Efficient Robust Parameter Identification in Generalized Kalman Smoothing Models

Jonathan Jonker, Peng Zheng, Member, IEEE, and Aleksandr Y. Aravkin, Member, IEEE

Abstract—Dynamic inference problems in autoregressive (AR/ARMA/ARIMA), exponential smoothing, and navigation are often formulated and solved using state-space models (SSMs), which allow a range of statistical distributions to inform innovations and errors. In many applications, the main goal is to identify not only the hidden state, but also additional unknown model parameters (e.g., AR coefficients or unknown dynamics). We show how to efficiently optimize over model parameters in SSM that use smooth process and measurement losses. Our approach is to project out state variables, obtaining a value function that only depends on the parameters of interest, and derive analytical formulas for first and second derivatives that can be used by many types of optimization methods. We illustrate this by implementing Newton, Gauss–Newton, and quasi-Newton algorithms on a numerical example. The approach can be used with smooth robust penalties such as Hybrid and the Student’s T, in addition to classic least squares. We use the approach to estimate robust AR models and long-run unemployment rates with sudden changes.

Index Terms—Maximum a posteriori estimation, optimization methods, parameter estimation, smoothing methods.

I. INTRODUCTION

The linear state space model (SSM) is widely used in tracking and navigation [6], control [1], signal processing [2], and other time series [10], [17]. The model assumes linear relationships between latent states with noisy observations

\[ \begin{align*}
    x_k &= G_k x_{k-1} + e_k^p, & k = 1, \ldots, N \\
    z_k &= H_k x_k + e_k^m, & k = 1, \ldots, N
\end{align*} \tag{1} \]

where \( x_0 \) is a given initial state estimate, \( x_1, \ldots, x_N \) are unknown latent states with known linear process models \( G_k \), and \( z_1, \ldots, z_N \) are observations obtained using known linear models \( H_k \). The errors \( e_k^p \) and \( e_k^m \) are assumed to be mutually independent random variables with covariances \( Q_k \) and \( R_k \). These covariances may be singular to capture standard autoregressive structures.

In many applications the models \( G_k, H_k, Q_k, R_k \) are specified up to model parameters \( \theta \). We restrict out attention to formulations where covariances \( Q_k, R_k \) are known, while \( G_k(\theta) \) and \( H_k(\theta) \) are \( \mathbb{C}^2 \) mappings of \( \theta \). This captures smoothing parameters in Holt–Winters c.f. [10], autoregressive and moving average parameters in ARMA c.f. [17], and unknown dynamic parameters in navigation models. In most of these models, \( G \) and \( H \) are affine functions of unknown parameters \( \theta \).

In this article, we focus on MAP estimators, optimizing the likelihood in both the state and the parameters. The appropriateness of MAP estimators always depends on the problem and context, e.g., joint MAP estimation of states and parameters can be biased [11], and the bias can be addressed through corrections or alternative formulations, including marginal likelihood [7]. Nonetheless, MAP estimates are the standard for many of the motivating applications, and the techniques developed here can be extended to other settings (such as marginal likelihood) as we discuss in the conclusions.

In all of these models, we can infer unobserved states \{\( x_k \)\} and parameters \( \theta \) using the maximum a posteriori, or MAP, analysis. The first equation in (1) is a prior on sequential differences in the state. Once we observe \( z_k \), we maximize the resulting likelihood to get the posterior estimates. Standard models assume the errors \( e_k^p \) and \( e_k^m \) are Gaussian, which gives rise to the least squares penalty in the MAP inference problem (see the red solid curve in Fig. 1). Changing the observation model to the Hybrid (blue dash) or Student’s T loss (violet dash dot) robustifies the MAP estimate to measurement outliers.

Fig. 1. Common smooth loss functions: Least squares (red solid), Hybrid (blue dashed), and Student’s T (violet dash dot).

Analogous changes to the innovations model allows the framework to track sudden changes.

Prior art for robust system identification includes adding a normalization term to the parameter estimate in a classic least square solver [14] and explicitly modeling noise as a mixture of heavy tailed pdfs [9]. ARMA systems specifically have been reformulated using SVMs to compute robust parameter estimates [15]. Partial minimization techniques have been used for range of inverse problems [4], [5]. Here, we consider SSMs with potentially singular covariance terms, and obtain MAP estimates for parameters and state variables simultaneously using second order methods.

We can design efficient algorithms if we (1) minimize with respect to the state sequence \{\( x_k \)\} and (2) obtain derivatives of the resulting value function, which depends on \( \theta \). Higher order derivatives enable...
second order methods to optimize this value function. We compute these derivatives for general models that cover motivating applications.

Example: Structural Unemployment Rate: We are interested in fitting parameters within structural unemployment rate models (see e.g., [13]). The state vector

\[ x_k = \begin{bmatrix} u_{k-1} \ u_{k-1} \ u_k \ u_k \end{bmatrix}^\top \]

tracks total (\(u\)) and “cyclic” (\(u^c\)) unemployment using the autoregressive model

\[ G_k = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -l_1 & 0 & 1 + l_1 & 0 \\ 0 & 1/2 - l_2 & 0 & l_2 \end{bmatrix}, \quad \epsilon^p_k = \begin{bmatrix} 0 \\ 0 \\ -l_1 \\ 0 \end{bmatrix} \quad \epsilon^m_k = \begin{bmatrix} 0 \\ 0 \\ -l_1 \\ 0 \end{bmatrix} \]

\[ H_k = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & \gamma/2 & 0 & \gamma/2 \end{bmatrix}, \quad \epsilon^v_k = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \]

Here, \(l_1\) and \(l_2\) are auto-regressive parameters while \(\gamma\) is an unknown measurement parameter. Unemployment rates can experience fast changes, so we need a heavy tailed model for innovations. To solve the full problem, we must:
1) estimate states \(\{x_k\}\) as well as parameters \(l_1, l_2, \gamma\);
2) account for the singular process covariance \(Q\);
3) use non-Gaussian losses (e.g., Hybrid and Student’s T) for \(u\) to track fast rate changes.

The article proceeds as follows. In Section II, we review optimization formulations for state and model parameter inference using singular and nonsingular covariance models, and introduce the value function which depends only on the model parameters, as e.g., \(l_1, l_2, \gamma\) above. In Sections III and IV, we look in detail at nonsingular and singular Kalman smoothing models, and obtain existence results and formulas for first and second derivatives of the value function. Finally, in Section V, we present use cases that show how to efficiently obtain structural parameters when using general losses and singular covariance structure.

II. DIFFERENTIATING IMPLICIT FUNCTIONS

In this section, we introduce a general theoretical result for calculating the derivatives for implicit functions in an optimization context. We then specialize this general theorem to nonsingular and singular SSMs in the following section.

Consider a \(C^2\)-smooth function, \(f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}\), where in SSM models we denote parameters by \(\theta\) in \(f(\theta, y)\) and the states together with any auxiliary variables (such as dual variables) by \(y\). The appropriate stationarity condition is given by

\[ H(\theta, y) := f_y(\theta, y) = 0. \]

For any given \(\theta\), the optimal estimate \(y(\theta)\) is obtained by solving the equation \(H(\theta, y) = 0\); in particular \(y(\theta)\) depends on \(\theta\) implicitly. When \(f\) is convex in \(y\), (5) is equivalent to global optimality. More generally, \(y(\theta)\) is any solution that satisfies (5), and the standard implicit function theorem (below) provides weaker conditions under which we know \(y(\theta)\) exists as a function locally. We introduce a variant of the implicit function theorem presented by [8] to characterize the structure of this implicit dependence. These developments do not require convexity of \(f\).

**Theorem 1:** (Implicit Functions and Derivatives): Suppose that \(\mathcal{Y} \subset \mathbb{R}^n\) and \(\mathcal{Y}' \subset \mathbb{R}^m\) are open, \(H : \mathcal{Y} \times \mathcal{Y}' \to \mathbb{R}^m\) is continuously differentiable. If there exists \(\overline{y} \in \mathcal{Y}\) and \(\overline{y}' \in \mathcal{Y}'\), such that \(H(\bar{\theta}, \bar{y}') = 0\) and \(H_y(\bar{\theta}, \bar{y}')\) is invertible. Then there exists (if necessary we choose \(\mathcal{Y}'\) and \(\mathcal{Y}\) to be small neighborhood of \(\overline{y}\) and \(\overline{y}'\) to guarantee the existence) a \(C^1\) mapping \(Y : \mathcal{Y}' \to \mathcal{Y}\) satisfying \(Y(\overline{y}') = \overline{y}'\), and \(H(\theta, Y(\theta)) = 0\) for all \(\theta\) in \(\mathcal{Y}'\).

Moreover, we have the formula

\[ Y_\theta(\overline{y}) = -H_\theta(\theta, Y(\theta))^{-1}H_\theta(\theta, Y(\theta)). \]

This variant of the implicit function theorem gives an explicit formula for \(Y(\theta)\), which we use in further development.

When the function \(Y(\theta)\) as above exists, we can define the value function

\[ v(\theta) = f(\theta, Y(\theta)), \quad \theta \in \mathcal{U}. \]

Our goal is to compute first and second derivatives of \(v\), which are summarized in the following corollary.

**Corollary 1:** (Derivatives of the Value Function (6)): Under the assumptions of Theorem 1, and using \(\overline{y}\) to represent the \(y\) obtained by evaluating \(Y(\overline{y}')\), we have

\[ v_{\theta}(\overline{y}) = f_\theta(\theta, \overline{y}), \quad v_{\overline{y}}(\overline{y}) = f_{\overline{y}}(\theta, \overline{y}) - H_{\theta}(\theta, \overline{y})^{-1}H_{\overline{y}}(\theta, \overline{y})^{-1}H_{\theta}(\theta, \overline{y}). \]

These derivatives are along the lines of those presented by Bell and Burke [8] and are mainly given here for a self-contained exposition.

We now compute analytic expressions of derivatives with respect to model parameters for both nonsingular and singular Kalman smoothing systems.

III. NONSINGULAR SSM

We use superscripts to distinguish process (\(p\)) and measurement (\(m\)) model variables and subscripts to represent partial derivatives (\(\theta, y, r, \ldots\)) and the index in the Kalman model (\(k\)). We also consider the following loss functions:

1) least squares: \(\ell(r) = \frac{1}{2} ||r||^2\);  
2) Hybrid: \(\ell(r; \nu) = \sum_i \sqrt{r_i^2 + \nu^2} - \nu\);  
3) Student’s T: \(\ell(r; \nu) = \sum_i \ln(1 + r_i^2 / \nu^2)\).

Consider the case where the covariance matrices \(Q_k\) and \(R_k\) in SSM are nonsingular. Prewhiten \(\epsilon^p_k\) and \(\epsilon^v_k\), the objective function of interest is given by

\[ f(\theta, y) = \sum_{k=1}^{N} \left\{ \ell^p_k \left( S_k^{1/2} (x_k - G_k(\theta)x_{k-1}) \right) + \ell^m_k \left( R_k^{1/2} (y_k - H_k(\theta)x_{k}) \right) \right\} \]

where \(y = x = [x_1; \ldots; x_N]\), \(\ell^p_k\) and \(\ell^m_k\) are the loss function corresponding to the distributions of \(\epsilon_k^p\) and \(\epsilon_k^m\). Here, we assume \(\ell^p_k, \ell^m_k\) are smooth; three key examples are least squares, Hybrid, and Student’s T losses. Objective (8) can be written compactly as

\[ f(\theta, y) = \ell^p \left( Q^{-1/2} (G(\theta)y - \zeta) \right) + \ell^m \left( R^{-1/2} (H(\theta)y - z) \right) \]

where \(G(\theta)\) and \(Q\) are given by

\[ \begin{bmatrix} I & 0 \\ -G_G(\theta) & I \end{bmatrix} \begin{bmatrix} Q_1 \\ \vdots \\ \vdots \\ -G_N(\theta) & I \end{bmatrix} \]
while \( H(\theta) \) and \( R \) are given by

\[
\begin{bmatrix}
H_1(\theta) & 0 & \cdots & 0 \\
0 & H_2(\theta) & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & H_N(\theta)
\end{bmatrix},
\begin{bmatrix}
R_1 \\
\vdots \\
R_N
\end{bmatrix}.
\]

Here, \( \zeta = [x_0; 0; \ldots; 0] \) and \( z = [z_1; \ldots; z_N] \).

The stationary condition in this case is given by

\[
\mathcal{H}(\theta, y) = f_y(\theta, y) = G(\theta)^\top Q^{-1/2} \mathcal{L}_p(r^p) + H(\theta)^\top R^{-1/2} \mathcal{L}_m(r^m)
\]

(10)

where \( r^p = Q^{-1/2}(G(\theta)y - \zeta) \) and \( r^m = R^{-1/2}(H(\theta)y - z) \).

In the least squares case, (10) is a linear equation solved by inverting the block-tridiagonal system

\[
G(\theta)^\top Q^{-1} G(\theta) + H(\theta)^\top R^{-1} H(\theta)
\]

for more general smooth penalties \( \mathcal{L}_p, \mathcal{L}_m \) a Newton method is needed to compute \( y(\theta) \).

By Theorem 1, existence and differentiability of \( Y(\theta) \) is guaranteed by the existence of the pair \((\overline{\theta}, \overline{y})\) such that, \( \mathcal{H}(\overline{\theta}, \overline{y}) = 0 \) and the partial Hessian below is invertible

\[
\mathcal{H}_y(\overline{\theta}, \overline{y}) = G(\overline{\theta})^\top Q^{-1/2} \mathcal{L}_p(r^p) Q^{-1} G(\overline{\theta}) + H(\overline{\theta})^\top R^{-1/2} \mathcal{L}_m(r^m) R^{-1/2} H(\overline{\theta}).
\]

When \( \mathcal{L}_p, \mathcal{L}_m \) are least squares or Hybrid, \( \mathcal{H}_y(\theta, y) \) is invertible for any \((\theta, y)\), and for every \( \overline{\theta} \) there exist a \( \overline{y} \) such that \( \mathcal{H}(\overline{\theta}, \overline{y}) = 0 \), since both penalties are strictly convex. In the case of the Student’s \( T \) the Hessian may fail to be positive definite at some pairs \((\theta, y)\) [3] and there is no absolute guarantee that the methodology will hold, as expected for a potentially nonconvex formulation. Practical behavior is another matter, and in numerical experience we see a positive definite Hessian at the minimizer, particularly when a descent method is used to solve the state space problem (i.e., the objective function is required to decrease at each iteration). We see that the derivative formulas hold.

We now compute remaining terms in Corollary 1, assuming for simplicity that \( G(\theta) \) and \( H(\theta) \) are affine functions of \( \theta \)

\[
\overline{r}_p := Q^{-1/2} \overline{\mathcal{L}}_p(r^p), \quad \overline{r}_m := R^{-1/2} \overline{\mathcal{L}}_m(r^m)
\]

\[
f_\theta(\theta, y) = (G(\overline{\theta})^\top Q^{-1/2} \mathcal{L}_p(r^p) + H(\overline{\theta})^\top R^{-1/2} \mathcal{L}_m(r^m)
\]

\[
f_{\theta y}(\theta, y) = (G(\overline{\theta})^\top Q^{-1/2} \mathcal{L}_p(r^p) Q^{-1} G(\overline{\theta}) + H(\overline{\theta})^\top R^{-1/2} \mathcal{L}_m(r^m) R^{-1} H(\overline{\theta})
\]

\[
\mathcal{H}_\theta(\theta, y) = (G(\theta)^\top \overline{r}_p) + G(\theta)^\top Q^{-1/2} \mathcal{L}_p(r^p) Q^{-1} G(\theta)^\top
\]

\[
+ (H(\theta)^\top \overline{r}_m) + H(\theta)^\top R^{-1/2} \mathcal{L}_m(r^m) R^{-1} H(\theta)^\top
\]

\[
= (G(\theta)^\top \overline{r}_p) + G(\theta)^\top Q^{-1/2} \mathcal{L}_p(r^p) Q^{-1} G(\theta)^\top
\]

\[
+ (H(\theta)^\top \overline{r}_m) + H(\theta)^\top R^{-1/2} \mathcal{L}_m(r^m) R^{-1} H(\theta)^\top.
\]

We now have, fully and explicitly, first and second derivatives of the value function \( v(\theta) \) in (7) for the nonsingular case. Though these results are straightforward, they do not appear in any smoothing literature we are aware of in this compact form, even for least squares losses.

\[\text{IV. SINGULAR SSM}\]

When covariances \( Q_k \) and \( R_k \) are singular, we rewrite (8) to include null-space constraints. A singular covariance matrix precludes any errors and innovations that are not in its range. We follow [12] in formulating this problem

\[
\min_{\theta, x, r^p, r^m} \mathcal{L}_p(r^p) + \mathcal{L}_m(r^m)
\]

s.t. \( Q^{1/2} r^p = G(\theta)x - \zeta \)

\( R^{1/2} r^m = H(\theta)x - z \)

(11)

and introduce the Lagrangian

\[
f(\theta, y) = \mathcal{L}(\theta, x, r^p, r^m, \lambda^p, \lambda^m)
\]

\[
= \mathcal{L}_p(r^p) + \mathcal{L}_m(r^m)
\]

\[
- (\lambda^p, Q^{1/2} r^p - G(\theta)x + \zeta)
\]

\[
- (\lambda^m, R^{1/2} r^m - H(\theta)x + z).
\]

(12)

When \( Q_k \) and \( R_k \) are invertible, we can solve for \( r^p, r^m \) in (11) and reduce the problem to (8), so nonsingular systems are a special case of (11). Formulation (11) can be solved for a variety of loss functions \( \mathcal{L}_p \) and \( \mathcal{L}_m \) (see [12]).

Here, we define the value function as a mini–max problem using the Lagrangian

\[
v(\theta) := \max_{\lambda^p, \lambda^m} \min_{x, r^p, r^m} \mathcal{L}(\theta, x, r^p, r^m, \lambda^p, \lambda^m).
\]

(13)

Just as in the nonsingular case, the difficulty with the Student’s \( T \) case is that it is not convex. It is therefore hard to evaluate \( v(\theta) \). However, the development below uses systems of equations that characterize stationarity, just as in the nonsingular case. The system we are interested in is given as follows:

\[
0 = \mathcal{H}(\theta, y) := f_y(\theta, y), \quad y := \{x, r^p, r^m, \lambda^p, \lambda^m\}
\]

that is, the system of equations that defines a saddle point of the Lagrangian. Explicitly, \( f_y(\theta, y) = 0 \) is given by

\[
f_y(\theta, y) = \begin{bmatrix}
G(\theta)^\top \lambda^p + H(\theta)^\top \lambda^m \\
\mathcal{L}_p(r^p) - Q^{1/2} \lambda^p \\
\mathcal{L}_m(r^m) - R^{1/2} \lambda^m \\
G(\theta)x - \zeta - Q^{1/2} r^p \\
H(\theta)x - z - R^{1/2} r^m
\end{bmatrix} = 0.
\]

(14)

\[\text{Y(\theta) is differentiable when } \mathcal{H}_y(\theta, y) = f_{yy}(\theta, y) \text{ is invertible}\]

\[f_{yy}(\theta, y) = \begin{bmatrix}
0 & 0 & 0 & G(\theta)^\top & H(\theta)^\top \\
0 & \ell_p(r^p) & 0 & -Q^{1/2} & 0 \\
0 & 0 & \ell_m(r^m) & 0 & -R^{1/2} \\
G(\theta) & -Q^{1/2} & 0 & 0 & 0 \\
H(\theta) & 0 & -R^{1/2} & 0 & 0
\end{bmatrix}.
\]

(15)

We state the following theorem.

**Theorem 2:** Invertibility of \( f_{yy}(\theta, y) \) is equivalent to invertibility of the so-called Hessian of the Lagrangian

\[
H(\theta)G(\theta)^{-1} Q^{1/2} (\ell_p(r^p))^{-1} Q^{1/2} G(\theta)^{-1} H(\theta)^\top + R^{1/2} (\ell_m(r^m))^{-1} R^{1/2}.
\]

When \( \mathcal{L}_p, \mathcal{L}_m \) are the least squares or Hybrid penalties, (15) is invertible if and only if

\[\mathcal{N}(R) \cap \mathcal{N}(QG^{-1}H^{-1}) = \{0\}.
\]

(16)
When Theorem 2 holds, we use Corollary (1) to get derivatives of \( v(\theta) \) in (13)

\[
v(\overline{\theta}) = f(\overline{\theta}, \overline{y})
\]

\[
v_{\theta}(\overline{\theta}) = f_{\theta}(\overline{\theta}, \overline{y})
\]

\[
v_{\theta\theta}(\overline{\theta}) = f_{\theta\theta}(\overline{\theta}, \overline{y}) - H_{\theta}(\overline{\theta}, \overline{y})^\top H_{\theta}(\overline{\theta}, \overline{y})^{-1} H_{\theta}(\overline{\theta}, \overline{y}).
\]

It remains only to compute \( f_0, f_{\theta\theta}, H_\theta, \) and \( H_\varepsilon_\theta \).

\[
f_0(\theta, \overline{y}) := (\langle \overline{x}, G(\theta)\overline{y} \rangle)_\theta + (\langle \overline{x}_m, H(\theta)\overline{y} \rangle)_\theta.
\]

When \( G, H \) are affine functions of \( \theta \), we have \( f_{\theta\theta} = 0 \). Finally

\[
H_{\theta}(\theta, \overline{y}) = f_{\theta\theta}(\theta, \overline{y}) = \begin{bmatrix}
(G(\theta)^\top K)^_\theta + (H(\theta)^\top K)^_\theta
0
0
(G(\theta)\overline{y})_\theta
(H(\theta)\overline{y})_\theta
\end{bmatrix}.
\]

A. Special Case: Invertible \( R \)

The structural unemployment model in the introduction has a singular \( Q \) but an invertible \( R \). In such cases, the derivative formulas can be written using only primal quantities, which significantly decreases the notational burden. In particular, using the optimality conditions we have

\[
\lambda^m = R^{-1/2} \epsilon^m = R^{1/2} \lambda^m
\]

\[
\lambda^p = -G(\theta)^\top H(\theta)^\top \lambda^m
\]

and so we get the explicit primal-only formula for \( v_\theta(\theta) \) by using these expressions

\[
v_\theta(\theta) = (\langle \overline{x}^0, G(\theta)\overline{y} \rangle)_\theta + (\langle \overline{x}_m, H(\theta)\overline{y} \rangle)_\theta.
\]

B. Special Case: Least Squares

If \( \epsilon^p(\cdot) \) and \( \epsilon^m(\cdot) \) are both given by \( \frac{1}{2} \| \cdot \|^2 \), the optimality conditions simplify substantially, and we have

\[
r^p = Q^{1/2} \epsilon^p, \quad r^m = R^{1/2} \lambda^m
\]

\[
Q^{1/2} r^p = Q \lambda_p = G(\theta)x - \zeta
\]

\[
R^{1/2} r^m = R \lambda_m = H(\theta)x - z
\]

\[
0 = G(\theta)^\top \lambda^p + H(\theta)^\top \lambda^m.
\]

Plugging these conditions back into the Lagrangian, we get the dual objective (i.e., the dual problem is to maximize the objective below with respect to \( (\lambda^p, \lambda^m) \))

\[
f^*(\lambda^p, \lambda^m) = -\frac{1}{2} \langle \lambda^p, Q \lambda^p \rangle - \frac{1}{2} \langle \lambda^m, R \lambda^m \rangle
\]

\[
- \langle \lambda^p \rangle^\top \zeta - \langle \lambda^m \rangle^\top z
\]

s.t. \( G(\theta)^\top \lambda^p + H(\theta)^\top \lambda^m = 0 \).

Using invertibility of \( G \), we eliminate \( \lambda_p \) and the affine constraint, obtaining an objective function in \( \lambda_m \) alone

\[
f^*(\lambda_m) = -\frac{1}{2} \langle \lambda^m, (HG^{-1}QG^{-T}H^\top + R) \lambda^m \rangle
\]

\[
- \langle \lambda^m \rangle^\top (z - HG^{-1} \zeta).
\]

In the least squares case, the dual solution \( \lambda^m \) that maximizes the above objective is unique exactly when the linear system

\[
HG^{-1}QG^{-T}H^\top + R
\]

is invertible, and then we have

\[
\lambda^m = (HG^{-1}QG^{-T}H^\top + R)^{-1} (z - HG^{-1} \zeta)
\]

a closed form solution. A simple sufficient condition for the invertibility of (18) is to have \( R \) itself invertible, as in special case A.

V. NUMERICAL EXAMPLES

We now apply the results of the previous sections to analyze two simple singular models with unknown states and parameters. In Section V-A, we present a state-space formulation for the AR-1 model, show how to robustify it to outliers in the data, and present explicit derivatives for the value function. We use these derivatives to design an efficient solver for both standard and robust AR models using Newton, Gauss-Newton, and quasi-Newton algorithms. In Section V-B, we apply the methods in this article to fit a structural model for unemployment rates that can track fast changes. While the structural unemployment model is currently used in the EU, in this article we only show results on simulated synthetic data where we know ground truth, leaving a data driven illustration to a future collaboration.

A. Robust AR Fitting

We examine the simplest case where a reformulated state space leads to singular \( Q_k \) and \( R_k \). This derivation illustrates how singular covariances can arise in more complex scenarios including all ARMA models. An AR-1 model is given by

\[
x_k = c + \varphi x_{k-1} + \epsilon_k^p
\]

\[
y_k = H_k x_k + \epsilon_k^m
\]

where \( \epsilon_k^p, \epsilon_k^m \) have covariances \( Q_k, R_k \) respectively, \( c \) is an unknown constant, and \( \varphi \) is a parameter to be estimated. To reformulate this model using (1) we create an augmented state

\[
\hat{x}_k^h = \begin{bmatrix} x_k \\ c \end{bmatrix}
\]

and use state equations

\[
\hat{x}_k^h = \begin{bmatrix} \varphi & 1 \\ 0 & 1 \end{bmatrix} \hat{x}_{k-1}^h + \epsilon_k^p
\]

\[
y_k = \begin{bmatrix} H_k & 0 \\ 0 & 0 \end{bmatrix} \hat{x}_k^h + \epsilon_k^m.
\]

Then

\[
\hat{Q}_k = \begin{bmatrix} Q_k & 0 \\ 0 & 0 \end{bmatrix}, \quad \hat{R}_k = \begin{bmatrix} R_k & 0 \\ 0 & 0 \end{bmatrix}
\]

This choice of \( \hat{Q}_k \) allows us to fit the constant \( c \) as part of the state while \( \hat{R}_k, k > 0 \) holds \( c \) constant across time. To compute the derivatives of the value functions we note that

\[
(G_1(\theta)\eta^\top )_\theta = \begin{bmatrix} \varphi & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \eta^p \\ \eta^m \end{bmatrix}_\theta = \begin{bmatrix} \eta^p \\ \eta^m \end{bmatrix} = \tilde{D}\eta
\]

(22)
TABLE I

| Table of Results When Run on Generated Data, The Second Row Indicates the Loss Functions That Were Used Where LS Stands for Least Squares (\(\frac{1}{2}||\hat{r}||^2\)), T Stands for Students T with \(\nu = 10\), and H Stands for Hybrid With \(\epsilon = .7\). |

|               | Nominal | Outliers | Large Jumps |
|---------------|---------|----------|-------------|
|               | ls/ls   | H/ls     | ls/ls       | ls/l     | ls/H     | ls/ls       | ls/l     | ls/H     | ls/ls       | ls/l     | ls/H     |
| Newton        |         |          |             |          |          |             |          |          |             |          |          |
| \(||\theta - \hat{\theta}||^2\) | 0.158   | 0.134    | 0.188       | 1.72     | 0.115    | 0.142       | 0.296    | 0.032    | 0.099       |          |          |
| Inner Iter    | 19      | 250      | 206         | 14.71    | 102      | 21          | 21       | 329      | 373         |          |          |
| Outer Iter    | 9       | 5        | 10          | 6        | 15       | 13          | 13       | 9        | 14          |          |          |
| Time          | 11.5    | 36.5     | 29.7        | 6.3      | 20.2     | 19.3        | 13.3     | 48.1     | 49.5        |          |          |
| L-BFGS        |         |          |             |          |          |             |          |          |             |          |          |
| \(||\theta - \hat{\theta}||^2\) | 0.158   | 0.137    | 0.184       | 1.72     | 0.115    | 0.141       | 0.296    | 0.034    | 0.099       |          |          |
| Inner Iter    | 19      | 228      | 207         | 19       | 86       | 88          | 32       | 375      | 519         |          |          |
| Outer Iter    | 10      | 18       | 10          | 9        | 17       | 12          | 16       | 14       | 15          |          |          |
| Time          | 6.2     | 33.9     | 26.5        | 6.3      | 19.4     | 14.0        | 10.6     | 51.4     | 67.5        |          |          |
| LM-Newton     |         |          |             |          |          |             |          |          |             |          |          |
| \(||\theta - \hat{\theta}||^2\) | 0.158   | 0.069    | 0.184       | 1.04     | 0.12     | 0.112       | 0.296    | 0.028    | 0.099       |          |          |
| Inner Iter    | 16      | 385      | 180         | 32       | 70       | 114         | 24       | 302      | 357         |          |          |
| Outer Iter    | 12      | 15       | 20          | 25       | 28       | 19          | 18       | 15       | 21          |          |          |
| Time          | 7.2     | 51.6     | 26.4        | 14.4     | 20.0     | 21.0        | 11.3     | 43.2     | 48.1        |          |          |

where \(\hat{D} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\). Let \(D\) be a block diagonal matrix with \(\hat{D}\) on the first subdiagonal, the rest zeros.

Then using the above, for the AR-1 model, we have

\[
(G(\theta)x)_\theta = -Dx, \quad (G(\theta)^T \lambda)_\theta = -D^T \lambda.
\]

We now have the expressions

\[
f_\theta(\theta, \bar{\theta}) := -\langle x_D, D \bar{x} \rangle, \quad f_{\theta\theta} = 0
\]

\[
\mathcal{H}_\theta(\theta, \bar{\theta}) = -\begin{bmatrix}
D_\theta^T & 0 \\
0 & 0 \\
D_\bar{\theta} & 0
\end{bmatrix}.
\]

When combined with the general results of Section IV, we get

\[
v(\bar{\theta}) = f(\bar{\theta}, \bar{\theta})
\]

\[
v_\theta(\bar{\theta}) = f_\theta(\bar{\theta}, \bar{\theta}) = -\langle x_D, D \bar{x} \rangle
\]

\[
v_{\theta\theta}(\bar{\theta}) = -\mathcal{H}_\theta(\bar{\theta}, \bar{\theta})^T \mathcal{H}_\theta(\bar{\theta}, \bar{\theta})^{-1} \mathcal{H}_\theta(\bar{\theta}, \bar{\theta}). \tag{23}
\]

**B. Fast Tracking of Unemployment Rates**

In this section, we apply the proposed approach to estimate parameters \((\ell_1, \ell_2, \gamma)\) for the structural unemployment model (2)–(4). To test the approach, we generate ground truth parameters and then create synthetic data in order to compare model performance and speed using different formulations and algorithms. The data is generated by fixing parameters to reasonable values similar to those observed in practice—namely, at \([\ell_1, \ell_2, \gamma] = [0.68, 1.41, -0.68]\), and applying the unemployment rate SSM (2)–(4) to generate the state as well as noisy observations. We then consider three cases—nominal errors, outliers in the observations, and jumps in the unemployment process. In the nominal cases, we use variance parameters known to the smoother. To generate outliers we randomly select 10% of measurements and add additional noise drawn from a \(N(0, 1)\) Gaussian distribution. To generate large jumps we add large deviations \(A = -2\) at indices corresponding to 25 and 65. Two Examples of the generated data are in Fig. 2. An example of the estimated using this data are in Fig. 3.

All algorithms are initialized at \([0 \ 0 \ 0]\), except for LM-Newton on T/ls in the nominal case, which is initialized at \([0 \ 0 \ 0.5]\), as the standard zero initialization leads to bad results for this (nonconvex) case.

In the first iteration, the state is initialized by propagating the initial \(x_0\) through the dynamics for all time. In subsequent iterations the state is always initialized using the previous state solution. In all methods, the full state at each iteration is computed using Newton’s method to find a saddle point of the augmented Lagrangian

\[
AL(y, \theta) = \ell^p(r^p) + \ell^m(r^m) - \langle \lambda^m, R^{1/2} r^m - G(\theta)x + \zeta \rangle
\]

\[
- \langle \lambda^m, R^{1/2} r^m - H(\theta)x + z \rangle
\]

\[
+ \frac{1}{2}||Q^{1/2} r^p - G(\theta)x + \zeta||^2
\]

\[
+ \frac{1}{2}||R^{1/2} r^m - H(\theta)x + z||^2.
\]

The value at an optimal point of (V-B) is the same as at an optimal point of (12). We can see (V-B) is better conditioned which leads to faster convergence in practice. For the outer iterations on the parameter space we compare a Newton method, L-BFGS, and a LM-Newton solver. The standard Newton and L-BFGS are from a standard python library. The LM-Newton solver is a quasi-Newton method where the
Hessian is boosted by a parameter that is updated adaptively based on model performance. The results are summarized in Table I.

All methods work well for convex models. In the nonconvex case, the algorithms become more sensitive. In particular when $\ell^p$ is Student’s T, we have to boost $\ell^p_{\nu}$ by a constant in order to make Newton’s method converge. In practice this constant must be tuned depending on the parameter $\nu$ in the Student’s T function. Convergence is therefore sensitive to the choice of $\nu$ and boosting constant but a good rule of thumb is to choose $1 \leq \nu \leq 20$ and boost just enough to make the Hessian positive definite.

VI. DISCUSSION

We presented a general approach for parameter estimation in singular and nonsingular Kalman smoothing models. In particular we showed how to compute first and second derivatives of the value function (optimizing over state) with respect to the hidden parameters for both singular and nonsingular cases, which captures a wide variety of models, including the motivating example. A simple numerical illustration shows how the computed quantities can be used by a variety of optimization methods. The examples also show that when working with structural parameters, it pays off to have convex subproblems within each iteration of the value function. While nonconvex losses such as Student’s T are always appealing from a modeling perspective, when the problem is to find both the state and parameters, the resulting models are more fragile than those that use convex losses. This observation opens the way to future research in both theory and algorithm design.

Finally, MAP estimates are only one approach to estimate parameters. Multiple work addresses bias in MAP estimates [7], [11]. These approaches can benefit from the techniques developed here, since they require solving auxiliary MAP-like problems. For marginal likelihood, these problems arise when the Laplace approximation is used [16]. How to best use the methodology developed here in these broader contexts is left open to future work.

VII. ACKNOWLEDGMENT

The authors would like to thank J. Nielsen (Economic Council of the Labour Movement) for pointing the use of Kalman smoothers in structural unemployment models, and for teaching about these models.

References

[1] B. D. Anderson and J. B. Moore, Optimal Control: Linear Quadratic Methods. Chelmsford, MA, USA: Courier Corporation, 2007.
[2] B. D. O. Anderson and J. B. Moore, Optimal Filtering. Englewood Cliffs, NJ, USA: Prentice-Hall, 1979.
[3] A. Y. Aravkin, J. V. Burke, and G. Pillonetto, “Robust and trend-following student’s t Kalman smoothers,” SIAM J. Control Optim., vol. 52, no. 5, pp. 2891–2916, 2014.
[4] A. Y. Aravkin, D. Drusvyatskiy, and T. van Leeuwen, “Efficient quadratic penalization through the partial minimization technique,” IEEE Trans. Autom. Control, vol. 63, no. 7, pp. 2131–2138, Jul. 2018.
[5] A. Y. Aravkin and T. Van Leeuwen, “Estimating nuisance parameters in inverse problems,” Inverse Problems, vol. 28, no. 11, 2012, Art. no. 115016.
[6] Y. Bar-Shalom, X. Rong Li, and T. Kirubarajan, Estimation with Applications to Tracking and Navigation. Hoboken, NJ, USA: Wiley, 2001.
[7] B. M. Bell, “The marginal likelihood for parameters in a discrete Gauss-Markov process,” IEEE Trans. Signal Process., vol. 48, no. 3, pp. 870–873, Mar. 2000.
[8] B. M. Bell and J. V. Burke, “Algorithmic differentiation of implicit functions and optimal values,” in Advances in Automatic Differentiation, Berlin, Germany: Springer, 2008, pp. 67–77.
[9] G. Bottegal, A. Aravkin, H. Hjalmarsson, and G. Pillonetto, “Robust EM kernel-based methods for linear system identification,” Automatica, Elsevier, vol. 67, pp. 114–126, 2016.
[10] R. J. Hyndman, A. B. Koehler, R. D. Snyder, and S. Grose, “A state space framework for automatic forecasting using exponential smoothing methods,” Int. J. Forecasting, vol. 18, no. 3, pp. 439–454, 2002.
[11] A. Isaksson, D. Törnqvist, I. Sjöberg, and L. Ljung, “Grey-box identification based on horizon estimation and nonlinear optimization,” In Proc. 41st ISCIE Int. Symp. Stochastic Syst. Theory App., 2010, pp. 1–6.
[12] J. Jonker, A. Aravkin, J. V. Burke, G. Pillonetto, and S. Webster, “Fast robust methods for singular state-space models,” Automatica, vol. 105, pp. 399–405, 2019.
[13] H. N. Mocan, “Structural unemployment, cyclical unemployment, and income inequality,” Rev. Econ. Statist., vol. 81, no. 1, pp. 122–134, 1999.
[14] R. R. Perez, V. F. Batlle, and L. S. Rodriguez, “Robust system identification of an irrigation main canal,” Adv. Water Resour., vol. 30, pp. 1785–1796, 2007.
[15] J. L. Rojo-Álvarez, M. Martínez-Ramón, M. de Prado-Cumplido, A. Artés-Rodríguez, and A. R. Riguerriz-Vidal, “Support vector method for robust ARMA system identification,” IEEE Trans. Signal Process., vol. 52, no. 1, pp. 155–164, Jan. 2004.
[16] Z. Shun and P. McCullagh, “Laplace approximation of high dimensional integrals,” J. Roy. Statist. Soc.: Ser. B (Methodological), vol. 57, no. 4, pp. 749–760, 1995.
[17] R. S. Tsay, Analysis of Financial Time Series, vol. 543, Hoboken, NJ, USA: Wiley, 2005.