Running Head: CLASSIFICATION LOSS FUNCTION

# Classification Loss Function for Parameter Ensembles in Bayesian Hierarchical Models.

Cedric E. Ginestet[ab],
Nicky G. Best[c], and Sylvia Richardson[c]

[a] King's College London, Institute of Psychiatry, Department of Neuroimaging
[b] National Institute of Health Research (NIHR) Biomedical Research Centre for Mental Health
[c] Department of Epidemiology and Biostatistics, Imperial College London

Correspondence concerning this article should be sent to Cedric Ginestet at the Centre for Neuroimaging Sciences, NIHR Biomedical Research Centre, Institute of Psychiatry, Box P089, King's College London, De Crespigny Park, London, SE5 8AF, UK. Email may be sent to cedric.ginestet@kcl.ac.uk

**Abstract**

Parameter ensembles or sets of point estimates constitute one of the cornerstones of modern statistical practice. This is especially the case in Bayesian hierarchical models, where different decision-theoretic frameworks can be deployed to summarize such parameter ensembles. The estimation of these parameter ensembles may thus substantially vary depending on which inferential goals are prioritised by the modeller. In this note, we consider the problem of classifying the elements of a parameter ensemble above or below a given threshold. Two threshold classification losses (TCLs) –weighted and unweighted– are formulated. The weighted TCL can be used to emphasize the estimation of false positives over false negatives or the converse. We prove that the weighted and unweighted TCLs are optimized by the ensembles of unit-specific posterior quantiles and posterior medians, respectively. In addition, we relate these classification loss functions on parameter ensembles to the concepts of posterior sensitivity and specificity. Finally, we find some relationships between the unweighted TCL and the absolute value loss, which explain why both functions are minimized by posterior medians.

KEYWORDS: Bayesian Statistics, Classification, Decision Theory, Epidemiology, Hierarchical Model, Loss Function, Parameter Ensemble, Sensitivity, Specificity.

# 1 Introduction

The problem of the optimal classification of a set of data points into several clusters has occupied statisticians and applied mathematicians for several decades (see Gordon, 1999, for a overview). As is true for all statistical methods, a classification is, above all, a summary of the data at hand. When clustering, the statistician is searching for an optimal partition of the parameter space into a –generally, known or pre-specified– number of classes. The essential ingredient underlying all classifications is the minimization of some distance function, which generally takes the form of a similarity or dissimilarity metric (Gordon, 1999). Optimal classification will then result in a trade-off between the level of similarity of the within-cluster elements and the level of dissimilarity of the between-cluster elements. In a decision-theoretic framework, such distance functions naturally arise through the specification of a loss function for the problem at hand. The task of computing the optimal partition of the parameter space then becomes a matter of minimizing the chosen loss function.

In spatial epidemiology, the issue of classifying areas according to their levels of risk has been previously investigated by Richardson et al. (2004). These authors have shown that areas can be classified according to the joint posterior distribution of the parameter ensemble of interest. In particular, a taxonomy can be created by selecting a decision rule $D(\alpha, C_\alpha)$ for that purpose, where $C_\alpha$ is a particular threshold, above and below which we classify the areas in the region of interest. The parameter $\alpha$, in this decision rule, is the cut-off point associated with $C_\alpha$, which determines the amount of probability mass necessary for an area to be allocated to the above-threshold category. Thus, an area $i$ with level of risk denoted by $\theta_i$ will be assigned above the threshold $C_\alpha$ if $\mathbb{P}[\theta_i > C_\alpha|\mathbf{y}] > \alpha$. Richardson et al. (2004) have therefore provided a general framework for the classification of areas, according to their levels of risk. However, this approach is not satisfactory because it relies on the choice of two co-dependent values $C_\alpha$ and $\alpha$, which can only be selected in an arbitrary fashion.

Our perspective in this paper follows the framework adopted by Lin et al. (2006), who introduced several loss functions for the identification of the elements of a parameter ensemble that represent the proportion of elements with the highest level of risk. Such a classification is based on a particular rank percentile cut-off denoted $\gamma \in [0,1]$, which determines a group of areas of high-risk. That is, Lin et al. (2006) identified the areas whose percentile rank is above the cut-off point $\gamma$. Our approach, in this paper, is substantially different since the classification is based on a real-valued threshold as opposed to a particular rank percentile. In order to emphasize this distinction, we will refer to our proposed family of loss functions as threshold classification losses (TCLs).

## 2 Classification of Elements in a Parameter Ensemble

We formulate our classification problem within the context of Bayesian hierarchical models (BHMs). In its most basic formulation, a BHM is composed of the following two layers of random variables,

$$y_i \overset{\text{ind}}{\sim} p(y_i|\theta_i, \boldsymbol{\sigma}_i), \qquad g(\boldsymbol{\theta}) \sim p(\boldsymbol{\theta}|\boldsymbol{\xi}), \tag{1}$$

for $i = 1, \ldots, n$ and where $g(\cdot)$ is a transformation of $\boldsymbol{\theta}$, which may be defined as a link function as commonly used in generalised linear models (see McCullagh and Nelder, 1989). The vector of real-valued parameters, $\boldsymbol{\theta} := \{\theta_1, \ldots, \theta_n\}$, will be referred to as a *parameter ensemble.*

### 2.1 Threshold Classification Loss

For some cut-off point $C \in \mathbb{R}$, we define the penalties associated with the two different types of misclassification. Following standard statistical terminology, we will express such misclassifications in terms of false positives (FPs) and false negatives (FNs). These concepts are formally described as

$$\text{FP}(C, \theta, \theta^{\text{est}}) := \mathcal{I}\left\{\theta \leq C, \theta^{\text{est}} > C\right\}, \quad \text{and} \quad \text{FN}(C, \theta, \theta^{\text{est}}) := \mathcal{I}\left\{\theta > C, \theta^{\text{est}} \leq C\right\}, \tag{2}$$

where $\theta$ represents the parameter of interest and $\theta^{\text{est}}$ is a candidate estimate. This corresponds to the occurrence of a false positive (type I error) and a false negative (type II error), respectively.

For the decision problem to be fully specified, we need to choose a loss function based on the sets of unit-specific FPs and FNs. The $p$-weighted threshold classification loss ($\text{TCL}_p$) function is then defined as

$$\text{TCL}_p(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}}) := \frac{1}{n}\sum_{i=1}^{n} p\,\text{FP}(C, \theta_i, \theta_i^{\text{est}}) + (1-p)\,\text{FN}(C, \theta_i, \theta_i^{\text{est}}). \tag{3}$$

One of the advantages of the choice of $\text{TCL}_p$ for quantifying the misclassifications of the elements of a parameter ensemble is that it is normalised, in the sense that $\text{TCL}_p(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}}) \in [0,1]$ for any choice of $C$ and $p$. Our main result in this paper is the following minimization.

**Proposition 1.** *For some parameter ensemble $\boldsymbol{\theta}$, and given a real-valued threshold $C \in \mathbb{R}$ and*

*$p \in [0,1]$, we have the following optimal estimator under weighted TCL,*

$$\boldsymbol{\theta}^{\text{TCL}}_{(1-p)} = \operatorname*{argmin}_{\boldsymbol{\theta}^{\text{est}}} \mathbb{E}\left[\text{TCL}_p(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}})|\mathbf{y}\right], \tag{4}$$

*where $\boldsymbol{\theta}^{\text{TCL}}_{(1-p)}$ is the vector of posterior $(1-p)$-quantiles defined as*

$$\boldsymbol{\theta}^{\text{TCL}}_{(1-p)} := \left\{Q_{\theta_1|\mathbf{y}}(1-p), \ldots, Q_{\theta_n|\mathbf{y}}(1-p)\right\}, \tag{5}$$

*where $Q_{\theta_i|\mathbf{y}}(1-p)$ denotes the posterior $(1-p)$-quantile of the $i^{th}$ element, $\theta_i$, in the parameter ensemble. Moreover, $\boldsymbol{\theta}^{\text{TCL}}_{(1-p)}$ is not unique.*

We prove this result by exhaustion in three cases. The full proof is reported in Appendix A. Note that the fact that $\text{TCL}_p$ is minimized by $\boldsymbol{\theta}^{\text{TCL}}_{(1-p)}$ and not $\boldsymbol{\theta}^{\text{TCL}}_{(p)}$ is solely a consequence of our choice of definition for the $\text{TCL}_p$ function. If the weighting of the FPs and FNs had been $(1-p)$ and $p$, respectively, then the optimal minimizer of that function would indeed be a vector of posterior $p$-quantiles.

## *2.2 Unweighted Threshold Classification Loss*

We now specialize this result to the unweighted TCL family, which is defined analogously to equation (3), as follows,

$$\text{TCL}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}}) := \frac{1}{n}\sum_{i=1}^{n} \text{FP}(C, \theta_i, \theta_i^{\text{est}}) + \text{FN}(C, \theta_i, \theta_i^{\text{est}}). \tag{6}$$

The minimizer of this loss function can be shown to be trivially equivalent to the minimizer of $\text{TCL}_{0.5}$. That is, we have

$$\operatorname*{argmin}_{\boldsymbol{\theta}^{\text{est}}} \mathbb{E}[\text{TCL}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}})|\mathbf{y}] = \operatorname*{argmin}_{\boldsymbol{\theta}^{\text{est}}} \mathbb{E}[\text{TCL}_{0.5}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}})|\mathbf{y}], \tag{7}$$

for every $C$, which therefore proves the following corollary.

**Corollary 1.** *For some parameter ensemble $\boldsymbol{\theta}$ and $C \in \mathbb{R}$, the minimizer of the posterior expected TCL is*

$$\boldsymbol{\theta}^{\text{med}} := \boldsymbol{\theta}^{\text{TCL}}_{(0.5)} = \left\{Q_{\theta_1|\mathbf{y}}(0.5), \ldots, Q_{\theta_n|\mathbf{y}}(0.5)\right\}, \tag{8}$$

*and this optimal estimator is not unique.*

The posterior expected loss under the unweighted TCL function takes the following form,

$$\mathbb{E}\left[\text{TCL}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}})|\mathbf{y}\right] = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{C} d\mathbb{P}[\theta_i|\mathbf{y}]\mathcal{I}\left\{\theta_i^{\text{est}} > C\right\} + \int_{C}^{+\infty} d\mathbb{P}[\theta_i|\mathbf{y}]\mathcal{I}\left\{\theta_i^{\text{est}} \leq C\right\}, \tag{9}$$

whose formulae is derived using $\mathcal{I}\{\theta \leq C, \theta^{\text{est}} > C\} = \mathcal{I}\{\theta \leq C\}\mathcal{I}\{\theta^{\text{est}} > C\}$. It is of special importance to note that when using the posterior TCL, any classification –correct or incorrect– will incur a penalty. The *size* of that penalty, however, varies substantially depending on whether or not the classification is correct. A true positive can be distinguished from a false positive, by the fact that the former will only incur a small penalty proportional to the posterior probability of the parameter to be below the chosen cut-off point $C$.

### *2.3 Relationship with Posterior Sensitivity and Specificity*

Our chosen decision-theoretic framework for classification has the added benefit of being readily comparable to conventional measures of classification errors widely used in the context of test theory. For our purpose, we will define the Bayesian sensitivity of a classification estimator $\boldsymbol{\theta}^{\text{est}}$, also referred to as the posterior true positive rate (TPR), as follows

$$\text{TPR}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}}) := \frac{\sum_{i=1}^{n} \mathbb{E}[\text{TP}(C, \theta_i, \theta_i^{\text{est}})|\mathbf{y}]}{\sum_{i=1}^{n} \mathbb{P}[\theta_i > C|\mathbf{y}]}, \tag{10}$$

where the expectations are taken with respect to the joint posterior distribution of $\boldsymbol{\theta}$. Similarly, the Bayesian specificity, or posterior true negative rate (TNR), will be defined as

$$\text{TNR}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}}) := \frac{\sum_{i=1}^{n} \mathbb{E}[\text{TN}(C, \theta_i, \theta_i^{\text{est}})|\mathbf{y}]}{\sum_{i=1}^{n} \mathbb{P}[\theta_i \leq C|\mathbf{y}]}, \tag{11}$$

where in both definitions, we have used $\text{TP}(C, \theta_i, \theta_i^{\text{est}}) := \mathcal{I}\{\theta_i > C, \theta_i^{\text{est}} > C\}$ and $\text{TN}(C, \theta_i, \theta_i^{\text{est}}) := \mathcal{I}\{\theta_i \leq C, \theta_i^{\text{est}} \leq C\}$. It then follows that we can formulate the relationship between the posterior expected TCL and the Bayesian sensitivity and specificity as

$$\mathbb{E}[\text{TCL}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}})|\mathbf{y}] = \frac{1}{n}\,\text{FPR}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}}) \sum_{i=1}^{n} \mathbb{P}[\theta_i \leq C|\mathbf{y}] + \frac{1}{n}\,\text{FNR}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}}) \sum_{i=1}^{n} \mathbb{P}[\theta_i > C|\mathbf{y}].$$

where $\text{FPR}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}}) := 1 - \text{TNR}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}})$ and $\text{FNR}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}}) := 1 - \text{TPR}(C, \boldsymbol{\theta}, \boldsymbol{\theta}^{\text{est}})$.

# 3 Conclusion

The fact that the posterior median is the minimizer of the posterior expected absolute value loss (AVL) function is well-known Berger (1980). That is, the posterior median minimizes the posterior expected AVL, where $\text{AVL}(\theta, \theta^{\text{est}}) := |\theta - \theta^{\text{est}}|$. One may therefore ask whether there is link between the minimization of the AVL function, which is an estimation loss and the classification loss function described in this paper. The proof of the optimality of the posterior median under AVL proceeds by considering whether $\theta^{\text{med}} - \theta^{\text{est}} \gtreqless 0$. This leads to a proof by exhaustion in three cases, which includes the trivial case where $\theta^{\text{med}}$ and $\theta^{\text{est}}$ are equal. Similarly, in the proof of proposition 1, we have also obtained three cases, which are based on the relationships between the $\theta_i$'s and $\theta_i^{\text{est}}$'s with respect to $C$. However, note that by subtracting $\theta_i^{(1-p)} \leq C$ from $C < \theta_i^{\text{est}}$ and ignoring null sets, we obtain $\theta_i^{(1-p)} - \theta_i^{\text{est}} < 0$, for the second case. Similarly, a subtraction of the hypotheses of the third case gives $\theta_i^{(1-p)} - \theta_i^{\text{est}} > 0$ for the third case, which therefore highlights the relationship between the optimization of the AVL and the TCL functions.

# Appendix A: Proof of TCL Minimization

*Proof of proposition 1 on page 3.*
Let $\rho_p(C,\boldsymbol{\theta},\boldsymbol{\theta}^{\text{est}})$ denote $\mathbb{E}[\text{TCL}_p(C,\boldsymbol{\theta},\boldsymbol{\theta}^{\text{est}})|\mathbf{y}]$. We prove the result by exhaustion over three cases. In order to prove that

$$\rho_p(C,\boldsymbol{\theta},\boldsymbol{\theta}^{(1-p)}) \leq \rho_p(C,\boldsymbol{\theta},\boldsymbol{\theta}^{\text{est}}), \tag{12}$$

for any $\boldsymbol{\theta}^{\text{est}} \in \boldsymbol{\Theta}$ with $\theta_i^{(1-p)} := Q_{\theta_i|\mathbf{y}}(1-p)$, it suffices to show that $\rho_p(C,\theta_i,\theta_i^{(1-p)}) \leq \rho_p(C,\theta_i,\theta_i^{\text{est}})$ holds, for every $i=1,\ldots,n$. Expanding these unit-specific risks,

$$\begin{aligned} &p\mathcal{I}\{\theta_i^{(1-p)} > C\}\mathbb{P}\left[\theta_i \leq C|\mathbf{y}\right] + (1-p)\mathcal{I}\{\theta_i^{(1-p)} \leq C\}\mathbb{P}\left[\theta_i > C|\mathbf{y}\right] \\ &\quad \leq p\mathcal{I}\{\theta_i^{\text{est}} > C\}\mathbb{P}\left[\theta_i \leq C|\mathbf{y}\right] + (1-p)\mathcal{I}\{\theta_i^{\text{est}} \leq C\}\mathbb{P}\left[\theta_i > C|\mathbf{y}\right]. \end{aligned} \tag{13}$$

Now, fix $C$ and $p \in [0,1]$ to arbitrary values. Then, for any point estimate $\theta_i^{\text{est}}$, we have

$$\rho_p(C,\theta_i,\theta_i^{\text{est}}) = \begin{cases} p\mathbb{P}[\theta_i \leq C|\mathbf{y}], & \text{if } \theta_i^{\text{est}} > C, \\ (1-p)\mathbb{P}[\theta_i > C|\mathbf{y}], & \text{if } \theta_i^{\text{est}} \leq C. \end{cases} \tag{14}$$

The optimality of $\theta_i^{(1-p)}$ over $\theta_i^{\text{est}}$ as a point estimate is therefore directly dependent on the relationships between $\theta_i^{(1-p)}$ and $C$, and between $\theta_i^{\text{est}}$ and $C$. This determines the following three cases:

**i.** If $\theta_i^{(1-p)}$ and $\theta_i^{\text{est}}$ are on the same side of $C$, then clearly,

$$\rho_p(C,\theta_i,\theta_i^{(1-p)}) = \rho_p(C,\theta_i,\theta_i^{\text{est}}), \tag{15}$$

**ii.** If $\theta_i^{(1-p)} \leq C$ and $\theta_i^{\text{est}} > C$, then,

$$\rho_p(C,\theta_i,\theta_i^{(1-p)}) = (1-p)\mathbb{P}[\theta_i > C|\mathbf{y}] \;\leq\; p\mathbb{P}[\theta_i \leq C|\mathbf{y}] = \rho_p(C,\theta_i,\theta_i^{\text{est}}), \tag{16}$$

**iii.** If $\theta_i^{(1-p)} > C$ and $\theta_i^{\text{est}} \leq C$, then,

$$\rho_p(C,\theta_i,\theta_i^{(1-p)}) = p\mathbb{P}[\theta_i \leq C|\mathbf{y}] \;<\; (1-p)\mathbb{P}[\theta_i > C|\mathbf{y}] = \rho_p(C,\theta_i,\theta_i^{\text{est}}), \tag{17}$$

Equation (15) follows directly from an application of the result in (13), and cases two and three follow from consideration of the following relationship:

$$p\mathbb{P}[\theta_i \leq C|\mathbf{y}] \gtreqless (1-p)\mathbb{P}[\theta_i > C|\mathbf{y}], \tag{18}$$

where $\gtreqless$ means either $<$, $=$ or $>$. Using $\mathbb{P}[\theta_i > C|\mathbf{y}] = 1 - \mathbb{P}[\theta_i \leq C|\mathbf{y}]$, this gives

$$\mathbb{P}[\theta_i \leq C|\mathbf{y}] = F_{\theta_i|\mathbf{y}}(C) \gtreqless 1-p. \tag{19}$$

Here, $F_{\theta_i|\mathbf{y}}$ is the posterior CDF of $\theta_i$. Therefore, we have

$$C \gtreqless F^{-1}_{\theta_i|\mathbf{y}}(1-p) =: Q_{\theta_i|\mathbf{y}}(1-p) :=: \theta_i^{(1-p)}, \tag{20}$$

where $\gtreqless$ takes the same value in equations (18), (19) and (20).

This proves the optimality of $\boldsymbol{\theta}^{(1-p)}$. Moreover, since one can construct a vector of point estimates $\theta_i^{\text{est}}$ satisfying $\theta_i^{\text{est}} \gtreqless C$, whenever $\theta_i^{(1-p)} \gtreqless C$, for every $i$, it then follows that $\boldsymbol{\theta}^{(1-p)}$ is not unique.

## Acknowledgments

This work was supported by a fellowship from the UK National Institute for Health Research (NIHR) Biomedical Research Centre for Mental Health (BRC-MH) at the South London and Maudsley NHS Foundation Trust and King's College London. This work has also been funded by the Guy's and St Thomas' Charitable Foundation as well as the South London and Maudsley Trustees.

## References

Berger, J. (1980). A Robust Generalized Bayes Estimator and Confidence Region for a Multivariate Normal Mean. *The Annals of Statistics*, **8**(4), 716–761.

Ghosh, M. (1992). Constrained bayes estimation with applications. *Journal of the American Statistical Association*, **87**(418), 533–540.

Gordon, A. (1999). *Classification*. Chapman and Hall, London.

Lin, R., Louis, T., Paddock, S., and Ridgeway, G. (2006). Loss function based ranking in two-stage hierarchical models. *Bayesian analysis*, **1(4)**, 915–946.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, 2nd ed.

Richardson, S., Thomson, A., Best, N., and Elliott, P. (2004). Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives*, **112**, 1016–1025.