

The asymptotic effect of tuning parameters

Ingrid Dæhlen^{1,2}, Nils Lid Hjort¹ and Ingrid Hobæk Haff¹

¹ Department of Mathematics, University of Oslo

² Norwegian Computing Center, Post box 114 Blindern, Oslo, 0314, Norway

Address for correspondence: Ingrid Dæhlen, Department of Mathematics, University of Oslo, Moltke Moes vei 35, 0851 Oslo, Norway, email: ingrdae@math.uio.no

Abstract

Tuning parameters are parameters involved in an estimating procedure for the purpose of reducing the risk of some other estimator. Examples include the degree of penalization in penalized regression and likelihood problems, as well as the balance parameter in hybrid methods. Typically tuning parameters are set to the minimizers of some estimator of the risk, a step which introduces additional randomness and makes standard methodology inapplicable. We derive precise asymptotic theory for this situation. Our framework allows for smooth, but otherwise arbitrary, loss functions and for the risk to be estimated by cross-validation procedures. Results include consistency of the optimal estimator towards a well-defined quantity and asymptotic normality after proper scaling and centring. We give explicit forms and estimators for the limiting variance matrix and results sharply characterizing the distance from the training error to the cross-validated estimator of the risk.

Keywords: cross validation, information criteria, large-sample theory, regularisation, two-stage estimators

1 Introduction

Suppose we have data $Z_1, \dots, Z_n$ and wish to estimate a parameter θ . Assume further that for each fixed value of some other parameter λ this can be done by maximising a

function $\Gamma_n(\theta, \lambda)$ defined in terms of the data. Examples include hybrid estimators where $\Gamma_n(\theta, \lambda) = \lambda\gamma_n^1(\theta) + (1 - \lambda)\gamma_n^2(\theta)$ and penalized likelihood methods, where for each λ , θ is estimated by the maximizer of $\ell_n(\theta) - \lambda R(\theta)$ for some function R . This latter situation covers the ridge regression. An example of the former case is the hybrid generative-discriminative estimator defined as the maximizer of $\lambda\ell_n^{\text{Gen}}(\theta) + (1 - \lambda)\ell_n^{\text{Disc}}(\theta)$ where ℓ_n^{Gen} and ℓ_n^{Disc} are the log-likelihoods in generative and corresponding discriminative models (see e.g. Bouchard and Triggs (2004)).

For each fixed λ , we can maximize $\Gamma_n(\theta, \lambda)$ with respect to θ to get the estimator $\hat{\theta}(\lambda)$. Furthermore, if Γ_n is sufficiently smooth in θ and a relatively standard set of regularity conditions is satisfied, $\hat{\theta}(\lambda)$ is consistent for a well-defined quantity $\theta_0(\lambda)$ and $\sqrt{n}\{\hat{\theta}(\lambda) - \theta_0(\lambda)\}$ converges in law to some known limit distribution. In practice, however λ is rarely fixed, and is instead set to minimize some estimated loss, usually approximated by means of an information criterion or cross-validation. There is therefore an additional randomness involved when working with tuning parameters. Because of this, the standard limit results for $\hat{\theta}(\lambda)$ are not immediately applicable to $\hat{\theta}(\hat{\lambda})$. In practice, this point is often ignored and inference about θ is made by treating λ as fixed at $\hat{\lambda}$ and by using pointwise results for $\hat{\theta}(\lambda)$. Using such a method ignores part of the fitting procedure and added randomness introduced by tuning λ , however. Because of this, the variance of $\hat{\theta}(\hat{\lambda})$ tends to be underestimated.

Figure 1 illustrates the ideas outlined above for a data set concerning female Pima Indians and prevalence of diabetes (publicly available in e.g. the R-package `mlbench`). In Section 6.1 we fit a penalized logistic regression model to these data where the tuning parameter dictating the size of the regularization term is set to the minimizer of the cross-validated Brier score. For each regression coefficient, Figure 1 shows histograms based on 2000 non-parametric bootstrap iterations, together with the densities in approximate distributions. The dotted curve corresponds to the distribution we get when using the pointwise results as described in the previous paragraph. The dashed lines correspond to the alternative approximate distributions derived in this article,

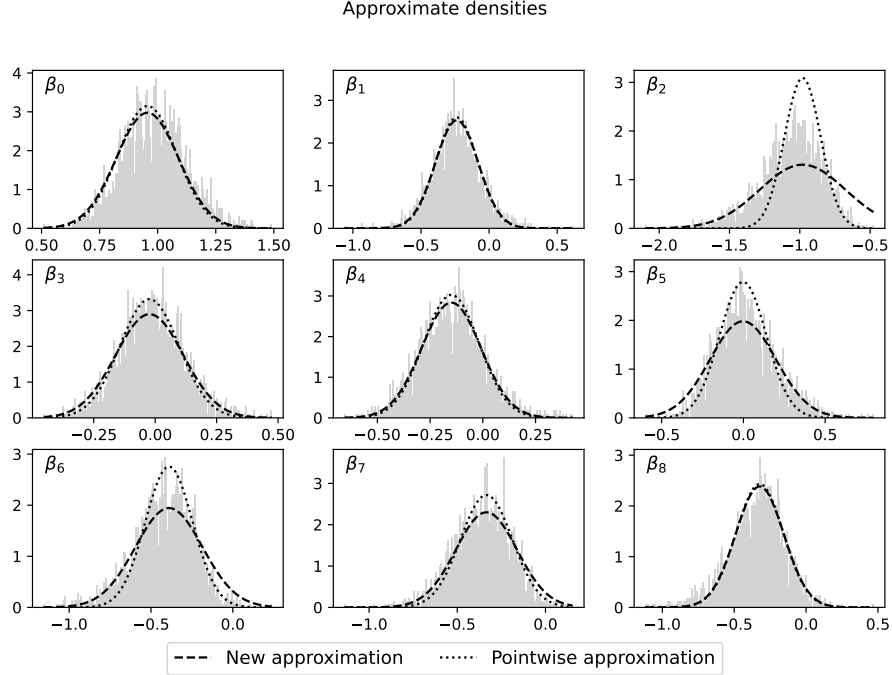


Figure 1: Histograms of draws from the bootstrap distribution of $\hat{\beta}(\hat{\lambda})_j$ for $j = 0, \dots, 8$ together with the density in the approximate distribution.

which take the tuning procedure into account. Note that the curves corresponding to the pointwise estimator seem a bit too narrow, while the curves corresponding to the new estimators are less sharp. This is a consequence of including and not including the effect of the tuning step in the limiting distribution.

In the present article, we fill a gap in the literature by deriving precise limit results for the estimator $\hat{\theta}(\hat{\lambda})$, allowing inference about θ to be made while taking the added randomness from tuning λ into account. We will start by deriving limit results when $\hat{\lambda}$ is estimated by minimising the relatively simpler training error (Section 2) or an information criterion (Section 3). We also derive estimators for the limiting variance matrix of $\hat{\theta}(\hat{\lambda})$ (Section 2.2). In Section 4, we show that the results of Section 2 hold true also when the cross validated error is used to estimate λ . This will require us to take a closer look at the cross validation procedure, and in Section 4.1 we show a result proving exactly how far off the training error is from the cross-validated counterpart. In Section 5, we consider a particular situation where the theory simplifies greatly

and in Section 6, we go through a couple of applications. In particular, we will return to the above analysis of diabetes in Pima Indians in Section 6.1. Lastly, Section 7 contains some concluding remarks.

1.1 Related work

To our knowledge, no results concerning the limiting distribution of $\hat{\theta}(\hat{\lambda})$ when $\hat{\lambda}$ is set to the minimizer of some general risk estimator exists in the literature. Arcones (2005) studies a similar situation, by deriving limit results for $\hat{\theta}(\hat{\lambda})$ when $\hat{\lambda}$ is set to the minimizer of some estimator of the limiting variance of $\hat{\theta}(\lambda)$. This result is closely related to our theorems, but Arcones' framework only works for one-dimensional parameters θ and λ . Furthermore, $\hat{\lambda}$ cannot minimize a general loss function or a cross-validated estimator of the risk.

Our results are different from those of Arcones (2005) in yet another way. Arcones (2005), as well as many others (e.g. Dodge and Jurečková (2000), Kato (2009), Germain and Roueff (2010) and (Spokoiny, 2025, Se. 10.2)), shows that $S_n(\lambda) = \sqrt{n} [\hat{\theta}(\lambda) - \theta_0(\lambda)]$ converges as a stochastic process to a limiting Gaussian $S(\lambda)$. Such results ensure that $\hat{\theta}(\lambda)$ behaves as expected for all values of λ and could in theory be used to prove our results where the asymptotic behaviour of $\hat{\lambda}$ is known. This is pointed out in (Dodge and Jurečková, 2000, Se. 7.2 and 10.2). In practice, however, the behaviour of the tuning parameter is not known, and very few attempts have been made at investigating this. Homrighausen and McDonald (2014) and Chetverikov et al. (2021) have done some work on the Lasso and Arcones (2005) considers a couple of specific tuning schemes, but apart from these papers, virtually no theoretical discussion on how $\hat{\lambda}$ behaves asymptotically and how this in turn impacts the limiting distribution of $S_n(\hat{\lambda})$ seem to be present in the literature. This is especially true when cross-validation is used to tune $\hat{\lambda}$. We therefore believe the current paper provides a significant contribution to statistical estimation theory. Consult the supplementary material Section S.2 for a discussion utilizing similar arguments as Arcones (2005)

when no natural limit for $\hat{\lambda}$ exists.

To investigate the joint limiting behaviour of $\hat{\theta}(\hat{\lambda})$ and $\hat{\lambda}$, we reformulate the tuning procedure as a two-stage estimation problem and show that $\hat{\theta}(\hat{\lambda})$ can be written as components of a Z -estimator (see e.g. Van der Vaart (2000, Ch. 5)). Many of our proofs are therefore based upon and inspired by results for two-stage estimators, see e.g. Joe (2005), Ko and Hjort (2019) or Dæhlen and Hjort (2025), and bears similarities with the framework of generalized estimating equations used frequently in econometrics, see e.g. Hansen (1982), Hall and Inoue (2003) or for a more thorough introduction, Shao (2003, Ch. 5).

To apply the two-stage arguments also in the context of cross-validation, we derive sharp results characterizing the difference between the cross-validated estimator of the risk and the training error in Theorem 6. This theorem can be seen as an extension of the result given in Stone (1977), where the author writes that the cross-validated likelihood is asymptotically equivalent to the AIC. Stone (1977) only writes down an informal argument, however, and such intuitive explanations are often repeated when the paper is cited, see e.g. Claeskens and Hjort (2008) and Konishi and Kitagawa (2008). To the best of our knowledge, no formal proof of the statement has ever been given, and we therefore believe that our Theorem 6 fills this gap in the literature. That being said, recently there has been some work related to approximations to the LOOCV (see Stephenson and Broderick (2020) for an overview), and the theoretical results concerning these approximations bear resemblance to our work (see in particular (Beirami et al., 2017, Le. 3)). We still believe our theorem to be novel as it is more general than that of Stone (1977) since it allows for general Z -estimators instead of only maximum likelihood estimators, and the functions used in the cross-validation procedure and estimation of θ need not be the same.

We also use Theorem 6 to discuss whether cross-validation is precise enough to capture the difference between different models, a topic which has been discussed by e.g. Shao (1993) or Shao (1997) in the context of linear models. Furthermore, our

theory allows us to characterize the optimism of the training error, something which is studied in e.g. Efron (1983), Efron (1986), Tibshirani and Knight (1999) or Efron (2004), albeit with a different approach.

Our methodology is in some sense related to the field of post-selection inference, see e.g. Berk et al. (2013), Lee et al. (2016), Bachoc et al. (2020) or indeed Kuchibhotla et al. (2022) for a recent review. In this field, the distribution of regression parameters is derived in a way that takes the model selection procedure into account. This is typically done by either conditioning on the model selection procedure or by creating confidence intervals by taking the supremum over all candidate models. We, on the other hand, include the selection of continuous tuning parameters into the fitting procedure itself. Because of this, the field of post-selection inference is related to the present article in spirit only. Furthermore, we provide explicit forms and estimators for the limiting variance of $\hat{\theta}(\hat{\lambda})$, something which is not obtained through the most popular post-selection inference techniques (see e.g. the aforementioned review article).

Another topic which bears some similarities with our work is the research concerning asymptotic optimality of different parameter tuning schemes. For a given prediction depending on some tuning parameter λ , these works investigate its distance to the “best” prediction we can possibly make, when λ is tuned according to some criterion. Craven and Wahba (1978), Li (1985) and Speckman (1985), Li (1986) and Li (1987) all investigate this problem in linear models, and more recently Mu et al. (2018) and Mu et al. (2021) have proven asymptotic optimality in a more advanced context. See also Mu et al. (2021) which contains a more comprehensive list of relevant literature. So far, however, the main concern of this field of research has been whether the tuning procedure ensures that the obtained and optimal predictions are close, rather than the asymptotic effect tuning parameters have on the limit distribution of $\hat{\theta}(\hat{\lambda})$. We believe this sets our results apart from the literature of asymptotic optimality.

1.2 Notation and definitions

Before we begin with the theory we will introduce notation which will be used throughout this article. For a given function $f(\theta)$ we will let $\nabla_{\theta=\theta_0}f(\theta)$ and $H_{\theta=\theta_0}f(\theta)$ be the gradient and Hessian matrix of f at θ_0 respectively, while $\partial_{\theta=\theta_0}f(\theta) = f'(\theta_0) = \partial f / \partial \theta(\theta_0)$ will denote the Jacobian matrix of f at θ_0 . When there is no room for misunderstanding, we will simplify notation and write $\nabla_{\theta_0}f$, $H_{\theta_0}f$ and $\partial_{\theta_0}f$ rather than the above. Furthermore, we will often identify matrices in $\mathbb{R}^{p \times q}$ with vectors in $\mathbb{R}^{pq}$ in the following way $A = (a_1, a_2, \dots, a_q) \in \mathbb{R}^{p \times q} \rightsquigarrow (a_1^T, a_2^T, \dots, a_q^T)^T \in \mathbb{R}^{pq}$ and hence, if A is a matrix, $\|A\| = (\sum_j \sum_k a_{jk}^2)^{1/2}$. We will also let $\xrightarrow{\text{Pr}}$ and $\xrightarrow{d}$ denote convergence in probability and distribution/law respectively.

We will let $Z_1, \dots, Z_n \in \mathbb{R}^d$ be independent and identically distributed (i.i.d.) variables from a distribution F and assume that for each fixed value of $\lambda \in \mathbb{R}^q$, $\theta \in \mathbb{R}^p$ is estimated by solving the equation $n^{-1} \sum_{i=1}^n \varphi(Z_i, \theta, \lambda) = 0$, for some function $\varphi: \mathbb{R}^{d+p+q} \rightarrow \mathbb{R}^p$. We will let $\hat{\theta}(\lambda)$ denote the solution. If $\hat{\theta}(\lambda)$ is a maximizer of an expression on the form $n^{-1} \sum_{i=1}^n \phi(Z_i, \theta, \lambda)$ for some function ϕ , this is satisfied with $\varphi(z, \theta, \lambda) = \nabla_{\theta} \phi(z, \theta, \lambda)$ under general conditions. In particular, if $\hat{\theta}(\lambda)$ maximises a penalized likelihood of the form $n^{-1} \sum_{i=1}^n \log f(Y_i, \theta) - \lambda P(\theta)$, $\varphi(z, \theta, \lambda) = u(Y_i, \theta) - \lambda P'(\theta)$ where u is the score function in the model. We will also comment that regression and classification models fit into this framework by considering the covariates as random and letting $Z = (Y, X^T)^T$ where Y is the response and X the predictor.

Lastly, fix a loss function $\psi: \mathbb{R}^{d+p} \rightarrow \mathbb{R}$. The training error, which will be denoted by $\text{TE}(\lambda)$, is equal to $n^{-1} \sum_{i=1}^n \psi\{Z_i, \hat{\theta}(\lambda)\}$. The leave-one-out cross validation (LOOCV) estimator of the risk is $n^{-1} \sum_{i=1}^n \psi\{Z_i, \hat{\theta}_{(-i)}(\lambda)\}$ and will be referenced by $\text{CV}(\lambda)$. We will assume that $\hat{\lambda}$ is set to the minimizer of some error estimator and refer to this process as “tuning” λ .

2 Two-stage tuning

We will start by investigating the asymptotic properties of $\hat{\theta}(\hat{\lambda})$ when $\hat{\lambda}$ is set by tuning with respect to the training error. If this function is convex in λ and the minimizer lies in the interior of the parameter space for λ , this is equivalent to solving $TE'(\lambda) = 0$. Hence, under these conditions, $\hat{\lambda}$ satisfies $TE'(\hat{\lambda}) = \sum_{i=1}^n \hat{\theta}'(\hat{\lambda})^T \nabla \psi[Z_i, \hat{\theta}(\hat{\lambda})]$. If $n^{-1} \sum_{i=1}^n \partial_{\hat{\theta}(\hat{\lambda})} \varphi(Z_i, \theta, \hat{\lambda})$ is invertible, the quantity $\hat{\theta}'(\hat{\lambda})$ exists by the implicit function theorem. Furthermore, it satisfies $n^{-1} \sum_{i=1}^n [\partial_{\hat{\theta}(\hat{\lambda})} \varphi(Z_i, \theta, \hat{\lambda}) \hat{\theta}'(\hat{\lambda}) + \partial_{\lambda} \varphi\{Y_i, \hat{\theta}(\hat{\lambda}), \lambda\}] = 0$ by the same result. Because of these relations as well as the definition of $\hat{\theta}(\lambda)$, we can express the vector $\hat{\alpha} = (\hat{\theta}(\hat{\lambda})^T, \hat{\lambda}^T, \hat{\theta}'(\hat{\lambda})^T)^T \in \mathbb{R}^{p+q+pq}$ as the solution to a set of equations. This implies that the full vector $\hat{\alpha}$ is what is called a Z estimator, and these are typically consistent for well-defined quantities and asymptotically normal after proper scaling and centring, see e.g. Van der Vaart (2000, Ch. 5). Since $\hat{\alpha}$ is Z-estimator by the previous arguments, we would expect the same to be true for $\hat{\alpha}$ (and in turn also $\hat{\theta}(\hat{\lambda})$). The following theorem shows that this is indeed correct.

Theorem 1. *For each fixed λ , let $\hat{\theta}(\lambda)$ be the unique solution to $n^{-1} \sum_{i=1}^n \varphi(Z_i, \theta, \lambda) = 0$. Furthermore, assume that $\hat{\lambda}$ is the unique solution to $TE'(\lambda) = o_p(1/\sqrt{n})$. Let $\alpha = (\theta^T, \lambda^T, D^T)^T \in \mathbb{R}^{p+q+pq}$ and $\hat{\alpha} = (\hat{\theta}(\hat{\lambda})^T, \hat{\lambda}^T, \hat{\theta}'(\hat{\lambda})^T)^T$. Define the function $\eta: \mathbb{R}^{d+p+q+pq} \rightarrow \mathbb{R}^{p+d+pq}$ as $\eta(z, \alpha) = (\eta_1(z, \alpha)^T, \eta_2(z, \alpha)^T, \eta_3(z, \alpha)^T)^T$ where $\eta_1(z, \alpha) = \varphi(z, \theta, \lambda)$, $\eta_2(z, \alpha) = D^T \nabla_{\theta} \psi(z, \theta)$ and $\eta_3(z, \alpha) = \partial_{\theta} \varphi(z, \theta, \lambda) D + \partial_{\lambda} \varphi(Z, \theta, \lambda)$. Assume the solution to $0 = \Psi(\alpha) = E\{\eta(Z, \theta, \lambda, D)\}$ is unique and equal to α_0 . Then $\hat{\alpha}$ is consistent for α_0 and furthermore,*

$$\hat{\alpha} = \alpha_0 - n^{-1} \Psi'(\alpha_0)^{-1} \sum_{i=1}^n \eta(Z_i, \theta_0, \lambda_0, D_0) + o_p(1/\sqrt{n}), \quad (1)$$

provided the function Ψ is continuously differentiable in a neighbourhood of α_0 , that $\Psi'(\alpha_0)$ is non-singular and that in a neighbourhood of α_0 the norm of $\eta(z, \alpha)$ and all first order partial derivatives of η with respect to α are bounded by functions $m_0(z)$ and $m_1(z)$ respectively with $E m_0(Z)^2, E m_1(Z)^2 < \infty$. In particular, $\sqrt{n}(\hat{\alpha} - \alpha_0)$ converges

in law to a $N(0, V_\alpha)$ distribution where $V_\alpha = \Psi'(\alpha_0)^{-1} \text{Var} \eta(Z, \alpha_0) \{\Psi'(\alpha_0)^{-1}\}^T$.

Remark. By Leibniz' integral formula, the function Ψ is smooth in α as long as the same holds true for η . Furthermore α_0 is the unique solution to $\Psi(\alpha) = 0$ as long as $\theta(\lambda)$, defined as the zero of $E \varphi(Z, \theta, \lambda)$, is well defined for all λ and there is only a single $\theta \in \{\theta(\lambda) \mid \lambda \in \Lambda\}$ minimizing the limiting risk, $E \psi[Z, \theta(\lambda)]$. For a discussion on what happens this condition fails to hold, see Section S2 in the supplementary material.

Proof. By the arguments preceding the theorem, $\sum_{i=1}^n \eta(Z_i, \hat{\alpha}) = 0$. Hence, consistency follows from the assumptions and Lemma S1.1 and Corollary S1.2 in the supplementary material. Equation (1) follows from Van der Vaart (2000, Th. 5.21). $\blacksquare$

In practice, we are rarely interested in the asymptotic behaviour of the full vector $\hat{\alpha}$, but rather in the limiting properties of $\hat{\theta}(\hat{\lambda})$ alone.

Corollary 1. *Assume the conditions of Theorem 1 hold true, and for each λ let $\theta(\lambda)$ be solution to $E \varphi(Z, \theta, \lambda) = 0$. Then $\hat{\theta}(\hat{\lambda})$ is consistent for $\theta_0 = \theta(\lambda_0)$ where λ_0 is the solution to $\nabla_\lambda E \psi\{Z, \theta_0(\lambda)\} = 0$. Let $D = \theta'(\lambda_0)$, $Z_1 = H_{\lambda_0} E \psi\{Z, \theta_0(\lambda)\}$, $Z_2 = H_{\theta_0} E \psi(Z, \theta)$, $J = -\partial_{\theta_0} E \varphi(Z, \theta, \lambda_0)$ and $b = \nabla_{\theta_0} E \psi(Z, \theta)$, then*

$$\begin{aligned} \hat{\theta}(\hat{\lambda}) = \theta_0 &+ A_1 \frac{1}{n} \sum_{i=1}^n \varphi(Z_i, \theta_0, \lambda_0) + A_2 \frac{1}{n} \sum_{i=1}^n \nabla_{\theta_0} \psi(Z_i, \theta) \\ &+ A_3 \frac{1}{n} \sum_{i=1}^n \{\partial_{\theta_0} \varphi(Z_i, \theta, \lambda_0) D + \partial_{\lambda_0} \varphi(Z_i, \theta_0, \lambda)\}, \end{aligned} \quad (2)$$

where $A_1 = J^{-1} - DZ_1^{-1}\{D^T Z_2 + W\}J^{-1}$, $A_2 = -DZ_1^{-1}D^T$ and $A_3 = -DZ_1^{-1}M$, where M is a $q \times pq$ block diagonal matrix where the diagonal blocks are each equal to $b^T J^{-1}$ and W is a $q \times p$ matrix which can be written as $W = ((b^T J^{-1} W^1)^T, \dots, (b^T J^{-1} W^q)^T)^T$

with

$$W^j = \begin{pmatrix} D_j^T H_{\theta_0} \mathbb{E} \varphi^1(Z, \theta, \lambda_0) \\ D_j^T H_{\theta_0} \mathbb{E} \varphi^2(Z, \theta, \lambda_0) \\ \vdots \\ D_j^T H_{\theta_0} \mathbb{E} \varphi^p(Z, \theta, \lambda_0) \end{pmatrix} + \partial_{\lambda_0^j} \partial_{\theta_0} \mathbb{E} \varphi(Z, \theta, \lambda),$$

where D_j denotes the j -th column of D . In particular, $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0(\lambda_0)\} \rightarrow N\{0, A^* K^* (A^*)^T\}$ where $A^* = (A_1, A_2, A_3)$ and $K^* = \text{Var} \eta(Z, \theta_0, \lambda_0, D)$ with η defined in Theorem 1.

Proof. This follows from Theorem 1 and formulas for the inverse of block matrices. $\blacksquare$

If $q = 1$, we have $M = b^T J^{-1}$ and $W = b^T J^{-1} W^1$ making all formulas a lot more readable. Furthermore, the expressions above often simplify greatly, see e.g. Section 5.

2.1 Truncated estimators

In Corollary 1 we assume that $\hat{\lambda}$ is found by solving $\text{TE}'(\lambda) = 0$. This is, however, not how λ is usually tuned in practice. Typically, $\hat{\lambda}$ is set to the minimizer of $\text{TE}(\lambda)$ in some compact set Λ . We will now discuss how Corollary 1 can be used to make inference also in this situation. For simplicity, we will take $q = 1$ and $\Lambda = [0, 1]$, but all results should be generalizable to higher dimensions and other parameter sets.

As long as TE is differentiable and $\hat{\lambda}$ lies in the interior of $[0, 1]$, minimising TE in Λ is equivalent to solving $\text{TE}'(\lambda) = 0$. Oftentimes, however, $\text{TE}(\lambda)$ might be increasing or decreasing with $\text{TE}'(\lambda) \neq 0$ for all values of λ . In such cases $\hat{\lambda}$ is set to 0 or 1 depending on which of the boundary points achieves the lowest value of $\text{TE}(\lambda)$, and the condition $\text{TE}'(\hat{\lambda}) = 0$ might not be satisfied. To study this more complicated estimator, we will work with a modified version of $\hat{\theta}(\lambda)$ defined in the following way:

$$\hat{\theta}_T = \hat{\theta}(0)I(\lambda \leq 0) + \hat{\theta}(\lambda)I(0 < \lambda < 1) + \hat{\theta}(1)I(\lambda \geq 1), \quad (3)$$

where I is the indicator function. This estimator agrees with $\hat{\theta}(\lambda)$ for all $\lambda \in [0, 1]$. Outside of the unit interval, however, the estimator is “truncated” to either $\hat{\theta}(0)$ or $\hat{\theta}(1)$ depending on which side of $[0, 1]$ λ lies. Hence, as long as $TE(\lambda)$ is convex and has a global minimizer $\hat{\lambda}_G \in \mathbb{R}$, $\hat{\theta}_T(\hat{\lambda}_G)$ will be equal to $\hat{\theta}(\hat{\lambda})$ where $\hat{\lambda}$ is the minimizer of $TE(\lambda)$ in $[0, 1]$. Furthermore, when the global minimizer of $TE(\lambda)$ lies in $[0, 1]$, $\hat{\lambda}$ is the global minimizer ensuring $\hat{\theta}_T(\hat{\lambda}_G) = \hat{\theta}(\hat{\lambda})$.

Theorem 2. *Assume the conditions of Corollary 1 hold true, and let $\hat{\theta}_T(\lambda)$ be defined as in (3). Furthermore, let $\hat{\lambda}_G$ be the global minimizer of $TE(\lambda)$. The following hold:*

- (a) *Assume $\hat{\lambda}_G \xrightarrow{\text{Pr}} \lambda_0 < 0$ and $\sqrt{n}\{\hat{\theta}(0) - \theta_0\} \xrightarrow{d} N$, where N is some zero mean normal distribution. Then $\hat{\theta}_T(\hat{\lambda}) \xrightarrow{\text{Pr}} \theta_0$ and $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}) - \theta_0\} \xrightarrow{d} N$.*
- (b) *Assume $\hat{\lambda}_G \xrightarrow{\text{Pr}} \lambda_0 > 1$ and $\sqrt{n}\{\hat{\theta}(1) - \theta_0\} \xrightarrow{d} N$. Then $\hat{\theta}_T(\hat{\lambda}) \xrightarrow{\text{Pr}} \theta_0$ and $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}_G) - \theta_0\} \xrightarrow{d} N$.*
- (c) *Assume $\hat{\lambda}_G \xrightarrow{\text{Pr}} \lambda_0 = 0$ and that $\sqrt{n}\{\hat{\theta}(\hat{\lambda}_G)^T - \theta_0^T, \hat{\theta}(0)^T - \theta_0^T, \hat{\lambda}\}^T \xrightarrow{d} (N_1^T, N_2^T, N_3)^T \sim N(0, V)$ for some matrix V . Then $\hat{\theta}_T(\hat{\lambda}) \xrightarrow{\text{Pr}} \theta_0$ and $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}) - \theta_0\} \xrightarrow{d} I(N_3 \geq 0)N_1 + I(N_3 < 0)N_2$.*
- (d) *Assume $\hat{\lambda}_G \xrightarrow{\text{Pr}} \lambda_0 = 1$ and that $\sqrt{n}\{\hat{\theta}(\hat{\lambda}_G)^T - \theta_0^T, \hat{\theta}(1)^T - \theta_0^T, \hat{\lambda} - 1\}^T \xrightarrow{d} (N_1^T, N_2^T, N_3)^T \sim N(0, V)$ for some matrix V . Then $\hat{\theta}_T(\hat{\lambda}) \xrightarrow{\text{Pr}} \theta_0$ and $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}_G) - \theta_0\} \xrightarrow{d} I(N_3 \leq 0)N_1 + I(N_3 > 0)N_2$.*

Remark. From Z-estimation theory, $\sqrt{n}\{\hat{\theta}(\lambda) - \theta(\lambda)\} = J(\lambda)^{-1}n^{-1/2}\sum_{i=1}^n \varphi\{Z_i, \theta(\lambda), \lambda\} + o_p(1)$ where $\theta(\lambda)$ is the root of $E\varphi(Z, \theta, \lambda)$ and $J(\lambda) = -\partial_{\theta(\lambda)} E\varphi(Z, \theta, \lambda)$ for $\lambda = 0, 1$. Combining this with (2) and the central limit theorem shows the joint convergence required for case (c) and (d).

Proof. We will only prove (a) and (c). The remaining cases are similar. Since $\hat{\lambda}_G \xrightarrow{\text{Pr}} \lambda_0 < 0$ the probability of $\hat{\lambda}_G < 0$ tends to 1. Furthermore, conditioned on this event, $\hat{\theta}(\hat{\lambda}_G) = \hat{\theta}(0)$. Hence, $\Pr\{\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}_G) - \theta_0\lambda_0\} \leq x\} = \lim_{n \rightarrow \infty} [\Pr\{\sqrt{n}\{\hat{\theta}(0) - \theta_0\} \leq$

$x\} \Pr(\hat{\lambda}_G < 0) + o(\Pr(\hat{\lambda}_G \geq 0))]$. The right hand side of this equation converges to $\Pr(N \leq x)$. This proves (a).

For (c) note that $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}_G) - \theta_0(0)\} = I\{\sqrt{n}(\hat{\lambda}_G - 0) \geq 0\}\sqrt{n}\{\hat{\theta}(\hat{\lambda}_G) - \theta_0(0)\} + I\{\sqrt{n}(\hat{\lambda}_G - 0) < 0\}\sqrt{n}\{\hat{\theta}(0) - \theta_0(0)\}$. Now since the function $g(x, y, z) = I(z \geq 0)x + I(z < 0)y$ has the following set of discontinuity points $D_g = \{(x, y, 0) \mid x, y \in \mathbb{R}\}$, and this set has measure zero in the multivariate normal distribution, the continuous mapping theorem guarantees that $\sqrt{n}\{\hat{\theta}_T(\hat{\lambda}_G) - \theta_0(0)\}$ converges in law towards $g(N_1, N_2, N_3) = I(N_3 \geq 0)N_1 + I(N_3 < 0)N_2$. This concludes the proof. $\blacksquare$

2.2 Estimators of the variance

Now that we know how $\hat{\theta}(\hat{\lambda})$ behaves asymptotically, we will define estimators of its limiting variance. We start by giving approximations to the quantities in Corollary 1.

Theorem 3. *Let $\hat{\lambda}$ be the minimizer of $TE(\lambda)$ and $\hat{\theta}(\hat{\lambda})$ the solution to $0 = \Phi_n(\theta, \hat{\lambda})$, where $\Phi_n(\theta, \lambda) = n^{-1} \sum_{i=1}^n \varphi(Z_i, \theta, \lambda)$. Define the following estimators: $\hat{Z}_1 = n^{-1} \sum_{i=1}^n H_{\hat{\lambda}} \psi\{Z_i, \hat{\theta}(\hat{\lambda})\}$, $\hat{Z}_2 = H_{\hat{\theta}(\hat{\lambda})} \Psi_n(\theta)$ and $\hat{b} = \nabla_{\hat{\theta}(\hat{\lambda})} \Psi_n(\theta)$, where $\Psi_n(\theta) = n^{-1} \sum_{i=1}^n \psi(Z_i, \theta)$. Furthermore, let $\hat{J} = -\partial_{\hat{\theta}(\hat{\lambda})} \Phi_n(\theta, \hat{\lambda})$, $\hat{D} = \hat{J}^{-1} \partial_{\hat{\lambda}} \Phi_n\{\hat{\theta}(\hat{\lambda}), \lambda\}$, $\hat{K}^* = n^{-1} \sum_{i=1}^n \hat{\eta} \hat{\eta}^T$, where $\hat{\eta} = \eta\{z, \hat{\theta}(\hat{\lambda}), \hat{\lambda}, \hat{\theta}'(\hat{\lambda})\}$ and*

$$\hat{W}^j = \frac{1}{n} \sum_{i=1}^n \left\{ \begin{pmatrix} \hat{D}_j^T H_{\hat{\theta}} \varphi^1(Z_i, \theta, \hat{\lambda}) \\ \vdots \\ \hat{D}_j^T H_{\hat{\theta}} \varphi^p(Z_i, \theta, \hat{\lambda}) \end{pmatrix} + \partial_{\hat{\lambda}_j} \partial_{\hat{\theta}} \varphi(Z_i, \theta, \lambda) \right\} \text{ for } j = 1, \dots, q,$$

with η as defined in Theorem 1. These estimators are consistent for the corresponding quantities, if there exists functions depending only on z that bound the norms of all partial derivatives of $\partial_{\theta} \varphi$, $\partial_{\lambda} \varphi$, $H_{\lambda} \psi$, $H_{\theta} \psi$, $\nabla_{\theta} \psi$, $(z, \theta, \lambda) \mapsto \eta(z, \theta, \lambda) \eta(z, \theta, \lambda)^T$, $H_{\theta} \varphi^j$ and $\partial_{\lambda_j} \partial_{\theta} \varphi$ for $j = 1, \dots, p$ with respect to $(\theta, \lambda, \theta')$ in a neighbourhood of $(\theta_0^T, \lambda_0^T, D^T)^T$.

Proof. This follows from standard theory, see e.g. Dæhlen et al. (2024, Le. 4, Ap. B). $\blacksquare$

We are now ready to define estimators of the limiting variance of $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0(\lambda_0)\}$.

Theorem 4. Assume we are minimising $TE(\lambda)$ in $[0,1]$ and the conditions of Theorem 3 and Corollary 1 hold true. Then, $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ is asymptotically normal with variance V which can be estimated consistently by $\hat{V}_1 = \hat{A}^* \hat{K}^* (\hat{A}^*)^T$ if $\hat{\lambda}$ lies in the interior of $[0,1]$, or $\hat{V}_2 = \hat{J}^{-1} \hat{K} (\hat{J}^{-1})^T$ when TE is minimised in λ with $TE'(\lambda) \neq 0$ for λ equal to 0 or 1. Here $\hat{K} = n^{-1} \sum_{i=1}^n \varphi\{Z_i, \hat{\theta}(\hat{\lambda}), \hat{\lambda}\}$, $\hat{A}^* = (\hat{A}_1, \hat{A}_2, \hat{A}_3)$ with $\hat{A}_1 = \hat{J}^{-1} - \hat{D} \hat{Z}_1^{-1} \{\hat{D}^T \hat{Z}_2 + \hat{W}\} \hat{J}^{-1}$, $\hat{A}_2 = -\hat{D} \hat{Z}_1 \hat{D}^T$ and $\hat{A}_3 = -\hat{D} \hat{Z}_1^{-1} \hat{M}$, where $\hat{M}$ is a block diagonal matrix with $\hat{b}^T \hat{J}^{-1}$ on each of its entries and $\hat{W} = ((\hat{b}^T \hat{J}^{-1} \hat{W}^1)^T, \dots, (\hat{b}^T \hat{J}^{-1} \hat{W}^q)^T)^T$.

Proof. This follows directly from Corollary 1 and Theorem 2, in conjunction with Theorem 3 and the continuous mapping theorem. ■

It could of course happen that λ_0 is exactly equal to either of the endpoints of Λ . Then, case (c) or (d) of Theorem 2 applies and $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ is not guaranteed to be asymptotically normal, so that approximating it by $N(0, \hat{V}_1)$ or $N(0, \hat{V}_2)$ will not be correct. In Section 5 we will, however, see that the distribution in cases (c) and (d) often simplifies to a central multivariate normal distribution also in this case, and that the variance can be estimated by either $\hat{V}_1$ or $\hat{V}_2$, defined in the preceding theorem.

3 Information criteria

The theory presented in the previous sections is not limited to cases where $\hat{\lambda}$ is set to the minimizer of the training error. The results also hold true when the tuning parameter minimises quantities which are “almost” equal to TE. The most obvious example of this is perhaps when $\hat{\lambda}$ is set to the minimizer of an information criterion.

Let $f_\lambda(z, \theta)$ be some parametric model and $\ell_{\lambda,n}(\theta)$ the corresponding likelihood. Furthermore, for each λ , let $\hat{\theta}(\lambda)$ be the maximizer of $\ell_{\lambda,n}(\theta)$. Then by definition Akaike’s information criterion (AIC), the Bayesian information criterion (BIC) and Takeuchi’s information criterion (TIC) take the following forms for each fixed λ : $AIC(\lambda) = -n^{-1} \ell_{\lambda,n}\{\hat{\theta}(\lambda)\} + n^{-1}p$, $BIC(\lambda) = -n^{-1} \ell_{\lambda,n}\{\hat{\theta}(\lambda)\} + n^{-1}p \log n$ and $TIC(\lambda) =$

$-n^{-1}\ell_{\lambda,n}\{\hat{\theta}(\lambda)\} + n^{-1}\text{Tr}\{\hat{J}(\hat{\theta}, \lambda)^{-1} \hat{K}(\hat{\theta}, \lambda)\}$, where p is the dimension of θ , the number of parameters in the model $f_\lambda(z, \theta)$ at a fixed λ . Hence, provided $\hat{J}$ and $\hat{K}$ are sufficiently regular, the above information criteria are all only $o_p(1/\sqrt{n})$ away from $n^{-1}\ell_{\lambda,n}\{\hat{\theta}(\lambda)\}$, allowing Theorem 1 to be applied after replacing φ by $\nabla_\theta \log f_\lambda(z, \theta)$ and $\eta_2(z, \alpha)$ by $(z, \alpha) \mapsto \partial_\theta \log f_\lambda(z, \theta)D + \partial_\lambda \log f_\lambda(z, \theta)$. Corollary 1 and Theorem 2 also hold after making similar modifications.

The fact that the results of this article can be applied also when λ is tuned by minimising the AIC, BIC or TIC follows more or less directly from the forms of the criteria. We will now discuss the focused information criterion (FIC), a newer criterion for which the result is less immediate. The FIC was introduced in Claeskens and Hjort (2003), and is an information criterion with a different aim than the classic criteria discussed above. Rather than ranking models according to how well they fit the data overall, the FIC evaluates models by how well they estimate some pre-specified parameter of interest. This parameter of interest is called the focus parameter and should be chosen to reflect the main goal of an analysis. If, for instance, the goal is to estimate the median in a population, the FIC prefers models for which the estimator of the median is good. The quality of an estimator is evaluated by its mean squared error (MSE), and hence, the FIC of a model is the estimated MSE of that model's estimator of the focus parameter. We will work with the relatively newer version of the criterion introduced in Jullum and Hjort (2017) and modified in Dæhlen et al. (2024).

As before, let $f_\lambda(z, \theta)$ be some parametric model and $\hat{\theta}(\lambda)$ the maximizer of $\theta \mapsto \ell_{\lambda,n}(\theta) = \sum_{i=1}^n \log f_\lambda(Z_i, \theta)$ for each λ . Let μ be the focus parameter with true value μ_0 and assume μ takes the value $\mu_\lambda(\theta)$ in the model $f_\lambda(z, \theta)$. The FIC of the model $f_\lambda(z, \theta)$ for each λ is the MSE of $\mu_\lambda\{\hat{\theta}(\lambda)\}$. Dæhlen et al. (2024) gives an estimator of this quantity: $\text{FIC}(\lambda) = \hat{b}_\lambda^2 + n^{-1}(2\hat{b}_\lambda\hat{c}_\lambda - \hat{\kappa}_\lambda + \hat{\tau}_\lambda)$, where $\hat{c}_\lambda$, $\hat{\kappa}_\lambda$ and $\hat{\tau}_\lambda$ are consistent estimators of certain population quantities defined in the paper, and $\hat{b}_\lambda = \mu\{\hat{\theta}(\lambda)\} - \hat{\mu}_0$ for some estimator $\hat{\mu}_0$, which is assumed to be consistent for μ_0 . Assume now that $\hat{\mu}_0$ is the solution to an equation on the form $0 = n^{-1} \sum_{i=1}^n \xi(Z_i, \mu)$. We

then have $\text{FIC}(\lambda) = n^{-1} \sum_{i=1}^n [\mu\{\hat{\theta}(\lambda)\} - \hat{\mu}_0]^2 + O_p(1/n)$, and hence, provided sufficient regularity, Theorem 1 can be applied by replacing $\theta(\lambda)$ by $(\theta(\lambda)^T, \mu)^T$, $\varphi(z, \theta, \lambda)$ by $(\varphi(z, \theta, \lambda)^T, \xi(z, \mu_0))^T$ and $\eta_2(z, \alpha)$ by $(\theta, \mu, \lambda, D) \mapsto 2\{\mu(\theta) - \mu\}(D^T \nabla_{\theta} \mu(\theta)^T, -1)^T$. In Section 6.2, we illustrate this in an example.

Remark. It is worth noting that we require the criterion used for fitting λ to be differentiable as a function of the tuning parameter. Because of this, our theorems do not cover situations where λ is discrete. In particular, our theorems cannot be applied when an information criterion is used to e.g. choose between a finite number of models.

4 Cross-validation

When λ is tuned by minimising the training error, $\hat{\theta}(\hat{\lambda})$ can be written as the components of a Z-estimator. Because of this the results of the previous section follow more or less directly from standard theory. In practice, however, λ is seldom tuned with respect to TE, but instead set to the minimizer of $\text{CV}(\lambda)$, the estimate of the risk function obtained by cross-validation. In this section, we will show that the difference between tuning with respect to TE and CV is negligible asymptotically and that all of the previously derived results can be applied also when $\hat{\lambda}$ is the minimizer of CV.

First, note that for each λ the vector $(\hat{\theta}(\lambda)^T, \hat{\theta}'(\lambda)^T)^T$ is the solution to equation $n^{-1} \sum_{i=1}^n \xi(Z_i, \theta, \lambda, D) = 0$ with $\xi(z, \theta, \lambda, D) = (\xi_1(z, \theta, \lambda)^T, \xi_2(z, \theta, \lambda, D)^T)^T$ where $\xi_1 = \varphi$ and $\xi_2(z, \theta, \lambda, D) = \partial_{\theta} \varphi(z, \theta, \lambda) D + \partial_{\lambda} \varphi(z, \theta, \lambda)$. This follows from definition of $\hat{\theta}(\lambda)$ and the implicit function theorem. Because of this, the vector $(\hat{\theta}(\lambda)^T, \hat{\theta}'(\lambda)^T)^T$ is a Z-estimator, and under weak conditions $(\hat{\theta}(\lambda)^T, \hat{\theta}'(\lambda)^T)^T$ is consistent for some vector $(\theta_0(\lambda)^T, D_0(\lambda)^T)^T$ and $(\hat{\theta}(\lambda)^T, \hat{\theta}'(\lambda)^T)^T \approx (\theta_0(\lambda)^T, D_0(\lambda)^T)^T + V(\lambda)^{-1} n^{-1} \sum_{i=1}^n \xi\{Z_i, \theta_0(\lambda), D_0(\lambda), \lambda\}$ where $V(\lambda) = -\partial_{\theta_0(\lambda), D_0(\lambda)} E \xi(Z, \theta, \theta', \lambda)$. See e.g. chapter 5 of Van der Vaart (2000) for proofs and sufficient conditions. Because of this $(\hat{\theta}(\lambda)^T - \hat{\theta}_{(-i)}(\lambda)^T, \hat{\theta}'(\lambda)^T - \hat{\theta}'_{(-i)}(\lambda)^T) \approx n^{-1} V(\lambda)^{-1} \xi\{Z_i, \theta_0(\lambda), \theta'_0(\lambda)\} = O_p(1/n)$. We formalize this idea in the following lemma.

Lemma 1. *Let the parameter sets Θ for θ , $\mathcal{D}$ for $\hat{\theta}'(\lambda)$ and Λ for λ be compact and $\xi(z, \theta, D, \lambda) = (\varphi(z, \theta, \lambda)^T, \xi_2(z, \theta, \lambda, D)^T)$ where $\xi_2(z, \theta, \lambda, D) = \partial_\theta \varphi(z, \theta, \lambda)D + \partial_\lambda \varphi(z, \theta, \lambda)$. Assume the conditions of Lemma S1.3 in the supplementary material hold true for this ξ , and in particular, that $m_0(z)$ bounds $\xi(z, \theta, D, \lambda)$ for all θ , D and λ . Then $\|\hat{\theta}_{(-i)}(\lambda) - \hat{\theta}(\lambda)\|, \|\hat{\theta}'_{(-i)}(\lambda) - \hat{\theta}'(\lambda)\| \leq a_{0,n}m_0(Z_i)$ where $a_{0,n} = O_p(1/n)$ uniformly in λ and does not depend on i .*

The proof is given in the supplementary material. With Lemma 1, we are ready to quantify the difference between $CV'(\lambda)$ and $TE'(\lambda)$.

Lemma 2. *Assume all conditions of Lemma 1 hold true with $E m_0(Z)^4 < \infty$. Then*

$$\frac{1}{n} \sum_{i=1}^n \hat{\theta}'_{(-i)}(\lambda)^T \nabla_{\hat{\theta}_{(-i)}(\lambda)} \psi(Z_i, \theta) = \frac{1}{n} \sum_{i=1}^n \hat{\theta}'(\lambda)^T \nabla_{\hat{\theta}(\lambda)} \psi(Z_i, \theta) + \delta_n(\lambda), \quad (4)$$

where $\|\delta_n(\lambda)\| = O_p(1/n)$ uniformly in λ , provided there exist square integrable functions p_1 and p_2 not depending on θ that bound respectively all first and second order partial derivatives of ψ with respect to θ .

Proof. By Lemma 1, $\hat{\theta}_{(-i)}(\lambda) = \hat{\theta}(\lambda) + R_n^i(\lambda)$ and $\hat{\theta}'_{(-i)}(\lambda) = \hat{\theta}'(\lambda) + S_n^i(\lambda)$ with $\|R_n^i(\lambda)\|, \|S_n^i(\lambda)\| \leq a_{0,n}m_0(Z_i)$. A Taylor expansion reveals that the left hand side of (4) is equal to $n^{-1} \sum_{i=1}^n \{\hat{\theta}'(\lambda) + S_n^i(\lambda)\}^T \{\nabla_{\hat{\theta}(\lambda)} \psi(Z_i, \theta) + H_{\theta_i^*} \psi(Z_i, \theta) R_n^i(\lambda)\}$, for some θ_i^* s on the line segments between $\hat{\theta}(\lambda)$ and $\hat{\theta}_{(-i)}(\lambda)$. Utilizing the existence of p_1 and p_2 , we get that $\|CV'(\lambda) - TE'(\lambda)\|$ is bounded by $a_{0,n} \hat{\theta}'(\lambda)^T n^{-1} \sum_{i=1}^n p_2(Z_i) m_0(Z_i) + a_{n,0} n^{-1} \sum_{i=1}^n p_1(Z_i) m_0(Z_i) + a_{0,n}^2 n^{-1} \sum_{i=1}^n p_2(Z_i) m_0(Z_i)^2$. Since $E p_1(Z)^2$, $E p_2(Z)^2$ and $E m_0(Z)^4$ exist, the means converge. This ensures $\|CV'(\lambda) - TE'(\lambda)\| = O_p(1/n)$ uniformly in λ as $\mathcal{D}$ is compact. $\blacksquare$

Since $CV'(\lambda)$ and $TE'(\lambda)$ are uniformly close as functions of λ by the above lemma, we would expect $\hat{\theta}(\hat{\lambda})$ to asymptotically behave similarly when tuning with respect to CV and TE. The following theorem confirms the intuition.

Theorem 5. *Let λ_{TE} be the minimizer of TE and λ_{CV} the minimizer of CV and assume that the conditions of Theorem 2 and Lemma 2 hold true. Then $\sqrt{n}\{\hat{\theta}(\lambda_{\text{TE}}) - \theta_0\}$ and $\sqrt{n}\{\hat{\theta}(\lambda_{\text{CV}}) - \theta_0\}$ are asymptotically equivalent. In particular, Theorem 1, Corollary 1 and Theorem 2 hold true for $\hat{\theta}(\lambda_{\text{CV}})$, and the limiting distribution of $\sqrt{n}\{\hat{\theta}(\lambda_{\text{CV}}) - \theta_0\}$ can be estimated as in Theorem 4.*

Proof. Let $\hat{\alpha}_{\text{CV}} = (\hat{\theta}(\lambda_{\text{CV}})^T, \hat{\lambda}_{\text{CV}}^T, \hat{\theta}'(\lambda_{\text{CV}})^T)^T$. Since $\text{TE}'(\lambda) = \text{CV}'(\lambda) + O_p(1/n)$ uniformly in λ , $\text{TE}'(\lambda_{\text{CV}}) = O_p(1/n)$. Hence, $n^{-1} \sum_{i=1}^n \eta(Z_i, \hat{\alpha}_{\text{CV}}) = O_p(1/n)$. Since $n^{-1} \sum_{i=1}^n \eta(Z_i, \hat{\alpha}_{\text{CV}}) = o_p(1/\sqrt{n})$ is all that is needed for Van der Vaart (2000, Th. 5.21), the conclusion of Theorem 1 follows for $\hat{\alpha}_{\text{CV}}$ as well. This ensures that $\sqrt{n}(\hat{\alpha} - \alpha_0)$ and $\sqrt{n}(\hat{\alpha}_{\text{CV}} - \alpha_0)$ are asymptotically equivalent. Theorem 5 follows from this fact. ■

4.1 A more precise approximation

Theorem 5 is a somewhat surprising result, as it tells us that asymptotically speaking it makes no difference if we tune with respect to TE or CV. Yet, in most applications, minimising TE and CV gives very different results. We will now take a closer look at what the difference between CV and TE really is. To make the derivations more readable, we will only show pointwise results for $\hat{\theta}(\lambda)$, and drop λ from the notation.

Arguing similarly as in the paragraph preceding Lemma 1 we can show $\hat{\theta}_{(-i)} - \hat{\theta} \approx -n^{-1} J^{-1} \varphi(Z_i, \theta_0)$ where θ_0 is the solution to $E \varphi(Z, \theta) = 0$ and $J = -\partial_{\theta_0} E \varphi(Z, \theta)$. This approximation is not particularly surprising. The function $z \mapsto J^{-1} \varphi\{z, \theta_0\}$ is called the influence function of $\hat{\theta}$ as it in some sense tells how much a single data point affects estimation of θ . See e.g. Huber (2009) for more details on influence functions. Since $\hat{\theta}_{(-i)}$ is the estimator you get instead of $\hat{\theta}$ if you remove the i -th data point, it is natural that the difference between $\hat{\theta}_{(-i)}$ and $\hat{\theta}$, is precisely the “influence” of Z_i .

Assuming ψ is sufficiently smooth in θ , a Taylor expansion shows $n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}) + n^{-1} \sum_{i=1}^n (\hat{\theta}_{(-i)} - \hat{\theta})^T \nabla_{\hat{\theta}} \psi(Z_i, \theta) \approx \text{CV}$. A further Taylor expansion of $\nabla_{\hat{\theta}} \psi(Z_i, \theta)$ around θ_0 , reveals $\text{CV} \approx n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}) - n^{-1} \sum_{i=1}^n (\hat{\theta}_{(-i)} - \hat{\theta}) \nabla \psi(Z_i, \theta_0) \approx n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}) -$

$n^{-2} \sum_{i=1}^n \varphi(Z_i, \theta_0)^T (J^{-1})^T \nabla_{\theta_0} \psi(Z_i, \theta)$. By the law of large numbers and properties of the trace operator, we hence get $CV \approx TE - n^{-1} \text{Tr}(J^{-1}C)$ where $C = E \varphi(Z_i, \theta_0) \nabla_{\theta_0} \psi(Z, \theta_0)^T$. The following theorem confirms these informal arguments. A full proof is given in the supplementary material.

Theorem 6. *Let $\hat{\theta}$ and $\hat{\theta}_{(-i)}$ the solutions to $\sum_{j=1}^n \varphi(Z_j, \theta) = 0$ and $\sum_{j \neq i}^n \varphi(Z_j, \theta)$ respectively. Assume further that $\hat{\theta}, \hat{\theta}_{(-i)} \in \Theta \subseteq \mathbb{R}^p$ where Θ is compact and that there exists functions m_0, m_1, p_1 and p_2 bounding the norms of φ , all partial derivatives of φ with respect to θ , ψ and all partial derivatives of ψ with respect to θ , respectively. Then, for a given function $\psi(z, \theta)$, we have*

$$n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}_{(-i)}) = n^{-1} \sum_{i=1}^n \psi(Z_i, \hat{\theta}) - n^{-1} \text{Tr}(J^{-1}C) + o_p(1/n) \quad (5)$$

where $J = -\partial_{\theta_0} E \varphi(Z, \theta)$ and $C = E \varphi(Z, \theta_0) \nabla_{\theta_0} \psi(Z, \theta)^T$, with θ_0 being the solution to $E \varphi(Z, \theta) = 0$, provided the eigenvalues of $\{\partial_{\theta} E \varphi(Z, \theta)\}^{-1}$ are bounded in Θ and $E m_0(Z)^4, E m_1(Z)^4, E p_1(Z)^4, E p_2(Z)^2 < \infty$.

When $\varphi(z, \theta) = \nabla_{\theta} \log f(z, \theta)$ for some density or probability mass function f , J is the Fisher matrix in the model. Furthermore, $C = K$ if $\psi = \log f$, where $K = \text{Var} \varphi(Z, \theta_0)$. Hence, Theorem 6 simplifies to $\sum_{i=1}^n \psi(Z_i, \hat{\theta}_{(-i)}) = \text{TIC} + o_p(1)$, where TIC is Takeuchi's information criterion (TIC), defined in Section 3. Additionally, if the model f is specified correctly, $K = J$ and $\sum_{i=1}^n \psi(Z_i, \hat{\theta}_{(-i)}) = \ell_n(\hat{\theta}) - p + o_p(1)$. The right hand side of this equation is easily recognized as Akaike's information criterion (AIC), leading Theorem 6 to simplify to the results of Stone (1977) when $\varphi(z, \theta) = \nabla_{\theta} \log f(z, \theta)$ and the parametric model is specified correctly.

Theorem 6 shows that CV and TE are more similar than one might initially think. In fact, the only thing separating the two quantities is a term of order $O_p(1/n)$, which is dominated by the size of TE itself, asymptotically speaking. In smaller samples, however, $n^{-1} \text{Tr}(J^{-1}C)$ might be quite large and will ensure that tuning with respect to TE and CV give different results. That being said, there are cases where $n^{-1} \text{Tr}(J^{-1}C)$

is not negligible and the difference between TE and CV will matter also in the limit. We will now turn our attention towards such a case.

4.2 When more precision is needed

For each parameter θ let $R(\theta)$ be some risk function, and let $\hat{\theta}$ be some estimator of θ based on data $Z_1, \dots, Z_n$. Assume further that $\hat{\theta} \xrightarrow{\text{Pr}} \theta_0$ and $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, V)$ for some matrix V . By Taylor expanding ψ around θ_0 and using the law of total expectation, one can show that, provided sufficient regularity,

$$E\{R(\hat{\theta})\} = R(\theta_0) + n^{-1}[\nabla_{\theta_0} R(\theta_0)c + (1/2) \text{Tr}\{H_{\theta_0} R(\theta_0)V\}] + o(1/n), \quad (6)$$

where c is equal to the limit of $n E(\hat{\theta} - \theta_0)$, see Dæhlen et al. (2024) for more on this parameter. If $\hat{\theta} \in \mathbb{R}$ and $R(\theta) = (\theta_0 - \theta)^2$, (6) reduces to the classical bias-variance decomposition of the MSE. With $R(\theta)$ equal to the error rate of certain classification procedures, (6) reduces to the expressions given in O'Neill (1980).

Note that (6) consists of three parts: a constant part, $R(\theta_0)$, a term of size $O(1/n)$ and a remainder term of smaller order. For two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ consistent for different values θ_0^1 and θ_0^2 respectively, the difference between their expected risk will asymptotically be dominated by how much $R(\theta_0^1)$ and $R(\theta_0^2)$ differ, making the $O(1/n)$ terms in (6) negligible asymptotically. If, on the other hand θ_0^1 and θ_0^2 are equal, the difference in expected risk for $\hat{\theta}_1$ and $\hat{\theta}_2$ will be dominated by the difference between the two corresponding terms of size $O(1/n)$ in (6). In this latter situation, we therefore need to estimate the $O(1/n)$ -effect from data in order to properly distinguish the risk of $\hat{\theta}_1$ from that of $\hat{\theta}_2$. Both TE and CV attempt to estimate $E R(\hat{\theta})$, but, as we will see, neither estimator is precise enough capture the $O(1/n)$ effects.

Let $R_n(\theta) = n^{-1} \sum_{i=1}^n \psi(Z_i, \theta)$ and $\hat{\theta}$ be some estimator consistent for θ_0 . Assuming

ψ is sufficiently smooth, a Taylor expansion shows

$$\text{TE} = R_n(\theta_0) + (\hat{\theta} - \theta_0)^T \nabla_{\theta_0} R_n(\theta) + (1/2)(\hat{\theta} - \theta_0)^T H_{\theta_0} R_n(\theta)(\hat{\theta} - \theta_0) + o_p(1/n). \quad (7)$$

Since $R_n(\theta_0)$ is a mean and $\hat{\theta} \xrightarrow{P} \theta_0$, we have $\text{TE} \xrightarrow{P} R(\theta_0)$, which means that TE does a good job, as far as estimating the constant part of (6) goes. Approximating the term of order $O(1/n)$ is, however, another story. The most obvious reason is perhaps that the remaining terms are nowhere near being consistent for $n^{-1}[\nabla_{\theta_0} R(\theta_0)c + (1/2) \text{Tr}\{H_{\theta_0} R(\theta)V\}]$, but even if they were, TE would not be precise enough to properly catch effects of size $O(1/n)$, as $R_n(\theta_0) = R(\theta_0) + O_p(1/\sqrt{n})$ by the central limit theorem. Since a term of order $O_p(1/\sqrt{n})$ will dominate $O(1/n)$ effects, the $O(1/n)$ -terms in (6) are too small to be captured by TE and will be washed out by the error of $R_n(\theta_0)$. As a result, TE is not well suited for evaluating estimators whose limiting risk differs only in the $O(1/n)$ -term. One might hope that this is a defect of the TE only, and that CV will be precise enough to capture the $O(1/n)$ effects. Sadly, Theorem 6 stops this idea dead in its tracks. Since the difference between CV and TE is a term of order $O_p(1/n)$, the added precision of CV is not sufficient to correct for the $O_p(1/\sqrt{n})$ error TE makes. This is pointed out in the context of model selection of linear regression models in Shao (1993).

The above might lead one to believe that the difference between CV and TE is negligible and that there is no reason to use CV rather than TE in analysis. This conclusion is, however, slightly too negative, as there is an important difference between CV and TE in expected value. To see this, assume that $E\psi(Z, \theta_0) = 0$ and that $\hat{\theta} = \theta_0 + J^{-1}n^{-1} \sum_{i=1}^n \varphi(Z_i, \theta_0) + o_p(1/\sqrt{n})$. Then, by the central limit theorem, $\sqrt{n}(\hat{\theta} - \theta_0)$ and $\sqrt{n}\{R_n(\theta_0) - R(\theta_0)\}$ converge jointly in law to a central normal distribution with variance matrix

$$\Sigma = \begin{pmatrix} J^{-1}K(J^{-1})^T & J^{-1}C^T \\ C^T(J^{-1})^T & \text{Var} \psi(Z, \theta_0) \end{pmatrix},$$

where J and C are as in Theorem 6 and $K = \text{Var } \varphi(Z, \theta_0)$. Hence, provided all necessary quantities are uniformly integrable (see e.g. Billingsley (1999, p. 31)), $E(\text{TE}) = R(\theta_0) + \text{Tr}(CJ^{-1}) + (1/2) \text{Tr}\{H_{\theta_0}R(\theta)J^{-1}K(J^{-1})^T\} + o(1/n)$ which misses $E R(\hat{\theta})$ by a term $n^{-1} \text{Tr}(J^{-1}C) + o(1/n)$. Because of this, TE is a biased estimator of $E R(\hat{\theta})$. CV, on the other hand, satisfies $\text{CV} = \text{TE} - \text{Tr}(J^{-1}C) + o(1/n)$ by Theorem 6. Hence, arguing similarly as for TE, we get $E \text{CV} = E R(\hat{\theta}) + o(1/n)$, ensuring that in expected value, CV is able to separate between estimators whose risk differ only in the $O(1/n)$ -term. We stress that this holds in expected value only, and that the above shows nothing more than the fact that the bias of CV is of order $o(1/n)$. As noted before, the error of CV will be of order $O_p(1/\sqrt{n})$ for any single sample, a term which is too large to properly capture effects of size $O(1/n)$.

The discussion given in the previous paragraphs has consequences for our theorems concerning the limiting behaviour of $\hat{\theta}(\hat{\lambda})$. By the arguments above, neither TE nor CV is able to separate estimators whose risk differs only in the $O(1/n)$ term. Hence, when tuning λ with respect to CV or TE, we should not expect that the criteria are able to identify the optimal value of λ in this $O(1/n)$ -regime. This also explains why tuning with respect to TE is asymptotically equivalent to tuning with respect to CV.

The fact that the theory breaks down when $\theta_0(\lambda)$, the solution to $E \varphi(Z, \theta, \lambda) = 0$, is constant as a function of λ , can also be seen from Theorem 1. When $\theta_0(\lambda)$ is constant as a function of λ , the equation $\Psi(\alpha) = 0$ has multiple solutions and $\Psi'(\alpha_0)$ will be singular. Because of this, the conditions of Theorem 1 do not hold true when $\theta_0(\lambda)$ is constant, making the theorem inapplicable in this case. This does not pose a problem for regularized estimators like in ridge regression, as $\theta_0(\lambda)$ will rarely be constant in this case, but for hybrid estimators, this can indeed be quite problematic. Take for instance the hybrid generative-discriminative classification model where $\hat{\theta}(\lambda)$ is set to the maximizer of the following expression $n^{-1} \sum_{i=1}^n \{\lambda \log f_{Y|X}(Y_i, \theta | X = X_i) + (1 - \lambda) \log f_{(Y,X)}(Y_i, X_i, \theta)\}$ for some parametric model $f_{(Y,X)}$ for (Y, X) and where $f_{Y|X}$ is derived from $f_{(Y,X)}$. In this case, $\hat{\theta}(\lambda)$ aims for the maximizer of

$\Gamma_\lambda(\theta) = \lambda \mathbb{E} \log f_{(Y,X)}(Y, X, \theta) + (1 - \lambda) \mathbb{E} \log f_{Y|X}(Y, X, \theta)$. If the model is misspecified, $\theta_0(\lambda)$ will rarely be constant as a function of λ . If, on the other hand, the model does hold and the true underlying distribution is $f_{(Y,X)}(\cdot, \theta_0)$ for some $\theta_0 \in \mathbb{R}^p$, $\theta_0(\lambda)$ will be constantly equal to θ_0 . Because of this, the difference in expected limiting risk will be on the order of $O(1/n)$ when the model is specified correctly and $O(1)$ otherwise. Hence, Theorem 2 and Theorem 5 can be applied in this case only when the model is misspecified. See Section S2 in the supplementary material for more discussions.

5 When the effect of tuning disappears

We will now go through a situation in which the expressions in Corollary 1 simplify greatly and the effect of the tuning procedure becomes negligible asymptotically.

Assume there exists a λ_1 such that $\varphi(y, \theta, \lambda_1) = \nabla_\theta \psi(y, \theta)$ for all θ and F -almost all y . Then, $\theta_0(\lambda_1)$ solves $\mathbb{E} \varphi(Z, \theta, \lambda_1) = \nabla_\theta \mathbb{E} \psi(Z, \theta) = 0$. Because of this, λ_0 , b and Z_1 defined in Corollary 1 are equal to λ_1 , 0 and J , respectively. Entering this into the expressions in the corollary, shows that the right hand side of (2) is equal to $n^{-1} J^{-1} \varphi(z, \theta_0, \lambda_1)$. This is the influence function of $\hat{\theta}(\lambda_1)$. Hence, $\sqrt{n}(\hat{\theta}(\hat{\lambda}) - \theta_0) = \sqrt{n}\{\hat{\theta}(\lambda_1) - \theta_0\} + o_p(1)$ in this case, and the additional randomness introduced by tuning λ vanishes asymptotically. In practice, this means that one can ignore the effect of tuning λ and use the classic approximation $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\} \xrightarrow{d} N\{0, J^{-1} K (J^{-1})^T\}$, where $J = -\partial_{\theta_0} \mathbb{E} \varphi(Z, \theta, \lambda_1)$ and $K = \mathbb{E} \varphi(Z, \theta_0, \lambda_1) \varphi(Z, \theta_0, \lambda_1)^T$.

A λ_1 such that $\nabla_\theta \psi(z, \theta) = \varphi(z, \theta, \lambda_1)$ actually exists in quite a few situations. In for instance ridge regression where λ is set to the minimizer of estimated MSE, we have $\varphi(z, \beta, \lambda) = \nabla_\beta \{(y - \beta_1^T x - \beta_0)^2 + \lambda \|\beta_{1:}\|^2\}$, where $\beta = (\beta_0, \beta_{1:}^T)^T$, ensuring $\nabla_\beta \psi(z, \beta) = \varphi(z, \beta, 0)$. More generally, for any model $Y_i = g(X_i, \theta) + \epsilon_i$ fitted by minimising $\sum_{i=1}^n \{Y_i - g(X_i, \theta)\}^2 + \lambda \|\theta\|^2$ and tuned according to the MSE, $\nabla_\theta \psi(z, \theta) = \varphi(z, \theta, 0)$. Even more generally, for a model fitted by minimising $\sum_{i=1}^n h\{Y_i, g(X_i, \theta)\} + \lambda \|\theta\|^2$ for some function h , and where λ is tuned by minimising the CV or TE

estimate of $E h[Y, g\{X, \hat{\theta}(\lambda)\}]$, $\nabla_{\theta}\psi(z, \theta) = \varphi(z, \theta, 0)$, ensuring that $\sqrt{n}\{\hat{\theta}(\hat{\lambda}) - \theta_0\}$ and $\sqrt{n}\{\hat{\theta}(0) - \theta_0\}$ are asymptotically equivalent. This includes many regularized likelihood models when cross validated Kullback-Leibler divergence is used to tune λ .

The above arguments might lead one to believe that there rarely is need to use Theorem 2 or Theorem 5 in practice. This is, however, not correct. To illustrate the potentially large effect tuning of λ can have on the limit distribution of $\hat{\theta}(\hat{\lambda})$, we performed a small simulation experiment. For a selection of λ -values, we fitted logistic regression in two variables with a ridge penalty term. This corresponds to setting φ equal to the gradient of $y(\beta_0 + \beta_1 x_1 + \beta_2 x_2) - \log(1 + \beta_0 + \beta_1 x_1 + \beta_2 x_2) - \lambda(\beta_1^2 + \beta_2^2)$ with respect to $\beta = (\beta_0, \beta_1, \beta_2)$. We chose $\hat{\lambda}$ by minimising $CV(\lambda)$ with $\psi(z, \theta) = [\exp(\beta_0 + \beta_1 x_1) / \{1 + \exp(\beta_0 + \beta_1 x_1)\} - y]^2$. For the true underlying distribution of the data, we chose the following. The binary variable Y was set to be Bernoulli distributed with equal probability for zero and one. Furthermore, for a selection of C -values, we used $X | Y = 0 \sim N(\mu, \Sigma)$ and $X | Y = 1 \sim N(-\mu, 2\Sigma)$ with $\mu = (1/2, 1/2)^T$ and Σ equal to the matrix with diagonal elements 2 and off-diagonal entries $-\sqrt{C/2}$.

For each selected C -value, we simulated $n = 100$ data points and computed $\hat{\lambda}_C$ and $\hat{\beta}_C(\hat{\lambda})$ based on the data set. We used cross-validation to tune λ . This procedure was repeated $B = 1000$ times for each C , and based on these B samples, we computed the empirical variance of $\sqrt{n}\hat{\beta}_C(\hat{\lambda}_C)$ for each C . We also estimated the variance using both the classic method and Theorem 5. The results are summarized in Figure 2, which shows the error of the two estimators as functions of C . From Figure 2 it is clear that the classic estimator is insufficient in this case. The variance of Theorem 5 is closer to the observed variability in almost all cases. Furthermore, we see that as $C \rightarrow 0$, the classic variance estimate decreases in effect. This is particularly clear for the estimate of the variance of the last component of $\sqrt{n}\hat{\beta}_C(\lambda_C)$, where the variance estimator provided by the classic method is a lot worse than the one proposed in this article and seems to increase even more as $C \rightarrow 0$. This simulation is of course only a toy example, but Figure 2 indicates that the error introduced by using the classic

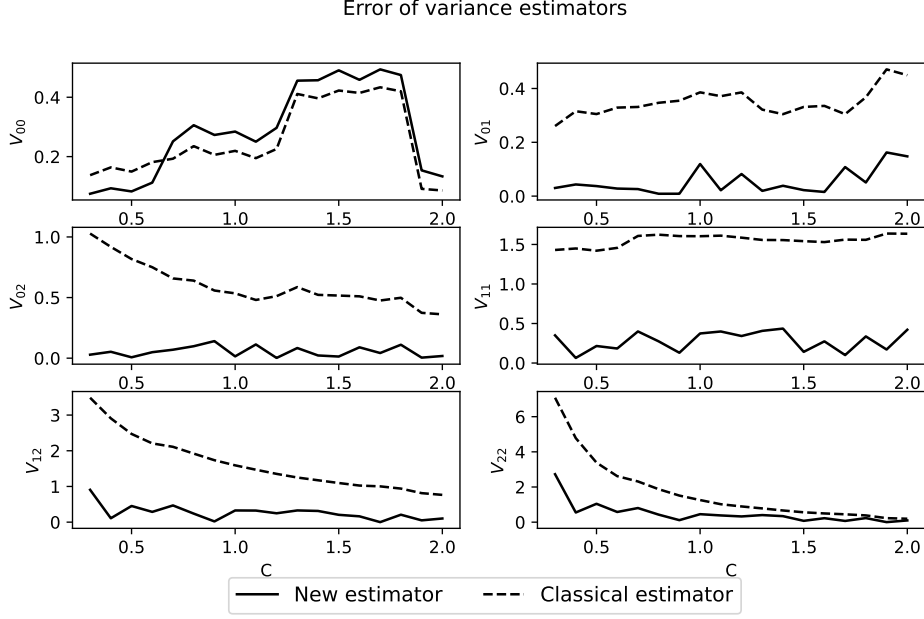


Figure 2: The plot shows the absolute error of the classic and new estimated covariance of $\sqrt{n}\hat{\beta}_C(\hat{\lambda}_C)$ compared to the observed variance based on a thousand simulated data sets with $n = 100$ data points. The variables V_{jk} refer to the covariance between the j -th and k -th component of $\sqrt{n}\hat{\beta}_C(\hat{\lambda}_C)$ for $j = 0, 1, 2$.

estimator of the variance potentially can be very large.

6 Examples

We will now go through a couple of applications illustrating our theoretical framework.

6.1 Diabetes in Pima Indians

The data set `pima.indians.diabetes`, publicly available in the R-package `mlbench` contains information about female Pima Indians and prevalence of diabetes. Denote the data set by $z_1, \dots, z_n$ where $z_i = (y_i, x_i^T)^T$ with x_i being a vector of covariates and y_i a binary variable encoding whether or not the individual has diabetes for $i = 1, \dots, n$. We fit a penalized logistic regression model to these data. For some

pre-specified value of λ , this corresponds to maximising the following expression

$$\ell_n(\beta, \lambda) = \sum_{i=1}^n [y_i \log p(x_i, \beta) + (1 - y_i) \log \{1 - p(x_i, \beta)\}] - n\lambda \|\beta_{1:}\|^2 \quad (8)$$

where $p(x, \beta) = [1 + \exp\{-\beta_0 - \beta_{1:}^T x\}]^{-1}$ when $\beta = (\beta_0, \beta_{1:}^T)^T$. To fit this model, we need to decide upon a value for the tuning parameter λ . One option is to choose λ minimising the Brier score defined as $\text{BS} = n^{-1} \sum_{i=1}^n [y_i - p\{x_i, \hat{\beta}(\lambda)\}]^2$ where $\hat{\beta}(\lambda)$ is the minimizer of (8). This ensures that the model has good predictive ability. In most cases, however, BS will underestimate the true error. To mitigate this, we will instead choose λ minimising the cross-validated estimate of the Brier score, defined as $n^{-1} \sum_{i=1}^n [1 - p\{x_i, \hat{\beta}(\lambda)_{(-i)}\}]^2$.

With the procedure described above, we find that the minimizer of the cross-validated Brier score is equal to $\hat{\lambda} = 0.0085$. Fixing λ at this value, we can now fit a penalized logistic regression model to the data by minimising (8) evaluated at $\lambda = \hat{\lambda}$. Furthermore, by classic asymptotic likelihood theory, we know that for each fixed λ , $\hat{\beta}(\lambda)$ is consistent for some well-defined quantity $\beta(\lambda)$ and that $\sqrt{n}\{\hat{\beta}(\lambda) - \beta(\lambda)\}$ converges in law to a $N\{0, J(\lambda)^{-1}K(\lambda)J(\lambda)^{-1}\}$ distribution where $J(\lambda)$ and $K(\lambda)$ are well-defined matrices, see e.g. Van der Vaart (2000, Ch. 5). By considering λ as being fixed at $\hat{\lambda}$, we can therefore approximate the limiting distribution of $\hat{\beta}(\hat{\lambda})$ by a $N\{\beta(\hat{\lambda}), J(\hat{\lambda})^{-1}K(\hat{\lambda})J(\hat{\lambda})^{-1}/n\}$ distribution, and use it to make inference about $\beta(\lambda)$. The resulting estimated marginal densities of each component of $\hat{\beta}(\hat{\lambda})$ are shown in Figure 1. We remark that the very small value of $\hat{\lambda}$ is a consequence of the penalization term being scaled by n , leading the “actual” size of the penalization to be $n\hat{\lambda} \approx 3.342$.

The above approximate distribution is based on pointwise results which do not take the randomness introduced by tuning λ into account. This can, however, be done with the results of this paper. Assuming all regularity conditions are in place, we can apply the theory of this article to $\hat{\beta}(\hat{\lambda})$ by setting $\varphi(z, \beta, \lambda) = \nabla_{\beta}[y \log p(x, \beta) - (1 - y) \log \{1 - p(x, \beta)\} - \lambda \|\beta_{1:}\|^2]$ and $\psi(z, \beta) = \{1 - p(x, \beta)\}^2$. By Theorem 5,

$\sqrt{n}\{\hat{\beta}(\hat{\lambda}) - \beta_0\}$ converges in distribution to a central normal distribution with variance as described in Corollary 1 and which can be estimated using the approximations of Theorem 4. For each $j = 0, \dots, 8$, densities of the marginal approximated distributions of $\hat{\beta}(\hat{\lambda})_j$ together with the densities obtained using the classic, pointwise results as well as histograms based on $B = 2000$ bootstrap samples of $\hat{\beta}(\hat{\lambda})_j$ can be found in Figure 1 which was discussed briefly in the introduction.

From Figure 1 we note that when taking the added randomness of the tuning procedure into account, we get larger estimated variances than when using the classic pointwise results. This is promising, as the estimated variance of $\hat{\beta}(\hat{\lambda})_j$ was closer to the empirical variance in the bootstrap sample for all but one value of j . For $\hat{\beta}(\hat{\lambda})_2$, the classic variance estimator performed better than what we get from Theorem 4. A quick glance at Figure 1, however, reveals that neither of the approximate distributions seem to fit well with the bootstrap distribution in this case. In fact, the histogram of this component is quite skewed, giving reason to believe that convergence has not been achieved for this particular parameter.

Although the results above are slightly in the favour of the theorems derived in the present article, the limiting distributions we get using our and the classic method are not astoundingly different. It is, however, important to remember that the classic result is a pointwise result concerning the limiting behaviour of $\hat{\beta}(\lambda)$ when λ is a fixed number. Because of this, the classic result really does not apply to $\hat{\beta}(\hat{\lambda})$ and cannot be used to say anything about the limiting behaviour when $\hat{\lambda}$ is set by tuning, as neither asymptotic normality nor consistency of $\hat{\beta}(\hat{\lambda})$ follows from standard theory. By using the classic pointwise method regardless of this fact, we are therefore really taking a gamble, hoping that the added randomness of the tuning procedure will be negligible. Theorem 5 and Theorem 4, on the other hand, do take the tuning procedure into account and can safely be applied to make valid asymptotic inference.

6.2 Hybrid likelihood

In the book *The better angels of our nature*, Pinker (2011) argues that the world is changing for the better and that violence is on the decline. This claim quickly stirred up heated debate in the academic community and since its publication a large number of articles have either attempted to support or refute Pinker’s claim, see e.g. Cirillo and Taleb (2016a,b) or Cunen et al. (2020). Recently, Dæhlen and Hjort (2025) investigated whether conflicts have become less violent over time. We will now use the theory derived in the present article to improve the analysis done in that paper.

Dæhlen and Hjort (2025) uses data from the Correlates of War data base (Sarkees and Wayman (2010)) concerning the number of battle deaths in the 95 most recent and concluded inter-state wars. The analysis is done by first splitting the data into “older” and “newer” wars with the Korean war as the cut-off point. To each of these data sets, they fit shifted log-normal models and use the result to make inference about the difference in the median number of battle deaths for older and newer wars. They also repeat the analysis for the difference in upper quartiles.

Further, to fit the parametric models, the framework of hybrid likelihood is used. We will not go into details concerning this method, but roughly speaking, they fit parametric models f_θ by minimising a convex combination of the log-likelihood function and a nonparametric counterpart called the log-empirical likelihood function, see e.g. Owen (2001). How much weight is put on either of these components is decided by a parameter called the balance parameter a taking values in $[0, 1)$.

Moreover, the FIC is used to select values for the balance parameters a_1 and a_2 . When investigating change in the median number of battle deaths, the median is used as focus parameter and the empirical median as the consistent estimator $\hat{\mu}_0$. When studying the difference in upper quartile, the 0.75-quantile and the empirical 0.75-quantile are used in place of the median and empirical median. By standard results, the empirical p -quantile satisfies $o_p(1/\sqrt{n}) = n^{-1} \sum_{i=1}^n \xi(Y_i, \mu)$ where $\xi(y, \mu) = \mu - \mu_p - \{p - I(y \leq \mu_p)\}/f(\mu_p)$ and μ_p is the p -quantile in the underlying distribution

of the data. Furthermore, by the results of Dæhlen and Hjort (2025), the estimators $\hat{\theta}_1(a_1)$ and $\hat{\theta}_2(a_2)$ can both be expressed as components of Z-estimators for each fixed value of a_1 and a_2 . Hence, by making modifications as explained in Section 3, we can use Corollary 1 to derive the limiting distribution of $\hat{\theta}_1(\hat{a}_1)$ and $\hat{\theta}_2(\hat{a}_2)$.

By combining the above with the estimators given in Theorem 4, we can approximate the variance of $g\{\hat{\theta}(\hat{a}_1)\} - g\{\hat{\theta}(\hat{a}_2)\}$, where $g(\theta)$ is the median in a log-normal distribution parametrized by θ , in a way taking the added randomness introduced by tuning a_1 and a_2 into account. Using this method, we get an estimated standard deviation of 18431. That is a lot higher than 3815, which is obtained using the pointwise results in Dæhlen and Hjort (2025). This is a consequence of Dæhlen and Hjort (2025) not incorporating the added randomness introduced by tuning a_1 and a_2 , leading to a much too small estimated standard deviation.

For the difference in upper quartiles, $\hat{a}_1 = 0$ and $\hat{a}_2$ is chosen as the largest accepted value of a_2 (in this case 0.99). Because of this, $\text{FIC}'_j(\hat{a}_j) \neq 0$ for $j = 0, 1$ and Theorem 2 leads us to approximate the limiting distribution of $\hat{\theta}_j(\hat{a}_j)$ by the pointwise limit of $\hat{\theta}_j(\hat{a}_j)$ for $j = 0, 1$. This leads to the same approximate distribution as what is used in Dæhlen and Hjort (2025), and hence, all estimated standard deviations agree. As a result, ignoring the effect of tuning of a_1 and a_2 becomes unproblematic in this case.

7 Concluding remarks

We have studied the effect tuning procedures has on the limiting distribution of $\hat{\theta}(\hat{\lambda})$. We covered multiple ways of tuning, including minimization of information criteria, the training error and the cross-validated estimator of the risk. In addition, we have defined consistent estimators for the limiting variance of $\sqrt{n}\hat{\theta}(\hat{\lambda})$ and proved a result sharply characterizing the distance from CV to TE. Lastly, we went through a simulation setting and applied the theory on two real data sets.

Our work does have some limitations. The perhaps most pressing issue with our

current theory is the inability to handle non-smooth functions φ and ψ . This excludes multiple interesting settings. Since, ψ must be smoothly differentiable, tuning with respect to the absolute error or error rate is not covered by our theory. Furthermore, smoothness of φ excludes many interesting models, like the Lasso or other models penalized by an L_1 -norm. This does, of course, make our theory less applicable, but the smoothness assumptions are crucial for our proofs, and we believe that extending our theory to cover non-smooth settings is far from trivial.

Further, we have assumed that for each λ , $\hat{\theta}(\lambda)$ is a Z-estimator. This assumption is necessary for the proof of Theorem 5 and Theorem 6, but might be more strict than what is really needed at least for the latter theorem. By the informal arguments preceding Theorem 6, one would expect that Theorem 5 and Theorem 6 hold true as long as $\hat{\theta}(\lambda)$ has an influence function IF, such that $\hat{\theta}(\lambda) = \theta_0(\lambda) + \text{IF}(Y_i, \theta, \lambda) + \epsilon_n(\theta, \lambda)$ with “regular enough” remainder term. Formalizing this intuition in a proper way would certainly extend our framework, but will require proving a more general version of Lemma S1.3 in the supplementary material.

Finally, we have focused on Theorem 2 and Theorem 5 as our goal was to study the limiting behaviour of $\hat{\theta}(\hat{\lambda})$ when $\hat{\lambda}$ is set by some tuning procedure. These results are, however, corollaries of Theorem 1, which gives the limiting distribution of the full vector $(\hat{\theta}(\hat{\lambda})^T, \hat{\lambda}^T, \hat{\theta}'(\hat{\lambda})^T)^T$. In particular, Theorem 1 can be used to study the limiting behaviour of $\hat{\lambda}$. This can, in turn, be applied to make inference about λ and create approximate confidence intervals and hypothesis tests for the tuning parameter. For instance, it might be used to formally test if a regularization is beneficial or if either of the two components in a hybrid model is significantly preferable.

Supplementary Materials

Proofs, lemmas and further discussions can be found in the article’s supplementary material.

Acknowledgments

This work is partially funded by the Norwegian Research council through the BigInsight center for innovation driven research (237718). Partial support was also given from Centre for Advanced Study, at the Academy of Science and Letters, Oslo, in connection with the 2022-2023 project Stability and Change, led by H. Hegre and N.L. The authors are indebted to insightful input from our reviewers, which helped improve our manuscript.

References

- Arcones, M. A. (2005). Convergence of the optimal m-estimator over a parametric family of m-estimators. *Test* 14, 281–315.
- Bachoc, F., D. Preinerstorfer, and L. Steinberger (2020). Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics* 48(1), 440–463.
- Beirami, A., M. Razaviyayn, S. Shahrampour, and V. Tarokh (2017). On optimal generalizability in parametric learning. In *31st Conference on Neural Information Processing System*.
- Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *The Annals of Statistics* 41, 802–837.
- Billingsley, P. (1999). *Convergence of probability measures*. New York: John Wiley & Sons.
- Bouchard, G. and B. Triggs (2004). The tradeoff between generative and discriminative classifiers. In *16th IASC International Symposium on Computational Statistics (COMPSTAT'04)*, Prague, Czech Republic, pp. 721–728.
- Chetverikov, D., Z. Liao, and V. Chernozhukov (2021). On cross-validated lasso in high dimensions. *The Annals of Statistics* 49(3), 1300–1317.

- Cirillo, P. and N. N. Taleb (2016a). The decline of violent conflicts: What do the data really say? *Nobel Foundation Symposium 161: The Causes of Peace*, 1–26.
- Cirillo, P. and N. N. Taleb (2016b). On the statistical properties and tail risk of violent conflicts. *Physica A: Statistical Mechanics and its Applications* 452, 29–45.
- Claeskens, G. and N. L. Hjort (2003). The focused information criterion. *Journal of the American Statistical Association* 98(464), 900–916.
- Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*. UK: Cambridge University Press.
- Craven, P. and G. Wahba (1978). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische mathematik* 31(4), 377–403.
- Cunen, C., N. L. Hjort, and H. M. Nygård (2020). Statistical sightings of better angels: Analysing the distribution of battle-deaths in interstate conflict over time. *Journal of Peace Research* 57(2), 221–234.
- Dæhlen, I. and N. L. Hjort (2025). Model robust hybrid likelihood. *Journal of Statistical Planning and Inference* (241), 106327.
- Dæhlen, I., N. L. Hjort, and I. Hobæk Haff (2024). Accurate bias estimation with applications to focused model selection. *Scandinavian Journal of Statistics* 5.
- Dodge, Y. and J. Jurečková (2000). *Adaptive Regression*. Springer.
- Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association* 78(382), 316–331.
- Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *Journal of the American statistical Association* 81(394), 461–470.

- Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *Journal of the American Statistical Association* 99(467), 619–632.
- Germain, J.-F. and F. Roueff (2010). Weak convergence of the regularization path in penalized m-estimation. *Scandinavian Journal of Statistics* 37, 477–495.
- Hall, A. R. and A. Inoue (2003). The large sample behaviour of the generalized method of moments estimator in misspecified models. *Journal of Econometrics* 114(2), 361–394.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the econometric society* 50, 1029–1054.
- Homrighausen, D. and D. J. McDonald (2014). Leave-one-out cross-validation is risk consistent for lasso. *Mach. Learn.* 97, 65–78.
- Huber, P. J. (2009). *Robust statistics* (2nd ed.). Hoboken: Wiley.
- Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of multivariate Analysis* 94, 401–419.
- Jullum, M. and N. L. Hjort (2017). Parametric or nonparametric: The FIC approach. *Statistica Sinica* 27(3), 951–981.
- Kato, K. (2009). Asymptotics for argmin processes: Convexity arguments. *Journal of Multivariate Analysis* 100(8), 1816–1829.
- Ko, V. and N. L. Hjort (2019). Model robust inference with two-stage maximum likelihood estimation for copulas. *Journal of Multivariate Analysis* 171, 362–381.
- Konishi, S. and G. Kitagawa (2008). *Information criteria and statistical modeling*. New York: Springer Science & Business Media.
- Kuchibhotla, A. K., J. E. Kolassa, and T. A. Kuffner (2022). Post-selection inference. *Annual Review of Statistics and Its Application* 9, 505–527.

- Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* 44, 907–927.
- Li, K.-C. (1985). From stein’s unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics* 13, 1352–1377.
- Li, K.-C. (1986). Asymptotic optimality of CL and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics* 14, 1101–1112.
- Li, K.-C. (1987). Asymptotic optimality for C_p , CL, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics* 15, 958–975.
- Mu, B., T. Chen, and L. Ljung (2018). On asymptotic properties of hyperparameter estimators for kernel-based regularization methods. *Automatica* 94, 381–395.
- Mu, B., T. Chen, and L. Ljung (2021). On the asymptotic optimality of cross-validation based hyper-parameter estimators for regularized least squares regression problems. *arXiv preprint*.
- O’Neill, T. J. (1980). The general distribution of the error rate of a classification procedure with application to logistic regression discrimination. *Journal of the American Statistical Association* 75(369), 154–160.
- Owen, A. B. (2001). *Empirical likelihood*. Boca Raton, FL: CRC press.
- Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. Toronto: Viking.
- Sarkees, M. R. and F. Wayman (2010). *Resort to war: 1816 - 2007*. Washington DC: CQ Press.
- Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American statistical Association* 88(422), 486–494.

- Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica sinica* 7, 221–242.
- Shao, J. (2003). *Mathematical statistics*. New York: Springer Science & Business Media.
- Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics* 13, 970–983.
- Spokoiny, V. (2025). Sharp bounds in perturbed smooth optimization. *arXiv preprint arXiv:2505.02002*.
- Stephenson, W. and T. Broderick (2020). Approximate cross-validation in high dimensions with guarantees. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, Volume 108. PMLR.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and akaike’s criterion. *Journal of the Royal Statistical Society: Series B (Methodological)* 39(1), 44–47.
- Tibshirani, R. and K. Knight (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 61(3), 529–546.
- Van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge, England: Cambridge University Press.