value to the perturbation in the corresponding observed value.”

As we know, the least squares (LS) regression may fail to produce a reliable estimator when the dataset subjects to heavy-tailed errors, and the least absolute deviations (LAD) estimator enjoys many good properties such as robustness to the outliers. Without any theoretic justification, some recent work used the concept of GDF directly in the selection of the optimal regularization parameter. For example, the GDF was used directly by Nychka et al. (1995), Yuan (2006), and Li et al. (2007) in quantile smooth splines, and Li and Zhu (2008) in $L_1$-norm quantile regression. In their work, a modeling procedure is amount to choosing an optimal regularization parameter, say $\lambda_{opt}$. Therefore, model selection procedures under the $L_1$ loss function deserve further investigation. The main goal of this manuscript is to justify the rationale of two definitions of GDF under the $L_1$ loss function: one based on the sensitivity of perturbation, and the other based on the covariance penalty. Furthermore, the connections between these two definitions, the application of GDF in correcting the bias, and the extension to M-estimation are discussed.

The remaining of the manuscript is organized as follows. In Section 2, we review some modeling procedures under the $L_1$ loss function. In Section 3, we give two definitions of GDF under the $L_1$ loss function, and investigate their connections. In Section 4, we propose the extended Akaike information criterion (EAICR), the adaptive model selection, and the extended generalized cross-validation (EGCV) under $L_1$ loss function. We extend the results to M-estimation in Section 5. In Section 6, we conduct some numerical studies. We conclude the paper with some discussions in Section 7. Some technical proofs are given in Appendix.

2. Review on some modeling procedures

Consider the following linear model:

\[ y_i = x_i' \beta + \epsilon_i, i=1, \ldots, n, \]

where $y_i$'s are independent observations, $x_i$'s are fixed $p \times 1$ predictors, $\beta$ is a $p \times 1$ coefficient vector, and $\epsilon_i$'s are assumed to be independently and identically distributed, with median 0 and a continuous, positive density $f(x)$ in a neighborhood of 0. For convenience, let the first component of $x_i$ be one and incorporate the intercept into $\beta$. Let $\alpha$ be any subset of $\{1, \ldots, p\}$ with $d_\alpha$ distinct elements. Let $\beta_\alpha$ and $x_{\alpha i}$ be the corresponding $d_\alpha$-vectors indexed by set $\alpha$. Let $M_\alpha$ be the sub-model, $y_i = x_{\alpha i}' \beta_\alpha + \epsilon_i$. Under the $L_1$ loss function, if a model $M_\alpha$ is fitted a priori, the LAD estimator of $\beta_\alpha$ is defined as

\[ \hat{\beta}_\alpha = \arg\min_{\gamma_{\alpha}} \sum_{i=1}^{n} |y_i - x_{\alpha i}' \gamma_{\alpha}|. \]
It is well-known that if $\varepsilon_i$ follows a double exponential distribution $DE(0,\sigma)$, $\hat{\beta}_a$ is actually the MLE of $\beta_a$, and the MLE of $\sigma$ is

$$ (3) \hat{\sigma}_a = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - X'_{ai} \hat{\beta}_a \right| / n. $$

Naturally, $\hat{\mu}_i = x'_{ai} \hat{\beta}_a$ provides an estimator for $\mu_i = x' \beta$ under model $M_a$. If $M_a$ is given a priori, then its complexity is measured by the DF, which simply equals $d_a$.

Any criterion can be treated as a modeling procedure when used to choose a model to fit the data. The most famous one is the Akaike information criterion (AIC, Akaike, 1973). For any model $M_a$, the AIC score is defined as

$$ AIC(\alpha) = -2L_\alpha + 2d_a, $$

where $L_\alpha$ is the log-likelihood under model $M_a$. The AIC modeling procedure is to choose the optimal model that minimizes $AIC(\alpha)$. When $\varepsilon_i$ follows a double exponential distribution,

$$ AIC(\alpha) = cn + 2n \log \left( \sum_{i=1}^{n} \left| y_i - X'_{ai} \hat{\beta}_a \right| \right) + 2d_a, $$

where $c_a = 2n \log(2/n) + 2n$. Later work generalizes AIC to the form of $-2L_\alpha + $penalty, where the penalty term penalizes the model complexity. For examples, the penalty term in BIC (Schwarz, 1978) is $d_a \log(n)$, in HQIC (Hannan and Quinn, 1979) is $2d_a \log\log(n)$, and in cAIC (Hurvich and Tsai, 1989) is $2(d_a + 1)(d_a + 2)/(n - d_a - 2)$. Especially, Ronchetti (1985) proposed a robust model selection criterion (AICR). That is, for a given $\lambda > 0$,

$$ (4) AICR(\alpha; \rho, \lambda) = 2 \sum_{i=1}^{n} p \left( \frac{y_i - x'_{ai} \hat{\beta}_a}{\hat{\sigma}} \right) + \lambda d_a, $$

where $\hat{\beta}_a$ is an M-estimator such that $\sum_{i=1}^{n} \psi \left( \frac{y_i - x'_{ai} \hat{\beta}_a}{\hat{\sigma}} \right) x_{ai} = 0$ with $\psi(u) = d\rho(u)/du$ and $\hat{\sigma}$ is a robust estimator of $\sigma$. The LAD estimator is an M-estimator by choosing $\rho(u) = |u|$. Thus for $\lambda = 2$, AICR under the $L_1$ loss function can be rewritten as

$$ (5) AICR(\alpha) = \sum_{i=1}^{n} |y_i - x'_{ai} \hat{\beta}_a| + d_a \hat{\sigma}. $$

3. GDF under $L_1$ loss function

In this section we give two definitions of GDF under the $L_1$ loss function for any modeling procedure, where a modeling procedure includes selecting an optimal sub-model $M^{\alpha^*}$, finding the LAD estimator $\hat{\beta}_a^*$, and then obtaining the fitted values $\hat{\mu}_i = x'_{ai} \hat{\beta}_a$.

3.1. GDF: on the sensitivity to perturbation

Under the Normal assumption and $L_2$ loss function, Ye (1998) defined the GDF as the sum of the sensitivity of each fitted value to perturbation in the corresponding value. We extend the definition to the $L_1$ loss function.
Definition 1

The GDF of a modeling procedure $M$ under the $L_1$ loss function is defined as

$$D(M) = \sum_{i=1}^{n} E \left[ \frac{\partial \hat{\mu}_i(y)}{\partial y_i} \right] = \lim_{\delta \to 0} \sum_{i=1}^{n} E \left[ \frac{\hat{\mu}_i(y+\delta e_i) - \hat{\mu}_i(y)}{\delta} \right],$$

where $\hat{\mu}_i(y)$ is the fitted value of $y_i$ based on data $y$ through the modeling procedure $M$, and $e_i$ is the $i$th column of the $n \times n$ identity matrix.

We can use the Monte Carlo algorithm proposed by Ye (1998) to estimate $D(M)$. The LAD estimator $\hat{\beta}$ may not be unique, but the estimator $\hat{\mu}$ is unique. As in Li et al. (2007), Li and Zhu (2008), we assume that data $(y_i, x_i), \ldots, (y_n, x_n)$ are in general positions such that the LAD estimator is unique throughout the manuscript.

First and foremost, we should verify that the GDF in Definition 1 is consistent with the DF, the number of predictors in model $M_\alpha$, if a model $M_\alpha$ is fitted a priori. Without this, any definition of GDF under the $L_1$ loss function seems risky. For a LAD estimator $\hat{\beta}$ in a given model $M$ with $d$ predictors, following the spirit of Li and Zhu (2008), we define the elbow set as $E_\mathbf{y} = \{1 \leq i \leq n : y_i - x_i' \hat{\beta} = 0\}$ and $N_\mathbf{y}$ as the set of $\mathbf{y}$ such that $|E_\mathbf{y}| > d$ for any given designed covariates $\{x_i\}'s$, where $|\cdot|$ is the cardinal value. We have the following result.

Theorem 1

Assume that model $M_\alpha$ fitted a priori. For any $y \in \mathbb{R}^n \setminus N_\mathbf{y}$, we have

$$\sum_{i=1}^{n} \frac{\partial \hat{\mu}(y)}{\partial y_i} = d_\alpha.$$  

In fact, $P(N_\mathbf{y}) = 0$ if $y_i$ has a continuous density. Thus, $D(M) = E[\sum_{i=1}^{n} \partial \hat{\mu}_i / \partial y_i] = d_\alpha$. That is, the GDF is consistent with the DF when $M_\alpha$ is fitted a priori. The proof of Theorem 1 is given in Appendix.

3.2. GDF: on the covariance penalty

Under the $L_2$ loss function, Efron (2004) used the covariance penalty to correct the bias while using the apparent error (err) to estimate the expected prediction error (Err). He considered a bigger class of loss functions, the $q$-class. However, the $q$-class does not include the $L_1$ loss function. To adopt the spirit of the covariance penalty to the $L_1$ loss function, we provide another definition of GDF under the $L_1$ loss function.

Definition 2

The GDF of a modeling procedure $M$ under $L_1$ loss function is defined as

$$D(M) = \sigma \sum_{i=1}^{n} E \text{sgn}(y_i - \mu_i)(\mu^\wedge_i(y) - \mu_i) = \sigma \sum_{i=1}^{n} \text{Cov}\{\text{sgn}(e_i), \mu^\wedge_i(y)\},$$

where $\text{sgn}(x) = 1, 0, -1$ when $x > 0, = 0, < 0$. 
We can use the parametric Bootstrap (Efron and Tibshirani, 1993) to estimate the GDF in Definition 2. Let \{ε_1^*,...,ε_n^*\} be a random sample of size \(n\) from the empirical distribution of the observed residuals \{\hat{ε}_1, ..., \hat{ε}_n\}, \(y^*\) be a Bootstrap sample of \{y_i^*=\hat{μ}_i+ε_i^*, i=1,...,n\}, \{\hat{μ}_1(y^*),...\hat{μ}_n(y^*)\} be the fitted values based on the Bootstrap sample, and \(F^*\) be the Bootstrap distribution of \(y^*\). Then, the covariance term in (6) can be estimated by \(\sum_{i=1}^{n} \text{Cov}_{F^*}\{\text{sgn}(ε_i^*), \hat{μ}_i(y^*)\}\).

Again, we should verify that the GDF in Definition 2 is consistent with the DF. In model (1), suppose \(\sum_{i=1}^{n} x_i x_i'\) has a positive definite square root \(V_n\) and \(\max_{1\leq i\leq n}|V_n^{-1}x_i|\to 0\) as \(n\to\infty\). By the asymptotic Normality of the LAD estimator shown in Pollard (1991), we have

\[
\frac{1}{\sigma} \sum_{i=1}^{n} \{\text{sgn}(ε_i)(\hat{μ}_i(y)-μ_i)\} = \left[ \sum_{i=1}^{n} \text{sgn}(ε_i)V_n^{-1}x_i \right] \left[ \sum_{i=1}^{n} \text{sgn}(ε_i)V_n^{-1}x_i \right] + oP(1) = d_n + oP(1).
\]

Therefore, the GDF defined in Definition 2 is almost to DF when a model \(M_α\) is fitted a priori.

For a modeling procedure, say AICR, we first select a “best” sub-model \(M_α^*\), then fit \(M_α^*\) to the data and get a LAD estimator. Let \(F_0\) be the distribution of future data \(y_0\). Under the \(L_1\) loss function,

(7) \(\text{Err} = E_{F_0}[\sum_{i=1}^{n} |y_i^0-x_i'\hat{α}_i\hat{β}|]\)

measures how well the modeling procedure predicts the future data \(y_i^0\), and

(8) \(\text{err} = \sum_{i=1}^{n} |y_i-x_i'\hat{α}_i\hat{β}|\)

measures only the goodness of fit. Since a more complex modeling procedure leads to a smaller err, we cannot use only err to compare the performances of modeling procedures. However, since Err reflects a trade-off between the goodness of fit and the complexity of a modeling procedure, we can use it to compare modeling procedures. The following result provides us a heuristic approximation to Err.

**Theorem 2**

Let Err and err be defined in (7) and (8). We have

(9) \(E\{\text{Err}\} = E\{\text{err}\} + E\{\sum_{i=1}^{n} \text{sgn}(y_i-μ_i)(\hat{μ}_i-μ_i)\} + E[h_0(y)],\)

where \(h_0(y) = oP(1)\).

Similar to Efron (1986), we can interpret \(\sum_{i=1}^{n} \text{sgn}(y_i-μ_i)(\hat{μ}_i-μ_i)\) in (9) as an “almost” optimism and \(σD(M)\) as the expected optimism. Thus, Theorem 2 provides an “almost” unbiased estimator of Err,

(10) \(\sum_{i=1}^{n} |y_i-x_i'\hat{α}\hat{β}| + σD(M)\).

3.3. Connections between two definitions

Here we investigate the connections between the above two definitions. Analogous to Lemma 1 in Stein (1981), we propose the following Lemma 1.
Lemma 1

Let \( Y \) be a random variable following \( \text{DE}(\mu, \sigma) \). For any function \( g: \mathbb{R} \rightarrow \mathbb{R} \) with derivative \( g' \) and \( \mathbb{E}|g'(Y)| < \infty \), we have

\[
\mathbb{E}[\text{sgn}(Y-\mu)g(Y)] = \sigma \mathbb{E}[g'(Y)].
\]

The proof of Lemma 1 is omitted since it is a special case of Lemma 3 in Section 5.

Theorem 3

GDFs defined in Definitions 1 and 2 are equivalent when \( e_i \) follows \( \text{DE}(0, \sigma) \).

Theorem 3 is a direct result of Lemma 1; it builds a connection between Definitions 1 and 2 under the double exponential assumption.

4. Some applications of GDF under \( L_1 \) loss functions

4.1. Extended AICR

The AICR criterion defined in (5) can only be used for comparing models instead of modeling procedures. If we estimate the GDF, \( D(M) \), by \( \hat{D}(M) \), then (10) provides us an extended AICR criterion (EAICR) to compare modeling procedures under the \( L_1 \) loss,

\[
(11) \text{EAICR} = \sum_{i=1}^{n} |y_i - \hat{\mu}_i| + \hat{D}(M) \hat{\sigma}. \]

The comparison between EAICR and AICR is analogous to one between EAIC and AIC proposed in Ye (1998) under the \( L_2 \) loss function.

4.2. Adaptive model selection

As an application of the concept of GDF, we propose a version of adaptive model selection under the \( L_1 \) loss function. This is motivated by Shen and Ye (2002), where a version of adaptive model selection was proposed under the Normal assumption and \( L_2 \) loss function, and Shen et al. (2004), where a version of adaptive model selection was proposed under Exponential-Family distributions and Kullback–Leibler loss function.

Under the \( L_1 \) loss function, a robust model selection criterion (4) can be simplified as

\[
(12) \text{AICR}_\lambda = \sum_{i=1}^{n} |y_i - \hat{\mu}_i(M_\alpha)| + \lambda d_\alpha \hat{\sigma},
\]

with respect to sub-models \( M_\alpha \). For a given \( \lambda \), the corresponding model selection procedure is amount to choosing the optimal model \( M_\hat{\alpha}(\lambda) \) to minimize \( \text{AICR}_\lambda \). Noting that for any given \( \lambda > 0 \), (12) is actually a modeling procedure

\[
\mathcal{M}_\hat{\lambda} : \mathbf{y} \rightarrow \text{selected model } M_{\hat{\alpha}}(\lambda) \rightarrow \hat{\mu}(M_{\hat{\alpha}}(\lambda)).
\]

Naturally, we attempt to choose an optimal \( \lambda \), which has the “best” performance. If the GDF of \( M_\lambda, D(M_\lambda) \), is estimated by \( D^\lambda(M_\lambda) \), then the optimal \( \lambda^\wedge \) is obtained by minimizing
(14) $\text{EAICR}_\lambda = \sum_{i=1}^{n} |y_i - \hat{\mu}(M\hat{\lambda}(\lambda))| + D^\wedge(M_\lambda) \hat{\sigma}_\lambda,$

with respect to $\lambda > 0$. Once a data-adaptive $\hat{\lambda}$ is obtained, we can select an optimal sub-model, $M\hat{\lambda}$ ($\hat{\lambda}$).

4.3. Extended generalized cross-validation

Under the $L_2$ loss function, the ordinary leave-one-out cross-validation (OCV) is a popular method (e.g., Stone, 1977) to provide an unbiased estimator of $\text{Err}$. Under the $L_1$ loss, it is natural for us to estimate $\text{Err}$ of model $M\alpha$ in (7) by

$$\text{OCV}_{M\alpha} = \sum_{i=1}^{n} |y_i - x'_{\alpha[i]} \hat{\beta}_{\alpha[i]}|,$$

where $\hat{\beta}_{\alpha[i]}$ is the LAD estimator of $\beta\alpha$ using data without subject $i$.

Following the spirit in Craven and Wahba (1979), we obtain the following leave-one-out lemma.

**Lemma 2**

For $k$ and $z$, let $\hat{\beta}_\alpha[k,z]$ be the solution of

$$\arg \min_{y\alpha} |z - x'_{\alpha[k]} + \sum_{i \neq k} |y_j - x'_{\alpha[j]}|.$$

Then $\hat{\beta}_\alpha[i,\tilde{y}_i] = \hat{\beta}_\alpha[i]$, where $\tilde{y}_i = x'_{\alpha[i]} \hat{\beta}_\alpha[i]$.

If we consider

$$y_i - x'_{\alpha[i]} \hat{\beta}_\alpha[i] = (y_i - x'_{\alpha[i]} \hat{\beta}_\alpha)/ (1 - h^*_{\alpha[i]}),$$

where $h^*_{\alpha[i]} = (x'_{\alpha[i]} \hat{\beta}_\alpha - x'_{\alpha[i]} \hat{\beta}_\alpha[i]) / (y_i - x'_{\alpha[i]} \hat{\beta}_\alpha[i])$, then from Lemma 2, $x'_{\alpha[i]} \hat{\beta}_\alpha[i,\tilde{y}_i] = x'_{\alpha[i]} \hat{\beta}_\alpha[i]$, where $y_i - x'_{\alpha[i]} \hat{\beta}_\alpha[i]$. Thus we have

$$h^*_{\alpha[i]} = \left( \frac{x'_{\alpha[i]} \hat{\beta}_\alpha[i,\tilde{y}_i] - x'_{\alpha[i]} \hat{\beta}_\alpha[i]}{y_i - \tilde{y}_i} \right) = h_{\alpha[i]} = \frac{\partial \hat{\beta}_\alpha}{\partial y_i}.$$

Therefore, we can approximate OCV as

$$\text{OCV}_{M\alpha} = \sum_{i=1}^{n} |y_i - x'_{\alpha[i]} \hat{\beta}_\alpha[i]| / (1 - h^*_{\alpha[i]}).$$

Furthermore, if we replace $h_{\alpha[i]}$ by $E \{ \Sigma_{j=1}^{n} h_{\alpha[j]} \} / n = D(M\alpha) / n$, we obtain EGCV defined by

(15) $\text{EGCV}_{M\alpha} = \sum_{i=1}^{n} |y_i - x'_{\alpha[i]} \hat{\beta}_\alpha[i]| / (1 - D(M\alpha) / n).$

Noting that $1 / (1 - x) = 1 + x$ when $x$ is small. Thus, the EGCV is similar to the EAICR. EGCV is also consistent to GACV proposed by Yuan (2006) if we apply Definition 1 of GDF in Section 3.1. The above connection also exists when we are interested in comparing different modeling procedures. For example, we consider the adaptive model selection procedure in Section 3.2. For a given $\lambda$, $M\lambda$ is a modeling procedure, and the OCV method provides an unbiased estimator of the prediction error for $M\lambda$. For given $\lambda$ and $i$, let $\alpha^* [i](\lambda)$ be the selected optimal model through
modeling procedure $M_\lambda$ using data without subject $i$, and let $\hat{\mu}_\lambda^{[i]} (=x'\alpha_i^{[i]} \hat{\beta}_\alpha^{[i]})$ be the fitted value of $y_i$ based on $M_\lambda^{[i]}(\lambda)$. Then OCV is defined as

$$OCV_\lambda = \sum_{i=1}^n |y_i - \hat{\mu}_\lambda^{[i]}|.$$ 

Following similar arguments, for a modeling procedure $M_\lambda$, we define

$$EGCV_\lambda = \sum_{i=1}^n |y_i - \hat{\mu}_\lambda|/(1-D(M_\lambda)/n),$$

where $\hat{\mu}_\lambda$ is the fitted value of $y_i$ using all the data, and $D(M_\lambda)$ is the GDF of the modeling procedure $M_\lambda$. Thus $EGCV_\lambda$ is similar to $EAICR_\lambda$.

5. GDF in M-estimation

In this section, we extend concepts of the GDF to M-estimation. In linear model (1), an M-estimator of $\beta_\alpha$ under any sub-model $M_\alpha$ is

$$\hat{\beta}_\alpha = \arg\min_{y_\alpha} \sum_{i=1}^n \rho(y_i - x'\alpha_i y_\alpha),$$

where $\rho(\cdot)$ is a known function of errors with the influence function $\phi(u)(=d\rho(u)/du)$. Also, $\hat{\mu}_\alpha (=x'\alpha \hat{\beta}_\alpha)$ provides us a robust estimator of $\mu_\alpha(=x'\beta)$ under the model $M_\alpha$. For any modeling procedure under M-estimation, say AICR in (4), we first select a “best” sub-model $M_\lambda$, and then fit the data to get an M-estimator $\hat{\beta}_\lambda$.

Consider a generalized $tp,q(\mu,\sigma)$ distribution with the probability density

$$f(x;\mu,\sigma,p,q)=cp,q\sigma^{-1}(1+|x-\mu|^p/(q\sigma^q))^{-q-1/p,i=1,\ldots,n},$$

where $c_{p,q}=(p/2)^{-1/q} B^{-1}(1/p, q)$. We have the following result

**Lemma 3**

Let $Y$ be a random variable following a generalized $tp,q(\mu,\sigma)$ distribution. Let $g:R \rightarrow R$ have derivative $g'$ with $E[g'(Y)]<\infty$. Then

$$E[\psi(Y-\sigma]\mu g(Y))]=\sigma E[g'(Y)],$$

where $\psi(u)=(p+1/q)|u|^{p-1}(1+|u|^p/q)-1\sgn(u)$.

Generalized $tp,q(\mu,\sigma)$ distribution is related to many distributions. For example, when $q \rightarrow \infty$, $Y$ in Lemma 3 has a generalized error distribution with probability density

$$f(x;\mu,\sigma,1,\infty)=p/(2\sigma\Gamma(1/p))\exp \{-|x-\mu|^p/\sigma^p\}.$$ 

Thus, Lemma 1 in Section 3 and Stein's lemma are two special cases of Lemma 3 when $p=1,q \rightarrow \infty$ and $p=2,q \rightarrow \infty$.

When $Y$ follows generalized $tp,q(\mu,\sigma)$, an M-estimator is obtained by choosing $p(u)=(p+1/q)\log(1+|u|^p/q)$. In the light of Lemma 3, we provide two definitions of the GDF for any modeling procedure $M$ in M-estimation.
Definition 3

Suppose $y_i$ in linear model (1) follows generalized $t_{p,q}(\mu,\sigma)$. The GDF in M-estimation is defined as

$$\Sigma_{i=1}^{n} E\left[\frac{\partial \hat{\mu}_i(y)}{\partial \mu_i} \right] = \frac{1}{\sigma} \sum_{i=1}^{n} E\left[\psi\left(\frac{y_i - \mu_i}{\sigma}\right)(\hat{\mu}(y) - \mu_i)\right] = \frac{1}{\sigma}$$

$$\Sigma_{i=1}^{n} \text{Cov}\left\{\psi\left(\frac{y_i - \mu_i}{\sigma}\right), \hat{\mu}_i(y)\right\},$$

where $\psi(u) = (p+1/q)p|u|^{p-1}(1+|u|^p/q)^{-1}\text{sgn}(u)$.

Definition 4

Suppose $y_i$ in linear model (1) satisfies $E[\psi((y_i - \mu_i)/\sigma)g(y_i)] = E[g'(y_i)]$. The GDF in M-estimation with influence function $\psi$ is defined as

$$\Sigma_{i=1}^{n} E\left[\frac{\partial \hat{\mu}_i(y)}{\partial \mu_i} \right] = \frac{1}{\sigma} \sum_{i=1}^{n} E\left[\psi\left(\frac{y_i - \mu_i}{\sigma}\right)(\hat{\mu}(y) - \mu_i)\right] = \frac{1}{\sigma}$$

$$\Sigma_{i=1}^{n} \text{Cov}\left\{\psi\left(\frac{y_i - \mu_i}{\sigma}\right), \hat{\mu}_i(y)\right\},$$

Thus, those results of the GDF can be extended to M-estimation. For example, we can use the defined GDF to correct the bias in M-estimation.

6. Numerical studies

In this section, we conduct simulation studies under the $L_1$ loss function to investigate: (1) the performance of GDF when a model is given a priori; (2) the performance of GDF for modeling procedures.

Consider linear models (1) with sample size $n=50$ for all the simulation studies. We generated the $p$-vector predictors $x_i$'s independently from a multivariate Normal distribution with mean $0_p$ and covariance matrix $\Sigma_x=(1-\rho)I_p+\rho 1_p 1_p'$; random errors $\epsilon$'s independently from three different types of distributions: (i) standard double exponential distribution $\text{DE}(0,1)$; (ii) standard Normal distribution $\text{N}(0,1)$; and (iii) standardized $t$-distribution $t(3)$. For simplicity, we only consider the zero intercept. We denote $aq0p-q$ as the vector with the first $q$ components being $a$ and the remaining being 0. Each simulation setting is repeated by 500 times.

We examine Theorem 1 under three residual distributions by both Examples 1 and 2. Meanwhile, we compare the Monte Carlo procedure in Definition 1 with the Bootstrap procedure in Definition 2 for computing the GDF in Example 1.

Example 1

Assume that $p=8$ and $\rho=0.5$ or 0.8. The true parameter $\beta_0=(3,1.5,0,0,2,0,0,0)'$. Five models, the truly most parsimonious model $M_0=\{1, 2, 5\}$, $M_1=M_0 \cup \{3\}, M_2=M_1 \cup \{4\}, M_3=M_2 \cup \{6\}$, and the full model $M_4$ are considered. GDFs are evaluated when each of the above five models is given a priori.
We compute the GDF using both the Monte Carlo procedure in Definition 1 and the Bootstrap procedure in Definition 2. We generate $B=500$ Bootstrap samples when the parametric Bootstrap procedure is applied.

Results of the averages and standard deviations of GDF from 500 iterations for Example 1. Those simulation results support Theorem 1 for all three residual distributions; that is, if a model is given a prior, the GDF in Definition 1 is equal to the number of predictors in the model. The Bootstrap procedure also provides us a good estimator of the GDF, especially when the given model is relatively close to the true one. However, the Bootstrap method generates more bias than the Monte Carlo method, especially when the given model is a full model. Although we only justify the connections of the two definitions under the assumption of the double exponential distribution, our simulation studies demonstrate the connections of these two definitions under all three distributions. Table 1 also shows us that the strength of the correlation among predictors does not affect the simulation results significantly.

**Example 2**

Assume that $p=20$, and the true model $\beta_0=25015$. The GDFs are evaluated for six models fitted a priori: under-fitted model $M_1=\{1\}$, correct model $M_{5,1}=\{1,\ldots,5\}$, wrong model $M_{5,2}=\{3,\ldots,7\}$, over-fitted model $M_{10,1}=\{1,\ldots,10\}$, noise-predictor-only model $M_{10,2}=\{6,\ldots,15\}$ and full model $M_{20}=\{1,\ldots,20\}$.

<table>
<thead>
<tr>
<th></th>
<th>$M_0$</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_3$</th>
<th>$M_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DF</strong></td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td><strong>DE(0,1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDF</td>
<td>2.99 (0.31)</td>
<td>3.98 (0.35)</td>
<td>5.00 (0.39)</td>
<td>5.95 (0.43)</td>
<td>7.96 (0.53)</td>
</tr>
<tr>
<td>GDF$_{BP}$</td>
<td>3.11 (0.10)</td>
<td>4.04 (0.13)</td>
<td>4.52 (0.15)</td>
<td>5.70 (0.16)</td>
<td>7.66 (0.18)</td>
</tr>
<tr>
<td><strong>N(0, 1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDF</td>
<td>3.00 (0.42)</td>
<td>4.00 (0.50)</td>
<td>5.00 (0.52)</td>
<td>6.06 (0.55)</td>
<td>8.07 (0.87)</td>
</tr>
<tr>
<td>GDF$_{BP}$</td>
<td>3.10 (0.11)</td>
<td>4.00 (0.13)</td>
<td>5.05 (0.15)</td>
<td>6.19 (0.16)</td>
<td>7.84 (0.20)</td>
</tr>
<tr>
<td><strong>t(3)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDF</td>
<td>3.00 (0.42)</td>
<td>4.04 (0.51)</td>
<td>4.97 (0.65)</td>
<td>6.00 (0.41)</td>
<td>8.02 (0.40)</td>
</tr>
<tr>
<td>GDF$_{BP}$</td>
<td>3.02 (0.13)</td>
<td>4.48 (0.15)</td>
<td>5.02 (0.15)</td>
<td>5.94 (0.16)</td>
<td>7.08 (0.19)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$M_0$</th>
<th>$M_1$</th>
<th>$M_2$</th>
<th>$M_3$</th>
<th>$M_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ρ=0.5</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DE(0,1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDF</td>
<td>3.00 (0.28)</td>
<td>3.99 (0.28)</td>
<td>5.02 (0.30)</td>
<td>6.03 (0.36)</td>
<td>8.00 (0.45)</td>
</tr>
<tr>
<td>GDF$_{BP}$</td>
<td>2.97 (0.12)</td>
<td>4.02 (0.14)</td>
<td>4.92 (0.14)</td>
<td>6.04 (0.18)</td>
<td>7.27 (0.20)</td>
</tr>
<tr>
<td><strong>N(0, 1)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDF</td>
<td>3.00 (0.54)</td>
<td>4.00 (0.63)</td>
<td>5.00 (0.67)</td>
<td>5.98 (0.48)</td>
<td>8.05 (0.76)</td>
</tr>
<tr>
<td>GDF$_{BP}$</td>
<td>3.01 (0.28)</td>
<td>4.17 (0.31)</td>
<td>4.60 (0.37)</td>
<td>5.57 (0.40)</td>
<td>7.67 (0.47)</td>
</tr>
<tr>
<td><strong>t(3)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDF</td>
<td>3.00 (0.31)</td>
<td>3.99 (0.34)</td>
<td>5.01 (0.31)</td>
<td>6.00 (0.32)</td>
<td>8.00 (0.35)</td>
</tr>
<tr>
<td>GDF$_{BP}$</td>
<td>3.06 (0.12)</td>
<td>4.08 (0.14)</td>
<td>4.94 (0.14)</td>
<td>5.42 (0.15)</td>
<td>6.93 (0.20)</td>
</tr>
</tbody>
</table>

*Note 1:* The mean (standard deviation) of GDF from 500 iterations are computed. *Note 2:* GDF is computed by Monte Carlo procedure with repeated time $T=200$. *Note 3:* GDF$_{BP}$ is computed by Bootstrap procedure with $B=500$ Bootstrap samplers.
This example shed more lights on Theorem 1. In this example, we check the validity of Theorem 1 when the model fitted a priori is partly wrong, over-fitted, under-fitted, and only include noisy predictors. Since Example 1 shows no influence of the correlation coefficients among all predictors, we only give the simulation output of $\rho=0.5$. Simulation results for Example 2 are summarized in Table 2. The results show that Theorem 1 is correct for all three residual distributions and all cases of models fitted a priori.

Table 2. GDF of models given a priori in Example 2, $\beta_0=(2\,0,15)$.  

<table>
<thead>
<tr>
<th>$\beta_0$</th>
<th>$\epsilon_i$</th>
<th>Model</th>
<th>$M_1$</th>
<th>$M_{5,1}$</th>
<th>$M_{5,2}$</th>
<th>$M_{10,1}$</th>
<th>$M_{10,2}$</th>
<th>$M_{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE(0,1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDF</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N(0,1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t(3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DE(0,1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GDF BP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N(0,1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>t(3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note 1: The mean (standard deviation) of GDF from 500 iterations are computed from Monto Carlo procedure. Note 2: The DF of $M_i$ or $M_{i,j}$ is $i$. Note 3: Under-fitted model $M_1=\{1\}$, correct model $M_{5,1}=\{1,\ldots,5\}$, wrong model $M_{5,2}=\{3,\ldots,7\}$, over-fitted model $M_{10,1}=\{1,\ldots,10\}$, noise-predictor-only model $M_{10,2}=\{6,\ldots,15\}$, full model $M_{20}=\{1,\ldots,20\}$.

In the following Example 3, we examine the performance of GDF as a measurement of the complexity of a modeling procedure by comparing EAICR and AICR, EGCV and GCV in model selection.

Example 3

Assume that $p=10$, $\rho=0$, 0.5 or 0.8. We study two true models: (1) $\beta_0=010$ and (2) $\beta_0=2307$.

We compare model selection performances by minimizing AICR, EAICR, GCV, and EGCV, respectively. In Table 3, we report the averaged numbers of predictors (NUM), the percentages of selecting the exact true model (CFR, correctly fitted ratio), the percentages of selecting models including the true model plus some redundant predictors (OFR, over-fitted ratio). As we expect EAICR performs much better than AICR, and EGCV works much better than GCV. In particular, the percentages of selecting the exact true model by EAICR and EGCV are much higher than one by AICR and GCV, respectively. Among all four criteria, EAICR performs best in model selection.

Table 3. Application of GDF to variable selection in Example 3.
<table>
<thead>
<tr>
<th>$\epsilon_i$</th>
<th>$\rho$</th>
<th>$\beta_0=0_{10}$</th>
<th>$\beta_0=2_{30}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUM</td>
<td>CFR (%)</td>
<td>OFR (%)</td>
<td>NUM</td>
</tr>
<tr>
<td>DE(0, 1)</td>
<td>0.8</td>
<td>AICR</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAIKR</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCV</td>
<td>2.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EGCV</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>AICR</td>
<td>1.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAIKR</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCV</td>
<td>2.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EGCV</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>AICR</td>
<td>1.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAIKR</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCV</td>
<td>2.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EGCV</td>
<td>0.78</td>
</tr>
<tr>
<td>N(0, 1)</td>
<td>0.8</td>
<td>AICR</td>
<td>2.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAIKR</td>
<td>0.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCV</td>
<td>2.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EGCV</td>
<td>1.58</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>AICR</td>
<td>2.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAIKR</td>
<td>0.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCV</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EGCV</td>
<td>1.72</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>AICR</td>
<td>2.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAIKR</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCV</td>
<td>2.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EGCV</td>
<td>1.48</td>
</tr>
<tr>
<td>t(3)</td>
<td>0.8</td>
<td>AICR</td>
<td>1.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAIKR</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCV</td>
<td>2.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EGCV</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>AICR</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAIKR</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GCV</td>
<td>2.35</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EGCV</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>AICR</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EAIKR</td>
<td>0.11</td>
</tr>
</tbody>
</table>
**Table**

<table>
<thead>
<tr>
<th></th>
<th>GCV</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2.25</td>
<td>11</td>
<td>89</td>
<td>4.61</td>
<td>23</td>
<td>77</td>
</tr>
<tr>
<td>EGCV</td>
<td>0.81</td>
<td>80</td>
<td>20</td>
<td>3.57</td>
<td>83</td>
<td>17</td>
</tr>
</tbody>
</table>

Note 1: NUM is the averaged number of predictors in the selected model. Note 2: CFR is the percentage of selecting the exact true model \(M_0\). Note 3: OFR is the percentage of over-fitted models by one additional predictor.

7. Discussion

Under the \(L_2\) loss function and Normal assumption, Stein (1981) developed his Lemma 1 based on which SURE estimator was proposed, Ye (1998) developed his concept of the GDF, and Shen and Ye (2002) proposed their version of adaptive model selection. In this manuscript, we obtain these three parallel results under the \(L_1\) loss function and provide two definitions of the GDF and justify their connections under the double exponential assumption. Following the steps of Ye (1998), we emphasize that we should realize the crucial difference between a model given a priori and a modeling procedure. Since a modeling procedure is the combination of both model selection and model fitting, the GDF works better than the DF while being used to adjust the bias in model selection. As a natural generalization, we also extend those results to M-estimation.

Acknowledgments

We thank the associate editor and the anonymous referee for constructive comments that substantially improved the presentation of the paper.

Appendix A.

**Proof of Theorem 1**

We aim to prove that (1) \(Nx\) is a finite collection of hyperplanes in \(R^n\) and (2) \(\hat{\beta} = \hat{\beta}(y)\) is a continuous function of \(y\) if \(\beta^*\) is unique.

If a LAD estimate \(\hat{\beta}^*\) implies \(|Ey|=k>d\), then there are \(k\) observations such that \(y_{ij} = x_{ij}^\prime \hat{\beta}^*\) for \(j=1,\ldots,k\). Stack them in matrix form, we have \(y_k = X_k \hat{\beta}^*\), where \(y_k\) is a \(k\)-vector and \(X_k\) is a \(k\times d\) matrix. Let \(\hat{\beta}^k = X_k^+ y_k\), where \(X_k^+\) is the unique Moore–Penrose generalized inverse of \(X_k\). Then \(y_k = X_kX_k^+ y_k\). Since \(\text{rank}(X_kX_k^*)<k\), \(y_k\) is in a hyperplane in \(R^n\). There are at most \(\sum_{k=d+1}^n \binom{n}{k}\) such hyperplanes. Thus (1) holds.

For any fixed \(y_0\), let \(y^m\) be a sequence converging to \(y_0\). We want to show that \(\hat{\beta}^m(y)\) converges to \(\hat{\beta}^*(y_0)\). Since \(\hat{\beta}^m(y)\) is bounded we only need to show that every converging subsequence converges to \(\hat{\beta}^*(y_0)\). Suppose that subsequence \(\hat{\beta}^m(y_{mk})\rightarrow \hat{\beta}^*\).

\[
\sum |y_{i0} - x_i^\prime \hat{\beta}^m(y_0)| + \sum |y_{i0} - x_i^\prime \hat{\beta}^*(y_0)| = \sum |y_{i0} - x_i^\prime \hat{\beta}^m(y_{mk})| + \sum |y_{i0} - x_i^\prime \hat{\beta}^*(y_0)| + \sum |y_{i0} - x_i^\prime \hat{\beta}^*(y_{mk})| + Ak,
\]

where \(Ak = \sum |y_{i0} - x_i^\prime \hat{\beta}^*(y_{mk})| + \sum |y_{i0} - x_i^\prime \hat{\beta}^*(y_{mk})| + \sum |y_{i0} - x_i^\prime \hat{\beta}^*(y_{mk})|\). Noting that \(Ak<2\sum |y_{i0} - x_i^\prime \hat{\beta}^*(y_{mk})| + \sum |y_{i0} - x_i^\prime \hat{\beta}^*(y_{mk})|\), we have \(\sum |y_{i0} - x_i^\prime \hat{\beta}^*(y_0)| \geq \sum |y_{i0} - x_i^\prime \hat{\beta}^*(y_{mk})|\). We have \(\hat{\beta}^* = \hat{\beta}^*(y_0)\) since \(\hat{\beta}^*(y_0)\) is the unique minimizer of \(\sum |y_{i0} - x_i^\prime \gamma|\). Thus (2) holds.

For any fixed \(y_0 \in R^n \setminus Nx\), let \(E_0\) and \(E_0^c\) denote the elbow set based on \(y_0\) and its complement. From (2), we can choose small enough \(\epsilon>0\) such that
for \( y \in \text{Ball}(y_0, \varepsilon) \subset \mathbb{R}^n \setminus N_x \) we have \( y_i - x_i \beta(y) \neq 0 \) if \( i \in E_{yc} \). Therefore, \( E_0 \subset E_{yc} \) and \( E_0 \supset E_y \). From (1), \( |E_y| = |E_0| = d \), and thus \( E_y \) is locally constant with respect to \( y \). This implies \( \sum_i = 1 n \partial \mu_i / \partial y = 1 \) if \( i \in E_y \) and \( \sum_i = 1 n \partial \mu_i / \partial y = 0 \) if \( i \in E_{yc} \). Thus Theorem 1 is obtained. □

**Proof of Theorem 2**

By expansion (for simplicity, one may think about Taylor expansion in the sense of Phillips, 1991) at \( \mu_i \), for new data \( y_i^0 \) generated from the same mechanism generating \( y_i \), we have

\[
\sum_{i=1}^{n} |y_i^0 - \bar{\mu}_i| = \sum_{i=1}^{n} |y_i^0 - \mu_i| + \sum_{i=1}^{n} \text{sgn}(y_i^0 - \mu_i)(\mu_i - \bar{\mu}_i) + f(0) \sum_{i=1}^{n} (\mu_i - \bar{\mu}_i)^2 + o_p(1),
\]

where \( f(0) \) is the density of \( \varepsilon_i \) at zero. Thus (9) holds because \( E_1 \{ \sigma_i = 0 \} \} = E_1 \{ \sigma_i = 1 \} \} = 0 \). □

**Proof of Lemma 2**

\[
|\tilde{y}_i - \tilde{x}_i \tilde{\beta}([n]) + \sum_{j \neq i} |y_j - \tilde{x}_j \tilde{\beta}([n])| < \sum_{j \neq i} |y_j - \tilde{x}_j \gamma_{zt}| \leq |\tilde{y}_i - \tilde{x}_i \gamma_{zt}| + \sum_{j \neq i} |y_j - \tilde{x}_j \gamma_{zt}|.
\]

**Proof of Lemma 3**

If \( Y \) has generalized \( t_{p,q}(0,1) \), then by integration by parts,

\[
E[\tilde{g}'(Y)] = g(Y)c_{p,q}(1 + |y|^p/q)^{-q-1/p} + \int_{-\infty}^{+\infty} g(Y)c_{p,q}(1 + |y|^p/q)^{-q-1/p} dy = \int_{-\infty}^{+\infty} g(Y)c_{p,q}(1 + |y|^p/q)^{-q-1/p} (1 + |y|^p/q)(pq^{-1}|y|^{p-1}) \text{sgn}(y) dy = E[\tilde{g}(Y)(q + 1/p)(1 + |y|^p/q)(pq^{-1}|y|^{p-1}) \text{sgn}(Y)].
\]

Thus we have \( E[\tilde{g}'(Y)] = E[\tilde{g}(Y)\psi(Y)] \) with \( \psi(u) = (q + 1/p)(|u|^{p-1})(1 + |u|^p/q)\text{sgn}(u) \).

If \( Y \) follows generalized \( t_{p,q}(\mu, \sigma) \), then \( \tilde{Y} = (Y - \mu) / \sigma \) follows \( t_{p,q}(0,1) \). Let \( \tilde{g} (y) = g(\mu + \sigma y) \). Then

\[
E[\tilde{g}'(Y)] = \frac{1}{\sigma} E[\tilde{g}'(\tilde{Y})] = \frac{1}{\sigma} E[\psi(\tilde{Y}) \tilde{g}'(\tilde{Y})] = \frac{1}{\sigma} E[\psi(\tilde{Y} - \mu / \sigma) g(Y)].
\]

Thus Lemma 3 is shown. □

**References**


M. Yuan. GACV for quantile smoothing splines, Computational Statistics and Data Analysis, 5 (2006), pp. 813-829