Trimming, optimal transportation, overfitting and statistical applications

Eustasio del Barrio

Trimming, optimal transportation, overfitting and statistical applications.

Eustasio del Barrio

Departamento de Estadística, Universidad de Valladolid

We present some ideas about *essential* model validation based on trimming and the minimal distance method. We argue about the interest of the Wasserstein metric in this sense and show the connection in this case of essential model checking to a problem of optimal incomplete transportation of mass. We describe also some unexpected overfitting effect caused by trimming and provide pointers to relevant literature in which this overfitting is used for some statistical applications.

Keywords: Model checking, trimming, optimal transportation, overfitting.

Trimming methods are a common tool in the design of robust statistical procedures. Trimming criteria are often based on some (implicit) spherical or elliptical assumptions for the model, removing, consequently, data which is far from the center of the sample in a symmetric way. Tha main drawback of this approach is that the possible contamination in practice is not only due to outliers. Among the proposed alternatives to overcome these difficulties we focus on those minimizing some distance, leading to the "impartial" trimming introduced in [7] and in greater generality in [6]. This impartial trimming methodology is based on the idea that the trimming zone should be determined by the data and has been successfully applied to different statistical problems including location estimation, regression problems, cluster analysis or principal component analysis.

This idea is very appropriate in goodness-of-fit or, more generally, in model validation, where the procedures are often based on minimizing distances. In fact, quite often the researcher is not really concerned about exact coincidence, but rather wants to guarantee that the random generator does not differ too much from the proposed model. The usual approach in the statistical literature to this 'not differ too much' consists of fixing a certain parameter related to the distribution of the random generator (possibly the distribution itself) and a distance in the parameter parameter space and cheching whether the 'distance' between the data and the model does not exceed a given threshold. To bo more precise, we observe $X \sim P$ and want to assess whether $P \in \mathcal{F}$. Often $P \in \mathcal{F}$ is not really important but rather $P \simeq \mathcal{F}$. We fix $\theta = \theta(P)$ and a metric, d. Rather than testing $H : \theta(P) \in \theta(\mathcal{F}) := \{\theta(Q) : Q \in \mathcal{F}\}$ we consider $H : d(\theta(P), \theta(\mathcal{F})) \leq \Delta$ vs. $K : d(\theta(P), \theta(\mathcal{F})) > \Delta$ or $H : d(\theta(P), \theta(\mathcal{F})) > \Delta$ vs. $K : d(\theta(P), \theta(\mathcal{F})) \leq \Delta$, the last choice of null hypothesis in agreement with the fact that the worst possible error is often accepting a model which is wrong.

In a recent series of papers ([1], [3], [4]) we propose a different approach to the problem with a motivation somehow influenced by robust statistics. A practitioner trying to fit, say, a regression model, would not discard the model if the deviation were due only to a few disturbing observations. She/he would rather remove these observations and consider the model as essentially correct. Allowing to remove an α fraction of the data for a better comparison to a pattern amounts to replace the empirical distribution,

$$\frac{1}{n}\sum_{i=1}^{n}\delta_{x_i} \qquad \text{by} \qquad \frac{1}{n}\sum_{i=1}^{n} b_i\delta_{x_i}$$

$b_i = 0$ for observations in the bad set, $b_i/n = \frac{1}{n-k}$ otherwise, where k is the number of trimmed observations, $k \leq n\alpha$ and $\frac{1}{n-k} \leq \frac{1}{n}\frac{1}{1-\alpha}$. Instead of keeping/removing we could increase weight in good ranges (by $\frac{1}{1-\alpha}$ at most) and downplay in bad zones, not necessarily removing

$$\frac{1}{n}\sum_{i=1}^{n} b_i\delta_{x_i}, \text{ with } 0 \leq b_i \leq \frac{1}{(1-\alpha)}, \text{ and } \frac{1}{n}\sum_{i=1}^{n} b_i = 1.$$

This motivates the following definition of α-*trimmings* of a probability measure.

Definition 1 *Given a measurable space (X, β) and probability measures, P, Q, on (X, β) we say that Q is an α-trimming of P if*

$$Q \ll P \quad \textit{and} \quad \frac{dQ}{dP} \leq \frac{1}{1-\alpha}.$$

We write $\mathcal{R}_\alpha(P)$ for the set of α-trimmings of P. We can allow $0 \leq \alpha \leq 1$ if $\mathcal{R}_0(P) = \{P\}$, $\mathcal{R}_1(P) = \{Q : \quad Q \ll P\}$. Note that $Q \in \mathcal{R}_\alpha(P)$ iff $Q \ll P$ and $\frac{dQ}{dP} = \frac{1}{1-\alpha}f$ with $0 \leq f \leq 1$. If $f \in \{0, 1\}$ then $f = I_A$ with $P(A) = 1 - \alpha$: trimming reduces to $P(\cdot|A)$. Our definition allows to play down the weight of some regions of the measurable space without completely removing them from the feasible set. This turns $\mathcal{R}_\alpha(P)$ into a well behaved set. Some nice properties of it are collected in the next Proposition.

Proposition 1 *For any probability measure, P,*

(a) $\mathcal{R}_{\alpha_1}(P) \subset \mathcal{R}_{\alpha_2}(P)$ *if* $\alpha_1 \leq \alpha_2$.

(b) $\mathcal{R}_\alpha(P)$ *is a convex set.*

(c) *For* $\alpha < 1$, $Q \in \mathcal{R}_\alpha(P)$ *if and only if* $Q(A) \leq \frac{1}{1-\alpha}P(A)$ *for all* $A \in \beta$.

(d) *If* $\alpha < 1$ *and* (X, β) *is a separable metric space endowed with its Borel* σ*-field then* $\mathcal{R}_\alpha(P)$ *is closed for the topology of weak convergence. If* X *is also complete then* $\mathcal{R}_\alpha(P)$ *is compact.*

Another remarkable fact about trimmings is that we can translate every model to a fixed reference probability to perform the trimming procedure on it. More precisely,

Proposition 2 *If T transports Q to P, then*

$$\mathcal{R}_\alpha(P) = \left\{Q^* \circ T^{-1} : \, Q^* \in \mathcal{R}_\alpha(Q)\right\}.$$

Here, be T trasporting Q to P we mean that T is a measurable map such that $Q(T^{-1}(A)) = P(A)$, see [2] for details. On the real line, for instance, we can take $Q = U(0,1)$, $P \sim F$, $T = F^{-1}$ to see that

$$\mathcal{R}_\alpha(P) = \{P_h : \, h \in \mathcal{C}_\alpha\},$$

where $\mathcal{C}_\alpha := \left\{h \in \mathcal{AC}[0,1] : \, h(0) = 0,\, h(1) = 1,\, 0 \le h' \le \frac{1}{1-\alpha}\right\}$ and P_h is the probability with d.f. $h \circ F$. For separable, complete $\mathcal{X}$ we could take as well $Q = U(0,1)$ and T the Skorohod-Dudley-Wichura map, but this would be of very limited use in applications. For $\mathcal{X} = \mathbb{R}^k$, it is more interesting to consider $Q \ll \ell^k$ and T the Brenier-McCann map, namely the unique cyclically monotone map transporting Q to P (see [2]). Once we fix our reference probability and our transportation scheme we can write

$$\mathcal{R}_\alpha(P) = \{P_R : \, R \in \mathcal{R}_\alpha(P_0)\}.$$

The idea of trimmings of a distribution can be used to assess whether the *core* of the distribution underlying the data can be assumed to follow a given model. As a measure of this essential fit we can use

$$\tau_{\alpha,1}(P, \mathcal{F}) = \inf_{R \in \mathcal{R}_\alpha(P_0), Q \in \mathcal{F}} d(P_R, Q_R),$$

where P_0 is some reference probability. We refer to the procedure of trimming in order to minimize the right-hand side above as *common trimming*. This way of trimming has been explored in [1] for the simple model consisting of a fixed probability measure as well as for two-sample problems. In [3] it has been applied to the assessment of essential normality of a data set. The method exhibits good behaviour, leads to tractable asymptotic distributions and is computationally convenient. We refer to [1] and [3] for details.

On the other hand, a more natural approach that need not consider some arbitrary reference is to measure essential model adequacy by

$$\tau_{\alpha,2}(P, \mathcal{F}) = \inf_{R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(\mathcal{F})} d(R_1, R_2),$$

$\mathcal{R}_\alpha(\mathcal{F})$ being the set of all α-trimmings of probabilities in $\mathcal{F}$. We refer to this way of trimming as *independent trimming*. Independent trimming is linked to the common idea that trimming elliminates contamination. In fact, if $P_2 = (1-\varepsilon)\, P_1 + \varepsilon\, Q$, then

$$(1-\alpha)\, P_1(A) \le (1-\varepsilon)\, P_1(A) + \varepsilon\, Q(A) \qquad \forall A.$$

Hence, $P_1 \in \mathcal{R}_\alpha(P_2)$ ($\alpha \geq \varepsilon$).

Independent trimming in the case of a simple $\mathcal{F} = \{Q\}$ leads to the idea of *best trimmed matchings*, namely, pairs $(P_\alpha, Q_\alpha) \in \mathcal{R}_\alpha(P) \times \mathcal{R}_\alpha(Q)$ such that

$$d(P_\alpha, Q_\alpha) = \min_{R_1 \in \mathcal{R}_\alpha(P), R_2 \in \mathcal{R}_\alpha(Q)} d(R_1, R_2)$$

and to the related *best trimmed approximations*, that is,

$$P_\alpha = \operatorname*{argmin}_{R \in \mathcal{R}_\alpha(P)} d(R, Q).$$

Compactness of $\mathcal{R}_\alpha(P)$ in the topology of weak convergence guarantees the existence of such a *best trimmed approximation* if d metrizes weak convergence (e.g., if d is the bounded Lipschitz or the Prokhorov metric). This best trimmed approximation need not be unique, but convexity of the set of trimmings ensures that the set of best approximations is a convex, compact set if d is a convex metric (meaning that $d(rP + (1-r)Q, R) \leq rd(P, R) + (1-r)d(Q, R)$ for all $r \in (0, 1)$). This holds, for instance, for the bounded Lipschitz metric. Nevertheless, the bounded Lipschitz or the Prokhorov metric are not easily computed in general and may not be the most interesting choice for applications. A very covenient choice is the Wasserstein metric $\mathcal{W}_p$, $p \geq 1$, defined by

$$\mathcal{W}_p^p(P, Q) = \inf_{\pi \in \mathcal{M}(P,Q)} \left\{ \int \|x - y\|^p d\pi(x, y) \right\},$$

where $\mathcal{M}(P, Q)$ is the set of Borel probability measures on $X \times X$ with marginals P and Q. W_p is a metric on the set $\mathcal{F}_p = \mathcal{F}_p(X)$ of probabilities with finite p-th moment provided X is a separable Banach space. Now, if P has finite p-th moment and $Q \in \mathcal{R}_\alpha(P)$ then

$$\int \|x\|^p dQ(x) \leq \frac{1}{1-\alpha} \int \|x\|^p dP(x).$$

This shows that $\mathcal{R}_\alpha(P) \subset \mathcal{F}_p$ if $P \in \mathcal{F}_p$. Further,

Proposition 3 *If $P \in \mathcal{F}_p$ then $\mathcal{R}_\alpha(P)$ is compact in the $\mathcal{W}_p$ topology.*

The problem of finding the best trimmed approximation in Wasserstein metric can be reformulated as a problem of *optimal incomplete transportation of mass* as follows. Assume we have some supply of mass (a pile of sand, some other good) located around X and a demand of mass needed at several locations scattered around Y. Assume further that the total supply exceeds the total demand (demand=$(1-\alpha)\times$ supply, $\alpha \in (0, 1)$) Then we don't have to move all the initial mass; some α- fraction can be dismissed. Find a way to complete this task with a minimal cost.

If we rescale to represent the *target distribution* by Q, a probability measure on Y and represent the *initial distribution* by $\frac{1}{1-\alpha}P$, P probability on X and $c(x, y)$ is the cost of

moving a unit of mass from x to y, an (Incomplete) transportation plan (a way to move part of the mass in $\frac{1}{1-\alpha}P$ to Q) is represented by π, a joint probability measure on $X \times Y$. Since the target distribution is Q and the amount of mass taken from a location in X cannot exceed available mass, the constrints to be satisfied by π are

$$\pi(X \times B) = Q(B), \quad B \subset Q \tag{1}$$

$$\pi(A \times Y) \leq \frac{1}{1-\alpha}P(A), \quad A \subset X \tag{2}$$

And the optimal incomplete transportation of mass problem becomes finding

$$\inf_{\pi \in \Pi(\mathcal{R}_\alpha(P),Q)} \int_{X \times Y} c(x,y) d\pi(x,y),$$

where $\Pi(\mathcal{R}_\alpha(P), Q)$ stands for the joint probabilities satysfing (1) and (2).

Now, if $X = Y$ is Banach separable and $c(x,y) = \|x-y\|^2$ then

$$\mathcal{W}_2^2(\mathcal{R}_\alpha(P), Q) = \inf_{\pi \in \Pi(\mathcal{R}_\alpha(P),Q)} \int_{X \times Y} c(x,y) d\pi(x,y)$$

The problem of finding the best α-matching can also be formulated as a *doubly incomplete transportation of mass* problem. A duality theory can be developed for both problems (see [5]). A remarkable consequence of the duality results there is that best α-matchings in $\mathcal{W}_2$ distance are unique under very mild assumptions.

Theorem 1 *If P or Q has a density then there exists a unique pair $(P_{\alpha_1}, Q_{\alpha_2}) \in \mathcal{R}_{\alpha_1}(P) \times \mathcal{R}_{\alpha_2}(Q)$ such that*

$$\mathcal{W}_2(P_{\alpha_1}, Q_{\alpha_2}) = \min_{R_1 \in \mathcal{R}_{\alpha_1}(P), R_2 \in \mathcal{R}_{\alpha_2}(Q)} \mathcal{W}_2(R_1, R_2)$$

provided $\mathcal{W}_2(\mathcal{R}_{\alpha_1}(P), \mathcal{R}_{\alpha_2}(Q)) > 0$.

A first simple (yet useful) consequence of this uniquenes result is that the empirical version of $\tau_{\alpha,2}(P, \mathcal{F})$, say, $\tau_{\alpha,2}(P_n, \mathcal{F})$, where P_n is the empirical measure based on i.i.d. data $X_1, ...X_n \sim P$, is a consistent estimator of $\tau_{\alpha,2}(P, \mathcal{F})$ (at least under several important models, see [2] or [4]).

A deeper fact is that trimming can produce some overfitting effect. We refer to [4] for details, but a simple formulation is as follows. The (quadratic) transportation cost from the empirical to the theoretical measure is (in dimension 1) of order n, that is, under some integrability assumptions

$$n\mathcal{W}_2^2(P_n, P) \to_w \gamma(P),$$

for some nondegenerate distribution $\gamma(P)$. If we write $P_{n,\alpha}$ for the best trimming of P_n to approximate P then

$$n\mathcal{W}_2^2(P_{n,\alpha}, P) \to_{\Pr.} 0,$$

provided that, after some α-trimming the common law of the data is equal to P. This overfitting effect is not dimension free (it does not show up for dimension $k \geq 3$ for the Wasserstein metric), but it can be exploited for interesting statistical applications with univariate data, see [4] for examples of this.

In summary, we have tried to present some introduction to the 'essential' model validation via consideration of trimmed probabilities and the minimal distance method. We have argued that the Wasserstein metric arising from optimal transportation enjoys nice features in this setup and have presented a quick overview to the key results on the topic. The interested reader can find more details on the theory as well as examples of applications in the cited references.

Bibliographie

[1] Alvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. and Matrán, C. (2008a). Trimmed comparison of distributions. *J. Amer. Statist. Assoc.*, **103**, 697-704.

[2] Alvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. and Matrán, C. (2008b). Similarity of probability measures through trimming. Submitted.

[3] Alvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. and Matrán, C. (2009a). Assessing when a sample is mostly normal. Submitted.

[4] Alvarez-Esteban, P.C.; del Barrio, E.; Cuesta-Albertos, J.A. and Matrán, C. (2009b. Over-fitting effects of trimming and statistical applications. Submitted.

[5] del Barrio, E. (2008). A duality based approach to optimal incomplete transportation of mass. Submitted.

[6] Gordaliza, A. (1991). Best approximations to random variables based on trimming procedures. *J. Approx. Theory*, **64**, 162–180.

[7] Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In W. Grossmann, G. Pflug, I. Vincze, y W. Werz (Eds.), in *Mathematical Statistics and Applications, Volume B*. Reidel, Dordrecht.