Reviewer Comments: Modifications in the manuscript appear in blue.

The authors would like to thank the associate editor for handling the paper and collecting the reviews. We have tried our best to revise the manuscript to accommodate all the reviewer comments. Following the suggestions of the reviewers and the comments of the AE we have made the following major changes:

- Based on the comment from Reviewer 2; we have modified the title.
- The extension to the general case (different means and different covariance matrices) has been added.
- Theorem 2 has been revised and modified according to the comments of the reviewers.
- Remarks for clarification of the matrix decomposition and the trade-offs of the proposed methods have been added.

Reviewer: 1

Recommendation: AQ - Publish With Minor, Required Changes

Comments: In this work the authors study sparse sampler design for Gaussian signal detection with correlated data using submodularity to facilitate efficient computations with guaranteed performance in some cases. I find their writing style and arguments clear and insightful, well motivated, and I believe this paper can be of interest to the wider signal processing community. We thank the reviewer for the positive comments on our paper.

Please find my comments to this manuscript below:

Q.1. Theorem 2. It seems that for a nearly diagonal covariance matrix $\mathbf{\Sigma}$ with the S matrix correspondingly becoming vanishingly small the constant C1 becomes very large and the bound on ϵ diverges. Can you comment on this, as it seems that in this case the SNR function should approach a modular set function with ϵ=0

R.1 Thanks for pointing this out. To make this more clear, following your recommendation and the ones from the other reviewers, Theorem 2 has been modified. We emphasize more the role of the Theorem in our work, since it explains why and when using the greedy heuristic directly on the SNR provides good results. However, as discussed in the manuscript, there might be instances where this bound does not provide much information on using the greedy heuristic. Therefore we introduce the submodular surrogate. In addition, as the SNR can be computed as a "by product" of our proposed surrogate, we can always select the one that performs the best.

Q.2. Eq. (29) suggests that maximizing the surrogate $f(\mathcal{A})$ is equivalent to maximizing $s(\mathcal{A})\gamma(\mathcal{A})$ where $s(\mathcal{A})$ is the SNR. I find the study of the term $\gamma(\mathcal{A})$ somewhat lacking depth, how do we know that $\gamma(\mathcal{A})$ does not behave in a completely erratic way or opposite to $s(\mathcal{A})$ (as the authors suggest). A discussion would be well warranted.

R.2 Indeed, that's true. In fact, $f(\mathcal{A})$ is defined as $\log\{s(\mathcal{A}\}\gamma(\mathcal{A})\}$ [cf. (28) and (30)], which due to the monotone non-decreasing property of the log function makes it equivalent to maximizing $s(\mathcal{A})\gamma(\mathcal{A})$. Despite this, in general, the only thing that can be said with certainty about $\gamma(\mathcal{A})$ is that this set function is monotone nondecreasing in the set size (therefore, does not behave opposite to $s(\mathcal{A})$), and that $f(\mathcal{A}) \neq s(\mathcal{A})$. Depending on which sensor is being selected, the projection of the standard basis, i.e., $\mathbf{e}_i = [0\,0\,\ldots\,1\,\ldots\,0\,0]^T$, onto the matrix $\mathbf{S}$ will define how it behaves. However, the optimization will try to align the selection to $\mathbf{S}$ in order to make the matrix $[\mathbf{S}^{-1} + a^{-1}\text{diag}(\mathbf{I}_{\mathcal{A}})]$ more positive definite and to reduce the SNR loss. That's why we treat $f(\mathcal{A})$ as a surrogate function. In addition, we have made the following modification in the

manuscript:

> *The set function $\gamma(\mathcal{A})$ is a monotone nondecreasing set function of the selected set size. In addition, rewriting the SNR expression as*
>
> $$s(\mathcal{A}) = \boldsymbol{\theta}^T \boldsymbol{S}^{-1}\boldsymbol{\theta} - \frac{1}{\gamma(\mathcal{A})}\boldsymbol{\theta}^T \boldsymbol{S}^{-1}\text{adj}(\boldsymbol{S}^{-1} + a^{-1}diag(\boldsymbol{1}_{\mathcal{A}}))\boldsymbol{S}^{-1}\boldsymbol{\theta}, \tag{31}$$
>
> *where $adj(\boldsymbol{A})$ is the adjugate of $\boldsymbol{A}$ defined as the transpose of the cofactor matrix of $\boldsymbol{A}$, we observe that in order to keep the nondecreasing property of the SNR with respect to the set $\mathcal{A}$, the growth rate of $\gamma(\mathcal{A})$ should be larger than growth rate of the quadratic form in (31). Hence, it is reasonable to consider (30) as a surrogate function for $s(\mathcal{A})$. Finally, we remark that maximizing (30) effectively maximizes a modified version of (25) where the inverse of $\boldsymbol{S}^{-1} + a^{-1}diag(\boldsymbol{1}_{\mathcal{A}})$ has been substituted by its adjugate [cf.(31)].*

Here, we reemphasize that the last sentence of the above extract, makes reference to the fact that in the expression:

$$s(\mathcal{A}) = \boldsymbol{\theta}^T \mathbf{S}^{-1}\boldsymbol{\theta} - \boldsymbol{\theta}^T \mathbf{S}^{-1}\left[\mathbf{S}^{-1} + a^{-1}\text{diag}(\mathbf{1}_{\mathcal{A}})\right]^{-1}\mathbf{S}^{-1}\boldsymbol{\theta},$$

the second term, which we refer to as loss in-signal-to-noise ratio, is monotone decreasing in the cardinality of the set $\mathcal{A}$, and relates to the $\gamma(\mathcal{A})$ function through the expression

$$s(\mathcal{A}) = \boldsymbol{\theta}^T \mathbf{S}^{-1}\boldsymbol{\theta} - \frac{1}{\gamma(\mathcal{A})}\boldsymbol{\theta}^T \mathbf{S}^{-1}\text{adj}(\mathbf{S}^{-1} + a^{-1}\text{diag}(\mathbf{1}_{\mathcal{A}}))\mathbf{S}^{-1}\boldsymbol{\theta},$$

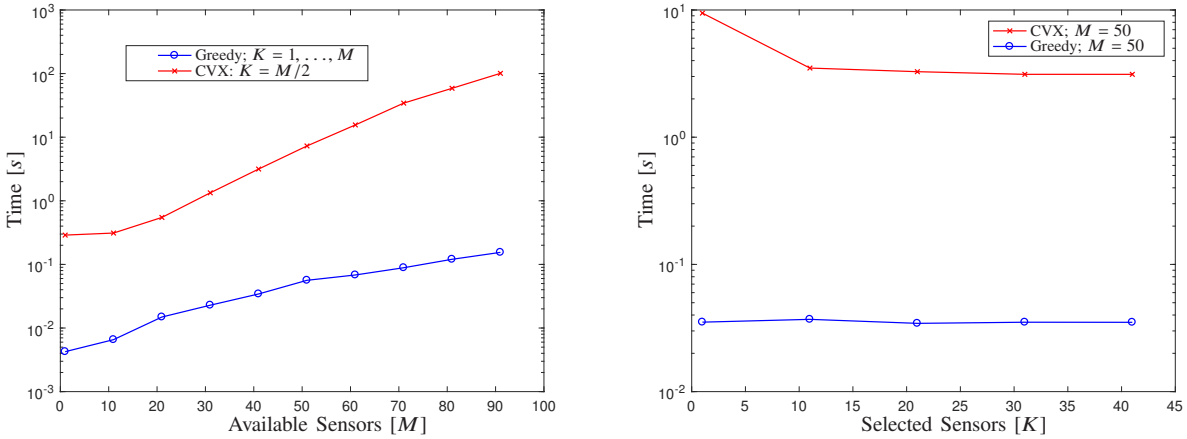
where for some matrix $\boldsymbol{A}$, $\text{adj}(\boldsymbol{A}) = \boldsymbol{C}^{\text{T}}$ denotes the adjugate matrix, and $\boldsymbol{C}$ is the cofactor matrix of $\boldsymbol{A}$. As the quadratic form in the above expression is nondecreasing on $\mathcal{A}$, the rate of increase of $\gamma(\mathcal{A})$, with respect to $\mathcal{A}$, should be larger in order to keep the monotonicity of the signal-to-noise ratio.

Q.3. Computational complexity analysis following Equation (35) , the authors find the computational complexity of their approach as $O(MK^3)$ compared to cubic complexity in M for the convex approach. However, as we have $K >> M$ it seems like the computational complexity of this approach is in fact larger as there is no fixed number of possible choices K in the convex formulation. Can the authors resolve this?

R.3 Thanks for this comment. We reemphasize that, in this work on sensor selection, the focus is on designing compression matrices. Therefore, we always have $K << M$.

Q.4. Section 4D, can the authors report algorithm running time when comparing their approach to the convex optimization to support their claim of improved running time?

R.4 If we consider instances where M is large the SDP formulation chokes (loading the whole SDP is already intensive), while the recursive formulation when selecting a small subset of sensors is still reasonable. In the following figure we show a small example to illustrate this. In the left figure, the size of the ground set changes, i.e., the number of available sensors M, and in the right figure the number of available sensors is kept fix but the number of selected sensors (K) increases. In the first example (left), the greedy method computes the solution for all possible selected sensors, i.e., $K = 1, \ldots, M$, while the convex method only solves it for $K = M/2$. In the second example (right), the time reported for the greedy method is for obtaining the total set of solutions for $K = 1, \ldots, M$.



In case the reviewer deems it necessary to include these pictures, we can either add these plots or report numbers from these plots in the final version of the manuscript.

Q.5. It would be helpful if the authors can calculate and discuss the sumodularity bound ϵ as determined according to their theorem 2 for all the specific settings for which they report numerical results.

R.5 Although computing this bound, despite being computationally intensive, might be useful for the example provided, we feel that plotting it doesn't add much value to the paper. The bound is used only to substantiate why greedy SNR works so well in some cases.

Q.6. Can the authors note on the possible extension of their method for applications where both means and covariance matrices are different for the two hypotheses?

R.6 Thanks for the recommendation. Indeed, an extension is possible and as shown in the added material in the manuscript, it reuses the solvers and methods proposed for the two particular cases as explained in Section V.B. Here the extract of the manuscript:

A. Uncommon Means and Uncommon Covariances

So far, only particular cases of the general hypothesis testing problem between two Gaussian distributions have been discussed. However, in the following, we show that the general case, i.e., distinct means and covariance matrices, can be solved using the same heuristics as discussed for the particular cases presented earlier in the previous sections.

To show that we can reuse the same machinery as in the uncommon covariances case, first let us consider the KL divergence:

$$\begin{aligned}\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) &= \tfrac{1}{2}\Big(\operatorname{tr}(\boldsymbol{\Sigma}_{0,\mathcal{A}}^{-1}\tilde{\boldsymbol{\Sigma}}_{1,\mathcal{A}}) - M + \log\det(\boldsymbol{\Sigma}_{0,\mathcal{A}}) \\ &\qquad - \log\det(\boldsymbol{\Sigma}_{1,\mathcal{A}})\Big).\end{aligned} \tag{43}$$

In the above expression we have rewritten the quadratic form in terms of a trace [cf. Table 1], and defined the matrix $\tilde{\boldsymbol{\Sigma}}_{1,\mathcal{A}} := \boldsymbol{\Sigma}_{1,\mathcal{A}} + (\boldsymbol{\theta}_{1,\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})(\boldsymbol{\theta}_{1,\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})^T$ which remains symmetric. From (43) it can be seen that the decomposition proposed before can directly be used by just replacing $\boldsymbol{\Sigma}_{1,\mathcal{A}}$ by $\tilde{\boldsymbol{\Sigma}}_{1,\mathcal{A}}$ in the $q_{\text{sub}}(\cdot,\cdot)$ set function. Therefore, for this measure there is no distinction between these two cases. In addition, as the J-divergence is a sum of KL divergences, the decomposition of the J-divergence follows immediately.

For the Bhattacharyya distance, the quadratic form that appears in the expression for the general case, i.e., the term related to the distances between the means, needs to be added to the decomposition given in (40). As this new term is not submodular, the surrogate function proposed for the uncommon means case can be directly employed to provide a submodular set function for the decomposition. As a result, the surrogate distance measure for the Bhattacharyya distance can be given through the decomposition:

$$\begin{aligned}\mathcal{B}_{\text{sub}}(\mathcal{H}_1\|\mathcal{H}_0) &:= g(\mathcal{A}) - h(\mathcal{A}); \\ g(\mathcal{A}) &= \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{\mathcal{A}}) + \frac{1}{8}\log\det(\mathbf{M}_{\mathcal{A}}); \\ h(\mathcal{A}) &= \frac{1}{4}(\log\det(\boldsymbol{\Sigma}_{0,\mathcal{A}}) + \log\det(\boldsymbol{\Sigma}_{1,\mathcal{A}})),\end{aligned} \tag{1}$$

which can be optimized using the approach discussed for the uncommon covariance case.

In summary, the greedy heuristic for the most general case of uncommon means and uncommon covariances can be develop using straightforward adaptations of the methods presented in Sec.V.A.

Typos:

Introduction: levering → leveraging

Beginning of Section III: use → used

Section IIIA: 0.63% → 63%

Immediately after (37): the last power in the topelitz matrix is $M/2-1$

R All the typos have been corrected. Thank you for proof reading this manuscript.

Reviewer: 2

Recommendation: RQ - Review Again After Major Changes

Comments: The paper deals with the very relevant problem of selecting sensors for Gaussian hypothesis testing with correlated data. Aiming to use greedy methods, the paper focuses on divergence surrogates for the probability of error. It then provides an additive approximation guarantee for SNR maximization in the case where the means are different and derives difference of submodular functions approximations for the divergences when testing signals with different covariances. Despite minor typos and clarity issues (especially in the technical derivations), the writing is clear and easy to follow.

Thank you for finding the writing clear and easy to follow.

Major issues

Q: The approximation guarantee for the SNR based on ϵ-submodularity is additive. As noted in the paper, the guarantee therefore depends on ϵ being small compared to the optimal value. However, it is not completely clear when Theorem 2 yields good guarantees. The text notes that it works better for well-conditioned Σ (small κ), i.e., when the data are not too correlated. This case, however, could be addressed before by from assuming the data are white (the mismatch in this situation is small). As the measurements become more correlated, the condition number increases and the guarantee worsens. A quick computation for the covariance (35) used in the simulations, for example, gives ϵ in the order of thousands, whereas $f(\mathcal{A}_{\text{opt}})$ is most likely on the order of tens. I would encourage a more extensive discussion of the results and plots with the value of the bound or the resulting guarantee for different K. As it is, one cannot judge the usefulness of the derived near-optimal result.

R: Thank you for raising this concern. We would like to emphasize that Theorem 2 has been modified to further improve the clarity. The provided bound is an informative upper bound on the ϵ constant. It characterizes the behaviour of the "loss" of submodularity of the set function in terms of the condition number of the correlation matrix. The introduced bound in Theorem 2 demonstrates why and when performing greedy selection over the SNR function works well. However, as remarked in the manuscript, ill-conditioned matrices that have an appropriate structure, e.g., diagonal matrices, can even lead to modular functions. The added text to the manuscript addressing this comment is the following:

Note that the characterization provided in Theorem 2, in terms of condition number of the matrix, excludes diagonal matrices. This is due to the fact that even if a diagonal

matrix is ill-conditioned, the resulting set function is still a modular set function. Therefore, here we focus on matrices that model well correlated errors.

Therefore, as discussed in the manuscript, we emphasize that performing greedy over the SNR function is not always a good option. However, the cases that lead to good results are mostly likely due to the insights provided by our bound.

Q: Algorithm 2 used for the different covariances case has no performance guarantee. Results in [42] only show that each step improves the performance and that it converges to a local optimum. However, it gives no approximation certificate. This should be made clear in the text.

R: Thanks for the comment. We have added the following in the text:

In addition, despite the fact that the optimization of the difference of submodular set functions is inapproximable [42], similar to CCP, the SupSub procedure is guaranteed to reach a local optimum of the set function when the procedure converges [42]

We added this since convergence only to a local optimum is guaranteed. So indeed we cannot guarantee near optimality in this case. Therefore, we have also modified the claim and the title accordingly.

Q: The cost function obtained from (26) appears to display a trade-off between numerical stability and accuracy. If $a \approx \lambda_{min}$, then S is almost rank deficient and M is very ill-conditioned. Moreover, S and Σ are no longer similar and the lower diagonal block of M is therefore not good approximation of the SNR. On the other hand, if $\beta \approx 0$, then $S \approx \Sigma$ and the cost function approximates the SNR well, but $1/a \to \infty$. Again M is ill-conditioned. These numerical issues should show up more in larger problems, so it would be important to show numerical results with larger ground sets (e.g., 1000). Whether this is a serious limitation or not should be noted in the text.

R: We would like to stress that, for a given $\mathbf{\Sigma}$, a can be chosen appropriately (i.e, choice of a is completely controllable, as long as $0 < a < \lambda_{\min}\{\mathbf{\Sigma}\}$). But as the reviewer observes, it is better to select a not too close to 0 or $\lambda_{\min}$ as this will lead to an ill-conditioned $\mathbf{M}$. In addition, the following remark has been added in the manuscript:

Remark 1. *We stress that for some choices of the parameter a, the matrix $\boldsymbol{M}_{\mathcal{A}}$ might be ill-conditioned. However, as the computation of the cost function is not performed directly on the matrix $\boldsymbol{M}_{\mathcal{A}}$ but through (35), we only required that the recursive*

inversion of the matrix in expression (35) is numerically well-conditioned. This can be achieved in practice by selecting the value for a far from both 0 and $\lambda_{\min}\{\mathbf{\Sigma}\}$, e.g., $a = 0.5\lambda_{\min}\{\mathbf{\Sigma}\}$. This approach avoids numerical problems that could arise due to the selection of the value of a.

Q: In (30) and forward, $s(\mathcal{A})$ is not the SNR unless $a = 0$. It is important to make this clear in the text and maybe use a different symbol to avoid confusion. Although this is not seriously detrimental to the contribution of the work, it appears in comments throughout the text.

R: There seems to be a misunderstanding. $s(\mathcal{A})$ is always the SNR, even for $a \neq 0$. The decomposition holds for any value of a which leads to an invertible matrix S. For any such value for a, (24) and (20) are equivalent. Only the final submodular surrogate will depend on the selected a.

Minor comments

Q: p.2, l. 50: "non-monotone separable constraints". Is this the case in this paper? It doesn't look like it, but I may be missing something. I cannot see why this is mentioned here at all.

R: Thanks for your comment. You are right, we don't make use of this, Therefore, we have removed this from the manuscript.

Q: The work repeatedly refers to the complexity of greedy as "linear in the size of the input set". The number of queries required by greedy is KM. It therefore depends polynomially on both the ground set and the number of sensors selected. Also, this is the query complexity of the algorithm. The complexity of each query is also a factor to be considered (which is the justification given for Section IV-C).

R: Thanks, for this comment. The distinction between "the number of cost function evaluations scales linearly" and "linear time complexity" has been clarified in the manuscript.

Q: p.3, l. 40, 2nd column: "the methods presented in this work can be easily extended to budget functions [...]" I understand this from an algorithmic point of view, but then what happens to the guarantees? What about complexity? Matroids can be hard to express efficiently.

R: Thanks for this comment. We have added a clarification as

However, by allowing an increase in computational complexity and degradation of the near-optimality guarantees, the methods presented in this work can be extended to budget functions, expressed as constraints, representable by other kinds of matroids [28]

Q: Definition 2: should it be $\epsilon > 0$?

R: We have changed this in the revised manuscript.

- Theorem 2: β appears to be an artifact of the proof. If so, why not minimize over β for the final result? Moreover, as it is, the statement and (22) are confusing, since a and C_1 depend on both β and λ_{min}. In fact, C_1 is not a universal constant. I suggest writing out the result in a more direct form.

R: That's a good observation. The proof and result have been restructured to consider only universal constants. β is introduced to state that a is a fractional part of $\lambda_{\min}$.

Q: p.7, l. 25: the (1-1/e) guarantee for stochastic greedy holds "in expectation" for submodular functions.

R: You are correct, a note about this has been added in the manuscript.

Q: What is the value of a used in the simulations of Section IV-D?

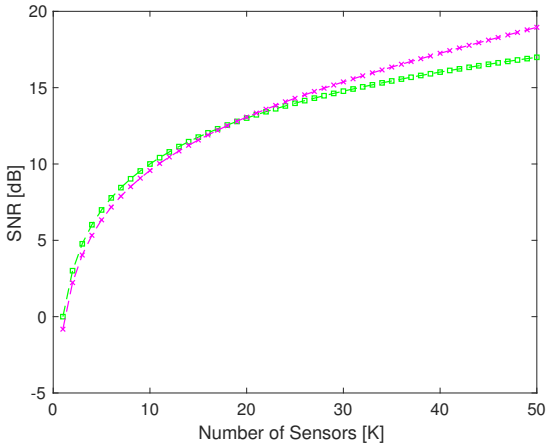
R: As we were free to select the parameter a, for our simulations we set $a = 0.5\lambda_{\min}\{\mathbf{\Sigma}\}$. We included this in the manuscript.

Q: The example in Section IV-E is very interesting and illustrative. It is also in line with the result of Theorem 2. Indeed, as $\rho \to 1$, $\epsilon \to \infty$ and the performance guarantee for the greedy search is no longer meaningful. This is indeed a situation where you would expect things to not work.

R: Thanks for the comment, indeed that's the case.

Q: Also in Section IV-E, Fig. 3 appears to show that for a small number of sensors, greedy SNR outperforms (or performs as well as) the submodular surrogate. Placing more than 25% of the ground set seems unlikely to happen, especially for large-scale problems. It would be interesting to see a detail of this region in the figure.

R: Indeed, for sparse sensing we will usually be looking at selections with $K << M$. Here the figure was meant to show that the submodular method reaches the maximal function value with less sensors, but yes, for a small number of sensors the selection leads to similar values as shown in the figure below.

[

In the case that the reviewer thinks that this figure has to be added to the final version of the manuscript, we will do so.

Q: I would suggest rewriting the proof of Theorem 2 to make the goals of each step clearer. As it is, the presentation is confusing.

R: Thanks for the suggestion. Following this, and some other review comments, the theorem as well as the proof have been rewritten.

In view of these issues, I recommend a major review of the paper. I believe the paper has enough novelty and contributions, but they are not completely clear in the current version of the manuscript. Currently, it is unclear that the paper uses and derives approximate and surrogate metrics. Moreover, the only clear near-optimal guarantee given in the text is for the submodular surrogate of the SNR (given the difficulty in judging the ϵ-submodularity guarantee mentioned above). At this point, it is unclear when the ϵ bound is meaningful

Thanks for your feedback, in the current version of the manuscript we have restructured Theorem 2 in order to convey in a better way its importance. However, as explained in the following part, the aim of the work is the introduction of submodularity as an alternative to convex optimization for sensor selection in detection. In doing so, we also justify the usage of the greedy heuristic for the SNR, and when it works good. In addition, as we show that we can jointly evaluate the true SNR and the surrogate function, the ϵ-submodularity guarantee for the SNR is meant for explaning why in certain instances using the true SNR leads to good results.

and the paper does not derive guarantees for Algorithm 2, which makes it hard to gauge the novelty of the manuscript. Reframing the paper and contributions (perhaps modifying the title) could clarify the position of the work. The text would benefit from copy-editing and the clarity

of some of the technical derivations could be improved.

Thank you. Considering your comment, we modified the title of our paper to:

Submodular Sparse Sensing for Gaussian Detection with Correlated Observations

The manuscript has been modified to lay out this point more clearly and editing the content has been carried out in order to improve the clarity of the proposed ideas.

Reviewer: 3

Recommendation: RQ - Review Again After Major Changes

Comments:

The paper proposes greedy methods for finding the most discriminative subset out of a given set of sensors for the purpose of event detection, under Gaussian distributions. Two cases are considered with respect to the measurements' distributions: equal mean values and distinct covariance matrices and equal covariance matrices and distinct means. In each of the cases, the paper finds submodular surrogate functions that replace the original error exponent metrics, Bhattacharya and Kullback Leibler distance, followed by the greedy algorithm that approximately optimizes the surrogates . The surrogate functions method is rooted in the Schur complement idea from authors' previous works on convex relaxation based sparse sensing, and it's nice to see that it applies in the context of greedy algorithms as well. However, judging overall, the method is of limited applicability as it does not lend itself to the realistic scenario when both the mean values and the covariance matrices differ under the two hypotheses. I would therefore strongly suggest an extension of the method to this general case. Alternatively, if the extension of the surrogate machinery proves challenging, it would be nice if we could at least have bounds on the induced suboptimality resulting from applying – e.g.– method 1 to the general case by ignoring the differences between close but unequal covariance matrices (and similarly for applying method 2 for distributions with close but unequal means).

Thanks for your comments. Your comment encouraged us to extend our results to the general case. It turns out that the tools that are applicable for the particular cases discussed in Sections *IV* and *V* can be directly reused for the general case. In the manuscript we have now added a subsection in which we explain how the general case can be treated by using the same decompositions proposed for the case of different covariance matrices. For instance, in the case of the KL divergence we can show that the decomposition does not change as the quadratic term, $(\boldsymbol{\theta}_1-\boldsymbol{\theta}_0)\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_0)^T$, can be rewritten using the trace function, i.e., $\text{tr}(\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_0)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_0)^T)$ and can be absorbed by the other trace in the divergence expression. In a similar fashion, this can be done for the other distance measures discussed in this manuscript. Details of this procedure are included in the revised version

B. Uncommon Means and Uncommon Covariances

So far, only particular cases of the general hypothesis testing problem between two Gaussian distributions have been discussed. However, in the following, we show that the general case, i.e., distinct means and covariance matrices, can be solved using the same heuristics as discussed for the particular cases presented earlier in the previous sections.

To show that we can reuse the same machinery as in the uncommon covariances case, first let us consider the KL divergence:

$$\begin{aligned}\mathcal{K}(\mathcal{H}_1\|\mathcal{H}_0) &= \tfrac{1}{2}\Big(\operatorname{tr}(\boldsymbol{\Sigma}_{0,\mathcal{A}}^{-1}\tilde{\boldsymbol{\Sigma}}_{1,\mathcal{A}}) - M + \log\det(\boldsymbol{\Sigma}_{0,\mathcal{A}}) \\ &\qquad - \log\det(\boldsymbol{\Sigma}_{1,\mathcal{A}})\Big).\end{aligned} \tag{2}$$

In the above expression we have rewritten the quadratic form in terms of a trace [cf. Table 1], and defined the matrix $\tilde{\boldsymbol{\Sigma}}_{1,\mathcal{A}} := \boldsymbol{\Sigma}_{1,\mathcal{A}} + (\boldsymbol{\theta}_{1,\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})(\boldsymbol{\theta}_{1,\mathcal{A}} - \boldsymbol{\theta}_{0,\mathcal{A}})^T$ which remains symmetric. From (43) it can be seen that the decomposition proposed before can directly be used by just replacing $\boldsymbol{\Sigma}_{1,\mathcal{A}}$ by $\tilde{\boldsymbol{\Sigma}}_{1,\mathcal{A}}$ in the $q_{\text{sub}}(\cdot,\cdot)$ set function. Therefore, for this measure there is no distinction between these two cases. In addition, as the J-divergence is a sum of KL divergences, the decomposition of the J-divergence follows immediately.

For the Bhattacharyya distance, the quadratic form that appears in the expression for the general case, i.e., the term related to the distances between the means, needs to be added to the decomposition given in (40). As this new term is not submodular, the surrogate function proposed for the uncommon means case can be directly employed to provide a submodular set function for the decomposition. As a result, the surrogate distance measure for the Bhattacharyya distance can be given through the decomposition:

$$\begin{aligned}\mathcal{B}_{\text{sub}}(\mathcal{H}_1\|\mathcal{H}_0) &:= g(\mathcal{A}) - h(\mathcal{A}); \\ g(\mathcal{A}) &= \frac{1}{2}\log\det(\boldsymbol{\Sigma}_{\mathcal{A}}) + \frac{1}{8}\log\det(\mathbf{M}_{\mathcal{A}}); \\ h(\mathcal{A}) &= \frac{1}{4}(\log\det(\boldsymbol{\Sigma}_{0,\mathcal{A}}) + \log\det(\boldsymbol{\Sigma}_{1,\mathcal{A}})),\end{aligned} \tag{3}$$

which can be optimized using the approach discussed for the uncommon covariance case.

In summary, the greedy heuristic for the most general case of uncommon means and uncommon covariances can be develop using straightforward adaptations of the methods presented in Sec. V.A.

Otherwise, I would recommend condensing the paper and submitting it as a letter (there are repetitions in the methodology which make potential shortening feasible).

We would like to emphasize that the current manuscript provides the following contributions

- a general framework for selecting sensors for Gaussian detection using the submodular machinery;
- efficient algorithms (using rank-one-like updates) leveraging submodularity that deliver comparable results with their convex counterparts, but significantly (orders of magnitude) faster.
- a technique based on the Schur complement to generate submodular functions from quadratic expressions, establishing a link with common convex approaches.
- we also address the questions of why and when doing greedy directly on SNR works well, while arguing that SNR is not always the best cost function to optimize greedily.

Considering all the above contributions combined with the fact that we extended our methods to the general Gaussian case, we strongly believe that the paper contains novel contributions and material for a TSP publication. Therefore, we feel that shortening this paper would not be appropriate.

Another suggestion for revision is to compare the presented method to the one from [16] based on Stieffel relaxation. In the case of equal means, the method from [16] has a closed form solution and the complexity comparable with the complexity of the method presented in the current paper. Therefore, it is instructive to make a numerical comparison between the two methods.

Similar to the approach in [16], the convex optimization method provides a continuous relaxation which can be solved exactly in the reals. However, both [16] and the convex approach require a projection onto a Boolean space. We opted to make a comparison of the submodular method with the convex method as its solution can be enforced to be sparse, differently from the solution obtained through a Stiefel manifold approach. In addition, the assumption made in [16] where $\mathbf{E}$ has to be fixed to meet $\mathbf{E}^T\mathbf{S}_0\mathbf{E} = \mathbf{I}$ is a little restrictive. Furthermore, the process of passing from the matrix in the Stiefel manifold (relaxation phase) to a Boolean matrix is not

clear. In contrast, with the convex approach, a sparse Boolean solution can easily be obtained, thus we compare the proposed methods with the convex optimization based solutions.