Multivariate bias reduction in capacity expansion planning

Marie-Liesse Cauwet, Olivier Teytaud

Multivariate bias reduction in capacity expansion planning

M.-L. Cauwet
O. Teytaud
TAO (Inria), LRI, UMR 8623
(CNRS - Univ. Paris-Saclay),
France

***Abstract*—The optimization of capacities in large scale power systems is a stochastic problem, because the need for storage and connections (i.e. exchange capacities) varies a lot from one week to another (e.g. power generation is subject to the vagaries of wind) and from one winter to another (e.g. water inflows due to snow melting). It is usually tackled through sample average approximation, i.e. assuming that the system which is optimal on average over the last 40 years (corrected for climate change) is also approximately optimal in general. However, in many cases, data are high-dimensional; the sample complexity, i.e. the amount of data necessary for a relevant optimization of capacities, increases linearly with the number of parameters and can be scarcely available at the relevant scale. This leads to an underestimation of capacities. We suggest the use of bias correction in capacity estimation. The present paper investigates the importance of the bias phenomenon, and the efficiency of bias correction tools (jackknife, bootstrap; combined with possibly penalized cross-validation) including new ones (dimension reduction tools, margin method).**

I. Optimization of power systems

A. Unit commitment & long term planning

A power grid consists of a transmission network, a distribution network, loads and power plants. Optimizing this power system means optimizing a given cost function, the *objective function*, under *constraints*.

The objective function includes economical, social and environmental costs. These costs take into account risks of failure [1], maintenance, risks for workers. The constraints are operational constraints of power systems, including demand satisfaction, maximum ramping rate, stock management constraints, start up costs, environment constraints such as minimum water flows (see e.g. [2]). Demand satisfaction is handled by production-side and demand-side management.

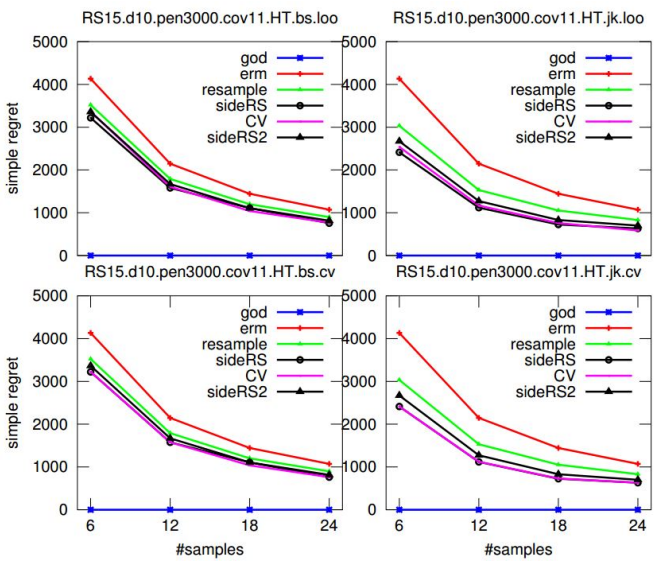
Solving this problem involves deciding which power plants are switched on/off; this is unit commitment [3], [4]. This also includes the dispatch, i.e. deciding the power output for each plant. This problem is multistage, stochastic and high dimensional. It is multistage due to coupling constraints between time steps, such as stock consistency and warm up costs. Stochasticity comes from the limited precision forecasts: e.g. demand [5] and inflows [6] are stochastic. Moreover, the recent years have seen an increase in production volatility, due to the raising use of renewable energies [7]. It is high dimensional also since large scale power systems deal with power grid at the continental scale.

Simulating and optimizing such power systems is crucial for testing the validity and cost of some scenarios. What are the costs of a purely renewable system ? Consider a limited budget (i.e. an upper bound on investments) over the next 50 years; what is the best investment planning ? What is the ecological/economical benefit, if we can relax the constraint of national independence ? What is the impact of a given gas supply cut-off / what is the best adaptation strategy to such a gas supply interruption ?

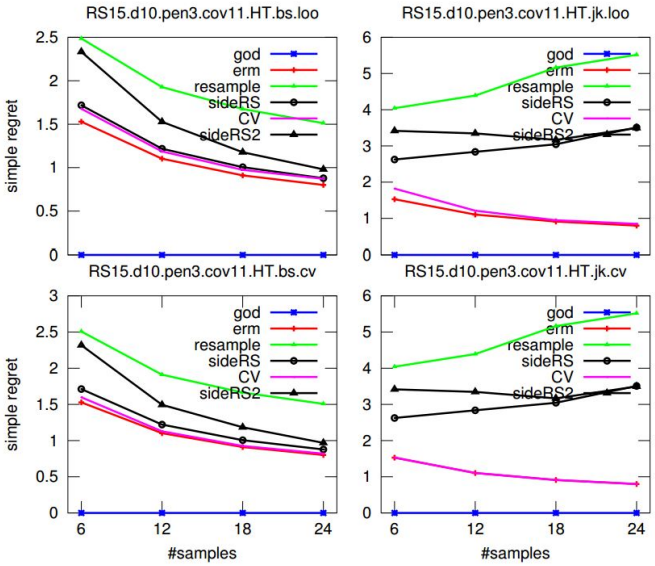
B. Optimization of power systems capacities

Quantifying the optimal connection capacities and storage capacities at the scale of a continent or more is an important optimization problem, with budget in dozens or hundreds of billions of euros. There are high level facts which are well known: in the European grid, conditions are better for wind power in the north, for solar power in the south, for additional hydroelectric storage in Scandinavia. Also, Africa is not that far, there are already connections between Europe and Africa [8], and increasing these connections is a possibility. This is especially relevant for the present work, since optimizing capacities implies optimizing against uncertainties [9]. Using optimization on average (or risk-aware variants, as well) over a probability distribution is a standard procedure for such a problem. The method heavily depends on random processes, modeling weather and consumption. Typically, histories are used, and the objective function is the average cost over this archive of possible weather scenarios [10]: such approaches are termed sample average approximation (Section II-A). This method is based on the assumption that the performance of a system can be reliably estimated by checking its performance on the finite set of available past data. They help for fine tuning power systems capacity expansion planning, because they use statistical effects, rather than hard $N-1$ constraints.

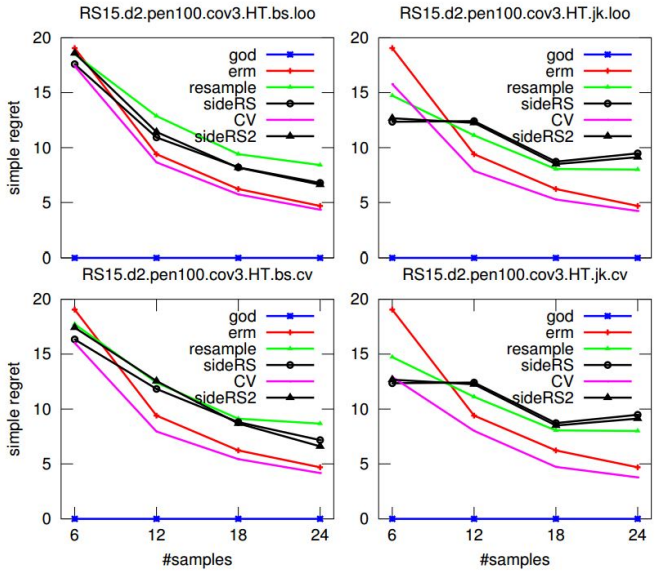
We discuss the shortcomings of this approximation (Section II-B) and propose alternate methods. Section III details some generic resampling tools, then model selection (Section IV) is considered in order to mitigate the instability of the resampling method. Section V presents experimental results.



(a) Dimension 10, $cov = 11$, $p = 3000$, 15 resamplings for the bias corrections, heavy tail. On this test case (i) L3 outperforms L1; (ii) jackknife outperforms bootstrap for bias-reduction; (iii) dimension reduction works well, in particular the absolute variant (Eq. 1).



(b) Dimension 10, $cov = 11$, $p = 3$, 15 resamplings for the bias correction, heavy tail. Compared to Fig. 1(a), the penalty p is small, which leads to a less skewed objective function, hence the bias is much smaller and ERM performs well.



(c) Dimension 2, $cov = 3$, $p = 100$, 15 resamplings for the bias reduction, heavy tail. In this case we see that model selection outperforms each of the models: this shows that the best estimate depends on the drawn sample, and that leave-three-out was able to "grasp" this effect.

For each subfigure: on the left (both top and bottom) **resample**, **sideRS** and **sideRS2** are performed with bootstrap; on the right (both top and bottom) **resample**, **sideRS** and **sideRS2** are performed with jackknife. On the top (both left and right) $L1$ model-selection; on the bottom, $L3$ model selection.

Figure 1. Simple regret in function of the sample size. The artificial problem is detailed in Section V-A.

II. M-ESTIMATORS AND BIAS

In all the paper, $\hat{\mathbb{E}}_{s\in S}k(s)$ is the average $\frac{1}{n}\sum_{i=1}^{n}k(s_i)$ if S is a sample $S=(s_1,\dots,s_n)$. $\hat{\mathbb{E}}_{N,\mathfrak{S}}k(s)$ is the empirical average $\frac{1}{N}\sum_{i=1}^{N}k(s_i)$ where $(s_1,\dots,s_N)$ is a random independently identically drawn sample from $\mathfrak{S}$. $\mathbb{E}_{\mathfrak{S}}$ denotes the expectation over the random process $\mathfrak{S}$. $|\Omega|$ denotes the cardinal of a set Ω.

A. Sample average approximation (SAA)

We denote by $f(s,x)$ the cost when choosing investment x and s is a realization of the random process $\mathfrak{S}$. We want to find x^* such that $\mathbb{E}_{\mathfrak{S}}f(s,x^*)$ is minimal. The *sample average approximation (SAA)* consists in tackle this optimization problem through the use of samples, as the random process is rarely available. We consider $\hat{x}(S)=\hat{x}$[1] minimizing:

$$x\mapsto\hat{\mathbb{E}}_{s\in S}f(s,x)=\frac{1}{n}\sum_{i=1}^{n}f(s_i,x),$$

with $S=(s_1,\dots,s_n)$ a sample of independent realizations of the random process $\mathfrak{S}$. Commonly S is an archive. This means that $\hat{x}$ is a M-estimator; it approximates a minimum over a finite sample. When it is obtained by minimizing an empirical estimate as above, it is termed an empirical risk minimizer (ERM) - but not all M-estimators are ERM. For power systems, we need detailed information, which is available for moderate values of n. n is typically between 5 and 100 depending on problems [11], [12]. 100 is optimistic as old data are less relevant due to climate change (though corrections are possible), erroneous measurements and missing values.

B. Bias & simple regret

Let us discuss the precision of the SAA in terms of quality of the obtained recommendation. The amount of data samples necessary to estimate properly the parameters of a system increases with the VC-dimension [13], which is linear in the number of parameters in smooth cases [14]. The amount of data requested for a given precision is termed the sample complexity. It also increases with the time constants of the problem; if the random processes are only approximately independent when they are 10 years apart from each other, the sample complexity might be multiplied by 10. The sample complexity also increases, typically, quadratically in the inverse precision. Hence, optimizing capacities (both generation capacities and network capacities) against a finite sample can lead to a *bias*. This bias is usually termed overfitting in machine learning [15], [16]. Typically, when SAA is applied, risks are underestimated, and therefore capacities dealing with uncertainties are underestimated, while uncertain assets are overestimated. SAA leads to invest too much in volatile production capacities, but not enough in network and storage capacities.

[1]When needed, the sample S on which the estimate is computed will be specified.

Let e be an estimate of a quantity x^*, depending on some stochastic random variable $\mathfrak{S}$. Then, the bias b of e is defined by:

$$b=\mathbb{E}_{\mathfrak{S}}e(\mathfrak{S})-x^*.$$

The stochastic random variable $\mathfrak{S}$ is the sample (i.e. S in Section II-A). In many cases, the bias of the ERM estimate $\hat{x}$ defined in Section II-A is significantly non zero, though it goes to zero asymptotically in the data size [14].

The concept of bias is less widely used in optimization, in particular in the field of large scale power systems, where the huge size of optimization problems makes deterministic optimization already quite hard. However it turns out that, in a renewable energy world, stochasticity really matters [7]. So this paper proposes to estimate the bias, and to take it into account in order to improve estimates of x^*.

We define a criterion that measures the quality of an estimate of x^*. The *Simple Regret* of an estimate e is:

$$SR_e=\mathbb{E}_{\mathfrak{S}}f(s,e)-\mathbb{E}_{\mathfrak{S}}f(s,x^*).$$

This is a random variable: the expectation operators operates on s with distribution $\mathfrak{S}$, and SR_e therefore depends only on the (possible) internal randomization of the estimator e. An estimate of the simple regret of estimator e will be denoted $\widehat{SR_e}$.

The ERM estimate $\hat{x}$ introduced in Section II-A is not necessarily the optimal one; the purpose of this paper is to propose some estimates with simple regret less than $SR_{\hat{x}}$. For this, we propose in Section III-A the use of bias correction tools, namely jackknife and bootstrap. In addition, we reduce the variance of these resampling estimators by a so called dimension reduction method in Section III-B.

III. BIAS REDUCTION

This section presents resampling estimates, i.e. tools for estimating the bias based on subsamples. Consider a sample $S=(s_1,\dots,s_n)$ of n realizations of a random process $\mathfrak{S}$. Resampling consists in splitting S into $\hat{S}$ and $\hat{S}'$, usually disjoint. Several such splits could be considered, leading to $\hat{S}_1,\dots,\hat{S}_N$, and their counterparts $\hat{S}'_1,\dots,\hat{S}'_N$.

A. Resampling estimates for bias reduction

Jackknife (JK) or Leave-One-Out (LOO). The jackknife resampling [17], also known as leave-one-out, uses the n subfamilies $\hat{S}_1,\dots,\hat{S}_n$ of cardinal $n-1$ defined by $\hat{S}_i=(s_1,\dots,s_{i-1},s_{i+1},\dots,s_n)$. The complementary family $\hat{S}'_i$ is $\hat{S}'_i=(s_i)=S\setminus\hat{S}_i$. When not all these S_i are required, we can consider a random sample. $\hat{S}$ is then randomly uniformly distributed among $\hat{S}_1,\dots,\hat{S}_n$ and $\hat{S}'$ is the complementary family. Let us consider $\hat{\hat{x}}=\hat{x}(\hat{S})$ and still $\hat{x}=\hat{x}(S)$. $\hat{\hat{x}}$ depends on $\hat{S}$; it is randomized. This means that in $\hat{\hat{x}}$, we consider the classical M-estimator, but applied to $\hat{S}$ instead of S. The proper bias-corrected estimator for jackknife $\hat{x}_{jk}$ is [17], [18]:

$$\hat{x}_{jk}=n\hat{x}-(n-1)\hat{\mathbb{E}}_{N,\hat{S}}\hat{\hat{x}}=n\hat{x}-(n-1)\left[\frac{1}{N}\sum_{j=1}^{N}\hat{x}(\hat{S}_j)\right],$$

where $\hat{S}$ is the random variable defined previously and $(\hat{S}_1, \ldots, \hat{S}_N)$ are N realizations of $\hat{S}$.

Bootstrap (BS). With n the cardinal of the sample S, bootstrap considers $\hat{S}$ a family of n points, randomly drawn in S, *with replacement*. Without replacement, this would not make any sense, as $\hat{S}$ would be equal to S; but with replacement, it is known that the difference between $\hat{S}$ and S can provide information on the difference between S and the original random process $\mathfrak{S}$ [18]. In bootstrap, $\hat{S}'$ can also be defined (it is the complementary family of $\hat{S}$ in S); however we do not need it in the present paper. In the case of bootstrap, the bias-corrected estimator is:

$$\hat{x}_{bs} = 2\hat{x} - \hat{\mathbb{E}}_{N,\hat{S}}\hat{\hat{x}} = 2\hat{x} - \left[\frac{1}{N}\sum_{j=1}^{N}\hat{x}(\hat{S}_j)\right],$$

where $\hat{S}$ is the random variable defined in the bootstrap resampling and $(\hat{S}_1, \ldots, \hat{S}_N)$ are N realizations of $\hat{S}$.

$\hat{x}_{jk}$ is usually a better estimate than $\hat{x}_{bs}$, though it is also sometimes mentioned that bootstrap is more versatile [19]. We will term these estimates *bias-corrected estimates*. However, the bias, after this correction, is not necessarily zero; it is just, in general, smaller than the bias of $\hat{x}$. The variance, on the other hand, is larger [20].

B. Dimension reduction for bias reduction

Let us consider $\hat{x}_r$ a bias-corrected estimate based on resamplings (r stands for resampling), either $\hat{x}_{bs}$ (bootstrap) or $\hat{x}_{jk}$ (jackknife). $\hat{x}_r$ has a smaller bias than the original estimate $\hat{x}$, but possibly a larger variance. In high-dimension, $\hat{x}_r$ might be very noisy, and, due to this, we might have $\mathbb{E}SR_{\hat{x}_r} > \mathbb{E}SR_{\hat{x}}$, where the expectation refers to the random sample S and to the internal randomization of the estimators, including the resampling step.

We define the *absolute dimension reduction* as follows:

$$\hat{x}'_r = \frac{\mu(\hat{x}_r)}{\mu(\hat{x})}\hat{x}, \tag{1}$$

with $\mu(v)$ the average of a vector v. This only makes sense if the different capacities have some sort of homogeneity: $(\hat{x})_1$,...,$(\hat{x})_d$ have the same unit and similar biases, where we denote by $(\hat{x})_i$ the i^{th} component of vector $\hat{x} \in \mathbb{R}^d$.

We also define the *relative dimension reduction* as follows:

$$\hat{x}''_r = \mu(\hat{x}_r/\hat{x})\hat{x}, \tag{2}$$

where $u = v/w$ denotes the componentwise division of vector v by vector w, i.e $(u)_i = (v)_i/(w)_i$ for $i \in \{1, 2, \ldots d\}$ if $v \in \mathbb{R}^d$ and $w \in \mathbb{R}^d$.

We will see in our experiments (Section V-C) that the absolute version works better than the relative one.

IV. Model selection (MS)

When several estimates are available, e.g. $\hat{x}_{A_1}(S), \hat{x}_{A_2}(S), \ldots, \hat{x}_{A_k}(S)$, it makes sense to "guess" which one is the best for the data at hand, in order to get a meta-estimate $\hat{x}_{meta}(S)$, which is, depending on some decision rule, equal to $\hat{x}_{A_{i^*}}(S)$ for a $i^* \in \{1, \ldots, k\}$. In our case, model selection can be used for determining which tool, between bias correction methodology and more classical tool such as ERM, should be preferred. We consider several variants for model selection: classical cross-validation (Section IV-A) and a recent modification of cross-validation, namely penalized CV (Section IV-B). Using these tools, we combine several estimates into a meta-estimate (Section IV-C). The margin method is then proposed for robustification purpose (Section IV-D).

As in Section III, we consider a sample $S = (s_1, \ldots, s_n)$ of n realizations of the random process $\mathfrak{S}$, a subfamily $\hat{S}$ of S and its complementary subfamily $\hat{S}' = S \setminus \hat{S}$.

A. Leave-k-out (Lk) for MS

Cross-validation in which the cardinal of $\hat{S}'$ is k is termed leave-k-out (Lk). Lk considers the random variable $\hat{S}$ uniformly distributed among the subfamilies of S of cardinal $n-k$. Leave-one-out is a special case of cross-validation, with $k = 1$.

For model selection, it is classical to use $\hat{S}'$ for testing the estimate built on $\hat{S}$. Given an estimate $e(S)$, depending on a sample S, we define the Lk cross-validation error estimate $\hat{\hat{f}}_{\mathrm{L}k}(e)$ by:

$$\begin{aligned}\hat{\hat{f}}_{\mathrm{L}k}(e) &= \hat{\mathbb{E}}_{\ell,(\hat{S},\hat{S}')}\hat{\mathbb{E}}_{s\in\hat{S}'}f(s, e(\hat{S})) \\ &= \frac{1}{\ell}\sum_{j=1}^{\ell}\frac{1}{|\hat{S}'_j|}\sum_{s\in\hat{S}'_j}f(s, e(\hat{S}_j)),\end{aligned}$$

This means that we randomly draw $\hat{S}$ and its complementary family $\hat{S}'$, ℓ times, and each time we build an estimate $e(\hat{S})$ which is tested on $\hat{S}'$. The average result is an estimate of $\mathbb{E}_{\mathfrak{S}}f(s, e(S))$.

For l large enough, increasing k in Lk reduces the variance of $\hat{\hat{f}}_{\mathrm{L}k}(e)$ as an estimate of the real loss $\mathbb{E}_{\mathfrak{S}}f(s, e(S))$, but increases the bias. Penalization (introduced in Section IV-B) is a tool for reducing the bias of cross-validation, with a moderate increase of the variance.

B. Penalized cross-validation for model selection (penk-F)

Cross-validation is classical; we present the penalized cross-validation method, which, interestingly, is a recent method necessary for making our tool effective in practice.

The cross-validation estimates $\hat{\hat{f}}_{\mathrm{L}k}(e)$ is biased since the training data set $\hat{S}$ is smaller than the real data set S. Penalized cross-validation [21] has been designed to counteract this effect. Informally, it consists in adding a penalization to the current estimated cost. The penalization pen$k$-F is built on S and $\hat{S}$. Given an estimate e, and a sample S, we define:

$$\hat{\hat{f}}_{\text{pen}k\text{-F}}(e) = \hat{\mathbb{E}}_{s\in S}f(s, e(S)) + C \cdot pen(S, e). \tag{3}$$

$$\begin{aligned}\text{with } pen(S, e) &= \hat{\mathbb{E}}_{\ell,\hat{S}}\left[\hat{\mathbb{E}}_{s\in S}f(s, e(\hat{S})) - \hat{\mathbb{E}}_{s\in\hat{S}}f(s, e(\hat{S}))\right], \\ &= \frac{1}{\ell}\sum_{j=1}^{\ell}\left[\frac{1}{|S|}\sum_{s\in S}f(s, e(\hat{S}_j)) - \frac{1}{|\hat{S}_j|}\sum_{s\in\hat{S}_j}f(s, e(\hat{S}_j))\right].\end{aligned}$$

$\hat{S}$ is a random variable as in the cross-validation section and C is an overpenalization constant. In other words, we randomly draw ℓ times a subsample (of size $n-k$) $\hat{S}$ from S, each time we build an estimate $e(\hat{S})$ which is tested both on S and $\hat{S}$. The average difference between these costs is the penalization, and the estimate of $\mathbb{E}_{\mathfrak{S}} f(s, e(S))$ is given by Equation 3; it is provably optimal in some simple cases [21].

C. Meta-estimate using model selection

Typically, we will define for example $\hat{x}_{meta,\mathrm{Lk}}(\hat{x}, \hat{x}_{jk})$

$$= \begin{cases} \hat{x} & \text{if } \hat{\hat{f}}_{\mathrm{Lk}}(\hat{x}) < \hat{\hat{f}}_{\mathrm{Lk}}(\hat{x}_{jk}) \text{ (i.e. } \widehat{SR_{\hat{x}}} < \widehat{SR_{\hat{x}_{jk}}}) \\ \hat{x}_{jk} & \text{otherwise} \end{cases}$$

We use the jackknife-corrected estimator if, for the Leave-k-out cross-validation, it is seemingly better than the simple M-estimator. More generally, given k estimators $\hat{x}_{A_1}, \ldots, \hat{x}_{A_k}$, and a model selection MS, $\hat{x}_{meta,MS}(\hat{x}_{A_1}, \ldots, \hat{x}_{A_k})$ is equal to the estimate $\hat{x}_{A_{i^*}}$ which is considered the best by the model-selection MS, i.e. such that $\hat{\hat{f}}_{MS}(\hat{x}_{A_{i^*}})$ is minimum.

D. The margin method

Let us consider the case in which we have k estimators $(\hat{x}_{A_1}, \ldots, \hat{x}_{A_k})$. Let us assume that A_1 is the default solution (ERM, in our case), that we wish to outperform with our new estimate. We use (penalized) cross-validation for selecting one of them. Let us call $\hat{x}_{meta}$ the resulting estimate, equal to $\hat{x}_{A_{i^*}}$, for a $i^* \in \{1, \ldots, k\}$, depending on the (penalized) cross-validation results. The result is satisfactory in most cases, but there are test cases in which $\hat{x}_{A_1}$ is better than $\hat{x}_{A_{i^*}}$ with $1 \neq i^*$, because the (penalized) cross-validation fails in finding the best among the k estimates. Then, we propose the following method, termed margin method: instead of comparing the estimated simple regrets $(\widehat{SR_{A_1}}, \ldots, \widehat{SR_{A_k}})$, of $(\hat{x}_{A_1}, \ldots, \hat{x}_{A_k})$ respectively, compare $((\mathbf{1}-\gamma)\widehat{SR_{A_1}}, \widehat{SR_{A_2}}, \ldots, \widehat{SR_{A_k}})$, for some $\gamma \in (0,1)$. Then, we expect the estimator $\hat{x}_{meta}$ to be more robust, in the sense that it is rarely worse than the original $\hat{x}_{A_1}$ - the $(k-1)$ others estimates $(\hat{x}_{A_2}, \ldots, \hat{x}_{A_k})$ are used only if the model selection considers $\hat{x}_{A_1}$ to be outperformed by far.

V. Experiments

Section V-A presents the experimentation framework, Section V-B lists the estimators considered and Section V-C presents the experimental results.

A. Test case

Let us consider an electric grid, connected to d distinct areas. An area $i \in \{1, \ldots, d\}$ is connected to the main grid only through one connection, with capacity $(x)_i$. The connection must be large enough so that the flow does not exceed the capacity, but larger connections are more expensive. Hence we should find a good compromise. The cost function, when the maximum consumption over the year is $s = ((s)_1, \ldots, (s)_d)$, for a non-negative $x = ((x)_1, \ldots, (x)_d)$, is

$$f(s,x) = p \times \left(\sum_{i=1}^{d} \mathbf{1}_{(s)_i > (x)_i} \right) + \sum_{i=1}^{d} (x)_i, \tag{4}$$

where

- p is a parameter: it is the penalty in case of fault, compared to the cost of 1 unit of network capacity.
- $(x)_i$ is the i^{th} network capacity, i.e. the capacity connecting area number i to the main grid.

Faults have long lasting consequence, far beyond the time during which the flow exceeds the capacities; hence the "binary" nature of the penalization. It is a common practice in power systems [11], [12] to consider the maximum over the year, and not the number of times or number of hours an overflow occurs. This is because overflow can lead to various problems, with an impact lasting long after the overflow itself, thus it does not make sense to consider that an overflow is less important just because it is short or occurred just once.

For this artificial experiment, the random process s is a discrete distribution (with support of cardinal 1500) generated as follows: cov standard centered Gaussian random variables are independently randomly drawn, in dimension d, with cov an integer. Let us define their covariance by V; hence, V is the identity if $cov \to \infty$ but might be far from identity when cov is small. s (resp. $\log(s)$ in the heavily-tailed case) is the d dimensional centered Gaussian with covariance V. The greater cov, the simpler the problem. Roughly speaking, cov large makes all areas $i \in \{1, \ldots, d\}$ more similar. Here, log refers to the logarithm with natural basis. $\log(x)$, when x is a d-dimensional vector, refers to $(\log((x)_1), \ldots, \log((x)_d))$.

B. Estimators

We compare the following estimators:

Name	bias correction	model selection	dim. reduction
$\hat{x}$	none	none	none
$\hat{x}_{bs}$	bootstrap	none	none
$\hat{x}'_{bs}$	bootstrap	none	absolute
$\hat{x}''_{bs}$	bootstrap	none	relative
$\hat{x}_{jk}$	jackknife	none	none
$\hat{x}'_{jk}$	jackknife	none	absolute
$\hat{x}''_{jk}$	jackknife	none	relative
$\hat{x}_{meta,L1}(\hat{x}, \hat{x}_{jk}, \hat{x}'_{jk})$	jackknife	L1	absolute
$\hat{x}_{meta,L1}(\hat{x}, \hat{x}_{bs}, \hat{x}'_{bs})$	bootstrap	L1	absolute
$\hat{x}_{meta,L2}(\hat{x}, \hat{x}_{jk}, \hat{x}'_{jk})$	jackknife	L2	absolute
$\hat{x}_{meta,L2}(\hat{x}, \hat{x}_{bs}, \hat{x}'_{bs})$	bootstrap	L2	absolute
$\hat{x}_{meta,L3}(\hat{x}, \hat{x}_{jk}, \hat{x}'_{jk})$	jackknife	L3	absolute
$\hat{x}_{meta,L3}(\hat{x}, \hat{x}_{bs}, \hat{x}'_{bs})$	bootstrap	L3	absolute

For the $meta$ versions (section IV-C), we also test the version with the penalization method (Section IV-B) and with the "margin" method (Section IV-D).

C. Experimental results

Results with Bias Reduction and Model Selection only. In this section, resamplings for model selection are based on $\#samples/k$ (see Fig. 1) random splits of the data into cross-validation folds of size k. Supplementary experimental results are presented in the extended material (www.lri.fr/~teytaud/rblong.pdf). We here provide a sample of results. Fig. 1

presents the simple regret (averaged over 200 independent runs) of the various estimators. **god** refers to the optimal solution; it has regret 0, by definition. **erm** denotes the classical M-estimator $\hat{x}$. **resample** refers to an estimate with bias reduction: $\hat{x}_{bs}$ (left) or $\hat{x}_{jk}$ (right). **sideRS** (resp. **sideRS2**) refers to $\hat{x}'_{bs}$ or $\hat{x}'_{jk}$ (resp. $\hat{x}''_{bs}$ or $\hat{x}''_{jk}$). **CV** refers to $\hat{x}_{meta,L1}$ (top) or $\hat{x}_{meta,L3}$ (bottom).

In Fig. 1(a) there is a strong bias in the M-estimator, due to the strong penalty (the loss function has a strong third derivative at the optimum). In this case, all tools work quite well: ERM with bias reduction outperforms vanilla ERM, leave-three-out succeeds in model selection, and dimension reduction (Eq. 1) improves the bias reduction. Then Fig. 1(b) presents a case with a small penalty; the situation is far less satisfactory, though leave-three-out successfully often selects the naive ERM estimator.

A detailed analysis shows that jackknife performs better than bootstrap for bias reduction when bias correction was already not that bad; on the other hand, it makes results much worse in some cases in which they were already poor. This is somehow consistent with the literature (see Section III-A). L3 performs better than L1 (see Section IV-A). The dimension reduction performs well in many cases. $\hat{x}'_{rs}$ outperforms the simple bias reduction $\hat{x}_{rs}$. So far, $\hat{x}''_{rs}$ is less efficient. This works even with cov small, i.e. inhomogeneous areas, as shown by Fig. 1(a). A small penalty $p = 3$ (Fig. 1(b)) makes it hard for any algorithm to outperform the simple $\hat{x}$ estimate.

Penalized CV & margin method. With the simple cross-validation, results are usually positive, but robustness is a main issue. The new method (i.e. bias reduction estimates) should never be significantly worse than the old one (i.e. ERM). Typically, we want to reduce the risk of something going wrong as in Fig. 1(b), where the best result is indeed the simple ERM. The penalized cross-validation (Section IV-B) seems to be a good candidate to perform a reliable selection among the estimates. Furthermore, to ensure that the bias reduction method is always better or equal to ERM, we propose a third ingredient in the selection method, so that it has a bias in favour of ERM in case of doubt - this is the "margin" methodology (Section IV-D). This will automatically disable the bias correction for problems which are less risk sensitive - i.e. for which the bias of ERM is lower.

Hence, three model selection methods are compared in the setting of Section V-A to choose among $\hat{x}$ (default ERM method), $\hat{x}_{jk}$ (ERM estimate corrected by jackknife), and $\hat{x}'_{jk}$ (ERM estimator corrected by jackknife with dimension reduction as described in Section III-B). These three selection method are (i) the classical CV (Section IV-A); (ii) the penalized cross-validation described in Section IV-B and (iii) the penalized-CV with margin described in Section IV-D.

There are 24 frameworks, combining dimension $d \in \{2, 3, 5, 10\}$; $cov = d+1$, $cov = 10(d+1)$, $cov = 3000(d+1)$; 15 or 150 resamplings for the bias reduction, where d is the number of capacities to be estimated. Each dot in Figures 2 corresponds to one of the 24 corresponding frameworks, with results averaged over the different sample sizes, namely 6, 12, 18 and 24 samples. Each figure corresponds to 1 (top), 4 (middle) or 16 (bottom) splits in the cross-validation associated to the "meta" part (model selection); and we distinguish $L1$ (left), $L2$ (middle) or $L3$ (right). Figure 2(a) displays results of the CV method versus the penalized-CV method. We use the default $C = 5/4 \times ((\#samples/k) - 1)$. Here $k \in \{1, 2, 3\}$ corresponds to the Lk considered. C is used as in Equation 3, overpenalization constant proposed in [21].

We see that the penalized cross-validation outperforms the standard cross-validation - but there are still cases in which the simple ERM is the best, in particular for intermediate values of the penalty, when it is difficult to know the best among ERM and the bias-corrected variants.

It ensues that the best method is the penalized-CV with 16 splits, $L3$, jackknife, with 10% margin. Additional results are displayed in Table I. These results indicate that ERM is the best for intermediate penalties. This fact is not surprising, these are the cases in which it is hard for CV methods to guess which estimator is the best. ERM is vastly outperformed in other cases (Fig. 2 and numbers in Table I).

VI. Conclusions

This paper is devoted to the bias correction in empirical risk minimizers, including the multivariate case. Many studies are dedicated to capacity expansion planning for power systems [22], [23], [11], [12], [9], [8]. To the best of our knowledge, bias correction has not been considered. Bias might be an overlooked serious issue in capacity estimation studies, which are a crucial part of power system optimization. We have investigated resamplings methods to reduce this bias.

In our experiments, jackknife performed better than bootstrap for correcting the bias. We improved the results thanks to a dimension reduction methodology. In high-dimensional cases, with homogeneous capacities to be estimated, averaging the bias correction over multiple capacities leads to a more efficient capacity correction than estimating each univariate correction alone. This technique provides an improved bias correction, and does not change the computational cost. The first variant of dimension reduction, termed absolute, was usually better in our experiments (Eq. 1, compared to Eq. 2).

For selecting estimators, penalized cross-validation outperformed the simple cross-validation. Furthermore, we developed the margin method for ensuring that the model selection

TABLE I. Performance of penalized-CV on various values of the penalty p in Eq. 4. The average normalized simple regret refers to the expected simple regret obtained by penalized-CV divided by the expected simple regret of the ERM. Results are averaged over all 24 experiments for each value of p (all possibilities with $d \in \{2, 3, 5, 10\}$, $cov \in \{(d+1), 10(d+1), 3000(d+1)\}$, 15 or 150 resamplings for the bias reduction). The standard deviations are at most 0.015.

penalty	0.1	1	3	10	30	100	3000
average normalized simple regret	.90	.98	1.03	1.04	.88	.66	.63

is almost always better than the sample average approximation method. Admittedly, this can reduce the average performance of the system; but it leads to the property that the meta-estimate is, in a stable manner, better or at least equal to the traditional estimate. We believe that such tricks are important for the acceptance of non-trivial statistical corrections.

Overall, statistical methods such as resampling can greatly increase the performance of capacity estimates - both for bias correction and for model selection. But there is a huge computational overhead. 100 samples for bias correction and 100 samples for model selection lead to a factor 10 000 on the computational cost. This is fortunately highly parallel, but the cost is far from being negligible.

Further work.

The bias correction methods we propose are adaptations, to optimization, of general principles. A mathematical analysis exists for these tools. On the other hand, the margin method and the dimension reduction methods are new. Dimension reduction methods need mathematical analysis; maybe there are better solutions than the two extreme cases (absolute, as in Eq. 1, and relative, as in Eq. 2), for instance by considering groups of related capacities. Mathematical analysis might help for understanding the bias/variance compromise in multivariate bias reduction of M-estimators (Eq. 1). The constant C in Eq. 3 is suggested only in a specific setting [21]; we did not try any optimization of this constant, so that our results are principled, but improvements might be possible.

Considering years, in our archive of data, as independent, is an approximation. This is a reasonable assumption for some parts of the world but not for others: studying the impact of this lack of independence is another important further work [24].

Additional experiments are part of the agenda, including high-dimensional cases with hundreds of capacities.

Acknowledgements.

We are grateful to the Ademe Post project for making this work possible (post.artelys.com).

REFERENCES

[1] C. Autorita per l'Energia Elettrica e il Gas, "Report on the events of september 28th, 2003 culminating in the separation of the italian power system from the other UCTE networks," 2004, accessed: 2010-09-09.

[2] D. Bertsimas, E. Litvinov, X. A. Sun, J. Zhao, and T. Zheng, "Adaptive robust optimization for the security constrained unit commitment problem," *Power Systems, IEEE Transactions on*, vol. 28, no. 1, pp. 52–63, 2013.

[3] N. Padhy, "Unit commitment-a bibliographical survey," *Power Systems, IEEE Transactions on*, vol. 19, no. 2, pp. 1196–1205, May 2004.

[4] G. Sheble and G. Fahd, "Unit commitment literature synopsis," *Power Systems, IEEE Transactions on*, vol. 9, no. 1, pp. 128–135, Feb 1994.

[5] RTE-ft, "Rte forecast team: Electricity consumption in France : Characteristics and forecast method," 2008.

[6] T. G. Siqueira, M. Zambelli, M. Cicogna, M. Andrade, and S. Soares, "Stochastic dynamic programming for long term hydrothermal scheduling considering different streamflow models," in *Probabilistic Methods Applied to Power Systems, 2006. PMAPS 2006. International Conference on*, June 2006, pp. 1–6.

[7] P. Pinson, "Renewable energy forecasts ought to be probabilistic," 2013, wIPFOR seminar, EDF.

[8] SYSTINT Workgroup, *European, CIS and Mediterranean Interconnection: State of Play 2006.*, Ucte–Eurelectric, 2007.

[9] S. Vassena, P. Mack, P. Rousseaux, C. Druet, and L. Wehenkel, "A probabilistic approach to power system network planning under uncertainties," in *IEEE Bologna Power Tech Conference Proceedings*, 2003.

[10] A. Shapiro, "Analysis of stochastic dual dynamic programming method," *European Journal of Operational Research*, vol. 209, no. 1, pp. 63–72, 2011.

[11] NERC, "Special reliability assessment: accomodating an increased dependence on natural gas for electric power phase ii: a vulnerability and scenario assessment for the north american bulk power system," North American Electric Reliability Corporation, Tech. Rep., 2013.

[12] A. Whitmarsh, S. Bojanowski, and W. Barber, "Gas security of supply report," Ofgem, Tech. Rep., 2012.

[13] V. N. Vapnik, *The Nature of Statistical Learning*. Springer Verlag, 1995.

[14] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic Theory of Pattern Recognition*. Springer, 1997.

[15] A. V. D. Vaart and J. Wellner, *Weak Convergence and Empirical Processes*. Springer series in statistics, 1996.

[16] V. Vapnik and A. Chervonenkis, "On the uniform convergence of frequencies of occurence events to their probabilities," *Soviet Mathematics-Doklady 9, 915-918*, 1968.

[17] M. H. Quenouille, "Approximate tests of correlation in time-series," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 11, no. 1, pp. 68–84, 1949. [Online]. Available: http://www.jstor.org/stable/2983696

[18] B. Efron, "The jackknife, the bootstrap, and other resampling plans," *CBMS-NSF Regional Conf. Series in Applied Mathematics*, vol. 38, 1982.

[19] J. Wellner, *Bootstrap and Jackknife Estimation of Sampling Distributions*, 2014.

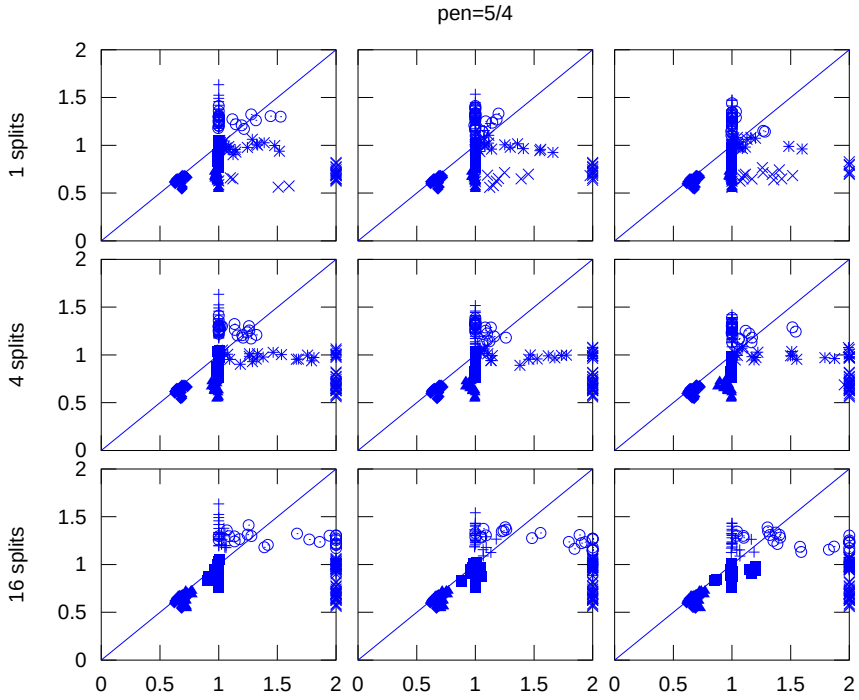
[20] F. Scholz, "The bootstrap small sample properties," Tech. Rep., 2007.

[21] S. Arlot, "V-fold cross-validation improved: V-fold penalization," Feb. 2008, 40 pages, plus a separate technical appendix. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00239182

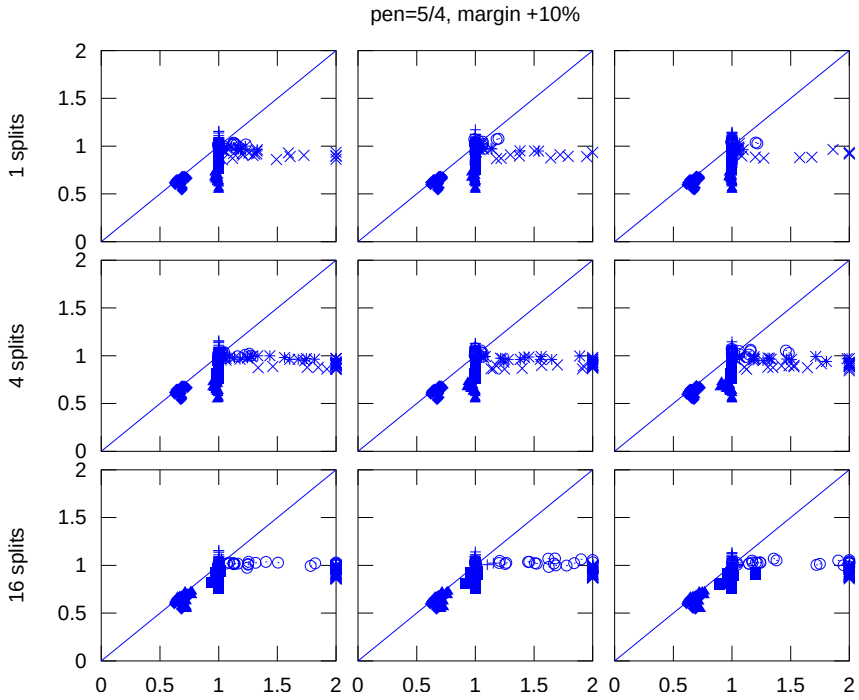
[22] P. Saisirirat, N. Chollacoop, M. Tongroon, Y. Laoonual, and J. Pongthanaisawan, "Scenario analysis of electric vehicle technology penetration in thailand: Comparisons of required electricity with power development plan and projections of fossil fuel and greenhouse gas reduction," *Energy Procedia*, vol. 34, no. 0, pp. 459 – 470, 2013, 10th Eco-Energy and Materials Science and Engineering Symposium. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1876610213010175

[23] M. Chaudry, N. Jenkins, M. Qadrdan, and J. Wu, "Combined gas and electricity network expansion planning," vol. 113, no. C, pp. 1171–1187, 2014. [Online]. Available: http://EconPapers.repec.org/RePEc:eee:appene:v:113:y:2014:i:c:p:1171-1187

[24] B. Yu, "Rates of convergence for empirical processes of stationary mixing sequences," *Ann. Probab.*, vol. 22, no. 1, pp. 94–116, 01 1994. [Online]. Available: http://dx.doi.org/10.1214/aop/1176988849



(a) Experiments on penalized cross-validation. X-axis: SR obtained by the CV method divided by the SR of the naive ERM. Y-axis: SR obtained by the penalized-CV method divided by the SR of the naive ERM. We see that difficult cases (Y-axis above 1) are $\circ$ and $+$, namely penalty 3 and 10 respectively: the classical method sometimes more than doubles the SR (X-axis is limited at 2). With the penalized CV, the worst cases are around 1.5. The fact that problems occur around these values is reasonable: they are the cases in which ERM and jackknife have comparable performance, so that CV might make a bad choice.



(b) Similar to Figure 2(a), but the CV gives a 10% bonus (i.e. "margin" method) to ERM (i.e $\gamma = 0.1$). We see that the obtained CV method (Y-axis) is almost always < 1 (hence beneficial), though there are still a few cases with SR more than in the ERM case. The margin is applied in both cases (CV, and penalized CV.)

Figure 2. Each dot corresponds to one test case, averaged over the different sample sizes (6, 12, 18, 24). The markers $\times$, $*$, $\circ$, $+$, $\square$, $\hat{}$, $\Diamond$, stand for penalties $p = 0.1, 1, 3, 10, 30, 100, 3000$ respectively.